Norm and Failure in Mind and Meaning

Akeel Bilgrami

Columbia University

Abstract

The paper first gives an argument for the Davidsonian thesis that norms constitute the human mind. Then it shows that that thesis is better formulated by Wittgenstein rather than by Davidson himself. And finally, it uses the Wittgensteinian formulation of the thesis to establish why Davidson was right to further claim that linguistic meaning was not normative despite the human mind being normatively constituted. Through this entire dialectic of the paper, the concept of failure is made central to the argument.

Keywords: Mind, Intentionality, Meaning, Language, Causality, Normativity, Value, Agency, Science, Psychology, Failure, Is-ought, Fact-value, Disposition.

1. Introduction

Donald Davidson was a pioneer among philosophers in arguing that normativity was central to understanding human behaviour and that it was what set apart the understanding and explanation of human behaviour from how we understand and explain all other phenomena. Though it is true that long before him, philosophers, in resisting the overreaching claims of positivism, had claimed that the social and human sciences were value-laden, those philosophers had not (or at least not explicitly) made value or norm constitutive of the human *mind*.[1] Davidson made this last claim central to understanding human behaviour and saw in it the roots of what made the study of human behaviour and society distinctive.

Having argued this, in an extremely surprising and paradoxical turn, Davidson went on to refuse the idea that normativity also constituted linguistic meaning. How is this possible without inconsistency? How is it that human linguistic behaviour *in particular*, i.e., the utterances of sounds with meaning, are not normative whereas human behaviour, *in general*, is to be understood as nor-

[1] Davidson has presented this idea in many papers of which the most detailed in presentation perhaps is "Mental Events" (1970).

ISSN 2465-2334

DOI 10.14275/2465-2334/20175.BIL

matively constituted? Davidson never addressed this last question, indeed never so much as raised it.

I will argue in this paper that though Davidson was right in his claim that the mind is constituted by normativity, he had a mistaken understanding of what that claim amounted to. I will try and show this to be so by contrasting Davidson's understanding of the claim with Wittgenstein's earlier way of elaborating it. This, of course, implies that Wittgenstein was the real pioneer in making the claim but my aim is not to judge who gets the prize for having made it first, but rather to assess who gets the claim more right. Having done so, I will then present at the end of the paper something like an argument for the conclusion that Davidson makes about *meaning* not being normative, a conclusion for which Davidson never gave an explicit argument but which is available to be given if one has a proper Wittgensteinian and not Davidsonian understanding of how and why the *mind* is constituted by normativity.

The concept of 'failure' will play a central role in the way I approach my dialectic to present these ideas and conclusions.

2. Failure, Norms, and Norms of the Mind

Failure, the very idea, presupposes a *norm,* by the lights of which it gets counted as such. And so, failure, I will argue, is essential to understanding the nature of norms.

But I begin with a qualifying restriction. There is frequent talk of failure that presupposes something less (or other) than a norm, as when we speak of 'heart failure' or 'engine failure'. What is presupposed in these latter expressions cannot—strictly—be a norm because these are breakdowns or cessations of a *mechanism* (whether natural or artificial). Mechanisms are defined by the presence of a *causal disposition* or *tendency* of nature or artifice. Hearts and engines are disposed to or tend to behave in certain ways under certain conditions. Under these conditions, if these natural or constructed tendencies proceed without interruption or obstacle, they are said to be functioning 'well', but the term 'well' here is not—again, strictly—a normative assessment; it merely registers that the causal disposition has been triggered and that the tendency is unhampered. Hence, strictly speaking, when a heart or engine fails, these failures presuppose only a descriptive notion of what is 'normal', which by a familiar sort of alchemy gets transformed in our careless understanding into something prescriptive or normative, a transformation whose genealogy has been illuminatingly studied by philosophers such as Foucault and Hacking under the label 'normalization'.[2]

I will not be looking at this sort of understanding of 'failure', only at failure, *strictly so called*, where the lights by which it is seen to be a failure is a 'norm' in the full and irreducible sense of the term, not a norm that reduces, in the end, to a descriptive *tendency* of nature or artifice while presenting itself on the surface as a prescriptive and evaluative standard. I described this second-class idea of norm as part of a 'careless' understanding of the term but the point of Foucault's and Hacking's analysis is precisely that it is not careless (nor second-class) at all but

[2] Hacking's brilliant elaboration of this may be found Hacking 1990 and Foucault's pioneering formulation of it is presented most explicitly first in Foucault 1977 and 1978: vol. 1, but is sophisticated in subsequent works such as, especially, Foucault 2003a, 2003b, and 2006.

part of a long institutional and social construction that affects the disciplining both of subjects of study and often (via such study) the structures of social and political domination through what Foucault calls 'bio'-power. I have no wish to deny what they say. But since my interests here, like Davidson's, are narrower and more purely methodological, I am concerned to distinguish this understanding of norms from norms that are not reducible in this way to causal dispositions and tendency. I do not even want to deny what is surely frequent—that once these causal and empirical tendencies get 'normalized' and erected into norms, many of these norms, so erected, get a life of their own and are not reducible, even eventually, to their genealogical ground in these causal and empirical tendencies. If so, that is a normativity that we bestow on them and that is not reducible to their genealogical basis in tendency. So understood, they are indeed norms, strictly so-called, in the sense that is this paper's topic. What I am certain of, however, is that the examples I gave, our talk of engine failure and heart failure, cannot, just as they stand, possibly presuppose norms in the properly strict and irreducible sense. When there is failure to live up to norms, in that strict sense of norms, we are potentially subject to criticism. But though we may, of course, criticize some*one* for not maintaining a heart's health or an engine's operability, we do so only because there are other values and norms—the value of life, perhaps, or of material productivity—that the functioning of the heart or machine respectively make possible. That does not imply that a heart's or engine's failure, qua failure of a mechanism of nature or artifice, in itself presupposes any norm in the strict sense.

