Chapter 5. The irreducibility of mind

Anomalism

When Donald Davidson's anti-reductionism about the mind is discussed, the focus usually is the thesis of the *anomalism of the mental*, which is often taken, seemingly also by Davidson himself, as the chosen formulation of the irreducibility thesis. As the word anomalism indicates, at centre stage is a claim to the effect that the mental does not enter into laws, or more precisely, that the mental does not enter into strict or exceptionless laws.

This is how he formulates the position of anomalism in "Mental Events":

> [T]here are no strict deterministic laws on the basis of which mental events can be predicted and explained. (Davidson 1970; 208)

In later papers Davidson stops speaking of 'deterministic' and replaces it by 'exceptionless'. But this update of formulation is unimportant for present purposes.

In this chapter I give what I think of as – broadly – an interpretation and defence of Davidson's anti-reductionism about the mind. But I am troubled by the emphasis on laws which I think threatens to take away the focus from other aspects of his anti-reductionism. Those aspects may be more defensible than anomalism. It is common to think that anomalism is meant to be established by an argument that involves an appeal to a certain view about the nature of the mental, in particular that the mental domain or mental vocabulary is governed by constitutive, synthetic a priori rationality principles. As I have already shown, I have reservations about the traditional view of what this involves. So one question is how the manner in which rationality principles are involved with the mental should be described. I think, however, that considerations on the role of rationality principles in interpretation make plausible a description of the nature of mind that can be said to be anti-reductionist even if the thesis of anomalism does not follow from it. There are some arguments that lead from

those principles to anomalism, but it seems to me that it is unclear whether they could be made to work. So my defence is qualified.

So, I think the focus on the existence of strict or exceptionless laws is wrong or unfortunate. The thesis is interesting in itself, however, and has its perhaps main point in its role in Davidson's argument for token-identity physicalism. In that context it is not right to argue that it gets the focus wrong. But the irreducibility thesis does seem to have independent importance for Davidson.

Problems with the focus on laws

In stating anomalism, does Davidson claim something that is in acute danger of being proved false tomorrow or the next day? No, this does not seem to be the problem. In fact there seems to be almost unanimous agreement among philosophers that there are not (and will not be) important examples of exceptionless laws in the special sciences. But the basis for this conviction is for most philosophers purely empirical, I presume. If we look at the historical record of the special sciences we find that they get along very nicely without exceptionless laws, and it just seems wildly implausible that they will ever produce truly exceptionless laws.

But Davidson clearly aims for a principled, a priori argument, and that is a different matter. Davidson reminds us that there is a form of hubris against which philosophers are often warned. Critics have tended to think that Davidson does make himself guilty of this form of hubris, guilty of poaching on the empirical preserves of science, since laws are very much the preserve of science.

A different problem with choosing anomalism as the formulation of the antireductionist position is that this feature does not distinguish the mental from the other special sciences such as biology, chemistry, geology, and the like. In fact it is certainly at best an open question whether there are truly exceptionless laws anywhere in science – including physics (Cartwright 1983, Cartwright 1999). Davidson admits that there are going to be truly exceptionless laws nowhere else than in basic physics. But then, if reductions require strict laws, the only place where there could be any reductions would be in basic physics. And this would certainly give an impoverished notion of scientific reduction. This leads for instance Jaegwon Kim to conclude:

"What this shows is that Davidson's idea of reduction is too narrow and unrealistic to be of much philosophical interest" (Kim 1998; 93). This verdict is similar to Richard Rorty's. He thinks there just is no philosophical interest in either reducing – or not being able to reduce – one idiom to another (Rorty 1999).

The possibility of other kinds of reduction than reductions based on strict laws has been discussed by many philosophers, and it may be insufficiently clear what Davidson's attitude is (or would be) about the whole spectrum of different kinds of reduction. Today we would probably say that Davidson's conception of reduction (as well as his conception of causation) is hampered by his reliance on the deductive-nomological model. I think there may well be versions of scientific reduction that should not be thought to be incompatible with the basic thrusts of Davidson's philosophy of mind. But I also think that whether or not anomalism follows from the arguments that Davidson offers, and whether or not the question of the existence of strict laws has great philosophical interest, there will be other forms of reduction and other forms of reductionism to which Davidson's general philosophical position – and the position of any interpretationist – should be opposed. This leaves us with the possibility that there is a distinctive antireductionism which is based on elements found in Davidson's philosophy, which nevertheless does not imply the non-existence of strict laws but has other implications. At any extent I think we would profit from finding a different statement as the key formulation of the thesis of antireductionism.

Survey of the chapter

This is the longest chapter in the dissertation. The meat of the chapter comes mainly in the section on Kim's reconstruction of the argument for anomalism and then especially in the following section on the uncodifiability of interpretation in which I present the kind of argument for irreducibility that I endorse.

The two sections that precede the discussion of Kim's argument are mainly preparatory. I begin with drawing some distinctions (primarily between non- and anti-reductionism). Then I turn to an attempt to "save" Davidson from anti-reductionism. Rorty has argued that there is nothing metaphysically special about the mental-physical distinction, and urged Davidson to give up his anti-reductionism. Bjørn Ramberg, however, tries to convince Rorty that the mental-physical distinction is

especially important. He thinks this is a distinction that we should do our best to entrench, and he apparently sees "scientism" as an influence that somehow threatens this distinction, or threatens the mental vocabulary, and talks about it as "oppressive". Still, he wants to hold on to Rorty's a-reductivism and to renounce anti-reductivism. Contrary to this I believe his talk of entrenchment and "oppression" requires a robust irreducibility argument. If such talk means anything it cannot be that the mental-physical distinction is *just* a very important distinction, and it cannot *just* be that the two vocabularies answer to different predictive and explanatory interests. It is sometimes thought that the irreducibility of mental concepts for Davidson simply consists in the fact that they are causal concepts designed to single out a special set of causes or causal conditions which interests us particularly among the totality of conditions which conspire to cause an event. But this, I think, is simply not adequate. Nevertheless, Ramberg's talk of interests that are served by a vocabulary and the idea that these interests could somehow be frustrated provides, I think, useful material for the irreducibility argument to be given later.

In the section on Kim I turn to one of the most influential readings of Davidson's anti-reductionism argument. I argue that this argument fails to be convincing and also that it fails as an interpretation of Davidson. Even in this case, however, there are useful elements of Kim's reconstruction that I build on later. The following main section contains the argument for irreducibility in the best shape that I have been able to give it. It builds heavily on William Child's reconstruction of Davidson's argument where the notion of the *uncodifiability* of interpretation takes a centre stage. Both Child's argument and my own development of it are more reconstructions of Davidson than faithful attempts at interpretation. In the last sections of the chapter I seek to further articulate the "methodological" interpretationism that comes together with this argument, and I trace out some of its implications. A particularly important point is that the anti-reductionism I support is considerably more modest than the one which is often attributed to Davidson (by Kim and Alexander Rosenberg, to mention two examples) and from which the impossibility of psychological explanation and the impossibility of psychology as a science seems to follow immediately. The immodest kind of anti-reductionism fits well with the adoption of the a priori theory conception of mind, and its immodesty constitutes one further argument against that conception of mentality.

Reductionism, non-reductionism and anti-reductionism

I will not try to characterise the many forms of reduction that could be relevant to the question of the relation between mind and body. One might be a reductionist, non-reductionist, or anti-reductionist with respect to either of these forms of reduction (perhaps in addition to other attitudes).

Under the label of a reductionist I will include someone who believes that reductions of a certain form will eventually be forthcoming, or that they are already taking place, or have already taken place. Thus, philosophers like Patricia Churchland and John Bickle are reductionists in this sense (and in other senses beside) in that they are preaching that the reduction of the mind to neuroscience is something that is currently going on (Churchland 1986) (Bickle 2003). A more modest faith in the likelihood of the success of reductive programs would be a more typical representative of reductionism in this sense. But, perhaps more importantly, reductionists include people who see themselves as *committed* to the success (or the in principle success) of a reductive program because of philosophical reasons. Churchland and Bickle may be included in this category too, but more paradigmatic representatives are David Lewis, Frank Jackson and Jaegwon Kim. According to Jackson, conceptual analysis, a kind of reduction, of non-fundamental concepts in terms of basic concepts must be possible if anything in the world should be describable by means of the non-fundamental concepts at all, and Lewis' position is similar. For Kim, the jigsaw pieces of his metaphysical picture simply won't fit together unless certain reductions can in principle be guaranteed. Unless mental properties can be reduced to physical properties mental causation and agency goes down the drain. This kind of reductionism, then, does not consist in *optimism* about the prospects for reductions (although one might be an optimist for different reasons), but in the perceived philosophical necessity of reductions. If one is a reductionist as well as a pessimist about the prospects for reduction, *eliminativism* could be a natural conclusion to draw.

One could be a non-reductionist in simply not endorsing reductionism in either of these senses. If one is agnostic about whether reductions (of a certain type or on a certain scale of magnitude) will be forthcoming, then one is non-reductionist in one sense. If one does not have a philosophical system in which reductions are required,

then one will be non-reductionist in that sense. In Richard Rorty's interpretation of Davidson (or the view he urges Davidson to accept) it is the non-reductionism of the second sort that is emphasised. Rorty sees in Davidson's position the potential for the liberation from the idea that philosophers should feel there is any kind of urgent need for providing reductions of any kind.

Rorty, then, opposes reductionism in the form of the claim that reductions are urgent philosophical business. Rorty, however, would also oppose *anti*-reductionism. On Rorty's view there is no particular philosophical interest either in reducibility claims or in irreducibility claims. His non-reductionism could accept irreducibility claims (or reducibility claims) for a fact, but would hold such facts to be philosophically insignificant. Anti-reductionism disagrees about that. The way I would define it, it could basically come in two forms. In the first place it is the combination of a certain irreducibility claim with the insistence that that fact is philosophically significant. In another form it is merely the opposition to reductions (in some form of reduction), without being a position on how *likely* it is that reductions will be forthcoming. Anti-reductionism could be a rather inarticulate sentiment, for instance the feeling that there would be something bad about reductions, perhaps the idea that what is special about the mind would be threatened or destroyed by reductions of certain kinds, or that reductions would eliminate or undermine certain aspects of our interpersonal practice.

Since the notion of reduction is kept open and there will be many different forms of it, it is possible to combine the attitudes of reductionism, non-reductionism and anti-reductionism for one person with respect to different forms of reduction. Rorty, as we have seen, urges that we should resist both reductionism (except the form in which it is merely the positive estimate of the probability of some future event) and anti-reductionism. There is no doubt that there are important strands in Davidson's philosophy of this kind of non-reductionist attitude, and that Rorty is justified in recruiting him as one of his heroes. At the same time it seems that Davidson is not entirely free from reductionism of the commitment type. I also think there are important elements in Davidson's philosophy that have to be described as anti-reductionist and not merely non-reductionist. Concerning reductionism, Davidson at some points describes himself as an ontological reductionist (Davidson 1985; 242f). More to the point, it may at least sometimes seem that he is motivated by considerations similar to those of Kim's about mental causation. Davidson's particular

form of physicalism, a token-identity theory of events, is forced on us, he thinks, by the recognition that mental events are causes. The way to "ensure" the causal efficacy of mental events is by identifying them with physical events. (I will discuss this in chapter 6.) Rorty thinks that the idiosyncrasies of Davidson's physicalism and the motivation behind it, not to mention Davidson's anti-reductionist elements, are best ignored. Tellingly, he describes Davidson's position not as *anomalous monism*, but as *non-reductive physicalism* (Rorty 1987).

I am to a certain extent sympathetic with this view of Davidson. 'Physicalism' is better than 'monism' because of the way Davidson understands monism[1], because the argument for monism fails, and because of the questionable motives that may at least seem to lie behind it. 'Non-reductive' is to some extent better than 'anomalous' in part because it captures what undoubtedly *is* an important kind of motive in Davidson's philosophy, namely the opposition to a certain traditional way of doing metaphysics. 'Anomalous' merely signifies a special form of irreducibility claim, and such claims are not very important, according to Rorty. (In addition, the argument for that special form of irreducibility may be faulty.) But I do think that reducibility and irreducibility claims are more interesting than Rorty allows (although I may agree that the importance of the specific irreducibility claim that *anomalism* is has been somewhat blown out of proportion), and I do think that there are important *anti*-reductionist elements in Davidson that should not be neglected.

To be an anti-reductionist without holding the corresponding irreducibility thesis seems to be dangerously close to irrational, because it seems to involve a form of fear of or resistance to something we can do nothing about. Resistance could only be rational if it is somehow *up to us* whether we are going to reduce something to something else. For most kinds of reduction this will certainly not be the case. In the overwhelming majority of cases too, reductions are not something to be feared or resisted, and anti-reductionism of that sort is consequently wrong-headed. Still, I will be proposing a form of anti-reductionism in the following pages. There are certain imaginable forms of reduction that would destroy the things they attempt to reduce, and which are therefore impossible in that they destroy what they attempt to account for and which should be resisted because those things are important. Those kinds of reductions have perhaps never been supported or claimed to be possible by any

[1] 'Monism' of course sounds more neutral than 'physicalism', but in Davidson's philosophy it means a specific form of physicalism, while 'physicalism' itself is very vague.

philosopher, and the importance of the corresponding irreducibility claim could well be doubted. I hope, however, that this very irreducibility claim can be used to throw some light on the nature of interpretation and communication and the nature of our mental concepts.

Two pragmatists on Davidson's mental-physical distinction

What is the mental-physical distinction according to Davidson? Among the things he emphasises is that rationality is "constitutive of the range of applications of such concepts as those of belief, desire, intention and action", whereas physical concepts have different constitutive elements. The constraints of rationality "cannot be stated in a purely physical vocabulary", they "have no echo in physical theory" (Davidson 1974b; 237, 239) (Davidson 1973; 259) (Davidson 1974b; 231). Although all concepts have their "norms", the norms of the mental are crucially different from those of the physical.

In "Davidson's Mental-Physical Distinction" (Rorty 1999)Richard Rorty argues that these differences are not fundamental. These differences of "constitutive principles" do not show that there are not many other distinctions to be drawn that are equally important (or non-important), hence there is no reason to call some principles constitutive rather than others. Pragmatism urges a principled openness to drawing distinctions in ever so many different ways we may find fruitful, and is opposed to exalting any features as excessively special, as this may segment old patterns of thinking. Fellow pragmatist Bjørn Ramberg, however, tries to convince Rorty that the mental-physical distinction is important (Ramberg 1999, Ramberg 2000). The interests that according to him lie *behind* the agency vocabulary, the interests which *have shaped* that vocabulary, are important. There are purposes that wouldn't be well served without this vocabulary, so we should do our bit to entrench it. Now, to the extent that certain forms of reduction might conflict with the pursuit of these interests, we could here have the basis for a certain form of anti-reductionism. Ramberg, however, does not give any concrete examples how our agency-vocabulary interests might be threatened, be it by reductions or other changes or discoveries. Ramberg says that for Rorty the right response to scientism is to be *a*-reductivist instead of being *anti*-reductivist, and endorses this himself. I think it is hard, though, to see how

this *a*-reductivism can be combined with the view that there is something *oppressive* (or potentially oppressive) about scientism. But this is what Ramberg says about scientism, and says that we should do our bit to entrench the agency vocabulary since those interests are especially important to us. On the other hand, it appears quite possible that a reductionist adherent to scientism would claim that a complete success of his program is fully compatible with pursuit of the interests behind the agency vocabulary, or that they would perhaps be even better served.

