Defining Mental Disorder

Jerome Wakefield and His Critics

The title-level DOI for this work is:

doi:10.7551/mitpress/9949.001.0001

23 Naturalism and Dysfunction

Tim Thornton

Introduction: Disorder and Naturalism

The concepts of illness, disease, and disorder all share a prima facie normative character. Even Robert Kendell, who defended a plainly factual account, conceded that appearance in his 1975 paper, "The Concept of Disease and Its Implications for Psychiatry." He writes,

> Before we can begin to decide whether mental illnesses are legitimately so called we have first to agree on an adequate definition of illness; to decide if you like what is the defining characteristic or the hallmark of disease. … By 1960 the 'lesion' concept of disease … had been discredited beyond redemption, but nothing had yet been put in its place. It was clear, though, that its successor would have to be based on a statistical model of the relationship between normality and abnormality. … But … [a statistical model] fails to distinguish between deviations from the norm which are harmful, like hypertension, those which are neutral, like great height, and those which are positively beneficial, like superior intelligence. (Kendell 1975, 309)

The normative aspect of disease is suggested in this passage by the dimension spanning harm to benefit. This is a distinction beyond mere degree of difference from a statistical norm. It is normative as opposed to merely (statistically) normal. But normative notions present a challenge for the philosophical program of placing complex concepts into a conception of nature, or "naturalizing" them as that project is usually known, especially given the most influential version of philosophical naturalism: reductionism. (In his book *Philosophical Naturalism,* David Papineau argues that its fundamental characteristic is "the thesis that all natural phenomena are, in a sense to be made precise, physical" [Papineau 1993, 1]. Hence, showing how concepts pick out real and natural features of the world involves, ultimately, reducing them to physical concepts.)

The reason that normativity presents a challenge to philosophical naturalism so understood is that, on an influential neo-Humean view, the natural world is not itself the source of normativity: thinking subjects are. As Hamlet says, on this view, "There is nothing either good or bad, but thinking makes it so." Thus, normative concepts

cannot be thought of as describing the natural world but as reflecting human subjectivity. This is clearest in cases where the normativity concerned, like Hamlet's line, takes the form of explicit value judgments.

Here is one such example in the philosophy of disorder. K. W. M. (Bill) Fulford defends an account of illness as an endogenously caused failure of ordinary doing (Fulford 1989). He argues against both Thomas Szasz, who contrasts mental and physical illness, and Robert Kendell, who assimilates them as value free, that mental illness and physical illness are *both* value terms (Szasz 1960). The idea that illness comprises an internally generated failure of ordinary doing explains its value-ladenness because the concept of *failure* itself suggests an ineliminable negative value judgment. But Fulford also argues that differences of opinion about the value judgments need not generally imply error because they do not answer to anything objective. Value judgments are projections of a subject's sentiments onto the world, and hence differences of opinions should be explored rather than corrected (e.g., Fulford 2004). Hence, the class of illnesses does not pick out anything objective. It is a reflection of both worldly facts but also subjective values about which there can be rational disagreement.

Fulford's account does not fit reductionist naturalism. Illness is not in that sense a "natural" concept but an alloy of worldly fact and human value with the latter underpinning the prima facie normative element of illness. To naturalize illness—at least in accord with the dominant reductionist reading of that term—would require some way to account for the normative dimension in value-free and naturalistic terms. That possibility is the subject matter of this chapter. To investigate its prospects, I will discuss Jerome Wakefield's influential harmful dysfunction model of disorder.

I. Wakefield's Harmful Dysfunction Model

To be clear from the start: Wakefield does *not* attempt to provide a value-free analysis of illness or disease or disorder (unlike others such as Christopher Boorse [1975] and Robert Kendell [1975], who do). (Note that although, perhaps influenced by Boorse, Wakefield talks of "disorder" rather than "illness" or "disease," he does not suggest any firm distinctions between them; he comments, "Some writers draw distinctions among *disorder*, *disease*, and *illness*. *Disorder* is perhaps the broader term because it covers traumatic injuries as well as disease/illness. I ignore these differences" [Wakefield 1992, 374]. I will follow his lead.) But he suggests that disorder can be analyzed as a conjunction of one specific value and a value-free medical science core.