In the little space I have, I cannot argue for this fundamental distinction between tendency and norm.[3] I will simply take it for granted. The distinction, though it has not gone without being contested, is intuitive and plausible. It seems to be everywhere evident in human thought and action. When we say of two beliefs, p and q, that 'q follows from p' we could mean by this two quite distinct things: 1) in subjects capable of belief, the belief that q *tends* to follow the belief that p (as a matter of *causal disposition*) and 2) subjects capable of belief *ought* to (as a matter of *rationality*) believe that q, given that they believe that p. There are very familiar distinctions at play here: between cause and reason, fact and norm, is and ought; and, as I said, I will only consider failures in those episodes of human thought or action that fall afoul of the latter in each of these pairs of distinctions—failures of reason, doing what one ought not or not doing what one ought, violation of norms in the full and irreducible sense of the term.

Before I move on from these preliminary clarifications and distinctions, let me declare one more restriction that I will impose on this paper's theme. When studying failure the primary interest in the social sciences has, quite understandably, been human behaviour that runs afoul of *social*, *political*, *legal*, and *ethical* norms since those are the norms that these disciplines are most concerned with—speaking too loudly, as it might be, or driving on the wrong side of the road, committing perjury, breaking a promise, etc. Though my concern in what follows is of indirect relevance to explanations in the study of society, my direct and primary interest, again like Davidson's, is not in such failures but in the

[3] There has, as is well known, been an enormous amount of writing on the subject among philosophers both insisting on and (ultimately) denying the distinction. I give an argument for the distinction in Bilgrami 2006: Ch. 5.

failures of thought and action of individuals by *their own* lights, whether those lights coincide with the lights of social, political and other such norms or not. (Of course, an individual's lights by which she assesses her own behaviour as amounting to failure, may very often be an internalization of social norms. How could it fail to be? But that is the genesis of her lights. It does not spoil the idea that they are *her* lights. And it is, *as such*, that I am concerned with norms and our falling afoul of them in failure.)

An example of failures by our own lights may be found in the domain of psychoanalysis or psychotherapy. We may be ostracized or sent to gaol if we fall afoul of social or legal norms but we feel guilt or go to analysts or therapists because we are, by our own lights, dissatisfied with our minds and actions. When I suffer from an anxiety that prompts me to seek therapeutic attention (I am putting aside here anxieties that owe exclusively to biochemistry and for which we turn exclusively to pharmacological treatment), it is by some lights of my own that I find my behaviour to be wrong—and it need not be a moral, social… wrong. I mention psychoanalysis and therapy only as examples. The phenomenon of failing by one's own lights is far more common and far more unremarkable than is suggested by these examples. It is ubiquitous and mundane and, on the face of it, often much less interesting than the cases that are of interest in psychoanalysis. But its study may have fundamental implications for how to understand what is distinctive about the explanation of human behaviour. This is the point that has emerged with much force ever since Wittgenstein and then later Davidson put it on centre-stage.

With these two restrictions of my thematic interests declared—a restriction to failures owing to *strict and irreducible norms* and to failures within *individual* psychology—let me turn now to expounding how norms figure in individual human behaviour and psychology.

3. Norms, Causes, and Reasons

Ever since Aristotle codified the explanation of individual human behaviour in the practical syllogism, a varied terminology has been deployed to identify the states of mind that go into such explanations: 'beliefs', 'desires', and 'intentions' are the terms philosophers primarily deploy, though in the behavioural and social sciences with the decision-theoretic sophistications of Aristotle's practical syllogism to include *degrees* of belief and desire, such terms as 'preferences' or 'subjective utilities' (for desires) and 'subjective probabilities' (for beliefs) have been coined. In what follows, for the sake of simplicity, I will speak only of 'beliefs' as an omnibus term standing in for the cognitive states that go into the explanation of human behaviour, 'desires' as such a term for the conative states, and 'intentions' as such a term for the more decisional states. Thus, (very roughly) following Aristotle's practical syllogistic schema, someone may be said to desire that her thirst be quenched, believe that drinking a glass of water is the best way to satisfy that desire, and form thereby an intention to drink a glass of water. These states of mind, if all goes without hindrance, result in her drinking a glass of water. (Aristotle himself leaves the intention out and goes straight to action from the desire and belief.)

Where is normativity supposed to enter into all this? It is familiar from Davidson's writing that it is as follows: the relation between the states of mind (the belief, desire, and intention) and the human behaviour (drinking the water) is said *not merely to be a causal* relation with the former causing the latter, but a relation that shows the latter, the behaviour, to be *rational*, given the presence of the former, the states of mind. (Here, as I said above, the notion of rationality or reason is not derived from social or legal or political norms, but from norms that govern the relations among an individual subject's own states of mind, i.e., norms of consistency, transitivity, coherence, etc., and the relations between those states of mind and her behaviour, i.e., the norms of practical syllogistic reasoning or, in its more sophisticated form, the norms of decision theory.)

Though both the element of causality and of rationality are thus in play, the relations between these two elements need much sorting out because, on the face of it, it is not clear how exactly they relate to one another; in fact, they seem to run up against each other and the rational element seems, at least at first sight, to cancel out the causal element. Failures are essential to understanding how this happens.