Ramberg does not see any important things turning on questions of ontology or reducibility, at least not where reducibility is understood as Davidson gets the notion from Ernest Nagel, in terms of bridge laws or strict correlations between predicates or properties. I will not comment on ontology here. About reducibility, Ramberg seems to be saying that Nagel-reductions would not in themselves threaten the agency vocabulary and the interests behind it. I think, in fact, that this is not very far from the truth. But this notion of reduction is not the only notion of reduction. Does Ramberg think that any reduction and any form of "reductive success" would be compatible with the continued pursuit of the interests which define the agency vocabulary? The trouble with allowing this, it seems to me, is that scientism is closely associated with reductionism, and that it is hard to identify scientism with anything but a belief in and a wish for a certain kind of reductive success. Merely highlighting the importance of our interests in rationality (and the rest) does not clinch the case against scientism. Even if it is clear that scientism is not in a particularly good position to do this highlighting, people could well ask why any legitimate concern could not be maintained within a scientific vocabulary, or when reductions are sought and found. We should want, therefore, if scientism renders us subject to certain forms of oppression, to know more about what this oppression consists in. We might then also need to know more precisely the concerns that are behind the agency vocabulary.

Do we need to entrench a particular vocabulary in order to retain the interests that are behind that particular vocabulary? A reductionist might disagree. If a vocabulary is shaped by the interest in a special set of features, patterns or aspects, must we stick to just that vocabulary in order to describe that set of features, patterns or aspects of reality? It is possible to disagree with this, and so we need an explanation of what it means to think there may be something oppressive about scientism and what the talk about entrenchment can come to.

This is how an adherent of scientism might reason. Aren't differences of interests behind different descriptions trivial? An art-restorer has different interests than an art historian in the description of a painting. (But their interests are overlapping; both would be interested in the type of canvas and paint that was used.) These differences of interests are perfectly compatible with both descriptions being incorporated into a larger encompassing picture. Both descriptions pick out features of the painting that is really there. Can an interest do more than cause us to highlight a certain set of features, and a different interest cause us to highlight different ones? If we just step back from our interests (or retain only the interest to give the full description) would we not ideally arrive at an interest-neutral full description? On this view of interests, what an interest does only seems to be to make us *disregard* certain features of the thing described; those features we are not interested in.

Interests are related to the ways we classify things and proceedings. Different terminologies (predicates) may cross-classify a subject-matter, and our interests determine which predicates we use. So, one set of interests can cause us to classify some subject-matter differently than another interest. If I am interested in solving polynomials, the distinction between algebraic and transcendental numbers is an important one[2]. The categories of algebraic and transcendental numbers partition the domain of real numbers. The categories of rational and irrational numbers partition the same domain differently (2 is both algebraic and rational, $\sqrt{2}$ is algebraic and irrational[3]). The fact that these differences of classification and the interests behind the different classifications are important does not mean that there is not a basic vocabulary in which all of these classes can be defined. (In fact our mathematical example illustrates just such a case.)

Let us compare the situation with Dennett's claim that there are patterns that are real but visible only from the intentional stance (Dennett 1987, Dennett 1991). The kinds of differences of interest we have discussed so far can not explain remarks such as this. They can only support the claim that the visibility of a pattern may be

[2] A polynomial is a function of the form $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$, where x is the variable and $a_0, a_1, \ldots, a_n$ are numbers.
An algebraic number is a number y which is a root of a polynomial in which $a_0, a_1, \ldots, a_n$ are integers (positive and negative whole numbers, i.e. … -2, -1, 0, 1, 2, …). (y is a root of a polynomial P(x) if y is a solution to the equation P(x) =0.)
A transcendental number is a real number which is not an algebraic number.

[3] 2 is a solution to the equation $x^2 - 4 = 0$, $\sqrt{2}$ is a solution to the equation $x^2 - 2 = 0$, so both are algebraic numbers.

lost when we adopt certain stances, or certain interests. They do not support the claim that there are patterns that are visible *only* from a certain stance or only when we approach a subject-matter with a certain interest in mind. Again, all that is supported is that interests (or stances) may cause us to disregard features. Of course, since we humans are limited beings, this in itself buys us some increase in perceptual abilities, in the weak sense in that it allows us to disregard complexities that would otherwise blur our view. But it does nothing to show those philosophers wrong who think that there is a point of view which is objective, impersonal, naturalistic, or physicalistic, from which everything and every pattern is in principle visible.

One might perhaps insist, against this, that the idea of an interest-neutral, all-encompassing point of view is a fantasy; and that at least science, or any particular science, does not provide such a point of view. And Ramberg does argue that for Davidson, physics and the natural sciences do have their own constitutive interests (for physics an interest in generality and exceptionlessness). So they are not interest-neutral, and therefore perhaps the intentional patterns will elude them? But why should this be so? The scientistic philosopher could at this point invoke the completeness of physics and deny that there are any patterns that elude basic science. He could even draw on the support of Davidson: "Explanation in terms of the ultimate physics, though it answers to various interests, is not interest relative: it treats everything without exception as a casue of an event if it lies within physical reach (falls within the light cone leading to the effect)" (Davidson 1987; 113). The reductionist could even grant for the sake of argument the opposite of what Davidson says about ultimate physics, namely that the natural sciences have special interests and are in that sense interest relative. It doesn't follow that this interest *must* cause blindness to the intentional patterns. (An art-restorer doesn't *have* to be blind to facts of art history that are irrelevant to his conservation interests. Often he wouldn't be.)

A hope characteristic of the Unity of Science movement was to have universal scientific language which was adequate to the identification of every causal factor and every pattern. Given such a language predicates could be defined in order to pick out any kind of phenomenon that we might be interested in. What we pick out as the cause in any given explanatory context, for instance, is governed by all sorts of pragmatic and context-dependent considerations. These explanatory interests bode for applying predicates that are designed to single out those particular circumstances that interest us. On the Unity of Science view the only thing that could be special about the

mental vocabulary is that it singles out some features of reality at the expense of other features. Sometimes the view that this is what is special about the mental vocabulary seems to be expressed by Davidson too.

> Mental concepts … appeal to causality because they are designed, like the concept of causality itself, to single out from the totality of circumstances which conspire to cause a given event just those factors that satisfy some particular explanatory interest. (Davidson 1991; 216)

The view that these concepts are designed to track patterns of rationality is central in Ramberg's interpretation of Davidson. But since all special science vocabularies answer to specific explanatory interests and are selective in the features of reality they pick out, there would seem to be nothing special about the mental-physical distinction as compared with other such distinctions. Could the point that Ramberg is urging against scientism be nothing other than that the mental vocabulary is not eliminable, given the interests that we have, and that those interests are especially important (more important than the interest to track geological patterns, say)? The point of scientism, however, would seem to be that all these explanatory interests could be pursued within a scientistic vocabulary (the universal language of science, perhaps). Of course Ramberg might point out that the dream of such a universal language is wholly unsupported, and that scientism, although it doesn't intend to eliminate the pursuit of any interests, would *in effect* lead to such elimination. If scientism is defined as a movement that would prefer to eliminate existing vocabulary (when the existing terms could not be defined in terms of the privileged vocabulary) rather than compromising the purity of a scientific vocabulary, then we could understand why it would be called "oppressive".

According to this line of thinking, then, eliminations are oppressive, but reductions are not. In this way it would be possible to be opposed to scientism and to be a-reductivist rather than anti-reductivist at the same time. One problem with this is that there may not always be a clear line between reduction and elimination. It is only the clearly *conservative* reductions that have no element of elimination. Some reductions may involve eliminations, and it is quite natural to call eliminations reductions in themselves, and this would seem to introduce the possibility of a form of scientism that is oppressive in favouring forms of reduction. My point here may

ultimately only derive from terminological differences, but the point is that I find it natural to describe opposition to scientism as anti-reductionism. Ramberg's talk about the need to entrench certain vocabularies and interests would also seem to deserve to be called anti-reductionist with respect to the question of the reduction of certain interests to others.

If the above explanation of what is oppressive about scientism is correct then it is also clear that the claim about the unavailability of a universal scientific language, something which is a kind of irreducibility claim, takes on a rather large significance. Ramberg's claim that "Scientism is not bad, I am sure Rorty would agree, because it gets the world wrong … but because it renders us subject to certain forms of oppression" (Ramberg 2000; 367), would then have to receive a certain qualification. If scientism *didn't* get the world wrong – if reductive definitions of mental terms in the scientific language had been possible, for instance – then it seems there wouldn't be anything oppressive about it. If this is the case, then irreducibility claims take on a greater philosophical significance than what at least Richard Rorty wants to acknowledge. It is precisely the truth of reducibility and irreducibility claims that determine whether a proposed reduction would be a conservative reduction or an eliminative one[4].

But even if we could identify something like an anti-reductionist motive in Ramberg, a-reductivism or non-reductionism is undoubtedly the position he finds it important to officially endorse. (Remember that because of the many different senses of reduction one can be a reductionist, non-reductionist and anti-reductionist at the same time.) If this non-reductionism consists in the view that among everything there is (i.e. everything physical in a broad sense) we can draw countless many distinctions according to all sorts of pragmatic interests, then so far this is something scientism could wholeheartedly endorse. This is so even if one stresses that within this domain there is a distinction between the mental and all the rest (the physical in a more narrow sense), and that that distinction is especially important.

In the end the bottom line for Ramberg against scientism seems to be to emphasise the importance of the interests behind the agency or person vocabulary,

[4] We might be interested in knowing why such an irreducibility claim holds. The explanation of what characterises mental concepts given above, namely that they are causal concepts designed to single out a special set of causal conditions from others, does nothing to explain this claim. In fact if that explanation were adequate the irreducibility on the contrary appears rather mysterious, since mental concepts then appear to be exactly the kind of concepts that should be definable in a scientific language.

and that to entrench the distinctness of the agency vocabulary is a good way to entrench the importance of these considerations. Scientism seems to be identified with a commitment to other interests; those of prediction and control, perhaps. Ramberg's position depends on the idea (which is natural in itself) that there could be conflict between different interests. But if the difference between those interests merely concerns which causal conditions we pick out among the totality of such conditions, then there is no basis for the idea of such a conflict or opposition. All the different cross-classifications could live happily side by side. If interests could conflict, or if interests could be threatened, we need to know more about how this is possible. Independently of this we should also like to know more about "vocabulary defining" interests if they are what defines what is special about the mental vocabulary.

We move a little closer to this when we ask what kind of interests we are talking about when we talk about the interests "a vocabulary is built to serve". That language has evolved in response to problems should be enough to answer the thought that interests may not matter much to the shape of vocabularies at all. Of course, we need not assume that a vocabulary is built to serve one unique interest, or one unique leading interest. Nor need we assume that it is built to serve some finitely specifiable list of interests, and we don't even need to assume that the interests that have shaped a vocabulary form a fully consistent set. If interests are what individuate vocabularies, there will thus be no sharp line of individuating them. We may sometimes speak of the vocabulary of biochemistry, sometimes of the vocabulary of natural science, sometimes of the vocabulary of public administration, and so on.

All this means that it is not entirely determinate what 'the vocabulary of agency' refers to[5]. This does not mean that we cannot usefully talk about it and single out some of the important interests behind it. So it should be possible to say for instance that a certain set of interests is not sufficient to mark the point of that vocabulary.

> … the vocabulary of agency marks a distinctive subject because it is built to serve interests that we pick out not simply by reference to the kinds of things we want to predict (people, as opposed to electrons or super-novas), but also by virtue of

[5] We probably don't need to map out which are the main- and subcategories, which categories overlap and how, of the following; the mental vocabulary, the personal vocabulary, the vocabulary of agency, the vocabulary of the propositional attitudes, the vocabulary of emotions, etc.

> features that are not merely predictive. What links special sciences together, if anything does, into a single contrastive entity with respect to the vocabulary of agency, is not an alleged reducibility … Rather it is a certain homogeneity of interest; that it can be characterized in terms of a purely predictive aspect. (Ramberg 2000; 366)

There is a question whether all the natural (or special) sciences really are homogeneous in interest, and also whether that interest is best described as "purely predictive". I think Ramberg would be the last to claim that there is such homogeneity; at this point he is merely stating the interpretation of Davidson. I will not pursue the question whether the natural sciences could be "reduced" to the pursuit of the interest in "pure prediction". What emerges as a possibility is a claim that the vocabulary of agency is not reducible to only this interest. Since this may be true of the sciences also, this may not yet be what we need to be able to make the mental-physical distinction really interesting, but one could think that however rich the scientific interests in fact are, the vocabulary of agency may still not be reducible to these interests. Ramberg notes; "With agency-vocabulary … we are characterizing a domain of kinds of objects (language-users) with a vocabulary not *just* geared toward prediction of the behavior of that kind" (Ramberg 2000; 366).

This alone does not get us very much closer to a characterisation of what must be added in order to describe the point of the agency vocabulary. Ramberg gives some hints (perhaps, for his purposes, he doesn't need to give more than hints). Immediately after the above quote where he says that the interests of the agency vocabulary aren't just predictive, he continues: "Or perhaps we should rather say that the predictive interests … are of a very peculiar sort – they turn on our revealing the kinds of traits that allow us to *recognize ourselves* in what we are talking about, and to bring to bear all those complicated considerations that we gesture at with the moral notion of a person. I would argue that charities and rationality-constraints, presented as a necessary methodological assumption of the ideal interpreter, really embody the very *point* of agency-talk, precisely because they are inextricably connected with this notion" (Ramberg 2000; 366).

I agree that charity and rationality constraints, and perhaps particularly the moral notion of a person, are important to the vocabulary of agency (or the mental vocabulary, or the personal one). Perhaps this characterisation is enough to convince

us that this vocabulary is important (if we ever doubted that). But if we think that these interests may contain the source of a certain irreducibility claim, then we should perhaps need a sharper description of them.

I think we get an indication that Ramberg might see certain reductions as more problematic than his a-reductivism would signal when he applies a metaphor. Rorty and Ramberg see vocabularies on an analogy with tools. Just as tools serve purposes and interests, vocabularies may do the same, and just as using one tool to solve a problem may interfere with, preclude or facilitate the use of another, certain uses of language may do so as well. ("the wrench may get the screw, but now the cylinder will be stripped – better to use the screwdriver and the rust-remover, since tomorrow we just might want to be able to reinsert a screw there" (Ramberg 2000; 365f).) If the analogy is a good one then the use of a vocabulary must have causal consequences, consequences which can either facilitate or frustrate other aims we may have (aims that are meant to be served by a different vocabulary). I think this is correct. But even if the use of a vocabulary obviously must have causal consequences (people might be brought to change their opinions, for instance), it is not obvious that it could have the right kind of consequences for the points at hand, in particular that it could affect the extent to which interests behind other vocabularies could be pursued. More needs to be said both about the consequences of (some kind of) application of a vocabulary, and about the interests that define a vocabulary and how they could be affected. What we need is to cash out on the tools metaphor employed by Ramberg and show in some detail how certain uses of language could have the consequence that it would no longer be possible to apply the mental vocabulary with the kinds of purposes we do have, in other words to show how scientism, or some of the reductions it craves, would be "oppressive".