On his account, the normative dimension is divided between two elements. It features in the value "harm," which forms one conjunct and helps encode the practical aims of medicine to intervene in only particular cases, the harmful ones. But it is also present in the concept of a dysfunction, which turns out to be "anchored in

evolutionary theory" (Wakefield 1999, 465). The resulting "harmful dysfunction analysis" contrasts with Fulford's analysis in that, in the latter, facts and values mingle "all the way down." By contrast, Wakefield's approach aims to characterize a purely descriptive *core* for medical science using the idea of biological functions. In other words, the normative component of any illness is divided into an irreducibly value-laden element of harm and into a deviation from a biological function, which is then reduced to, or naturalized via, descriptive biological theory. The class of biological dysfunctions is thus a natural class even if the broader class of disorder is not. The focus here is that narrow class.

The challenge of giving a descriptive, nonnormative, or nonevaluative account of function to explain this core element of disorder goes hand in hand with giving an account of *failure* of function. Only if an account can be given of what a *divergence* of the behavior of a system from its function comprises has the notion of a function that could be successfully *or* unsuccessfully executed been substantiated. But if such an account of divergence of function can be given in value-free, descriptive terms, then it would ipso facto successfully account for failure of function. Thus, it would be a mistake to assume that in characterizing disorder partly as a *failure* of function, Wakefield has already conceded the game to value theorists because "failure" is an evaluative concept as Fulford seems to suggest (Fulford 1999, 2000). If function and divergence from it can be analyzed in descriptive terms, then so can "failure" of function: it is any divergence from function. An apparently normative or evaluative concept would be reduced to a value-free descriptive analysis.

It may still seem that, in the case of function, a nonnormative descriptive account is a hopeless nonstarter precisely because the distinction between *success* and *failure* surely cannot be reduced to a purely factual or descriptive vocabulary. But just such descriptivist accounts of natural function have been proposed elsewhere as part of the wider legacy of Darwin. In the philosophy of language and thought, for example, Ruth Garrett Millikan proposes that the intentionality or "aboutness" of thoughts and beliefs can be naturalized using the notion of biological functions (see especially Millikan 1984). She argues that even conscious human purposes—paradigm instances of genuine teleology—are susceptible to this form of reductionist naturalism (Millikan 1998, 309).

Much has been written on the definition of function. Two broad approaches are perhaps most influential: the views of Cummins and Wright. Rachel Cooper summarizes their differences thus:

> Of the best known positions, those who adopt Cummins-type views (Cummins 1975) claim that the function of a sub-system is whatever it normally currently does that contributes towards the goals of a larger system. On such an account the function of the heart is to pump blood around the body, as this is what hearts currently normally do that contributes to the

> organism surviving and reproducing. On the other hand, those who favour Wright-style approaches (Wright 1973) think that the function of a sub-system is fixed by its history. In the biological domain, the Wright-function of a sub-system is whatever it was naturally selected to do. (Cooper 2007, 30–31)

Elsewhere, Cooper (2002, 268) suggests other options and suggests that there are prima facie difficulties with all of them for the analysis of disorder or disease.

For the function of *X* to be *Z*, any of the following might be considered necessary:

1. *X* was originally selected because it does *Z*.
2. In the recent past, selection has been responsible for maintaining *X* because it does *Z*.
3. Currently, selection is responsible for maintaining *X* because it does *Z*.
4. At all times, *X* has been selected because it does *Z*.

It is difficult to choose between these options as each is associated with potential problems. One of the problems that Cooper highlights is that if one opts for the original selective advantages of some trait that, as a matter of fact, now also prima facie serves another function, then failure of that current prima facie function will not count as disease. But if recent history is taken to be key, then, because human societies and technologies now affect actual reproduction, traits that might seem prima facie to be dysfunctional but that are compensated for through human intervention cannot count as diseases. I will ignore these particular difficulties here.