Take the Aristotelian practical syllogism above as a causal claim. If it is a causal claim comparable to the causal claims of natural science, it will not restrict itself to just that particular claim about that person drinking water on that occasion but aspire to greater generality. Let us transform it to a more general claim as follows:

> The desire that one quench one's thirst, the belief that drinking water is the best way, all things considered, to quench one's thirst, and so the intention that one drink water, *cause* one to drink water.

Now, like all causal generalisations, whether in natural science or about human behaviour, this one too will have to be qualified by ceteris paribus clauses to rule out spoiling conditions that have the effect of the generalization *failing* to hold. So we will have to prefix such a ceteribus paribus clause to the causal claim above: "*All things being equal*, the desire that__ the belief that__ and the intention that__ cause__".

But now there is a problem. We have no idea of a general *sort* about what is being held steady (or equal) in the ceteris paribus clause of such a causal claim. This is because we cannot really gather the different things that *spoil* such a causal claim into general categories or sorts of spoiling conditions. On one day, someone may not drink the water because he prefers just then to drink orange juice even though he knows water is better for him (for his health, say), on another day he may not drink it because he is feeling lazy, on yet another day, he may not drink it because he gets distracted…and so indefinitely on and on. There is no common element in these spoilers that can be informatively stated *ex ante*. We wait for the failures due to one or other of an indefinite number of such spoilers and the 'all things being equal' clause rules each out *only ex post*. There really is no informative thing we can say in advance about what sorts or kinds of things cause the failure of the causal generalization to hold. In advance, at best we can say something completely *un*informative such as: If she believes__ desires__ and intends__ then, *if she is rational*, she will do__.

But that really only shows that the causal claim, qua causal claim, has no empirical weight or punch, no informative, explanatory strength and power. The 'all things being equal' clause (now taking the form of an assumption of her rationality) tells us nothing that carries *empirical* information about what will spoil the causal claim and therefore what has to be ruled out. It has a 'whatever it takes' quality to it, and the mention of rationality is an admission of the completely normative nature of these allegedly causal generalisations, merely pretending on the surface to causal explanatory power. Compare such generalisations, say, to the law of falling bodies and you immediately get a sense of the contrast between the two sorts of explanations. When, it comes to the laws of natural science, we have a relatively clear *ex ante* idea of what has to be held steady by the ceteris paribus clause. That is to say, the kinds of things that would cause the generalization to *fail* to hold are things that we have a clear understanding of in advance of those failures. We know and can state in advance the *sorts* of things that spoil the generalization from holding and so we can state informatively in advance the conditions that have to be held equal. Both the terms 'ex ante' and 'sorts' are important here in understanding failure. In explanations of human behaviour conditions under which causal claims fail cannot be *sorted* into general kinds of conditions and stated *ex ante*. They can only be observed after the fact because there is nothing that they share in common that we have a grip on in advance as a general *sortal* claim that is informative. That is why we simply appeal—a waving of the hand, as it were—in the ceteribus paribus clause to the agents' rationality to save the causal claim, but in doing so it ceases to be a causal claim with any empirical import and the normative element (of an assumption of rationality) in human behaviour replaces the causal element on centre-stage. That gives a preliminary hint of the ineliminable normativity in our understanding of human minds and behaviour.

4. Wittgenstein vs Davidson on Norms of Mind

At this point, there is an interesting disagreement that might arise between two positions on how exactly to understand this normative element that so dominates the individual human mind and what constitutes failure by the its lights.

One position was made familiar by Davidson.[4] Our states of mind such as beliefs, desires, and intentions are *causal* dispositions to behave in certain ways. But unlike the causal dispositions studied by natural science (the solubility of salt, the fragility of glass, which result, under the relevant triggering conditions—the placing in water, being struck by a rock—in salt dissolving and glass shattering), such dispositions of the mind which result in human behaviour have a further feature: they are *answerable to the norms* of logic and decision theory. So, normativity enters human mentality in the form of *general principles of rationality* (the principle of non-contradiction, the principle of transitivity, etc.) to which the mental states that are disposed to cause human behaviour are answerable.

A quite different view owes to Wittgenstein. Wittgenstein claimed[5] that states such as beliefs, desires, and intentions are *themselves* primarily normative states, not causal and dispositional states. So normativity is not restricted to general principles of rationality (to which our mental dispositions such as beliefs, desires, intentions, are answerable). It is more widespread. Our beliefs, desires, intentions are, each and all, themselves normative states. What does it mean to say that mental states are themselves normative? Wittgenstein gives the following sort of example to illustrate what he has in mind. Take intentions. If I

[4] See again Davidson 1970.

[5] See his extensive discussion of both 'intentions' and 'expectations' in Wittgenstein 1953 for this view of mental states.