The claim that I am going to defend later in this chapter is that some reductions that are at least in some sense conceivable (for instance some forms of reductive definitions) would be choices about how our vocabularies should develop and hence choices that determine conceptual change. In this kind of situation certain reductive interests could be in conflict with others, in particular with the application of rationality assumptions as regulative ideals in interpretation.

Kim on psychophysical laws

If the irreducibility of mind or the anomalism of the mental is not to be a trivial consequence of a much more general phenomenon, then it must derive from something that is special about mind and mental properties, and in the context of Davidson's philosophy it is natural to assume that this must be their rational character. We will look at two interpretations that emphasise this link in different ways in this chapter. First I will look rather briefly into the interpretation given by Jaegwon Kim, and then later in a little more detail I will look at William Child's interpretation of Davidson's argument.

Recently, Steven Yalowitz (Yalowitz 1997) has been willing to give an interpretation according to which the argument for anomalism does not depend on anything that is peculiar to the mind. Anomalism on his reading follows fairly directly for any discipline that uses causal or dispositional concepts. Strict, exceptionless laws have no *ceteris paribus* or normality conditions built into them, but causes and dispositions are picked out against a background of normal conditions. All causal citations and explanations that are interest relative in the sense of picking out some elements rather than other causal conditions are anomalous in this sense. Yalowitz makes it quite clear that this anomalism applies broadly. Rationality plays no part at all in the argument for anomalism as he sees it. As for the normativity of mental descriptions, there is no special sense in which this enters either, except in the general sense in which this reveals the causal character of those concepts. "Whether something is an electron, and is thereby covered by the set of laws which refer to electrons, depends on whether it acts in the ways typical of electrons. An object cannot fail, in general, to act in these ways without ceasing to be an electron. Therefore, normative elements enter into physical explanation in similar ways that they do in psychological explanation; in neither case is there a sharp distinction between the two" (Yalowitz 1997; 247). Yalowitz admits that the concept of an electron could not figure in a strict law. At this point, though, when the irreducibility of mental terms is none other than the irreducibility of a term like 'electron', I think it begins to be hard to see what the interest in such a conception of anomalism consists

in[6]. Yalowitz allows that rationality may have a part to play in showing why the mental causal concepts are not *eliminable*, and says that Davidson may have confused the question of irreducibility (which affects all causal terms) and non-eliminability (which perhaps only concerns some of them). But since even 'electron' turns out irreducible on this conception, there hardly seem to be any scientific concepts that could be reducible. This is clearly not a very useful concept of reduction. On the other hand I think it is not unreasonable to expect of an irreducibility claim about the mind that we should be capable of learning something about the *mind* from it. For that reason I think we should turn to attempts to connect the irreducibility claim with something that is special for mentality, and the attempts to connect it with rationality do that.

Kim's reconstruction of the argument for anomalism turns on the idea that the mental and physical domain are governed by disparate constitutive and essential features, and that laws linking mental and physical properties would violate or disrupt the essential features of each. Kim sees Davidson's argument for the anomalism of the mental as proceeding from a lemma, which Kim calls psychophysical anomalism:

> *Psychophysical Anomalism:* There are no psychophysical laws, that is, laws connecting mental and physical phenomena. In fact there *cannot* be such laws. (Kim 1985; 196)

A striking thing about this formulation that we should note immediately is that Kim makes no explicit restriction to strict laws, as Davidson does. So Kim seems to think that the argument, if successful, would establish the non-existence of *ceteris paribus* laws as well, or perhaps he assumes that laws must *ipso facto* be strict for Davidson. The argument for psychophysical anomalism has a *reductio ad absurdum* structure. If we can assume bridge laws of the form

(1) Necessarily, a person believes p if and only if he is in state B_1.
(2) Necessarily, a person believes q if and only if he is in state B_2.

[6] In fairness to Yalowitz, he thinks the point of this consists in its role in the argument for monism, and that is certainly not an unreasonable interpretation. My focus, however, is whether irreducibility could teach us something about the nature of mind.

Then, given a psychological law of the form

(3) If S were to believe p, S would also believe q (where p logically, and obviously, implies q).

We would be able to conclude

(4) If S were in state B_1, he would also be in state B_2.

If this is true it will be a physical law or an instance of a physical law (B_1 and B_2 may be neurological brain states, for instance). Alternatively, given (4), we would be able to infer (3). Bridge laws like (1) and (2) would bring the mental and the physical into close relations that would enable us to move from one to the other.

The reason why this is an unacceptable situation is the different constitutive principles of the two domains, the principles to which each must pay allegiance unless they are to lose their autonomy and integrity. Bridge laws would allow the special features of the one to be "transmitted" into the other domain and thereby infringe upon the latter's integrity and autonomy.

Rationality is the constitutive principle of the mental. Any psychological principle should be based on normative considerations about what rationality requires. (3), Kim says, is grounded in the principle of rationality maximisation or optimisation. If we attribute a belief p to a person and q is an obvious logical consequence of p, then this rule says that we should attribute the belief that q to the person. (3) is thus a normative law that holds of necessity. Physical laws are descriptive and are supposed to be contingent. But if (4) can be derived from (3), we will have established a dependency between physical states on the basis of reflections of a conceptual kind. This will infringe on the contingent status of physical laws[7]. If (3) is what grounds (4), Kim says that this would be an intolerable intrusion on the closedness and

[7] Kim doesn't say that it is part of the constitutive features of physics that its laws are only contingent, which is understandable given Davidson's silence on the point. Kim only infers that the constitutive features includes the absence of rationality. But this way of rendering the argument seems close to what Kim intended, as thinks Steven Yalowitz (Yalowitz 1997). Davidson himself waited until (Davidson 1995) by giving his best considered views on physics.

comprehensiveness of physical theory. This would be absurd, and we must reject the 'only if' direction of the bridge laws (1) and (2).

If we were to derive (3) from (4) on the other hand, this would mean that the role of the rationality considerations constitutive for the mental had been pre-empted. Hence the 'if' part of the biconditionals must be rejected as well. The basic problem with such bridge principles is that they would allow us either to derive a *necessary* principle (a psychological law) from a merely *contingent* law (the physical law), or else they would allow us to derive a *contingent* law from a *necessary* principle. And both possibilities would be absurd.

I will not go into great detail about the merits of this argument[8]. As Yalowitz has noted, unless (3) and (4) are strict and exceptionless, the argument will not work. If (3) only expressed a *ceteris paribus* connection, for instance, there would already be an element of contingency in it, and the empirical contingency of the physical law (4) would thus not "infect" it.

Kim's reconstruction immediately lands him with an interpretive problem. The principle of psychophysical anomalism is meant to be a step towards establishing the anomalism of the mental. But psychophysical anomalism only establishes that there are no psychophysical laws from which mental phenomena can be predicted and explained. It doesn't establish that there are *no* laws from which mental phenomena can be predicted and explained. The argument has a lacuna that Kim thinks needs to be filled by the following principle:

> *Psychological Anomalism:* There are no purely psychological laws, that is, laws connecting psychological events with other psychological events, which can be used to explain and predict these events. (Kim 1985; 210)

There are two problems with this for Kim. The first is that he doesn't find in Davidson's writings any explicit argument in favour of this principle[9]. The second problem is that Kim's reconstruction of Davidson's argument for psychophysical anomalism requires the existence of necessary psychological principles; and what are

[8] See for instance (Yalowitz 1997) and (Tiffany 2001) for more detailed discussion, and rejection, of the argument.

[9] As Yalowitz notes, the argument that there are no purely psychological *strict* laws is trivial and explicit in "Mental Events". Because mental events interact with physical events, any truly exceptionless laws would have to incorporate the effects of physical causes (blows to the head and the like), so they would at least have to be psychophysical.

these principles if not psychological laws? Not only would they be psychological laws, they would as Yalowitz has shown have to be *exceptionless* laws as well. Kim's solution to these problems is that there is a distinction between descriptive and predictive/explanatory laws on the one hand and normative laws on the other.

Before considering further problems with Kim's interpretation we should note a nice feature of his reconstruction. This is that it explains why Davidson's position and argument does not rule out true generalisations. In "Mental Events" Davidson explicitly allows that there may be true (but probably immensely complicated) universal generalisations stating co-extensions between mental and physical predicates. Kim's reconstruction would explain why these correlations are not inconsistent with anomalism. Such brute correlations don't have the power to transmit the constitutive properties of one domain into the other, but laws are strong enough for that. For instance, laws would be strong enough to establish evidential channels between the mental and the physical domain, and this would mean that each domain's "allegiance to its proper source of evidence" could be compromised, but mere generalisations (*ex post facto*) do not constitute such evidential channels.

Importantly, however, Kim thinks that Davidson's argument is meant to exclude *all* psychophysical (or descriptive psychological) laws, and thus that if there are any laws that are not strict then they must be excluded as well. Kim reiterates his view that if Davidson's argument succeeds in banning strict psychophysical laws, then it will also banish non-strict laws as well, in a paper from 1993 (or that, at least, an explanation is required if it doesn't). "I have always thought that the power of the Davidsonian argument for mental anomalism is seen in the fact that, if it works at all, it should work against laws of all kinds – for example, statistical laws as well as deterministic ones … Remember: non-strict laws, whatever they are, are supposed to be laws!" (Kim 1993; 25) This, as I will discuss further in chapter 6, would create big problems for explanation in psychology and perhaps also for action explanation. I will argue in later sections of this chapter that Kim is wrong in thinking that Davidson's argument excludes non-strict laws. (But I think Kim is partly right in thinking the distinction between what the argument excludes and what it doesn't is not between strict and non-strict laws. I will instead suggest that reductions of the offending type could be based on non-strict laws, but also that it could be the case that reductions of the offending type do not automatically follow from strict laws.)

Kim's interpretation fits very well with the a priori theory conception of the mind. The sharp distinction between descriptive and normative laws requires a sharp distinction in the epistemology for such laws, and since descriptive laws must obviously be empirical, the source of the normative laws is presumably rational reflection or conceptual analysis. It would require an extremely strong form of conceptual analysis to yield strictly exceptionless psychological principles, though, and neither this conception of conceptual analysis nor the existence of such exceptionless principles have any plausibility at all. This also holds for the idea that the psychological principles are not descriptive. This would probably mean that they could not be sensitive to counterexamples, and that is simply false. The major problem for Kim's account considered as an interpretation of Davidson is that Kim ignores Davidson's insistence that he is only denying *strict* laws[10]. Kim's interpretation therefore cannot be fully correct.

Another problem with Kim's reconstruction of the argument for anomalism and its reliance on an insulated body of psychological laws is its relation to programs of functionalist reduction in the spirit of David Lewis. The very existence of a sharply defined body of psychological principles is a prerequisite for such reductions. Lewis argued that given such a theory, the possibility of explicit definition and reduction of its terms follow. I will not pursue this question here, but this connection might be worrisome for an interpretation of an irreducibility argument.

There is one element in Kim's reconstruction that I find useful and which is independent of his view of psychological principles as necessary, exceptionless and non-descriptive. This is when Kim discusses the difference in *conditions of attribution* for neural descriptions and mental descriptions. A neural state N has its conditions of attribution, Kim says; conditions under which their attribution to an organism is warranted. These conditions are probably very complex and difficult to articulate, but what matters "is only that the ascertaining of whether they hold in a given situation is regulated by the constitutive rules and principles of physical theory, not by those of the mental" (Kim 1985; 205). In the same way, a mental state M has its own conditions of attribution that answer to the constitutive principles of the mental. Now, if we assume that there is a psychophysical law that connects N with M, say; it is a

[10] In fairness to Kim we should note that in "Mental Events" Davidson only distinguishes between strict laws on the one hand and (strict or non-strict) generalisations on the other. It is only much later (and I think in no work that precedes Kim's paper) that Davidson begins to talk about non-strict *laws*.

law that if N obtains, then M obtains, then these conditions of attribution are in danger of loosing their purity. Such laws, Kim says, "would permit us to attribute intentional mental states independently of the rationality maximization rule; at least, they would force this rule to share its jurisdiction over mental attributions" (Kim 1985; 206). Also in this argument mere psychophysical correlations would not create the kind of tension Kim thinks he sees. Something with the modal force of laws is required in order for one kind of attribution condition to "infect" those of another domain.

I think that Kim is here on the track of something correct. But the argument cannot work as it is now stated. One thing that is wrong with it is that it seems to require that descriptions have very fixed (even if complicated and hard to articulate) conditions of attribution. That seems very doubtful. Why couldn't those conditions develop and grow as our knowledge grows, for instance? And why would it not be possible that the rationality maximisation rule should share its jurisdiction over mental attributions? Our discussion in chapter 3 led us to the conclusion that rationality considerations only play *a* part in the attribution of mental states, not that they hold exhaustive sway over such attributions. The kind of interference that psychophysical laws would have over attribution conditions is not clear. Another question that I will merely note and put to the side for the moment arises if we assume that there is a way in which 'attribution conditions' of mental descriptions could be infected or disturbed, and it is whether empirical laws in themselves are strong enough to infect or disturb them in this manner.

I now turn to another interpretation of Davidson's anti-reductionism, one which in many ways builds on and goes deeper into the details of the kinds of considerations discussed in the last couple of paragraphs.

The uncodifiability of interpretation

In his book from 1994, *Causality, Interpretation and the Mind*, William Child gives a reconstruction of Davidson's argument. At centre stage is the claim that the norms of rationality, and also the norms of interpretation, are not codifiable. This formulation gives, I think, a much better focus for discussion of Davidson's anti-reductionism. It shows the intimate connection with interpretationism about the mind and with views about the nature of interpretation.

Some of the themes from Davidson that Child builds his interpretation or reconstruction around are the following: Rationality is "constitutive of the range of applications of such concepts as belief, desire, intention and action". But, "Physical concepts have different constitutive elements". So, "Standing ready, as we must, to adjust psychological terms to one set of standards and physical terms to another, we know that we cannot insist on a sharp and law-like connection between them" (Davidson 1974b; 237, 239). Davidson says that the constraints of rationality cannot "be stated in a purely physical vocabulary" (Davidson 1973; 259). They have "no echo in physical theory" (Davidson 1974b; 231), (cited in (Child 1994; 56f)). What supports the idea that the norms of rationality have no echo in physical theory, according to Child, is the thesis that rationality is uncodifiable.

Child doesn't define the notion of codification, but he explains it with a reference to a discussion by John McDowell on practical rationality. A codification of practical rationality would be a formulation of principles from which, given a description of any set of particular circumstances, we can deduce a specification of what to do in that situation. But this is not something that is possible, McDowell argues.

> If one attempted to reduce one's conception of what virtue requires to a set of rules, then, however subtle and thoughtful one was in drawing up the code, cases would inevitably turn up in which a mechanical application of the rules would strike one as wrong – and not necessarily because one had changed one's mind; rather, one's mind on the matter was not susceptible of capture in any universal formula. (McDowell 1979; 336)

I will try to give a sharper definition of codification and return to whether we should accept this below.

I think that the connection between Davidson's anti-reductionism and interpretationism about the mind is often not given enough emphasis by commentators, and is sometimes entirely missed. The interpretationism that I think is relevant to this question is the fairly modest methodological interpretationism which has the following sort of slogan:

> Interpretation has an ineliminable role in understanding people and is irreducible to other methods.