Both Wakefield and Millikan favor a historical approach (like Wright's) connecting functions to actual evolutionary selective histories, and as will become clearer, this is most apt for the reductionist project in question. Roughly speaking, the biological or proper function of a particular trait of an organism is what explains the evolutionary success and survival value of that trait. (In fact, Millikan defines functions within an account of reproductively established families, but the details of her theory will not matter here.)

Crucially, for the purposes of capturing the prima facie normativity of disorder (Wakefield) and intentionality (Millikan), biological functions are distinct from dispositions. The biological function of a trait and its dispositions can diverge. Engineering limitations might cause the actual behavioral dispositions of a trait to diverge from the biological function it thus only partially exemplifies. Further, the divergences might themselves be life threatening and play no positive part in explaining the value of the trait. The best explanation of the survival of that organism and those like it cites the function that helped propagation or predator evasion, for example, and not those aspects of its behavioral dispositions that diverged unhelpfully from it.

This point is sometimes put by saying that what matters is not which traits or dispositions are selected but what function they are selected *for*. The distinction between "selection of" and "selection for" can be illustrated by the example of a child's toy

(Sober 1984). A box allows objects of different shapes to be posted into it through differently shaped slots in the lid. The round slot thus allows the insertion of balls, for example. It may be that the actual balls allowed through or "selected" in one case are all green. But they are selected *for* their round cross section and not their green color. Millikan stresses the fact that the biological function of a trait may be displayed in only a minority of actual cases. It is the function of sperm to fertilize an egg, but the great majority of sperm fails in this regard (Millikan 1984, 34). Since biological functions can diverge from mere dispositions, they have extra resources necessary for accounting for the idea of *failure* of function. The distinction between success and failure of a system, organism or organ can be defined by reference to its functioning in accord with its biological function.

Wakefield's work on disorder is in the same tradition: providing a reductionist account of a problematic concept by appeal to evolutionary theory. He offers an initially distinct, but eventually similar, account of natural functions. Drawing on essentialist accounts of natural kinds, such as water or gold, he suggests that natural functions likewise have an underlying essence. Thus, natural functions are defined as sharing whatever the initially unknown essential process is, which explains prototypical nonaccidental beneficial effects such as eyes seeing. This is a surprising claim given that what unites natural kinds such as gold or water are first-order physical properties. No first-order physical properties unite natural functions. But Wakefield goes on to invoke explanation and natural selection in a much more standard way:

> A natural function of a biological mechanism is an effect of the mechanism that explains the existence, maintenance or nature of the mechanism via the same essential process (whatever it is) by which prototypical nonaccidental beneficial effects … explain the mechanism which cause them. … It turns out that the process that explains the prototypical non-accidental benefits is natural selection acting to increase inclusive fitness of the organism. (Wakefield 1999, 471–472)

Thus, like Millikan, Wakefield relies on an account of natural function drawn from explanation within evolutionary theory to distinguish those dispositions that accord with a system's naturally selected function from those that do not.

However, despite this connection between the prima facie normativity of the concept of disorder and the normativity, albeit rooted in evolutionary history, of biological functions, there remains an ambiguity between two possible reductionist aims for such an account. In the next section, I will clarify this through a detour into the philosophy of thought or mental content. I will then argue, in the following section, that an objection derived from Wittgenstein's discussion of rules threatens one version but not the other. In the final section, I will consider which aim is appropriate to naturalizing mental disorder.

II. What Kind of Reductionist Naturalism?

In this section, I will distinguish between two aims for reductionist naturalism that can be compared to two horns of the Euthyphro dilemma. But to make this more concrete, I will characterize the difference using the actual aims of two competing, but reductionist, approaches in the philosophy of content: those of Millikan, already mentioned, and of Jerry Fodor.