intend to take an umbrella when I go to work in the morning, then, *just by forming that intention*, I have generated a norm. This norm will be the basis of an assessment of my future behaviour. If I take the umbrella, I have acted in *accord* with the norm that is my intention. If I do not take it, I have *failed* to live up to that norm. And 'accord' and 'fail', as I said at the very outset, are terms that presuppose norms. To express this normativity one could, so long as one is clear that it is not a moral ought but a broader ought of rationality, say that if I intend to do something, then I ought to do it. Failing to do it is to fail of the rationality demanded by a state of mind such as an intention. We may say similar things about desires. A desire, being less decisional (or being pre-decisional in the course that practical reason takes) than intentions, may be overridden by other desires, so the norm it generates is only a prima facie norm, a prima facie ought. If I desire that I help the poor, then prima facie (unless it is overridden by other desires) it will generate a norm that will assess my future behaviour as being in *accord* with it or not. If I give money to Oxfam, for instance, I will have acted in accord with the desire that is my norm. If I do not, nor do any similar thing, I will have *failed* to live up to the norm that is—or is generated by—my desire. (About desires in particular, I should add a caveat: the term 'desires' in ordinary talk is not always used in this normative sense but rather to stand for *urges*. When it does so, it is not in the realm of normativity. Urges are precisely and primarily tendencies and causal dispositions and lack the primary normative sense that Wittgenstein has in mind. So, the term 'desire' is ambiguous. When one says, I have a desire to smoke a cigarette, I could mean either that I have an urge or I have a commitment to smoke. I can, of course, have both an urge and a commitment to smoke, but when I do, I have two distinct states of mind, not one. In what follows when I speak of desires I am restricting myself, like Wittgenstein, to the normative rather than the dispositional sense of the term. That is only to be expected since my interest in this paper is in the phenomenon of failure and, therefore, in the normatively inflected explanation of human behaviour.) Beliefs too are norms. If I believe that there is a table in front of me, that is a norm in the sense that I ought to, it commits me to, believing and not believing a range of other things: it commits me to believing that there is something in front of me, it commits me to believing that if I run very fast into it, I will likely hurt myself,[6] it commits me to not believing that there is nothing in front of me; and so on. If I do believe (and refrain from believing) these other things, then I am in accord with the norms that are generated by my belief; if not, I have failed to live up to those norms.

I have just used the word 'commits'[7] in describing the normativity of beliefs and desires. The term is a useful one and can be wielded generally to describe

[6] Here, obviously, I am committed to this only if I have some other beliefs as well, such as that tables are standardly made of a hard substance, that hard substances on impact of sufficient velocity hurt one's body, etc. All that shows is that mental states such as beliefs and desires are possessed not singly as nuggets but are holistically and inferentially linked.

[7] The idea of commitment in the study of the human mind was first elaborated in any detail by Levi 1983 in the context of a highly sophisticated and original theory of belief revision. Since then, Brandom 1994 has also deployed it in his *Making It Explicit* within a framework of 'score-keeping' and 'entitlements', and I have discussed it along quite dif-

the normativity of all such states of mind. When I have these states of mind, they are like commitments I undertake, as it were within myself (and not within the framework of some external or social contract). They are commitments in the sense of being '*internal* oughts'. Intentions commit us to certain actions, desires commit us prima facie to certain actions, and beliefs commit us to certain other beliefs.

And so, a useful way to put the difference between Davidson's and Wittgenstein's views might be this: Davidson thinks that our states of mind such as beliefs, desires, and intentions are not themselves commitments but rather (causal) dispositions which are answerable to the only real commitments we have—to the principles of rationality such as consistency, transitivity, etc. By contrast, Wittgenstein thinks that our states of mind are themselves commitments or internal norms. Our very possession of such states of mind are *commitments* to doing and thinking various things. So for Wittgenstein, the mind is cluttered with far more norms or commitments than it is for Davidson. Davidson's view of norms is an austere one (the only commitments we have are our commitments to the principles of rationality such as consistency, transitivity, etc., all other states of mind are dispositions). Wittgenstein's is a more bloated view (each and every belief, desire, and intention is itself a commitment).

Who is right? Much may turn on the answer to this question.

Both views presuppose that *norms* are irreducible to the causal and physical states of nature as the natural sciences study them. This presupposition is common ground for both of them and it is a view that I have taken for granted too, as I said at the very outset of this essay. Both views moreover claim that the *mental states* human beings possess are also irreducible to the causal and physical states studied by the natural sciences. But, on Davidson's view these (the irreducibility of norm to nature and the irreducibility of mental states to nature) are *two different* irreducibilities, whereas on Wittgenstein's view they are the *same* irreducibility. This is because for Wittgenstein beliefs and desires and intentions are *themselves* norms or commitments on a par with legal, political, ethical norms, only restricted to individual mentality. So the irreducibility of these mental states is just a special case of the general irreducibility of norms. For Davidson, on the other hand, beliefs, desires and intentions are not themselves norms. They are causal dispositions. The only norms of mind there are, are the principles of rationality. However, our mental states, despite being causal dispositions are governed by or answerable to these principles of rationality in a way that the dispositions studied by the natural sciences are not, and that is why they are not reducible to the latter. Hence, for Davidson, there are two distinct irreducibilities—of norm to nature and of mental states to nature.

One might think Ockham's razor should be sufficient to make us favour Wittgenstein's view over Davidsons, but I think there is a deeper reason to do so. And *failure* of mind is a good way to bring out the deeper reason. Let me illustrate this with an example or scenario of failure.