Of course, Davidson wants to argue for a much stronger and general thesis, namely that interpretation is necessary for propositional thought in general. But that is not our topic here. The uncodifiability claim relates to this methodological interpretationism. Codification of interpretation would, as I see it, constitute a reduction of interpretation to other methods.

Davidson does, at least in some places, connect his anti-reductionism with the method of interpretation. This is how he concludes "Belief and the Basis of Meaning":

> It is … the methods we must invoke in constructing theories of belief and meaning that ensures the irreducibility of the concepts essential to those theories. Each interpretation and attribution of attitude is a move within a holistic theory, a theory necessarily governed by concern for consistency and general coherence with the truth, and it is this that sets these theories forever apart from those that describe mindless objects, or describe objects as mindless. (Davidson 1974a; 154)

Here is perhaps the most relevant passage from Davidson where he states that strand of argument that I will be struggling to reconstruct. This is the last paragraph of the section aiming to establish anomalism in "Mental Events":

> The point is … that when we use the concepts of belief, desire, and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory. An arbitrary choice of

> translation scheme would preclude such opportunistic tempering of theory; put differently, a right arbitrary choice of a translation manual would be a manual acceptable in the light of all possible evidence, and this is a choice we cannot make. We must conclude, I think, that nomological slack between the mental and the physical is essential as long as we conceive of man as a rational animal. (Davidson 1970; 223)

As with almost any argument from Davidson this is very dense. In particular, I think Davidson moves too quickly in drawing the conclusion. I will argue that a methodological kind of anti-reductionism, as well as definitional and conceptual anti-reductionism appears much more plausible, but that nomological anti-reductionism is a much taller order.

There are many different notions of reduction, and for philosophers of mind the question of reduction has perhaps primarily come to mean inter-theoretic, nomological, definitional or ontological reduction, and the key notions involved have been those of theories, laws, properties and concepts. The scientific paradigms that have been upheld are either those of theoretical reductions (reduction of thermodynamics to statistical mechanics) or ontological (or identity) reductions (water and H_2O). But there is another paradigmatic form of reduction in the sciences; reduction of methods, or reduction of problems. One example is the reduction of relativity theory to Newtonian physics. When philosophers have had inter-theoretic reduction in mind, they have insisted that the reduction must be in the other direction (i.e. that Newtonian physics, because it is approximately true only in special cases, must be reduced to relativity theory which holds in all cases), but in science it is more common to see the reduction as being partial and as occurring in the other direction. This is because in situations when we are not dealing with extreme masses or velocities, the difficult problems of calculating within relativity theory reduces to the much simpler problem of calculating within Newtonian mechanics. We have the same way of speaking in mathematics. In favourable situations the problem of calculating the volume under a graph reduces to iteration of integration.

This second form of reduction is important because the irreducibility of interpretation is in the first place related to it. A basic irreducibility claim will be that the method of ascribing mental states (the method of interpretation) does not reduce to non-interpretational methods.

At some places Davidson express himself a little more cautiously when it comes to nomological reduction. In the reply to Rorty in the Library of Living Philosophers volume, for instance, he says only that there are not "empirical laws linking [mental concepts] with physical phenomena in such a way as to make them disposable" (Davidson 1999; 599). His primary concern in this place therefore appears to be conceptual anti-reductionism, and his reason for opposing nomological reductionism is that he fears that that might have conceptual reductionist consequences as well.

The argument from the uncodifiability of rationality

Child says that if we accept that rationality is uncodifiable, two things follow. "First, that there is no set of general principles from which, together with a specification of any agent's physical properties, we can derive a complete and detailed specification of her mental properties. … Second … that there can be no system of strict laws on the basis of which actions and other mental phenomena could be exactly predicted and explained" (Child 1994; 60).

The way Child sets out his argument the first of these conclusions appears as a lemma from which the second follows as the main conclusion. Let us have a look at the structure of his argument.

The argument for the lemma is short and sweet.

> If rationality is uncodifiable, there is no system of principles from which we could derive, given a specification of an agent's physical properties, a statement of what it would be rational for her to believe, desire, and do. If there are no such principles for deriving a statement of *what would be rational for S*, there are no such principles for deriving a statement of *which attributions of mental properties make best sense of S*; and if there are no principles for deriving that, there are no principles for deductively deriving a specification of *S's mental properties*. (Child 1994; 60)

The major conclusion is supposed to follow rather easily from the lemma. Because mental states sometimes have non-mental causes, any system for exactly predicting or

explaining mental phenomena must include (or be) psychophysical principles or laws. But this is just what is excluded by the lemma.

For convenience's sake, here is the structure of the argument:

The argument from the uncodifiability of rationality

(1) Rationality is uncodifiable → There is no system from which we can derive, from a specification of an agent's physical properties what it would be rational for her to believe, desire, and do.

(2) There is no such system[11] → There is no system for deriving a statement of which attributions of mental properties make best sense of the agent.

(3) There is no such system[12] → There is no system for deducing the mental properties of an agent.

Lemma/ First Conclusion: There is no set of general principles from which, together with a specification of any agent's physical properties, we can derive a complete and detailed specification of her mental properties.

(4) Because mental states sometimes have non-mental causes, any system for exactly predicting or explaining (any) mental phenomenon must include (or be) psychophysical principles or laws.

(5) From the lemma it follows that there cannot be laws or principles from which a detailed and complete specification of the mental properties of an agent can be derived, and so there are no laws for exactly predicting or explaining mental states.

Major/ Second Conclusion: There is no system of strict laws on the basis of which actions and other mental phenomena could be exactly predicted and explained.

[11] There is no such system as mentioned in the consequent of (1).
[12] There is no such system as mentioned in the consequent of (2).

I think we can predict that someone who rejects interpretationism will not think of this as a valid argument. One way to see this is to focus on the phrase "making best sense" which might be held to be ambiguous. In one sense one could say that the truth always is the ascription that "makes best sense" – despite how surprising it might be. But if this is how we take it then the phrase does no work in the argument, and could be dropped.

In another sense what "makes best sense" is a subjective notion. It could be taken simply to mean what merely appears to someone most reasonable or rational. And, someone might argue, even if there is no system for deducing what ascription would *appear as most rational* to a group of people, or for deducing what ascription that would *strike them as making best sense*, nothing about what ascription of mental properties that is true follows from this. There are also other possible ambiguities in the argument.

But whether or not there is a reading of this which makes it a valid argument, I propose to pass that question by. There are at least two readings of both premise (1) – the uncodifiability of rationality – and of the lemma or First conclusion. I'll call them the normative and the extensional reading. I happen to think that the claims are most plausible on the normative reading, but I am less sure about the extensional reading. I also think that there might be arguments for the two normative theses, premise (1) and the lemma, that are largely parallel, and therefore that if you want to establish the lemma on the normative reading it is not clear that we need the argument above. Finally, I think that if one should hope to establish the extensional thesis, the interesting action will be in an inference from the normative reading of the lemma to the extensional.

The uncodifiability theses

So, let us consider whether the theses are acceptable. One who is not impressed is Frank Jackson.

> … it is sometimes said that rationality is *uncodifiable*. What is certainly true is that we cannot, as of now, write down in a natural language necessary and sufficient conditions for being rational. (Though we can say something useful and to the point – whatever the defects of the inductive logic sections of textbooks and extant discussions of experimental design, they are very far from useless.) What would be incredible, in my view, would be if there were no story to be told constructible from our folk-classificatory practice: we are finite beings; we do not work by magic; we give useful information to each other by means of the word 'rational'. There must, therefore, be a story to be told (extracted). And when it is told (extracted), rationality will have been codified. (Jackson 1998; 67)

I think one can certainly feel that there is something, to put it politely, heroic about Jackson's position here. A very natural thought, it seems, is that rationality or interpretation couldn't be codified simply because of the enormous complexity of such an undertaking.

But we need to be careful about our reasons for holding that claim. If we think complexity is the reason, then the reason is not very principled, and there seems to be nothing that separates interpretation from many other subjects. Compare the task of codifying the norms of when it is correct to describe something as a fish. This too may be just too complex a task to be possible for human beings.

If we return to our two theses in the first part of the argument; premise (1) and the first conclusion, we can see that they have at least two different readings. They can be read with two different kinds of emphases. On the one hand, these claims refer to systems of general principles, or systems of deduction, or systems of strict laws, and a natural point at which to lay the emphasis is on the claim that there are no systems or principles of this kind. But on the other hand, these statements make reference to us, in the claims that *we* cannot derive, or *we* cannot deduce, or *we* cannot predict or explain this and that, and that is another possible point at which to lay the emphasis.

Let us see if we can give formulations that capture these different readings and emphases.

Rationality; Extensional version:

(RE) There does not exist a system from which what it would be rational for an agent to believe, desire, and do, deductively follows from a specification of her physical properties.

Rationality; Normative version:

(RN) We should not (mechanically) use any system from which claims about what it would be rational for an agent to believe, desire, and do, deductively follow from a specification of her physical properties, to derive what it would be rational for her to believe, desire, and do.

Mental properties; Extensional version:

(ME) There does not exist a set of general principles from which, together with a specification of any agent's physical properties, a complete and detailed specification of her mental properties deductively follows.

Mental properties; Normative version:

(MN) We should not (mechanically) use any set of general principles or a deductive system to deduce a (complete and detailed) specification of an agent's mental properties (from a specification of her physical properties).

These formulations suggest ways in which we can make the notion of codification sharper. On one way of doing this, a definition of codification or codifiability only has to make use of extensional vocabulary. On the other way of doing it, the definition uses deontic terms. In some cases these definitions may overlap, but in other cases the distinction will be important. In either case a codification will involve the formulation of explicit rules or statements such as 'when so-and-so obtains, P is to be done' or 'when so-and-so obtains, P is true'. Apart from the explicitness requirement, there will be requirements against circularity and triviality. We should not want to allow rules of the form 'whenever you are in a situation in which A is the thing to do, then you should do A' as providing a codification of virtue. I do not know how to formulate this requirement, but roughly speaking we should want the antecedent to be independent and simpler or conceptually more primitive than the consequent. It

should be possible to verify that a situation is such that the antecedent obtains without understanding the meaning of the consequent. I will just call this requirement the *independence* requirement, and assume that it is intuitively clear.

On the extensional definition, a certain set of descriptions or predicates are codified if conditions are stated (in a different vocabulary) such that if they hold, then the codified descriptions hold too. Any correct non-circular definition will thus be an example of a term that has been extensionally codified. A more interesting example would be the Ramseification of a theory (see (Lewis 1970)). We can think of the "new" theory (the theory that is Ramseified) as having been codified when the Ramseification provides conditions formulated only in "old" terms that are sufficient for the application of a "theoretical" or "new" term.

Let us attempt an explicit definition:

> A set of descriptions or predicates P is **extensionally codified** if there is another set of *independent* descriptions or predicates Q, and there is a set of true explicit statements R, such that for any situation s and any condition p in P, there is a conditional statement r in R with an antecedent a that gives a description of s in the terms from Q, and r is either $(a \rightarrow p)$ or $(a \rightarrow \neg p)$.

The intuitive idea behind extensional codification of a set of descriptions is that for any description in that set, one only needs to look to the descriptions in another independent set in order to check whether it obtains or not. This definition requires the set of statements R to be explicit, that is, that we actually possess them. It is therefore quite difficult to satisfy. But a set of descriptions or predicates are **extensionally codifiable** if such true statements *exist* even if they haven't yet been found.

The other conception of codification, which I think is more fundamental, concerns norms and rules. It concerns the formulation of a complete set of explicit rules that are to be followed without exception of any kind. The only understanding required for applying the rules must be that one understands that a situation occurs in which a specific rule *is* to be applied. After that, following the rule can be done mechanically and quite without thinking. An algorithm is the paradigmatic example of codification. Codification in this sense will be defined over a set of practical questions. A practical question can concern all questions about what we should do in

all sorts of circumstances. It can also concern whether we *should* call something by a certain name; hence there is a connection with interpretation.

> A set of practical questions P has been **codified** if there is a set of *independent* descriptions or predicates Q, and a consistent set of explicit rules R *has been laid down*, such that for any situation s and any question p in P, there is a rule r in R with an antecedent a that gives a description of s in the terms from Q, and r is either *if a do p*, or *if a do not do p*.

Since codification in this sense does not require truth but only consistency, it is in a sense much easier to achieve. It only requires that someone has laid down a set of explicit enough rules. The rules might even be stupid. We could of course define codification such that the rules had to state what we in fact ought to do, that they had to be *good* rules. The reason why I don't want to define it thus is that I want to be able to claim later that there are codifications that we should resist. In particular that we should resist the codification of interpretation. This means that in a certain sense, every practical question will be codifiable. That only requires that someone draws up a stupid rule for it. But in another sense I will say that a set of practical questions are not codifiable if they could not be codified without formulating rules that we ought not to accept. And in this sense I will claim that interpretation (and practical reason) is not codifiable.

On the basis of these definitions we could give other formulations of principles that are related to (RE), (ME), (RN) and (MN).

> (EUR) Rationality (and cognate predicates) is extensionally uncodifiable.
> (EUM) Mental properties are extensionally uncodifiable.
> (UR) (The uncodifiability of rationality) What is rational is not codifiable.
> (UI) (The uncodifiability of interpretation) Interpretation is not codifiable.

One difference with the earlier principles is that there is no special mention of *physical* properties here. This buys us some greater generality.

For Child it seems to be important that any proposed codification of interpretation would have to be extensionally false, in other words that (ME) and (EUM) are true.

Perhaps this is correct and perhaps this is plausible. But I would argue that we would do well to start with normative uncodifiability; the claim about how interpretation should not proceed.

But why should we accept this?

Someone who objects to (UI) might grant that the way interpretation actually does work, there are no algorithms for interpreting people, and that in the present circumstances at least we shouldn't accept any algorithms. But, this objector might say, the uncodifiability of interpretation is only a present limitation, it derives only from our limited knowledge. It is because we know so few conclusive clues for arriving at mental descriptions of people that we have to treat the clues that we have with such caution and hermeneutic prudence.

So, (UI) must presuppose that we cannot come to know the relevant correlations or bridge laws, because, the objector might say, if we did know such correlations or laws, then it would be all right to use a deductive system that incorporated them in arriving at mental descriptions of people. The normative claim, therefore, cannot be used in an argument that aims to establish that there cannot be, or that we could not know strict psychophysical laws.

I admit I am uncertain whether (UI) could be used to establish (EUM). But I think it is wrong that an argument for (UI) must presuppose that we don't know any strict psychophysical correlations, or in other words that it must presuppose (EUM). An interpreter considers what, in the light of available evidence, is the interpretation that makes best sense of the subject. The objector we are considering, however, seems to have to think that it is possible in principle, at least, that we should come to know something that would make this feature of interpretation eliminable, or that interpretation is eliminable when it comes to understanding people (if this feature is taken as essential in interpretation). This then, raises the question whether we could come to know something that would justify us in codifying interpretation. That question has at least two aspects. One question is whether we could *come to know* something like that. Another question is if we were to know something like that, whether that would be *sufficient* for codifying interpretation. We shall return to those questions below.