In the lengthy appendix to his book *Psychosemantics*, Jerry Fodor articulates a general argument for reductionism in the philosophy of content:

> I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of *spin*, *charm* and *charge* will perhaps appear upon their list. But *aboutness* surely won't; intentionality simply doesn't go that deep. It's hard to see … how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and intentional are real properties of things, it must be in virtue of their identity with (or supervenience on?) properties that are *neither* intentional *nor* semantic. If aboutness is real, it must be really something else. (Fodor 1987, 97)

The promise of the program is that intentionality itself will be fitted into a conception of nature—or "naturalized"—by a reduction to properties that are not essentially or intrinsically intentional or semantic. Since the latter are supposed to constitute the former (through identity or supervenience), it is not that they are not intentional or semantic, but they are not essentially so. Thus, the concepts in the reduction base can be understood independently of grasp of the concepts to be understood, thus serving a project of philosophical naturalism.

The work of the rest of the book looks at first sight to be a contribution to this task through the articulation of what Fodor calls a "representational theory of mind." This comprises a "language of thought" (LOT) to explain the relationships between mental representations construed as internal vehicles of mental content combined with a version of a causal theory of reference (an asymmetric dependence theory) connecting those internal vehicles to the world.

But, on reflection, while, if successful, the representational theory of mind would be a step toward a reduction of intentionality, it is not that the actual aim of the representational theory of mind as set out in *Psychosemantics* is quite as radical as the argument in the appendix. Consider the argument that mental representations must have a structure to map the structure of mental contents.

> Practically everybody thinks that the *objects* of intentional states are in some way complex: for example, that what you believe when you believe that … P & Q is … something composite, whose elements are—as it might be—the proposition that P and the proposition that Q.

> But the (putative) complexity of the *intentional object* of a mental state does not, of course, entail the complexity of the mental state itself. It's here that LOT ventures beyond mere Intentional Realism … LOT claims that *mental states*—and not just their propositional objects—*typically have constituent structure*. (Fodor 1987, 136)

The aim of the account seems to be to explain how it is possible for thinkers to think thoughts with the right systematic relations to other thoughts. It *is* possible if there are inner vehicles of thoughts with an isomorphous structure to the structure of thought and, in turn, if the syntactic properties of those vehicles mirror their semantic relations and are suitably connected to their causal properties. Fodor attempts to show that it is not mysterious—that it is natural—that creatures like us can think the thoughts we can. If I may use the phrase the "space of reasons" to stand for the rational relations between thought contents and between them and the world, it seems that Fodor takes his question to be:

- Given the space of reasons, how is it possible for creatures like us to respond to it?

His answer is a piece of a priori engineering design. We must be creatures with an innate conceptual repertoire carried by a language of thought. This is still a form of reductionist naturalism. The fact that, medical limitations aside, humans can grasp a potential infinity of thoughts and chart the rational relations between them can seem puzzling and call for philosophical attention. The representational theory of mind is an attempt to make that less mysterious by showing how it would be possible for suitably engineered creatures to have those characteristics. But it is less radical than it might be because, within the main body of the text at least, it takes the conceptual connections themselves for granted.

Millikan, by contrast, aims to do something more ambitious with her evolutionary, or teleosemantic, theory of mental content. As summarized above, she deploys a tool that seems more promising than a causal theory of reference to account for mental content because it is itself an apparently normative notion: biological or proper function. A biological function is normative because it sets a standard against which the actual behavior or dispositions of a biological trait or subsystem can be compared. She deploys this idea not just aim to explain how possessing mental content is the proper function of some cognitive system. Rather, particular representational contents are supposed to be explained in this way. The contents carried by inner vehicles are specified via the proper functions of those vehicles: that for which they are selected. Hence, the selective advantages conferred must be characterizable in nonintentional terms. The meaning or content carried must drop out of the evolutionary theory rather than be presupposed in specifying the advantage.