Suppose I believe that p. And suppose also that, in a fit of distraction, when asked, I assent to something that implies not-p. Or to keep things simpler, sup-

ferent lines in an analysis of intentionality in Bilgrami 2006: Ch. 5. But the basic idea really goes back, as I have said above, to Wittgenstein 1953.

pose that when asked I assent to not-p. I have certainly in some sense violated a principle of rationality, the principle that demands consistency, what Aristotle called the principle of non-contradiction. Now, according to Davidson that is the only failure on my part. But is it? If it were the only failure on my part, the only instruction to me would have to be: "You are inconsistent, so either give up the belief that p or withdraw your assent to not-p". But, in the scenario as I have described it, that is altogether the wrong instruction to give me. The right instruction to give me is just simply: "Withdraw the assent to not-p." Why? Because it is my belief that is a commitment whereas, just by the way the scenario unfolds, the assent to not-p is not the expression of a commitment, so not really an expression of belief in the normative sense. *It was made in distraction*, so no commitment was really made by the assent to not-p in the way that the belief that p amounts to a commitment. The example helps to bring out the sense in which beliefs themselves are commitments. In the example there is more than a failure of consistency, more than a failure of adhering to a principle of rationality that demands consistency. There is also *a failure to live up to a commitment that is present in the very existence of the belief itself.* The inconsistency involved is not between two commitments but between a commitment and an assent that does not express a commitment, an assent that is inconsistent with the only commitment that I have, which is the belief. In other words, Davidson misunderstands the nature of the failure of mind here. The scenario demands a different understanding of my failure than his instruction to me would suggest. But his position *requires* him to give me the wrong instruction to remedy my failure since his position has no other normative resources than the principles of rationality. His position therefore cannot accurately capture the real nature of my failure. It underdescribes my failure. The weakness of his position and the strength of Wittgenstein's position emerges in the fact that the latter alone would provide the right instruction in this scenario, and that would, in turn, bring out the sense in which beliefs themselves are norms or commitments. Similar examples can be run on desires and intentions.

5. Norms, Intentionality, and Explanation

There is much significance that follows from the special nature of this failure as I have just expounded it and what it reflects about the nature of the norms of our mentality.

Let me turn to that significance by first clarifying a little bit more what is meant by saying that beliefs, desires, and intentions, that go into the explanation of individual human behaviour are norms or commitments of this kind.

What is it to have such a commitment? I have said that it is not to have a causal disposition since it is normative and norms cannot be merely tendencies of nature. What this implies, then, is that one can have a commitment (a desire that one help the poor, say) and not be disposed, even prima facie, to act as the commitment requires. When that is so, one is chronically failing to live up to a commitment, but it does not mean that one lacks the commitment—if it did, we would lose the contrast between commitment (norm) and disposition.

So, a hallmark of a commitment, as of all norms, is that they do not cease to be what they are in the presence of failure. Indeed failures are essential to them in the sense that if there is no *possibility* of failure, it is doubtful that it is a norm we are speaking about. *Norms are norms only if there is a possibility of failing* *by their lights*. If we were monsters of rationality and goodness, we would not have the normative concepts of rationality and morality.

It may seem that here we have a problem. If one can have a commitment and not be disposed to act on it, then it may seem that we cannot distinguish between someone who has a commitment but does not act on it and someone who does not have the commitment at all. In answering this difficulty, something important about the nature of commitments (or norms of mind) comes to the surface. The difficulty is removed only when we point out that all commitments—as a matter of definition—require *a conditionally formulated second-order disposition or dispositions*, which may be characterized as follows: when one has a commitment and fails to act on it, one is disposed to feelings of guilt, one is disposed to self-criticism, and disposed to efforts to try and do better by, for instance, cultivating the first-order dispositions to act in accord with the commitment. To define a commitment as requiring this second order disposition(s) does not reduce the commitment to a mere disposition or tendency because the second-order disposition cannot so much as be characterized except in terms of that commitment. So there is no elimination of the normative element, by bringing in a disposition of this kind at a higher order. Such a disposition is entirely parasitic on the existence of the commitment. (It is only if commitments were defined in terms of their corresponding *first*-order dispositions that we would be under the threat of reducing commitments to dispositions. But we have already said that there is no reducing commitments to first-order dispositions since the desire that I help the poor can exist without there existing any even prima facie disposition to help the poor.) And once we point out that commitments have this second order disposition built into them, we have answered our difficulty. We have a way of distinguishing someone who has a commitment and someone who lacks it. The subject who lacks the commitment lacks this second-order disposition possessed by the subject who has the commitment.

Following Wittgenstein, I have said that the states of mind (beliefs, desires, intentions…) that go into the explanation of what is distinctively (individual) human behaviour are normative states or commitments, and I have distinguished these from the more purely causal states or dispositions that human beings also possess. Does this mean that commitments have *no* causal power? Are they epiphenomenal, making no difference to the world, to the causation of human action. Are they merely relevant to answering the question "Is what someone did rational by her lights?" and not relevant at all to answering the question, "Why did someone do what she did?"

It would be an implausible limitation to impose on human subjectivity to make normative states of mind entirely epiphenomenal in this way. They cannot altogether fail to have some causal point. But if there is some causal relevance that commitments have, if they do make a difference to what occurs in the world, it had better not be the same sense of 'cause' that is present in the causal relations that are studied in the natural sciences since we know those to be causes in the merely (first order) dispositional sense. Does this mean we have two different notions of cause, one that is present in the explanations of physical behaviour studied by the natural sciences and a different notion that figures in the normatively inflected explanations of human behaviour? I cannot see any way of avoiding saying so.

If commitments have some causal effect, if they do make a difference to what individual subjects do, then statements of the form 'Her desire (understood as commitment, not an urge) that she__caused her to__' (or more specifically, for instance, "Her desire that she help the poor caused her to give money to Oxfam") make perfectly good sense, just as much sense as statements like "Dipping the blue litmus paper in acid caused it to turn red". But they do not make the same sense. What is the difference in sense in these two uses of 'cause'?[8]

Understanding the nature and the implications of failure, as I said, is crucial to answering this question, crucial to capturing the distinctive form of causality that is present in the explanation of individual human behaviour. The natural sciences systematize the dispositions in nature, bringing them under generalizations and laws. To state what is obvious and well-known, the objects, whose causal, dispositional properties they regiment in this way are not agents and subjects in the way that the human objects of study are. Another way to put this utterly familiar point is to say that, unlike the explanations of human behaviour, they do not study phenomena which possess a richly configured 'first person point of view'.