Consider (UR), the uncodifiability of rationality, and let us have a look at the quote by McDowell again. Why would a mechanical application of "rules of conduct"

inevitably strike us as wrong? Here is one kind of reply: Someone who acted that way would more resemble a machine than a human being. Suppose someone said: "I have my principles. I have my rules. I think they are well justified, and I am going to abide by them no matter what." But this is just a declaration not to use his moral judgment in the future, and if the man would go on to put this into practice we would say he had put his moral judgment out of work. There are duties. But one duty is to apply our moral judgment and sensitivity with fresh eyes to concrete cases, and not to be *blinded* by what duty (our explicit rules) says. Even if this man had justified belief in his principles, and those principles were in addition true, such a procedure would be wrong of him to follow. Applied to the case of interpretation, this reasoning gives the following thought: Somebody who merely used an algorithm to interpret another would in so doing demonstrate a dehumanising attitude towards the other.

On this way of looking at things, uncodifiability rests on a moral imperative; not to allow one's practical deliberation or one's interpretation of others to be codified.

> But his kind of reply is not enough. If this was all, it might merely point to a tragic condition of human life. We might possess a method that we were convinced was the best in terms of knowing someone's mind, but this method might be banned, at least in its pure form, for moral reasons. If we, after we had applied the method that we were convinced was the best method, for moral reasons had to perform some ceremony of "looking with fresh eyes", the situation would only be tragicomic.

> So, unless we have an argument why we could not have this kind of confidence – this kind of knowledge, this reply is seems insufficient. We need a way of turning the moral imperative into an epistemic one, one that says that we would be at an epistemic disadvantage if we allowed interpretation to be codified. But perhaps, however, this reply suggests another possibility for arguing for this, namely that the kind of rough and ready justification we might have for psychophysical generalisations or correlations, the kind of knowledge we could have of them, would not – or should not – be sufficient for codifying interpretation and practical deliberation, that is, it would not be sufficient for applying rules based on them *mechanically*.

So, what are the central questions we have?

- There is the question about the existence of psychophysical correlations or laws – or, more generally (i.e. without special concern for physical properties), the question about the extensional uncodifiability of mental properties (EUM).
- Then there is the question whether we could know that there are such correlations.
- And there is the question whether such knowledge would or should affect a codification of interpretation (or another form of conceptual reduction).

Davidson says something that could be interpreted as a claim that there may be correlations, but that knowledge of them is unattainable by us. He says: "If by absurdly remote chance we were to stumble on a nonstochastic true psychophysical generalisation, we would have no reason to believe it more than roughly true" (Davidson 1970; 216).

Section II of "Mental Events" opens by noting that there "could" be coextensive mental and physical predicates. In fact this is implied by the full position of anomalous monism, as emphasised by for instance Mark Johnston (Johnston 1985). So Davidson's position is that there could be (or will be) true general statements, but he says (in "Mental Events") that they will not be lawlike. And he also says that ruling a true or false general statement lawlike is a priori. The suggestion here seems to be that psychophysical generalisations would merely be *brute correlations*, where brute correlations are generalisations that could not be supported or known by ordinary inductive means.

But the notion of lawlikeness, which supposedly should explain the difference between a law and a brute correlation, is a difficult notion. And in later papers Davidson seems to prefer to replace the distinction between lawlike and non-lawlike generalisations with the distinction between exceptionless or strict – and rough generalisations. And this makes it appear that the original denial of the status of laws to the generalisations stating the coextension of mental and physical predicates turns

only on the enormous complexity such statements would have. But this is unattractive for two reasons. In the first place, the degree of complexity does not seem like an a priori matter at all, and in the second place it is not clear that strictly unexceptional physical laws or generalisations (at least concerning observable macro-level phenomena such as the lighting of a match, for instance) would be any less complicated than a psychophysical one.

Let us admit, at least for the sake of argument, that it is an empirical question how strict a correlation we may establish between a mental predicate or phenomenon and a physical one, and that it is an empirical question how simple or complex the physical predicate or phenomenon in question is.

Davidson says that if we were to find such correlations, then we would have no reason to think that they were more than approximately true. This way of stating the point can be too strong, or at least misleading. Remember that there are two questions. One is what we could know concerning psychophysical (or other) correlations. Another is whether whatever knowledge we could attain is sufficient for codifying ascription of mental predicates (or something slightly weaker than codification – that of accepting something as a criterion for the ascription of mental predicates). Davidson's remark is too strong or misleading if it is taken to mean that we could have no inductive justification for psychophysical generalisations whatsoever. Say that it is an empirical question how strict and how simple the correlation is. What I think Davidson should have said is that no matter how strict or rough the correlation was found to be, it is still a different and a further question whether we should decide to accept the physical phenomenon or predicate as a *criterion* for the application of the mental predicate, and thereby determining its extension. Science cannot dictate such a decision for us.

I must explain the special use I am going to make of the notion of a criterion. In ordinary use, a criterion is a standard on which a judgment or decision may be based, and there could be many logically independent criteria for the same phenomenon. I don't think the possibility that in some cases only some of the criteria obtain is excluded in the ordinary notion of a criterion. In that case, I am not certain that ordinary usage excludes the possibility that a 'criterion' C for a phenomenon A could obtain, and it yet be possible to judge that A did not obtain, perhaps because none of the most important criteria obtained in that case. The way I am going to use

the notion of a criterion this possibility is excluded. A criterion for the application of a predicate or description is related to the notion of (normative) codification. I can allow there to be more than one criterion for the application of a predicate, but something will not count as a criterion if exceptions are allowed as conceivable. The same independence condition as in the case of codification will apply here as well. If a condition C and a predicate P are independent, then C is **accepted as a criterion** for the application of P if it is accepted as an unexceptional rule that if C obtains in a situation, P should be applied in that situation. Accepting a criterion for a predicate is a partial codification of application of that predicate[13].

To begin to justify why scientific knowledge could not dictate a decision to accept something as a criterion for a mental predicate, consider the following. Say there is a general statement, or an algorithm or an inference procedure – could we not just acquire justified confidence or knowledge that it gives the correct result, by ordinary inductive means? Let us say, at least for the sake of argument, that inductive support could be acquired this way.

This would be a kind of confidence that was built up gradually. Consider that process. So long as we were in the process of building up this inductive evidence, we would have to check, by using our interpretive skills independently of the algorithm, whether we found the result generated by the algorithm in particular cases to be the one which made best sense of the agent. Suppose we had performed a large number of these tests and none of them falsified the algorithm or the general statement. What next? Let us grant that some kind of "confidence" or justified belief could be built up this way. Something like this happens all the time. People are often quite confident about what a particular dress code, or visual appearance, or a particular manner of speech, all mean. And we often rely on such cues in rash descriptions of each other, without pausing to consider, conscientiously, whether that description really is the one which makes best sense of the other person (or whether we know enough to make a judgment). We could certainly have built up confidence in our algorithm in the same fashion.

[13] Accepting a criterion C for P and a criterion B for ¬ P is still weaker than the codification of applications of P, since one may wonder whether P in a situation where neither C nor B obtains. If C is a criterion for P and ¬ C is a criterion for ¬ P, then application of P has been codified.

But what happens if we base a judgment on the algorithm and our judgment is challenged? If we respond by reverting to our informal interpretive skills, discuss and argue with our fellow interpreters what makes best sense, independently of the algorithm, then we do in effect agree that it would be wrong to rely on the algorithm or inference procedure mechanically. If we respond by checking the steps in the algorithm, or by double-checking whether the physical condition obtains, and *that alone* is what we do, then something more fundamental than our level of confidence has changed. We now see the algorithm, or the physical condition mentioned in the general statement, as providing the criterion for the application of the mental description.

It is important to note that this is a change in the use of language, and therefore a change on our part. In the case of physical predicates, a similar choice is determined, I presume, by pragmatic considerations of utility and simplicity of theory. The important point here is not exactly what the determinants are in conceptual change in science, only that this is a step, and that there are considerations that go into determining whether or not to push through such a step. Such a choice (to adopt a "lower-level" phenomenon as the criterion for a "higher-level" one) does not require that the connections were initially found to be exceptionless. Reductions based on non-strict correlations are the rule rather than the exception in the history of science.

Whether or not a reduction in science is optional (and the talk of "choice" here can be taken literally), the role of the considerations that go into determining conceptual changes indicates that it might not be the same considerations that operate in different contexts.

I am inclined to describe as drastic the change involved in starting to take our imagined algorithm as the criterion for the application of a mental description. In the terminology of Strawson (Strawson 1962), this begins to look as the adaptation of a purely objective attitude at the expense of the personal attitudes, or an attitude of involvement. That is something that requires argument, of course. But if it is correct that this would be a (drastic) change of interest and change in use of language, it could be described as Davidson does as a change of subject.

We have now introduced two different reductions of the attribution of mental properties; accepting a criterion and codifying a set of descriptions. Since acceptance of a criterion is much weaker than codification, it would seem to require a stronger

argument to show that we should not accept criteria for (any) mental predicates than to show that we should not codify interpretation. This has in part to do with the fact that codification concerns sets of predicates while criteria concern single predicates. (The set being codified could according to the definition contain only a single predicate, but in interesting cases, as with interpretation, it will typically be a large set).

Codification of the application of a set of descriptions, as I have described it, is a fairly drastic way to streamline language. One could certainly argue that apart from a few trivial cases, codification is *never* justified in science, and that the uncodifiability of interpretation therefore marks nothing special about the mind. But, as Davidson said about the question of the non-existence of strict laws, the reason might differ in different cases. If this is true, then even if codification is an idealisation of a form of reduction or conceptual change to which few real examples correspond, this might nevertheless constitute an argument that different considerations are at stake when reductions of mental and physical concepts are discussed.

I have said that an inductively supported correlation between A and B cannot dictate the acceptance of A as a criterion for B. But in the scientific case, if A is independent of B and is also in some sense of a "lower level" (being, for instance, more directly observable), then it can at least sometimes be just a matter of convenience whether to take A as a criterion for B. In the case of interpretation this cannot be just a question of convenience, and this suggests that other kinds of considerations are at stake.

One way of approaching these matters would be to try to show that there are considerations against the codification of interpretation that are different than the considerations against codification of large sets of physical predicates. When it comes to criteria I think we shall see a difference in that they are much more acceptable in the physical sciences than they would be for mental concepts. The most modest type of argument that might still not be insignificant is to show that the considerations that ought to determine "choices" about conceptual development of our mental vocabulary are (or should be) different than the pertinent considerations regarding the vocabulary of the (other) sciences.

Uncodifiability must be the norm and not the exception. Codification of small sets of predicates may perhaps be possible and justifiable, but when it comes to large sets this becomes increasingly implausible. Child, we may note, talks about "complete and detailed" attributions of mental properties. When we are talking about sets of predicates that are this large, we may be pretty sure that it could never be justified to codify them. The question is whether we have here struck on something that is special for the mental domain. Child had argued that physical properties are codifiable but mental properties are not. I think we could at most hope to show that there are different reasons why they are uncodifiable.

When it comes to sets of physical descriptions one reason why it would be wrong to codify them is simply that we don't have the knowledge that we would need for this to be a useful step to take. When it comes to sets of descriptions comparable in size to the mental descriptions covered by interpretation, we could also safely say that such knowledge is forever practically unattainable. Because of these limitations on our part and our fallibility there is a great epistemological advantage in maintaining a rather generous degree of flexibility in which kinds of revision of physical theory we must make given any recalcitrant observation.

There is thus an epistemological motive for maintaining some degree of holism in a physical vocabulary. Codification, and to a lesser degree acceptance of criteria, would significantly compromise such holism. If one set of predicates were codified, then the flexibility about theory revisions in face of unexpected observation would have to be confined to the vocabulary into which that set had been codified. But although there is thus an epistemological motive for holism, there are also other, counteracting epistemological motives towards simplification. The concepts of science develop together with our knowledge. Often, when our knowledge grows, things get more and more complicated. And that, in turn, gives reasons for favouring simplifications of language, or reductions, at certain points. We want a manageable vocabulary with descriptions that we know how to attribute in concrete cases.

In some cases, such simplifications are just a question of the degree of convenience of one choice rather than another, and reductions can be based on correlations that are (prior to the reduction) less than strict or exceptionless. There is thus no absolute ban against criteria for physical vocabulary; in fact, there seems in some cases to be a presumption in favour of it. Medical scientists are, for instance, on the search for a definition of cancer. This would, it seems, involve reaching an

agreement on what all the criteria for cancer are, and thus for expanding and completing their current list of criteria. There is no presumption against the addition of criteria to their list, if such an addition proves to be convenient in the light of treatment and diagnosing of illnesses. I will argue now that the threshold for accepting criteria for mental vocabulary is significantly higher than what it is for non-mental terms.

In the first place we have a much stronger interest in the mental case in maintaining a very holistic vocabulary. In the second place there is in the mental case much less opportunity for accepting criteria for some descriptions without this implying codification of the whole vocabulary. The feature of criteria that is required for their epistemological acceptability is that they are relatively local. It does not seem possible that there could be criteria for the attribution of propositional attitudes one by one, nor for a cluster of propositional attitudes. If there were criteria for a subset of the attitudes, that would mean that the attribution of attitudes from this subset was no longer sensitive to which attributions of propositional attitudes outside of this set that were made, and that would just be absurd. Acceptance of criteria for some propositional attitudes could not remain a local phenomenon; it would have to amount to a complete codification of the attribution of propositional attitudes. For non-mental vocabulary there seems in contrast to be much greater opportunity for this kind of regimentation of the vocabulary, and that is part of what makes scientific reductions possible and laudable.

If the application of a set of descriptions P has been codified into an independent set of descriptions Q or a criterion from Q has been accepted for a description from P, that does not imply that someone does not have to exercise their sense of judgment at all, but only that this exercise of judgment now can be confined wholly to whether the Q descriptions are to be applied. It is of course very rare that this is possible even in the case of physical descriptions, but actual cases may in a certain sense approach this ideal type. When a physical term is applied on the basis of evidence, as a result of an inference and as the result of an exercise of judgment, then even if it is not the mechanical derivation on the basis of a codification, this process still is different from a process of interpreting people in an important respect. There is typically in the physical or in the scientific case a willingness to shift this need for exercise of judgment to a lower level.

Take as an example a doctor trying to arrive at a correct diagnose of a patient. True, we often say of such a doctor that he "interprets" the symptoms of the patient. What he does is to exercise his judgment in order to form the most plausible hypothesis about the cause of the symptoms. But this is not interpretation in the personal or rationality-sensitive sense. If the doctor comes by the presence of what he takes to be criteria for a certain disease, he is more than happy to abandon interpretation. It is not an end in itself for the doctor to interpret, therefore the discovery of or the decision to treat something as a criterion, and thereby abandoning interpretation, would not count as a change of subject. If he did discover what he took to be criteria for the presence of the disease, it is not that he would no longer need to exercise his good sense of judgment. But it would have been shifted to a lower level: he would now need to exercise his sense of judgment in determining whether the criteria were present in the patient.

A *constitutive ideal* of physical vocabulary is that its descriptions should not be sensitive to considerations of what is good, right and beautiful. The description of the orbits of celestial objects should, for instance, be *insensitive* to the view that circular motion is the perfect and most divine kind of motion. The coming of age of science was the recognition that physical descriptions should be unaffected by considerations of rationality and purpose; they should be interpretation *in*-sensitive.

The willingness of the doctor in our example to shift levels is not, and should not, be present in the case of interpreting people. The evidence in interpretation is interpretation sensitive, and a willingness to shift the need for exercise of judgment to a level lower than this would mean a willingness to exercise our judgment only about evidence that is not sensitive to our overall rationality assessments.