The aim of this analysis is thus more ambitious than Fodor's project because Millikan aims to naturalize the structure of conceptual connections or "the space of reasons" itself. Assuming that logic charts the rational connections between contents, it

is significant that Millikan claims that given a teleosemantic account, logic itself will become "the first of the natural sciences" (Millikan 1984, 11). So her key question is something like:

- Given our biological natures, how is it possible for creatures like us to respond to what we take to be the space of reasons, whatever it is?

The difference in actual ambition between Fodor and Millikan is akin to the Euthyphro dilemma. Given a suitable theology, the following biconditional would be true:

- For any act *X*: *X* is pious if and only if *X* is loved by the gods.

The dilemma stems from considering the "order of determination," in Crispin Wright's phrase, of this biconditional (Wright 1992). Is the pious loved by the gods because it is pious, or is it pious because it is loved by the gods? Fodor's and Millikan's projects take opposing views. In effect, Fodor adopts the first horn and derives a priori engineering constraints on the gods (or thinkers) given that we know that they are able to track piety (or the space of reasons), antecedently understood. Millikan, by contrast, adopts the second horn and aims to explain piety (or the space of reasons) by describing the engineering of the gods (or thinkers) in independent (of piety or the space of reasons) evolutionary terms.

This distinction matters to the force of an objection that can be raised against Millikan's program, which I will now summarize.

III. A Wittgensteinian Objection to Millikan's Project

There is a familiar objection to Millikan's program based on Wittgenstein's rule following considerations. It is tempting to think that meaning or mental content needs some sort of vehicle such as sign or symbol. Any such sign can, however, be interpreted in an unlimited number of ways and thus needs to be coupled with the correct interpretation. This point is often emphasized by commentators by suggesting interpretations of even extended demonstrations by example of the meaning of words or of mathematical series that are consistent with the examples given but that deviate or are "bent" in some future application (e.g., Blackburn 1984). But if mental content is explained as a mental sign that stands in need of the correct interpretation, then the content of the interpretation will also need to be similarly underpinned. And this initiates a vicious infinite regress.

> "But how can a rule shew me what I have to do at *this* point? Whatever I do is, on some interpretation, in accord with the rule."—That is not what we ought to say, but rather: any interpretation still hangs in the air along with what it interprets, and cannot give it any support. Interpretations by themselves do not determine meaning. (Wittgenstein 1953, §198)

The same goes for accounting for understanding written or spoken signs or symbols. In the absence of any coherent account of a final interpretation that somehow blocks

the regress, any account of mental content that depends on an interpretation faces a challenge.

Millikan's teleosemantic account of mental content is a form of interpretation-based theory. Past behavior is a set of signs to be interpreted. Like the interpretation of signs, such behavior is consistent with an unlimited number of possible functions or rules including both continuations that seem natural and logical and an unlimited number of other "bent" rules that deviate in unnatural ways. The normativity of a function implies that not every aspect of the behavior of a trait or subsystem needs to match the function: what the trait is for. A trait may fail to match the function for which it was selected. What ensures the determinacy of biological function—what selects just one of the possible rules—is a particular *explanation* of the persistence of trait over evolutionary time. If the potential rewards of a trait are sufficiently great, it may be that actual behavioral dispositions of previous instances of the trait only rarely match the function that explains the trait's persistence. Hence the potential gap between actual past performance and the appropriate functional explanation. But the lesson from the discussion of "bent" rules is that finite past behavior could be explained as exemplifying many different or "bent" functions, all of which would have been equally successful in the past but that would diverge in the future. (Note also that this worry is not merely a kind of Quinian marginal indeterminacy akin to the difference between rabbits and undetached rabbit parts. Competing bent rules might be utterly different in future applications. By what principle is just one selected? I will return to this thought at the end.)