A first person point of view is a property possessed by creatures with sentience and the property is often described with the omnibus term 'consciousness'. Being omnibus, that term is meant to capture a wide variety of phenomena. Some creatures with sentience, however, possess a first person point of view that consists not only of consciousness in the sense of what is 'given' to their senses (this is sometimes described with phrases like 'what it's like to be') but a wider phenomenology that includes the normative states we have been discussing such as beliefs and desires and intentions. These are richer (if that is the right word) since they bring within consciousness a complex element of 'self-consciousness', as was evident when I defined these normative states or commitments as requiring *second-order* dispositions to feelings of guilt, to self-criticism, and to efforts at trying to do better to live up to the commitments. There is no attributing these normative states or commitments to a subject without also attributing these higher-order dispositions. All this may seem obvious to anyone who has reflected even momentarily on what makes the behavioural and social sciences stand apart from the natural sciences.

But what we can infer, once the obvious is recorded, is something less obvious—the following distinction about causation and failure.

Suppose we have, on the basis of observation and theory, come to some causal generalization in the natural sciences—the one about acids and blue litmus, being an example. And now suppose that in future observation this generalization begins to fail to hold, we begin to observe that acids are not causing blue litmus to turn red. We cope with this failure by loosening, perhaps even eventually losing, our confidence in the initial causal generalization, in the causal power that we once thought acids to possess. In short, failure presents a certain form of crisis. It forces refutation of initially made generalizations. That, at any rate, is a model of how natural science proceeds, and Karl Popper was, of course, its most explicit philosophical theorist. As is well-known, this model was powerfully questioned first by Duhem, then Quine, and most influentially by Kuhn who described the crisis that is generated by failures of this kind in quite

[8] I am frankly claiming here that if we take normativity seriously in the study of mind we will have to introduce a distinctive notion of cause. This is denied by Davidson and also explicitly by others who have made normativity central such as John McDowell.

different terms—not as one that forces refutation, but one that forces adjustment in theory, the addition of auxiliary hypotheses, to 'save the phenomena', as Duhem put it. These theoretical adjustments were said by Kuhn and others (such as Feyerabend) to change the meanings of the terms that went into formulating the initial generalizations, so though nothing was refuted, what transpired was a change of subject. Theories about some phenomenon were not *improved* by a process of refutation of an earlier theory and the formulation of a new theory since talk of improvement requires comparability (commensurability) of the earlier theory with the later theory. But comparability, in turn, presupposes constancy of the meanings of terms in the passage from the earlier refuted theory to the later improved theory; rather, on the Kuhnian view, the later theory was no longer theorizing about the same phenomenon. Thus, 'mass' in Einstein's physics did not mean what it meant in Newtonian mechanics. This is a quite different conception of the crisis generated by failure and a quite different fate for the causal claims that fail than is suggested by the idealized model that Popper had celebrated.[9]

However, the crisis that is generated by failure in the causal claims that traffic in irreducibly normatively constituted notions of belief, desire, intention, etc., are entirely different from *either* of these conceptions of crisis because the phenomenon being studied is possessed of a first person point of view and because the normative states in question are defined in terms of second-order dispositions of the kind I mentioned earlier. Suppose, then, that we have come to see a subject as possessing a commitment—a desire, as it might be, or an intention that she help the poor—one that we expect will cause her to do certain sorts of things: give money to Oxfam, as it might be, or to panhandlers on the street, … We have this expectation because the idea of such a commitment brings with it a causal generalization, a statement of the causal power of the commitment: "commitment__causes actions such as__". Suppose, however, that (as with the failure of blue litmus to turn red when dipped in acid) she does none of those actions. What crisis does this generate? The point of these being normative states that the causal claims traffic in, is precisely to say that nothing is refuted about the correct attribution of the commitment to the agent in question. As I said, a commitment does not cease to be a commitment if it is not lived up to, that being the nature of normative states (i.e., since the *possibility* of failure is *defined* into the kinds of state they are, the fact of failure cannot cancel the idea that a commitment or normative state of mind exists). So failure refutes nothing. Does failure lead to some more Kuhnian adjustment being made by the theorist? No, that too is not what is demanded by failure. The difference between Popper and Kuhn on what transpires upon failure is a Trotskyite difference within a shared understanding that one is dealing with phenomena that possess *no* first person point of view. What does failure force, then, when phenomena with a first person point of view are the objects of study—that is, when the objects of study are *subjects* and in particular subjects with mental states not merely understood in dispositional terms but normative terms?

[9] The classic works in this familiar territory are, Popper 1959, Kuhn 1962, Quine 1953, and Duhem 1954. Feyearbend's *Against Method* (1975) is a spirited contribution to the debate, taking a sustained polemical stance against Popper's view.

The answer to this question lies in the way we have characterized the normative nature of beliefs, desires, intentions as commitments. If, as I have insisted, these states are defined in terms of certain second-order dispositions that I had elaborated above, the failure to act in accord with a commitment, disposes the subject who has the desire or commitment to be self-critical and to try and do better by way of living up to the commitment. So the point really is that in the explanation of individual human behaviour we record that commitments do *cause* actions that are in accord with the commitments—it is just that we get a sense of the distinctness of the notion of cause that is operative here, when we also record that in the cases where they *fail* to cause such behaviour, no question of the refutation of the existence of the commitment even so much as arises, since when there is failure *the phenomenon in question* from its first person point of view *itself* strives to improve the causal power of the commitment. There is simply no analogue to this in the phenomena that the natural sciences explain. Neither Popper's nor Kuhn's account of what follows upon the crisis brought about by failure are therefore relevant.