If it is true, as I argued in chapter 3, that rationality should be a regulative ideal in interpretation, then we should accept the following principle:

> *(Rationality as a regulative ideal):* We should consider, *after* all evidence has been noted and all rules of thumb, shortcuts, conventions, inference procedures, etc. have been applied, whether the resulting interpretation makes *best sense* of the subject.

From this it follows automatically that interpretation should not be codified. Codification would imply that the evidence for these concepts could no longer be interpretation sensitive, and that rationality could no longer be a regulative ideal for their application.

There seems therefore to be a deeper reason for why mental vocabulary is uncodifiable than why physical vocabulary is uncodifiable. The point of interpretation lies at least in large part in its connection with communication. To communicate with someone involves among other things to ask the other person for reasons for his views, and also to take the other person's views as potential reasons for views for oneself, for instance as reasons to change one's own views. The information in which the views of the other person is given to us will then be most interesting to our purposes if it has incorporated the sensitivity to norms of overall rationality. But it is not just that we must be ready to adjust our description of other people according to how they conform to our ideal of rationality. We must stand prepared to adjust our ideal of rationality by the examples of other people; when we manage to find an action intelligible, for instance, we may have to adjust our conception of rationality in the process. This is a point emphasised by John McDowell in "Functionalism and Anomalous Monism" (McDowell 1985). Davidson says that when we are applying the vocabulary of the propositional attitudes we "aim to discover rationality in the phenomena" (Davidson 1991; 215). This tells us something about what kind of information that is the point of interpretation.

I will take it as obvious that the norms of rationality should not be codified, since this would be equivalent to the decision not to allow the slightest bit of flexibility or fallibility in one's views about what is a reason for what. As Davidson has emphasised, it is hard to conceive of someone as a thinker at all if he does not even recognise the possibility of error on his part. Communication has a central place in our way of dealing with our fallibility and in our endeavour to get things right. Codification of interpretation would seriously undermine or pre-empt this role for communication, and this makes it hard to see how someone could codify interpretation without at the same time having codified rationality.

If the argument given above that acceptance of criteria for the attribution of propositional attitudes has to expand into a full codification of interpretation, then we have also given an argument for:

> *(No criteria):* In interpretation, no data should be elevated to the status of criteria.

We can think of certain modifications to this principle, though. The argument I gave concerned the propositional attitudes. But I argued in the chapter on emotions that we should think of the mental domain in wider terms than this. Thus there is a question whether the argument extends beyond the propositional attitudes. I made an analogy with Strawson's point about free will. His point was that there is no sense of rational in which it could be rational to completely abandon the personal attitudes or the attitude of involvement in favour of the objective attitude. The parallel point about interpretation would be that we should not want to completely abandon the possibility of interpreting by totally mechanising or codifying the ascription of mental predicates. This would not rule out that *some* mental predicates could get their extension fixed by our deciding to take a physical condition as a necessary and sufficient condition for it. Imagine this happening for the notion of *depression*, for instance. One suggestion for what characterises the mental domain beyond merely the propositional attitudes (or a suggestion for what *the mark of the mental* is) is that the question of its interpretation can arise. I believe depression is currently such a concept. But perhaps this does not need to signify an in-principle irreducibility of this concept. The argument concerning the propositional attitudes might not rule out that we could come to decide to redraw the map of the mental and to eliminate certain phenomena from the mental realm. But these remarks also suggest a form of argument for not reducing some of our concepts; those concepts that we, for whatever reason, want to keep in the family with the propositional attitudes; those concepts that we have an interest should remain interpretation sensitive concepts. I discuss this issue further in the section on Paul Griffiths' program for reducing (or eliminating) emotion concepts.

What, then, about the question whether this form of methodological anti-reductionism can be used to support something like the extensional uncodifiability of mental properties or the anomalism of the mental? I shall have to leave that question for

another occasion. I think it is *possible* that further interpretationist premises could establish that conclusion; perhaps theses about the *adequacy* of rational interpretation for the understanding of people, or theses that link the *nature* of mental concepts to these principles of interpretation or see these principles as *constitutive* for those concepts, but I will not attempt this here.

I think that the irreducibility claims I have defended are rather modest. For example; that data should not be elevated to the status of criteria certainly doesn't mean that it couldn't be used as evidence in interpretation, and the same is true for inductively confirmed generalisations or laws.

Strong supervenience

Brian McLaughlin (McLaughlin 1985) argues that Davidson should not accept strong supervenience, only weak supervenience, and Jaegwon Kim has also argued that strong supervenience would give us a reduction of mental predicates to physical predicates, contrary to the anomalism of the mental. William Child has, however, argued that strong supervenience is perfectly compatible with the kind of irreducibility that the mental has.

Strong supervenience yields sufficient conditions for a subject to have a given mental property. If we allow infinite disjunctions it will also yield necessary conditions for any given mental property. So we must accept modalised biconditionals for every mental property of the form: Necessarily for all x, x has M iff x has P1 or P2 or P3 or …

Strong supervenience will threaten anomalism only if the biconditionals (or conditionals) that it yields count as laws, Child explains. Many people have argued that infinite disjunctions are somehow illegitimate in laws, but this is not Child's line of attack. He says:

> Suppose we are prepared to treat these conditionals as laws. The laws will be principles saying that in a completely specified set of physical circumstances, with every aspect of context fixed, a subject will have a given set of mental properties: Necessarily, $(x)(P^*x \rightarrow Mx)$. That principle links a set of mental characteristics with a

> single, completely specified set of physical characteristics. But it is not part of a system of laws by reference to which one could derive mental characterizations from physical ones; it does not, for example, tell us what mental change in S would be brought about by a given physical change. Even if we knew all the supervenience conditionals derivable by considering every subject in the history of the world, we would not have the resources, in a new case, to derive a subject's mental properties directly from a specification of all the physical circumstances of the case (for any two actual subjects must differ in some physical respect); rather, in each new case we must make a new judgment about the application of the norms of rationality in that case. (Child 1993; 231f)

Child's point here seems to be that the conditionals do not give an answer about new cases; that they do not form a *system* of laws. He says that the conditionals may be "allowed, in some sense, to be lawlike; they support counterfactuals, for example". But they do not form a system of laws (or are not part of a system of laws), and "because they do not form a system of laws for the precise prediction and explanation of particular mental phenomena, they are no threat to the anomalism of the mental" (Child 1993; 232). What we have, then, are basically atomic principles, that do not allow any automatic application to new cases. This non-systematicity seems to derive from the fact that the antecedents of the conditionals must contain a "completely specified set of physical characteristics". Child thinks that such a "complete specification" will have to be so specific that no new case will count as similar.

It is true that strong supervenience does not imply that there need be any system among the different conditionals that makes it up. It is consistent with strong supervenience that having uncovered a large number of such conditionals should give us nothing to go on whatsoever in deriving a mental description from a physical one in a new case. Strong supervenience does not imply the existence of laws that could be useful to us in predicting mental phenomena. For this reason Child notes that strong supervenience is consistent with the anomalism of the mental. (Child still uses 'anomalism' as a description of Davidson's anti-reductionism. For Child, 'anomalism' seems to mean the non-existence of laws that could be used to derive mental descriptions from physical descriptions, so that reliance on interpretation was no longer necessary.)

But Child needs more than the fact that strong supervenience does not imply the existence of a system of conditionals, because strong supervenience clearly does

not imply that the conditionals are not systematic either. It is consistent with both possibilities, and Child seems to need a separate argument that there will in fact be no system, which he gives by saying that the antecedents of the conditionals must be completely specific.

If Child is right that the antecedents must be so specific that application to new cases is ruled out, I would grant that this is an important conclusion. But this claim could surely be doubted. Functionalists certainly have the hope that it should be possible to specify the physical states which realize a given mental state without going this specific, and certainly without having to mention absolute spatiotemporal localisation. It may be that Child has a good argument against the functionalists in the sense that it is highly *unlikely* that the antecedents could be so simple that they would be of any use at all for anyone predicting or deriving finely individuated mental descriptions. But this question, it seems to me, is just, or at least mainly, an empirical question. Davidson's argument for the irreducibility of the mental is relatively a priori. This is not what the irreducibility claim fundamentally hangs on.

Whether the principles must contain completely specified physical characteristics is something of a red herring, and has the unfortunate consequence that we do not see the full extent to which Davidson's anti-reductionism may be consistent with strong supervenience. For surely it is not plausible that the conditionals implied by the supervenience thesis should be *completely* unsystematic and unrelated to each other.

What I complain about is the order in which Child presents his points, and the relative emphasis he gives them. I would not fuzz about the form that the antecedents in the principles given by supervenience must take. The crucial movement, as I see it, is indicated in the first sentence in the quote by Child above: "Suppose we are prepared to treat these conditionals as laws". Now, my question is; what would it mean to treat such principles (never mind the form which the antecedents take) as *laws*? Remember, first, that the laws in question are bridge laws, i.e. synchronic laws, and assume, what the functionalists hope and what Child denies, that the conditionals did have such a form that it was possible to apply them to new cases. Assume that they are as beautifully systematic as you please. To start treating such principles as laws is a change on our part. The functionalist would probably say that the only change on our part was the acquisition of new knowledge (or belief). But I think that in our context 'treating such principles as laws' could mean either of two things. The

one thing that it could not plausibly mean is that one started to regard the principle as *indefeasible*. (Belief in, or even knowledge of, laws never amounts to that.) But there is a question as to what kind of considerations that one recognises as possible defeaters of the law. If one recognises as possible defeaters situations in which we would judge that despite the occurrence of the physical condition, the mental description can not be attributed because it is not the interpretation that makes the best overall sense of the person, then I think there can be no objection to treating such principles as laws. But if treating the principles as laws means that one recognises no sense in asking after the derivation from the principles was made whether this ascription of mental state really was the one that made best sense from an overall interpretation of the subject, then I would say, with Davidson, that by doing this we would have changed the subject. Such a change would mean to abandon the enterprise of understanding another as essentially an interpretive project.

Like Child, I maintain that this form of anti-reductionism does not imply that the conditionals or biconditionals yielded by the assumption of strong supervenience must be false. On the contrary, they may be true. We could also apply them; and base attributions of mental descriptions could be justified on the basis of them. It is not the case that our attributions of descriptions are unjustified unless they are given a "total" justification. So long as we continued to regard the mental description as something that ultimately answers to overall interpretation, though, we couldn't adopt the physical condition as a criterion for the mental one.

Antony on the anomalism of the mental

Louise Antony is a critic of Davidson who has very ably spotted the central points of his position, I think. But she blows up these aspects too much. This is what enables her to criticise anomalous monism as being an inadequately realist philosophy of mind. Her exaggerations provide a useful foil against which to contrast my own understanding of these points, and that is why I discuss her treatment of Davidson.

On her interpretation, the basic point for Davidson is that it should not be an empirical question whether people are rational. According to Antony, Davidson rejects lawlike connections between the physical and the mental because they would

establish evidential conduits that would "put the enterprise of rational interpretation at empirical risk" (Antony 1994; 236).

Antony importantly thinks that Davidson must mean to reject not just strict, exceptionless psychophysical laws, but also the psychophysical analogues of rough *ceteris paribus* laws common to the non-basic natural sciences. As she notes, lawlike connections, of any strength, are evidential conduits. "Lawlike connections, however strict, are evidential conduits, and it is the flow of evidence from the physical to the psychological that Davidson is afraid of" (ibid.). "Allowing the possibility of any lawlike connections between the mental and the physical *is* as Davidson recognises, equivalent to allowing the possibility that we are not, as a matter of empirical fact, rational creatures" (Antony 1994; 237).

But, Antony asks; is any of this up to us? "If there are tight connections, there are tight connections – what can we do about it?" Antony herself clearly thinks that there is nothing we can do about this, and that it is surely not up to us, but she thinks that Davidson disagrees on exactly this point. "But that is just it – in the end, I think, *Davidson does* believe it is up to us. In the end, Davidson mounts a pragmatic appeal to accept Brentano for the sake of an edifying self-conception" (Antony 1994; 237).

This way, Antony takes very seriously Davidson's claim that "The limit thus placed on the social sciences is set not by nature, but by us when we decide to view men as rational agents" ((Davidson 1974b; 239) – cited in (Antony 1994; 237)). She takes it literally that "we decide". There cannot be psychophysical laws because we decide that there shall not be.

I agree with Antony in placing the emphasis here; that the irreducibility of the mental does derive from us and our interests that come with our mental vocabulary. The irreducibility does derive from the impossibility for rational creatures to accept a certain kind of conceptual development. But it is, in the first place, an overstatement to speak literally about decisions here. Sometimes such overstatements are perhaps harmless. It is an overstatement to speak of decisions when a society "decides" to change the criteria of application for a certain predicate. It was not a decision, in any literal sense, when "fish" was denied application to whales, but it was a conceptual change that was motivated by interests of ours. I won't try to describe those interests, but they probably include our interests in having a simple classificatory system, and one that tracks biologically important features of organisms. It is harmless, I think, to speak of decisions to change our concepts in situations like this. Because there was a

way to use the word prior to the change and a different way after, the situation is similar to that of a genuine choice in that there are two alternatives that are shown to be possible, in a sense. It may be more misleading to speak of a "decision" *not* to change our conceptual practice, because there may be no real option (except a weakly conceivable option) to change. I mention this only to forestall a certain possible misunderstanding. The Davidsonian irreducibility of the mental derives, according to my interpretation, from a claim that we should "decide"[14] to retain a certain feature of our conceptual practice, namely that we do not allow any short-cuts or criteria for ascribing mental predicates to replace overall judgments of what makes best sense of people in interpretation.

Antony interprets Davidson's anti-reductionism as stemming from a view of psychology as intrinsically non-scientific (rather than that the dependence is the other way around). Central in her interpretation is a reading of the analogy (as well as the disanalogy) Davidson draws between measurement of physical magnitudes and interpretation of people.

The measurement analogy

Davidson begins his discussion of length in "Mental Events" with the claim that confidence that a statement is "homonomic, correctible within its own conceptual domain, demands that it draw its concepts from a theory with strong constitutive elements" (Davidson 1970; 220). The discussion of length is introduced as an example of this. This creates a little puzzle, however, for the obvious candidate (L, below) for being a strongly constitutive principle seems to be excluded from having this role by Quinean holism.

Unless the relation *longer than* holds as transitive, we cannot easily make sense of the concept of length. The following principle, governing the relation *longer than* is therefore a candidate for being a constitutive principle for the concept of length.

(L) $L(x,y)$ and $L(y,z) \rightarrow L(x,z)$

[14] The talk of decision should not be taken literally here. It is a question whether it really is in any conceivable sense "up to us" to choose whether to treat people as rational or not. Cf. (Strawson 1962).