Millikan considers and responds to this objection in "Truth Rules, Hoverflies, and the Kripke-Wittgenstein Paradox." Male hoverflies spend their time hovering and waiting for female hoverflies to pass by at which point they accelerate in pursuit. "The geometry of motion dictates that to intercept the female the male must make a turn that is 180° away from the target minus about 1/10 of the vector angular velocity (measured in degrees per second) of the target's image across his retina" (Millikan 1993, 219). Millikan calls this the "proximal hoverfly rule" and suggests that male hoverflies are genetically programmed to follow it. That is, they have some internal mechanism "of a kind that historically proliferated in part because it was responsible for producing conformity to the proximal hoverfly rule, hence for getting male and female hoverflies together" (Millikan 1993, 219). So the biological function of the mechanism is to follow that rule.

But the behavior of actual hoverflies may not accord with just that. One possibility is that hoverflies have some optical blind spots such that a female arriving in the blind spot of a male provokes no reaction. Such a possibility, however, would not be part of what explains the continued existence of the mechanism in the fly population. It would be noise rather than signal. The more worrying possibility is a rule that accords with all past successful fly-on-fly action but that diverges in the future. What in the

evolutionary historical record could rule that out? Millikan dismisses such possibilities as follows:

> [The "bent" rule] is not a rule the hoverfly has a biological purpose to follow. For it is not because their behaviour coincided with that rule that the hoverfly's ancestors managed to catch females, and hence to proliferate. In saying that, I don't have any particular theory of the nature of explanation up my sleeve. But surely, on any reasonable account, a complexity that can simply be dropped from the explanans without affecting the tightness of the relation of explanans to explanandum is not a functioning part of the explanation. (Millikan 1993, 221)

This is rather a brisk response. A key element of it is that the bent rule contains a complexity that Millikan's preferred explanation lacks. Her explanation is simpler. But from what perspective is her explanation simpler?

In the case of trying to naturalize mental content, her explanans is the meaning or content of particular vehicles of content. But she cannot invoke our prior grasp of what such contents seem more natural—of what it would be natural to *mean*—since that is what is supposed to drop out of, rather than being presupposed by, her analysis. And, of course, the content of the proper function is not just a matter of looking to behavioral dispositions but selecting a function that best explains them. But without presupposing the pattern that meaning imposes, the pattern of the space of reasons, what other principle is there to say what makes for a simpler explanation?

Millikan's response would be legitimate for an attempt to answer the question I have suggested that Fodor attempts (despite his appendix) to answer:

- Given the space of reasons, how is it possible for creatures like us to respond to it?

Fodor's implicit question presupposes, rather than seeks to explain, the pattern of normative liaisons between mental contents. Answering that question, it is *not* illicit to deploy a notion of simplicity that presupposes a prior grasp of the space of reasons because that is not what the question seeks to answer. But Millikan's question is more ambitious, and hence it is illicit in attempting it to presuppose a notion of simplicity that is based on prior grasp of the space of reasons.

IV. Biological Function and Illness, Disease, and Disorder

What, then, is the reductionist aim of appealing to biological functions in the philosophy of mental disorder? Two options can be articulated by translating from Fodor's and Millikan's questions in the philosophy of content. I suggested that Fodor's question was:

- Given the space of reasons, how is it possible for creatures like us to respond to it?

Translated into the context of naturalizing disorder gives something like:

- Given the concept of disorder, how is it possible for creatures like us to suffer it?

Unlike the parallel question in the philosophy of content, however, this question does not seem worth an a priori answer. I do not mean to presuppose an articulated theory of what is, and isn't, worth philosophical attention. But there seems no pressing need to articulate a general theory of a failure to meet a normative standard by contrast with the felt need in the parallel semantic case to articulate a general theory of how it might be possible to meet one. Thus, this question is not an appropriate way to model reductionism about the concept of mental disorder. What of the other option?

Millikan's question was:

- Given our biological natures, how is it possible for creatures like us to respond to what we take to be the space of reasons, whatever it is?

Translated into the philosophy of mental disorder gives a question something like:

- Given our biological natures, how is it possible for creatures like us to suffer what we take to be disorder, whatever it is?