It is not as if natural scientific explanation is not norm or value laden. How many times have we heard that when we have two equally efficacious natural scientific explanations, we choose the simpler of two natural scientific theories, and simplicity is a value! And no doubt there are more interesting forms of value than simplicity that are deployed in the natural sciences. But none of this affects the nature of the causality that figures in the natural sciences. What is distinctive about the normativity of the explanation of individual human behaviour is that, for the reasons I have been elaborating, the states of mind that go into the explaining relate causally to the behaviour in a very distinctive way, and thus they explain it in a way that has no echo at all in the natural sciences.

Though my focus has been on individual human psychology, the consequences of this distinctive form of causality and explanation has and is bound to have wide consequences for how to understand social phenomena. Drawing those consequences in detail must remain a task for another occasion.

6. Norms of Mind vs Norms of Language

I have been arguing that failure illuminates the idea of a norm and I have been looking at the distinctive way it does so in the explanation of individual human behaviour, in particular. What has been key in the analysis I have offered is that the states of mind that go into the explanation of such behaviour are normative in a sense that was first illuminatingly suggested by Wittgenstein. Beliefs, desires, and intentions are themselves normative states or commitments and this makes a vital difference to the distinctiveness of how they cause behaviour and, therefore, how we understand and explain individual human behaviour.

Wittgenstein's own example of this form of normativity was given in his account of intention which has it that an intention generates a norm, dividing actions into those are in accord with the intention and those which fail to be in accord with it (accord and failure being normative notions). Failure, then, is essential to understanding norms and vice versa. The way I have put this point is to say that *there could be no norm if there was no possibility of failing to live up to it*. I want to conclude this paper with what I think is a startling and highly revealing *exception* to this otherwise impeccable claim. I want to argue very briefly that though Wittgenstein is certainly right that intentions are normative states in the way he presents, the intentions with which individuals *mean* things with their words (in one perfectly good sense of the word 'mean' or 'meaning') *cannot fail* to be fulfilled. In short, though there can and must be failures to live up to one's intentions *in general*, there can be no failures to live up to one class of intentions, the intentions to mean something with what we say. These are a degenerate form of intention. Or if that sounds pejorative, they are a 'limiting case' of intention.

It is this last point that allows for Davidson's conclusion that *meaning* is not normative. But Davidson, who did not give any explicit argument for this quite correct conclusion, cannot really help himself to this last point since it is formulated in terms of a notion of normativity of *mind* that owes not to him, but to Wittgenstein's idea that intentions are themselves normative states, something that Davidson denies. The point, as I am making it here, is that this is correct in general of intentions, but not in particular of meaning intentions, which are a degenerate case of intentions.

Let us explore this by asking, what might amount to a failure of a meaning intention?

Perhaps something like this. I am walking down a path with a friend. I point to something and say "That's a snake" with the intention of getting him to believe that there is a snake in our path. But what is in the path is really a rope, not a snake. So I have made a mistake. A failure. But is it a failure of meaning? That cannot be right. It is a failure of other sorts. A failure of perception. A failure to utter a true statement. A failure, therefore, to communicate the facts. But did I fail to communicate the meaning I intended? No, because my meaning intention was not the intention to communicate the facts, it was not even the intention to get my friend to believe that there was a snake in our path (though I did intend that, that was not my *meaning* intention); rather my meaning intention was to use the words 'That's a snake' to *mean something in particular by my words*, viz., that there is a snake in our path, and *that* intention did not fail to get fulfilled. The fact that there was a rope in our path, not a snake, does nothing to spoil the meaning intention from getting fulfilled. The intention was impeccably fulfilled, and in fact my utterance would not have amounted to the falsehood and miscommunication of facts that it was, if my meaning intention had not been fulfilled.

And so the question arises. Can one *ever* fail to fulfill a meaning intention? What could possibly count as a failure to fulfill a meaning intention? I think, for one perfectly natural understanding of what 'meaning' is, these questions must be answered by saying 'no' and 'nothing', respectively.

Here again, as throughout the paper, it should be obvious that by 'meaning', I am interested in the actions (in this case linguistic actions) of individuals and not interested in the meanings of words as they occur in sociolects, i.e., meanings that dictionaries attempt to specify in each of their entries. If dictionaries captured all that there was to meaning, then one might contrive to say that we sometimes fail to live up to our meaning intentions. Thus, it is certainly true that when I speak a word of English, and I do not fully have a grip on what the word's meaning, *as it is given in the dictionary*, is, and my intention when I use the word is described as the intention to use that English word (*as it is elaborated in that dictionary entry*) then it can happen that I fail to fulfill that intention, so described. I do not have a full grip on what I intend. In other words I do not know what I am talking about. But even as that happens, I do have *something* in mind when I speak those words, even if it does not coincide with what the dictionary defines them to be. And if my intention were described as meaning *that*, rather than what I do not have a grip on (the dictionary meaning of the word), then that intention is fulfilled. It is *this* notion of meaning, I am saying, that cannot fail.