This principle obviously does not exhaust the content of the concept of being longer than, since it is common to all transitive relations. We must suppose that there is some empirical content which distinguishes the concept from other transitive relations, and on the basis of which we may assert that one thing is longer than another, Davidson writes. So, he says, to imagine that this empirical content is partly given by the predicate O(x,y) and the following 'meaning postulate':

(M) $O(x,y) \rightarrow L(x,y)$

(L) and (M) imply that we can never find objects such that O(x,y), O(y,z) and not O(x,z). But given that O(x,y) is an empirical predicate, Davidson admits, we *cannot* have such a guarantee. If we *think* we observe a counterinstance to transitivity, there are many possible moves and it is not clear that there is exactly one thing we should give up. "It is better to say the whole set of axioms, laws, or postulates for the measurement of length is partly constitutive of the idea of a system of macroscopic, rigid, physical objects" (Davidson 1970; 221). One could ask why this holism doesn't destroy the distinction between constitutive and non-constitutive elements. For it is not as if (L) and (M) together with the rest of the statements that make up this system are constitutive in the sense that they couldn't be falsified by experience. Each case in the measurement of length, Davidson remarks, "tests a theory and depends upon it" (Davidson 1970; 221).

We might remark that Davidson does not say that this system is constitutive of the concept of length, but that it is partly constitutive of the idea of a system of macroscopic rigid physical objects. I don't know if this difference is significant. He does say, however, that the concept of length would have no application if we counted (L) false, that if we say (M) gives a wrong test for length it would be unclear what we thought was the content of *longer than*, and similar remarks which perhaps justify the claim that (L) (and (M)) are (partly) constitutive of the concept of length.

When Davidson connects "confidence that a statement is homonomic" with the presence of strong constitutive elements, and when he says "I suggest that the existence of lawlike statements in physical science depends upon the existence of constitutive (or synthetic a priori) laws like those of the measurement of length"

(Davidson 1970; 221), then one would perhaps be led to expect that psychology would be distinguished by the absence of constitutive elements.

But, of course, this is not what he says. What, then, does he think are the constitutive elements of psychology?

Analogously to (L), there is a principle of decision theory stating the transitive nature of relative preference. It may be instructive to compare this principle with (L).

(P) $P(x,y)$ and $P(y,z) \rightarrow P(x,z)$

Davidson says that the cases are parallel in (Davidson 1974b; 236f), that the satisfaction of conditions of consistency and rational coherence may be viewed as constitutive of the range of applications of such concepts as those of belief, desire, intention and action. (And he explicitly cites the principle of the transitivity of preference as something without which we couldn't make clear sense of attributions of preference.)

Davidson had a short career as an experimental psychologist testing variants of decision theory. One thing he found was that "as time went on, people became steadily more consistent; intransitivities were gradually eliminated" (Davidson 1974b; 235f). There was no feed-back or reward during the experiment. Describing why he gave up that career, Davidson says "I found it impossible to construct a formal theory that could explain this, and gave up my career as an experimental psychologist" (Davidson 1974b; 236). A page later he writes about the significance of this experience (he writes 'experiment', but it is surely the experience he had with it that was important) that it shows how easy it was to interpret choice behaviour so as to give a consistent and rational pattern.

> When we learn that apparent inconsistency fades with repetition but no learning, we are apt to count the inconsistency as merely apparent. When we learn that frequency of choice may be taken as evidence of an underlying consistent disposition, we may decide to write off what seem to be inconsistent choices as failures of perception or execution. My point is not merely that the data are open to more than one interpretation, though this is obviously true. My point is that if we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behaviour, then we are committed to finding, in the pattern of behaviour, belief, and desire, a large degree of rationality and consistency. (Davidson 1974b; 237)

This passage is not very easy to understand. It seems natural to connect this with Davidson's repeated, and notorious, claim that psychology somehow falls short of a "serious science". Davidson reports how easy it was to interpret the data in such a way as not to conflict with principles such as (P). But that in itself may not be different from contexts of experimentation in other sciences. It has been a lesson at least since Duhem that reinterpretation of the data may block falsification of the theory to be tested. Did Davidson feel that there was something different in the way he and his co-experimenters insulated principles such as (P) from falsification? Was there something *unscientific* in the reinterpretations that fell so easily? If this *was* how Davidson saw things, I would urge that it signifies a far too strict conception of science. I argued in chapter 3 that the pattern in how the data could legitimately be reinterpreted is different in the psychological case than the pattern of scientifically justified reinterpretation in an experimental context in the physical sciences. But it is also different from cases in the physical sciences where the reinterpretation was found to be *ad hoc* and therefore an *unscientific* manner of insulating the theory. In no way does it follow that psychology is not an experimental science at all, or that psychological hypotheses cannot be experimentally tested.

On Antony's interpretation, psychology cannot be an experimental, empirical science in any sense at all for Davidson. The presence of constitutive principles for the measurement of length does not imply that the resulting theory of rigid physical objects is not empirical. She thinks that this marks the disanalogy with the constitutive principles in psychology. This is how she states it. (The emphasis is hers.) "*It is an empirical matter whether our theory of physical objects applies, and it is not an empirical matter whether our theory of rationality applies*" (Antony 1994; 235). In support of this claim, she quotes Davidson: "It is not merely, as with the measurement of length, that each case tests a theory and depends upon it …" (Davidson 1970; 221). The reason why it cannot be an empirical matter is, she says, that Davidson refuses to allow our "theory" of rational agents to be empirically contingent. It cannot be contingent whether there are any rational agents. The justification for imputing to the physical world a formal structure that obeys the constitutive principles of measurement is empirical, but in the case of rationality it is pragmatic. If we want to be able to view ourselves as rational, then we must interpret ourselves so that no observations or experiments are allowed to falsify that assumption.

Although I think Antony admirably spots the important points in Davidson's argument, she clearly exaggerates them. For one thing, she forgets the 'merely' in the quote "It is not merely, as with the measurement of length, that each case tests a theory …" and reads it "It is not, as with the measurement of length, that each case tests a theory …"

For Antony, what she sees as merely the decision to treat ourselves and others as rational agents buys us the autonomy of the mental, but this is a Pyrrhic victory because it happens at the cost of the reality of the mental. She thinks that a naturalistic mentalist has to allow the possibility that we are not as a matter of empirical fact rational creatures. She says "what I want to know is whether I *really am*" [an autonomous, rational agent] (Antony 1994; 237), that is, not to have this important question just declared by fiat. But what question exactly does Antony think Davidson settles by fiat? Certainly, it is an empirical question to what degree any given person is rational, just as it is empirical whether or not that person believes a specific proposition. It is an empirical question how strong a person's inferential skills are, whether he is prone to the gambler's fallacy, prone to wishful thinking, self-delusion and so on. To claim that the answers to questions of this order are declared by fiat by Davidson would be to say that there is no role for interpretation in the ascription of mental properties. Davidson's position might certainly have its problems, but it doesn't contain absurdities of this order. Perhaps Antony is not thinking of particular claims such as this, but rather a question such as whether a person is rational, or whether there are rational agents at all. But I am not sure that we can give good sense to these questions, nor that Davidson needs to maintain that they do make clear sense. We could say that the question whether any given person (or creature with an animal- or human-like appearance) is rational at all is empirical in the sense that we may start to doubt this if its behaviour resists all our attempts at interpretation and if we have success at interpretation we tend to declare the creature rational. The first person case is different of course, but we don't need to consider its special complications here.

Perhaps Antony thinks that "our theory of rationality" or "folk psychology" is not empirical in Davidson's view. But what are these things? Whatever they may be they would presumably consist of a body of general statements. Any person has his views about what people of a certain kind tend to do in certain situations, and he has his views about what people ought to do and think. There is no given generalisation he believes and no given norm that he upholds that must be shared by the next man.

Now, some statements about the tendencies of people to act in certain ways are so vague that it is unclear what kind of experience might contradict or confirm them, and in that sense there are problems with thinking of them as empirical, but apart from that, any given generalisation that is sufficiently precise can certainly be refuted or confirmed by observation, and there is no need for Davidson to deny this. But the vaguer statements are empirical in the sense that they have connections with the more precise ones.

Our norms are not empirical in the following sense; they shouldn't change merely because we encountered people who acted or thought contrary to our norms. But that doesn't mean they never do change. They could certainly be changed by experience and our encounters with others. As I've already remarked, it is a central point of John McDowell's (McDowell 1985) that we should be open to adjusting our standards of what is rational by our encounters with others. We should be open to letting our norms be educated by the example of others, and thus the observation of new, perhaps better, forms of exhibited rationality would change our views about what is rational.

If Davidson is right, then our norms do have an ineliminable role to play in the interpretation of people, and this, if you will, means that we could say that there's a non-empirical element in the ascription of mental properties to people. We *could* say that. But of course interpretation depends on what people say and do, and so is empirical. Antony says that "One thing that clearly worries Davidson is the possibility of competition between two sources of evidence – the physical and the behavioural/interpretative" (Antony 1994; 236). Apparently she thinks that it is only the flow of evidence from the physical to the psychological that Davidson is afraid of, and that therefore perhaps her claim is that it is only the physical evidence that Davidson, by fiat, declares to be irrelevant for psychological ascriptions. Of course some evidence will be of such a kind that we don't know how to make any use of it for the purpose of particular interpretations. For some evidence it may be that it is only *as of now* that we do not know how to make use of it. But this doesn't give us any reason for distinguishing between "two sources of evidence". There is just evidence, and it is physical and empirical.

In an earlier paper, Antony says that for Davidson "in psychology, non-empirical 'interpretive' concerns trump all others" (Antony 1989; 178). As I think of Davidson's anti-reductionism, it is almost precisely the opposite claim that is made. In

interpretation, no data and no evidence can trump all the others. The claim is not that some empirical evidence is blocked from any influence on interpretation. The position on the contrary encourages the view that *any* evidence *may* be relevant. The point is that after new evidence has been noted, an interpreter must consider which ascription makes best sense of the person, and that is in part an evaluative question which involves the norms the interpreter has.

If we accept that no data can be elevated to the status of criteria, there is another way we could think about the measurement analogy. The difference between the two cases may not reside so much in a difference between (L) and (P). But look at the principle (M); $O(x,y) \rightarrow L(x,y)$. Read strictly, this principle says that the relation O(x,y) is a criterion for L(x,y). For physical predicates or descriptions (though I doubt that this is the case for length) such a decision may be justified on the basis of pragmatic considerations and the degree to which the correlation is inductively supported, but this is not so for mental predicates.

The view of Kim and Antony that psychology could not be an empirical science for Davidson is shared by Alexander Rosenberg. In his case we can also see very clearly how this view is fuelled by the attribution of the a priori theory conception of mind to Davidson. Rosenberg notes that sciences do not need to begin with 'closed' theories and 'homonomic' generalisations. The crucial question is whether its theories and generalisations can be tested. This is not a problem for the special sciences (except psychology), Rosenberg argues, because there are ways of measuring the variables of these generalisations independently of the theories. The terms and generalisations of Mendelian genetics are far from strict, for instance, and type-type reductions may not be forthcoming. "Nevertheless, we may improve and correct Mendelian claims, systematically explain their predictive failings and successes, because we can identify and measure the values of the causal variables of Mendelian theory independently of that theory, through the facilities of molecular biology" (Rosenberg 1985; 405).

This, in contrast, is not possible with regards to psychological generalisations, Rosenberg argues. Why not? For Rosenberg the 'anomalousness' of psychological terms means precisely that they have no linkage with the rest of science. Our only grasp of psychological terms is via their implicit definition by the generalisations in

question. Rosenberg explains that "the implicit definitions intentional generalizations provide are all we have to go in characterizing and actually identifying propositional attitudes" (Rosenberg 1985; 405). To have this view of mental terms is possible only if one accepts the a priori theory conception of mind. Rosenberg clearly does so, or rather attributes this view to Davidson; "our confidence in the truth of intentional claims of any generality is based not on their predictive or explanatory power, but on the fact that they represent propositions as close to analytic in their grounds as any propositions may be" (Rosenberg 1985; 404).

I think we have seen, though, both in this chapter and in chapter 3 that there is no reason why an interpretationist should take this view of psychological generalisations. There are *myriad* ways in which psychological terms are linked to other terms (both psychological and others), and there is an *overabundance* of types of evidence that are relevant to the testing of interpretations. And it is not as if psychological explanation must apply something like a monolithic theory where the individuality of its separate claims disappears completely. Psychological generalisations are not in themselves any different from generalisations in the other special sciences. They are perfectly testable, even if the experimental logic of cases where we deal with rational subjects presents its own challenges, as explained in chapter 3.

Paul Griffiths on emotion concepts and conceptual dynamics

In this section I give an illustration of one instance of a form of reductionist policy concerning conceptual development to which methodological interpretationism is opposed. I use this example in order to draw this contrast more clearly.

I have been defending a form of interpretationism which upholds the irreducibility of interpretation; that the norms or rules of interpretation are uncodifiable. If we were presented with a set of precise rules for ascribing mental predicates to people on the basis of evidence of whatever kind (behavioural, neurophysiological, or whatever), a mechanical application of these rules would be very different from the process of interpretation. To interpret means to ask *after* any clues or hints or shortcuts have been used to arrive at an ascription of a mental predicate, whether the result makes overall best sense of the person. This means that

whatever we can say about which data that were present or absent in previous cases of successful interpretation, this experience cannot be applied mechanically to a future case; an interpreter would have to use his interpretive skills anew. Interpretation has an essential degree of openness. In interpretation, no data can be elevated to the status of criteria.

This openness in principle to new considerations opens up to a potential battlefield between old and new methods of determining mental ascriptions, between, for instance laboratory tests and clinical diagnoses and ordinary communication, and between old and new kinds of evidence; what people say and do and hormones and neurotransmitters and patterns of neuronal firings. Many forms of considerations are part of determining the shifts of relative importance these methods and elements acquire. There can, no doubt, be reasons for favouring simplifications, or reductions, at certain points. I now turn to a discussion of Paul Griffiths' claim that a certain form of naturalistic reduction (and partly elimination) of emotion concepts would be an epistemic gain. This is in clear tension with interpretationism because (if followed to its ultimate conclusion) it threatens to eliminate the role of interpretation in psychological understanding. Interpretationism insists that as long as we are in the business of making sense of people there is a real epistemic gain in allowing certain concepts to remain undetermined by any fixed amount and kind of evidence. The present section aims to show one example of a philosophical policy concerning conceptual development to which this interpretationism is opposed.

Paul Griffiths presents and endorses what he calls the causal homeostatic theory of concepts. "A category brings together a set of objects with correlated properties. The category has causal homeostasis if this set of correlations has some underlying explanation that makes it projectable. A successful category captures … a causal homeostatic mechanism – something which means that the correlations can be relied on to hold up in unobserved instances" (Griffiths 1997; 188). Another word for 'successful category' in this sense is 'natural kind'. The concept of causal homeostasis is related to that of the essence of a category. According to Griffiths, we can think of the 'essence' of a category as any theoretical structure that accounts for the projectability of the category. This includes micro-structural essences, which are the causal homeostasis mechanisms of fundamental chemical kinds, but these are not the only kinds of essence; "biological taxa, the other classic example of natural kinds,

turn out to be united by external forces. Biological taxa at all levels of the taxonomic hierarchy form projectable categories because their members are descended from a common ancestor … The causal homeostatic mechanism is descent" (Griffiths 1997; 189). Even artefacts may be said to have an essence, even though they are towards the nominal end of the continuum between real essences and nominal essences.

This seeming liberalism about what essences can be does not preclude Griffiths from firmly opposing description theories of meaning. Such theories do not adequately explain conceptual dynamics. They cannot explain why the intension and extension of concepts should change as a result of scientific development, and they cannot explain why concepts can retain their identity despite radical changes of theory. The causal homeostasis theory explains these phenomena. "The use of a concept for explanation and induction commits its user to the *project* of having a category with causal homeostasis. The pursuit of this project is what causes revision of extension and intension" (Griffiths 1997; 193).