Millikan's looks the better model question for the philosophy of medicine. It makes questioning the nature of the concept of illness itself central rather than something presupposed.

It may be objected, however, that this cannot apply to Wakefield's analysis of disorder as harmful dysfunction because he relies on an approach that he describes using the phrase "black box." The notion of a natural function is defined as sharing whatever the initially unknown essential process is that explains prototypical nonaccidental beneficial effects such as eyes seeing. It merely transpires—it comes as an a posteriori discovery—that the process that explains the prototypical nonaccidental benefits is natural selection acting to increase inclusive fitness of the organism. And hence, the objection might run, this cannot be used to shed light on what is meant by the kind of dysfunction that underpins—when conjoined with harm—the idea of a disorder.

That objection goes too quickly, however. Although it is true that Wakefield does not engage in traditional conceptual analysis to underpin his conception of function, and hence dysfunction, that fact does not rule out the use of the supposedly empirically derived conception to shed light on the nature of function itself. It does not imply that he has to restrict himself merely to the kind of a priori engineering that Fodor attempts.

Furthermore, there is positive reason to think that he does more. One of the virtues that Wakefield claims for his analysis is that it is able to hold contemporary psychiatric taxonomy to account. In his persuasive coauthored book (with Allan Horwitz) *The Loss of Sadness*, for example, he argues that depression is overdiagnosed because "normal sadness"—that is, sadness that is in accord with human emotional biological function—is mistaken for genuine, pathological depression (Horwitz and Wakefield 2007). But since, he argues, mental disorder presupposes an underlying biological dysfunction, however unpleasant grief is, for example, it cannot be a disorder or an illness

providing it is serving its (presumed) biological function. Given that the analysis is used to explain the very idea of disorder (via the underlying notion of dysfunction), the form of reductionist naturalism belongs to the more radical second horn.

But if so, because it shares the task of naturalizing the normativity of pathology, Wittgenstein's objection is a serious objection. That is, a biological teleological account cannot rule out wildly divergent accounts of the functions in play, functions that explain the presence of traits. And if so, we need a better version of naturalism for the philosophy of disorder.

V. A Quinian Response?

My attempt to shed light on the nature of, and hence prospects for, reductionism in the philosophy of disorder has turned on an analogy between it and the philosophy of content and the prospects for reductionism about semantics. But it might be objected that there is a key disanalogy between the two cases.

Consider an argument often compared with Wittgenstein's discussion of rules and meaning: Quine's argument for the indeterminacy of translation (Quine 1960. (Since the relation between indeterminacy of translation, inscrutability of reference, and holophrastic indeterminacy is subject to interpretation and debate within Quine scholarship, a rough summary of one aspect of the argument will suffice.) Quine approaches the study of meaning via a particular methodological constraint.

> In psychology one may or may not be a behaviourist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behaviour and having his own faltering verbal behaviour observed and reinforced and corrected by others. … There is nothing in linguistic meaning, then, beyond what is to be gleaned from overt behaviour in overt circumstances. (Quine 1987, 5)

This reflects a commitment to the idea that facts about meaning are public and shareable. But Quine goes further in limiting the kind of facts available to fix the facts about meaning to those that fit a particular scientistic worldview. The project is constrained by connecting prompted ascent to sentences with environmental stimuli physicalistically described. These facts, however, underdetermine the translation of sentences. Since they are the only relevant facts, this implies that sentence translation is indeterminate.

With this argument in place, one might object that the comparison deployed so far between meaning and the concept of disorder is inappropriate. While Quinian indeterminacy is revisionary of our prephilosophical concept of meaning, the concept of disorder can more readily tolerate some such slack. So on the assumption that the Wittgensteinian argument outlined in previous sections sufficiently matches the Quinian argument just sketched, and on the assumption that some degree of indeterminacy is

no threat to the medical concept of disorder, the claim that the project of reducing the concept of disorder to naturalistic terms introduces an element of indeterminacy is no threat to the project.