The fact is that I and everybody else frequently uses words in this way, in a way that departs from dictionary definitions; and moreover frequently I (and those others) are perfectly well understood as meaning what I (they) mean by them rather than what the dictionary says about what they mean. That is to say, I mean something and I am understood to mean that, even if it does not square with what the dictionary gives as the meaning. So the notion of meaning is not exhausted by the sociolectical understanding of language. A great deal of meaning is idiolectical and is understood as such in conversation as well as in writing. It is this notion of meaning that I am concerned with since it is what is most closely tied to the intentions with which individuals speak words.

Just so as to get away from the tyranny of the dictionary, let us take an example of a slip of the tongue. Suppose I utter "I am going towndown". There is no such term as 'towndown' in the dictionary. So we have here a case of meaning that is not, not even on the surface, possessed of a dictionary meaning. Still, I mean something by that utterance because I intend to mean something by it. I intend to mean by those uttered words that I am going downtown. And suppose (plausibly, as so often happens in cases of slips of the tongue) that that is exactly what I am understood to mean by my hearers. So I have both meant something in a non-sociolectical (taking dictionaries to elaborate sociolectical meanings) sense of meaning and been understood quite correctly to mean that. And moreover my utterance is not a metaphor or a figure of speech. I *literally meant* by "I am going towndown" that I am going downtown because that is what I *intended* to literally mean by my utterance. What is true is that I did not intend to utter the *sounds* I did. I intended to utter different sounds. So it is misspeaking, but *not* a mismeaning. Such failure as there was, was a failure of vocalization not a failure of a meaning intention being fulfilled.

It would be foolish to deny that there is this notion of meaning (and literal meaning at that), tied to a speaker's intentions, which is independent of what words mean in dictionaries. Even if our words, meaning what they are intended by us to mean, coincide to a large extent (as they surely will) with what the dictionaries say they mean—this is a contingent fact. It is not conceptually required if we are to mean things with our words that we chime perfectly with the entries of dictionaries.

We have then a notion of meaning around which there cannot be any failure. That is to say when I intend to mean something by the sounds I utter—for example mean that I am going downtown, by the sounds "I am going towndown"—there cannot be any failure to fulfill the intention. I succeed in meaning just what I intend and often will be understood as meaning exactly that, even if—as in this case (which is why I picked it, to show the irrelevance of dictionaries to this notion of meaning)—the word does not even exist in the dictionary.[10]

[10] Davidson 1986 discusses examples that are in dictionaries when he discusses malapropisms. And because Davidson, unlike Wittgenstein, does not think that intentions are

This kind of intention, however, is unique among intentions. All other intentions are normative, as I have argued over the preceding sections of this paper, because there is always the possibility of our failing to fulfill them. If the possibility of failure did not exist, it would be wrong to think that we are in the region of norms. But intentions to mean something by our words cannot fail to be fulfilled. Does this reveal that we have a counter example to the thesis that intentions—being norms—presuppose the possibility of failure? No, what it reveals rather is that these intentions to mean something are a degenerate case of intentions. The Wittgensteinian claim to normativity of intentions is impeccable. It just runs up against a limiting case when it comes to meaning. Failures of meaning, in one perfectly good sense of 'meaning', are impossible, even as failures of mind are ubiquitous.

How do we diagnose this unique and peculiar property of intentions when they attach to meaning? I will not be able to answer this question at any great length as I close a paper that is already too long. But I will say just this by way of a cryptic hint of diagnosis. In the ordinary case when we intend something (when we make a commitment) and we fulfill the intention or commitment, there are two acts. The intention (the commitment) and the fulfillment of it. But in the degenerate case of the intention to mean something with what we say, the intention (commitment) and its fulfillment are not two acts, but one. These are deep waters and I regret that I cannot elaborate this diagnosis more fully in the space I have here.

References

Bilgrami, A. 2006, *Self-knowledge and Resentment*, Cambridge, MA: Harvard University Press.

Brandom, R.B. 1994, *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Cambridge, MA: Harvard University Press.

Davidson, D. 1970, "Mental Events", reprinted in his *Essays on Actions and Events*, Oxford: Clarendon Press, 1980, 207-27.

Davidson, D. 1986, "A Nice Derangement of Epitaphs", reprinted in his *Truth, Language, and History*, Oxford: Oxford University Press, 2005, 89-108.

Duhem, P. 1954, *The Aim and Structure of Physical Theory*, Princeton: Princeton University Press (first published 1914).

Hacking, I. 1990, *The Taming of Chance*, Cambridge: Cambridge University Press.

Feyerabend, P. 1975, *Against Method: Outline of an Anarchistic Theory of Knowledge*, London: New Left Books.

Foucault, M. 1977, *Discipline and Punish: The Birth of the Prison*, New York: Pantheon Books.

Foucault, M. 1978, *The History of Sexuality*, New York: Pantheon Books.

Foucault, M. 2003a, *Abnormal: Lectures at the Collège de France*, 1974-1975, New York: Picador.

themselves normative states, he does not raise the question about intentions and meaning that I am raising here.

Foucault, M. 2003b, *Society Must be Defended: Lectures at the Collège de France*, 1975-1976, New York: Picador.

Foucault, M. 2006*, Psychiatric Power: Lectures at the Collège de France*, 1973-74, New York: Palgrave Macmillan.

Kuhn, T.S. 1962, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Levi, I. 1983, *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*, Cambridge, MA: MIT Press.

Popper, K. 1959, *The Logic of Scientific Discovery*, London: Routledge (first published 1934).

Quine, W.V.O. 1953, "Two Dogmas of Empiricism", in his *From a Logical Point of View*, Cambridge, MA: Harvard University Press.

Wittgenstein, L. 1953, *Philosophical Investigations*, Oxford: Blackwell.