Griffiths discusses two challenges to the causal homeostasis theory of concepts that are raised by work by Ian Hacking. The first challenge is raised by the possibility of categories that are what Griffiths calls "substantially socially constructed". "The existence of a particular emotion in individuals may be the result of the existence of the corresponding emotion concept in the culture" (Griffiths 1997; 197). Fear is *not* an example. People would presumably get frightened even if they completely lacked fear-concepts. Hacking suggests that post-Freudian hysteria and multiple personalities syndrome may be phenomena that would not and could not exist without the patients being familiar with the corresponding concepts and lists of typical syndromes. (The phenomena apparently do not exist – that is, we cannot find them – in cultures lacking the relevant concepts.) Now, it is true that Griffiths' theory of concepts needs to acknowledge that some categories may have causal homeostasis through these sorts of reflexive mechanisms, but since the existence of the concept does seem to qualify as (part of) such a homeostasis mechanism, I don't think that this is a very deep objection to the theory. The second challenge is more serious.

Griffiths acknowledges that one must recognise the full range of dynamics at work in conceptual change, brought to light by Hacking's work. The causal homeostasis view "predicts that concepts will evolve to maintain and increase the causal homeostasis of the categories to which they refer. Hacking reminds us that concepts are also used for social and political ends. They are used to condemn things,

to promote attention to one aspect of a situation rather than another, and to induce conformity with certain norms of behaviour" (Griffiths 1997; 198).

These ends and agendas cannot always be compatible with each other. They are often historically conditioned and partial. Griffiths mentions different definitions of indigeneity as an example of how a concept can be used to serve the claims of one group of people as opposed to another. Although there can be real conflicts over which scheme of classification to adopt within "the realm of the purely epistemic", these conflicts are of different sorts, Griffiths says. When one such scheme threatens to displace another, the explanatory and predictive aims of the threatened scheme are taken over or subsumed by the new one. Conflicts between conceptual schemes that are tied to social or political goals usually cannot be resolved in this way, because the different agendas may be genuinely incompatible.

Griffiths reasonably says that philosophers would do well to couple their analytical efforts with sociological and historical examination of the role of emotion concepts. About philosophers engaging in traditional conceptual analysis, he says.

> At present these philosophers suppose that they are uncovering the true nature of emotion as revealed a priori in vernacular emotion concepts. In fact they are picking apart the beliefs about emotion that have become prevalent through the interaction of the various dynamics which affect emotion concepts. This places the philosopher in the unsatisfactory position of trying to understand emotion by looking at a picture of emotions which often deliberately misrepresents them. (Griffiths 1997; 201)

How can Griffiths say that such pictures of emotion misrepresent them? No doubt this is often true, but my question is really what Griffiths can mean when he says this. When we say that a concept misrepresents reality, what kind of criticism of the concept is that? Can concepts be criticised because their agendas are non-epistemic? Is *that* why they misrepresent, according to Griffiths? Or is it rather that we criticise these concepts and agendas *morally*, as e.g. being *shallow* (which can be the case for some conceptions of ethnicity or manliness or of being a good Christian), or as having other undesirable consequences, such as entering into the kind of self-reflexive and self-perpetuating mechanism claimed by Hacking of hysteria, or of being a hindrance for coping with traumas or treating them (where Griffiths' discussion of hysteria and child abuse might be examples). But if the concepts must somehow have moral shortcomings for it to be right to say that they misrepresent, then there will be a

contrast to those concepts that are (morally) good. Do these concepts also misrepresent? There seems to be no ground for saying that. A *good* concept of courage, for instance, should be deep in the sense that it captures what is important about that virtue, and it seems right to describe such a concept as truthful.

Griffiths does not say that some concepts (the scientific ones) have only purely epistemic agendas. He says that to the extent that science is driven by non-epistemic agendas, then the causal homeostasis theory will not adequately predict conceptual change, but he says that the categories that have genuine causal homeostasis are the categories that *should* be adopted by a scientific psychology of emotion, because its goals are induction and explanation (Griffiths 1997; 200). This claim might be all right, but the stronger claim that concepts which have other or additional aims than "explanation and induction" must by nature misrepresent reality is clearly wrong.

By emphasising conceptual dynamics Griffiths hopes to show that philosophies attempting conceptual analyses of emotion concepts will fail because they only consider a concept at one specific stage of its development. The interpretationism I defend also emphasises dynamical aspects of emotion concepts (although different ones) and is equally sceptical of conceptual analysis (for different reasons). But when Griffiths emphasises the interest to track causal homeostasis mechanisms (CHMs) as the *sole* interest that should guide the development of emotion concepts, his position conflicts with interpretationism. Griffiths' view of concepts is dynamical only so long as we have not identified CHMs. It seems that if a CHM is identified Griffiths will recommend adopting it as a criterion for the application of an emotion term. This particular emotion term would be removed from the stock of terms that are sensitive to the overall interpretation of the subject. When decisions to treat CHMs as criteria for mental concepts are limited and local, this does not in itself contradict the dictum that interpretation must be ineliminable in understanding people. But for Griffiths the considerations that ought to decide what to treat as the extension of an emotion term should only be what is conducive to the scientific projects of 'explanation and induction'. He explicitly argues for eliminating the concept of emotion (at least for scientific purposes). The proviso "for scientific purposes" might seem to take the sting out of this claim, but Griffiths does speak of the aims of explanation (and induction) in terms as if they were unique to science (or that to the extent a discipline does have these aims, Griffiths' theory of concepts and his consequent eliminativism should

apply). He also more or less explicitly reserves for science to have the aim as well as to have the ability of providing understanding. "Vernacular concepts are involved in a whole range of nonepistemic projects", he says. "But as far as understanding ourselves is concerned the concept of emotion … can only be a hindrance" (Griffiths 1997; 247). The eliminativism is not, after all, of a moderate type.

He thinks the category of emotion must be replaced by at least two more specific categories. One is the category of affect programs. He is quite confident that this category will be scientifically useful. He also believes that some category answering to the "higher cognitive emotions" might be forthcoming, but we know much less about what the causal homeostasis mechanism of this category might be. In any case, there is little reason to subsume these two types of phenomena under a single heading, Griffiths thinks, and since identifying 'emotion' with either one of the categories would constitute a radical conceptual revision, he opts for elimination.

Interpretationism on the contrary emphasises that concepts can be useful when they are allowed to be sensitive to overall rationality considerations. As long as we "decide to view men as rational agents with goals and purposes, and as subject to moral evaluation" (Davidson 1974b; 239) then we cannot expect mental concepts and the social sciences "to develop in ways exactly parallel to the physical sciences" (Davidson 1974b; 230). This doesn't mean that the interests and considerations that are added to those Griffiths have in mind are non-epistemic. Alertness to the project of making sense of each other is not a hindrance to understanding ourselves, but an epistemic gain.

Let us briefly return to the discussion of emotions in chapter 4. Peter Goldie's emphasis on narratives and narrative structure is very much compatible with methodological interpretationism. The claim that interpretation is irreducible to other methods involves as central points the dynamical nature and openness of interpretation, the point that no kind of evidence can assume the status as criteria, i.e. shortcut the need to consider whether the resulting description of the person is the one that makes best sense of him or her, according to normative rationality considerations of the most inclusive sort. All sorts of episodes and elements could potentially be important in the project of making sense of a person. This flexibility of elements and structure exactly parallels the flexibility of a narrative. The sense-making project is

not limited to actions and propositional attitudes. Emotions and their expressions can be made sense of by embedding them in a narrative of the person.

Interpretation is also an ineliminable part of explaining emotional phenomena by embedding them in a narrative. The need for this kind of explanation and understanding of emotions has been obscured by many emotion theorist's conflation of the distinction between an emotion and an emotional episode. Emotions (love or jealousy, for instance) may last for years, involving (in their narrative) particular emotional episodes or experiences. This is contrary to the claim by Ekman and others that emotions typically last only seconds or minutes, and are the so-called "basic emotions". Goldie argues that to avoid confusion, these short-term episodes should not be called emotions (and therefore not basic emotions). They should be given some other name if emotional episode will not do, perhaps the term affect-program response, which both Ekman and Griffiths have used (Goldie 2000; 104ff).

The narrative structure of emotions means that important aspects of them cannot be studied in the laboratory. A typical scientific study of emotion involves experiments with one antecedent event or situation and an emotional response. The requirements of replicability ensure that the experimental situation cannot be too complex. Novels and other literature yield a different dimension of understanding of people (their thoughts and feelings) by telling a complex story. A good narrative, one which properly selects which episodes to mention and which aspects to describe and how to arrange and juxtapose the elements, cannot be reduced to a single formula. What we learn about the emotions of the people involved in the narrative cannot be summarised without loss in formulas such as 'x experienced E', 'y felt F towards x', or in combinations of them. Sometimes we would be disinclined to use words that name emotions at all. After quoting at length from passages from Tolstoy's *War and Peace*, Goldie notes that there would be something wooden about any attempt to state exactly what the hero, Tushin, felt in the described situation by naming an emotion. It would be better to tell the whole story again with all its details (Goldie 2000; 71). Goldie even notes that when judgment tests (tests of how subjects classify facial expressions of emotions) are free (i.e. there is not a fixed set of emotion terms from which subjects must choose; they are asked to freely choose a single word to describe a face), subjects are often disinclined to give a single emotion word in response, preferring instead to tell a story (Goldie 2000; 90). Articulating an emotion can often best be done by telling a story.

Telling the story, or seeing and comprehending the story, means in part to be able to see what should be elements of the story and what should be left out. It is to see which elements make sense. It is to see how these elements make sense together. Telling or grasping the story is to interpret. It is to see which composition and which selection makes the best sense of a person.

Summary

I have argued for a type of anti-reductionism that I think is appropriately moderate and which accommodates scientific psychological research in a more plausible manner than what interpretationism based on the a priori theory conception of mind does. The irreducibility claims for which I have argued are compatible with the existence of psychological laws and as far as I can see with reasonable forms of scientific reduction. This contrasts with the kind of anti-reductionism that authors like Kim, Rosenberg and Antony attribute to Davidson, and which effectively eliminates the possibility of psychological laws and explanations, or in short the possibility of a science of psychology.

This modesty is necessary in its own right in order to make interpretationism into a credible position. But it also seems to be required for an adequate account of action explanation. If interpretationism is not adequate in this respect, it would falter even closer to home, so to speak.

References

Antony, L. (1989). "Anomalous Monism and the Problem of Explanatory Force" *The Philosophical Review* 98 (2): 153-187.

Antony, L. (1994). "The Inadequacy of Anomalous Monism as a Realist Theory of Mind" in G. Preyer, F. Siebelt and A. Ulfig (eds.) *Language, Mind and Epistemology*, Kluwer Academic Publishers, 1994

Bickle, J. (2003). *Philosophy and Neuroscience*, Kluwer Academic Publishers

Cartwright, N. (1983). *How the Laws of Physics Lie*, Oxford University Press

Cartwright, N. (1999). *The Dappled World*, Cambridge University Press

Child, W. (1993). "Anomalism, Uncodifiability, and Psychophysical Relations" *The Philosophical Review* 102 (2): 215-245.

Child, W. (1994). *Causality, Interpretation and the Mind*, Oxford, Clarendon Press

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*, MIT Press

Davidson, D. (1970). "Mental Events" in *Essays on Actions and Events*, Oxford University Press, 1980

Davidson, D. (1973). "The Material Mind" in *Essays on Actions and Events*, Oxford University Press, 1980

Davidson, D. (1974a). "Belief and the Basis of Meaning" in *Inquiries into Truth and Interpretation*, Oxford University Press, 1984

Davidson, D. (1974b). "Psychology as Philosophy" in *Essays on Actions and Events*, Oxford University Press, 1980

Davidson, D. (1985). "Replies" in B. Vermazen and M. B. Hintikka (eds.) *Essays on Davidson: Actions and Events*, Oxford University Press, 1985

Davidson, D. (1987). "Problems in the Explanation of Action" in *Problems of Rationality*, Oxford University Press, 2004

Davidson, D. (1991). "Three Varieties of Knowledge" in *Subjective, Intersubjective, Objective*, Oxford University Press, 2001

Davidson, D. (1995). "Laws and Cause" *Dialectica* 49 (2-4).

Davidson, D. (1999). "Reply to Jennifer Hornsby" in L. E. Hahn (ed.) *The Philosophy of Donald Davidson*, Open Court, 1999

Dennett, D. (1987). *The Intentional Stance*, MIT Press

Dennett, D. (1991). "Real Patterns" *The Journal of Philosophy*: 27-51.

Goldie, P. (2000). *The Emotions*, Oxford University Press

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

Jackson, F. (1998). *From Metaphysics to Ethics*, Clarendon Press

Johnston, M. (1985). "Why Having a Mind Matters" in E. Lepore and B. McLaughlin (eds.) *Actions and Events*, Blackwell, 1985

Kim, J. (1985). "Psychophysical Laws" in *Supervenience and Mind*, Cambridge University Press, 1993; Originally published in E. Lepore and B. McLaughlin (eds.) *Actions and Events*, Basil Blackwell,1985

Kim, J. (1993). "Can Supervenience and 'Non-Strict Laws' Save Anomalous Monism?" in J. Heil and A. Mele (eds.) *Mental Causation*, Oxford University Press, 1993

Kim, J. (1998). *Mind in a Physical World*, MIT Press
Lewis, D. (1970). "How to Define Theoretical Terms" in *Philosophical Papers, Vol. I*, Oxford University Press, 1983
McDowell, J. (1979). "Virtue and Reason" *The Monist* 62: 331-350.
McDowell, J. (1985). "Functionalism and Anomalous Monism" in E. Lepore and B. McLaughlin (eds.) *Actions and Events*, Blackwell, 1985
McLaughlin, B. (1985). "Anomalous Monism and the Irreducibility of the Mental" in E. Lepore and B. McLaughlin (eds.) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Blackwell, 1985
Ramberg, B. (1999). "The Significance of Charity" in L. E. Hahn (ed.) *The Philosophy of Donald Davidson*, Open Court, 1999
Ramberg, B. (2000). "Post-ontological Philosophy of Mind: Rorty versus Davidson" in R. Brandom (ed.) *Rorty and His Critics*, Blackwell, 2000
Rorty, R. (1987). "Non-Reductive Physicalism" in *Objectivity, Relativism and Truth*, Cambridge University Press, 1991
Rorty, R. (1999). "Davidson's Mental-Physical Distinction" in L. E. Hahn (ed.) *The Philosophy of Donald Davidson*, Open Court, 1999
Rosenberg, A. (1985). "Davidson's Unintended Attack on Psychology" in E. Lepore and B. McLaughlin (eds.) *Actions and Events*, Blackwell, 1985
Strawson, P. F. (1962). "Freedom and Resentment" in G. Watson (ed.) *Free Will*, Oxford University Press, 1982; Originally published in *Proceedings of the British Academy* **48** 1-25, 1962
Tiffany, E. C. (2001). "The Rational Character of Belief and the Argument for Mental Anomalism" *Philosophical Studies* (103): 285-314.
Yalowitz, S. (1997). "Rationality and the Argument for Anomalous Monism" *Philosophical Studies* (87): 235-258.