Such a response fails, however, because the supposed parallel between Wittgenstein and Quine is misleading. Quine accepts a degree of indeterminacy in his positive account of meaning. He thinks that the evidence that fixes meaning only goes so far. This is, of course, because Quine builds in an assumption that the evidence has to be physicalistically described. Against that background, meaning would be indeterministic. But what, aside from scientism, justifies that restriction?

Wittgenstein, by contrast, does not think that meaning is indeterministic. The contexts that play a role in constraining it are described in intentional terms. The apparent parallel just sketched between Quine and Wittgenstein is not part of the latter's positive account but rather a reductio ad absurdum of reductionism. More significant, however, is that Wittgenstein's negative argument does not deliver merely a domesticated indeterminacy but rather no shaping of content in the future at all and hence undermines the very possibility of radical reductionist naturalism in the case of disorder too.

Conclusion

I suggested that if the Wittgensteinian argument against the more radical reductionist aim is successful while there is no rational reason to pursue the more modest aim, then the philosophy of disorder needs a better conception of philosophical naturalism. Fortunately, there are other approaches that can still justifiably claim to be forms of philosophical naturalism.

One can, for example, sketch the broader context in which apparently puzzling concepts are used in such a way as to make their use clear. One example of this is the way that Daniel Dennett attempts to demystify the mental by describing the "intentional stance" within which mental properties are characterized (Dennett 1991). Dennett's aim is, by describing how the stance works and by suggesting that it is merely one of many possible stances for making sense of the world, to show how the properties so ascribed are perfectly natural even though they cannot be reduced to the properties deployed in other stances, such as the physical stance.

Dennett's approach is one version of naturalism without reductionism. It helps to highlight the key assumption behind Fodor's argument for reductionism quoted earlier in this chapter. The passage starts with an appeal to the shape of a future, completed physics. That serves as the benchmark of the really real and hence prompts the challenge for puzzling concepts. If they do not appear on the ultimate list, then either they mark an unreal property or they must be reducible to concepts on the list. The

challenge, however, presupposes without justifying the assumption that the physicists' list is such a benchmark and that assumption can be contested.

John McDowell, for example, has suggested that nature is not restricted to what can be described using the vocabulary of the physical sciences. Criticizing a reductionist construal of naturalism, McDowell, for example, says,

> What is at work here is a conception of nature that can seem sheer common sense, though it was not always so; the conception I mean was made available only by a hard-won achievement of human thought at a specific time, the time of the rise of modern science. Modern science understands its subject matter in a way that threatens, at least, to leave it disenchanted. … The image marks a contrast between two kinds of intelligibility: the kind that is sought by (as we call it) natural science, and the kind we find in something when we place it in relation to other occupations of 'the logical space of reasons,' to repeat a suggestive phrase from Wilfrid Sellars. If we identify nature with what natural science aims to make comprehensible, we threaten, at least, to empty it of meaning. By way of compensation, so to speak, we see it as the home of a perhaps inexhaustible supply of intelligibility of the other kind, the kind we find in a phenomenon when we see it as governed by natural law. It was an achievement of modern thought when this second kind of intelligibility was clearly marked off from the first. (McDowell 1994, 70–71)

McDowell commends a different response to the prospects of a failure of reductionism. Rather than regarding this as impugning the reality of the properties or concepts concerned, it may merely show that reductionists have started with an impoverished conception of the real or of nature. Central to McDowell's picture is the possibility of undermining a dualism of normativity and nature. Nature itself may contain norms; it may be "fraught with ought" in Sellars's phrase. It is not restricted to what fits within the "realm of law" articulated by the physical sciences but also includes those emergent patterns and properties that have to be fitted within a different pattern of intelligibility: the "space of reasons." This phrase marks in McDowell's work (following Sellars) the rational pattern of intentional states (broadly construed to include ethical demands; it is thus comparable with but broader than Dennett's intentional patterns [Dennett 1991]). But an analogous conclusion could be drawn for other, nonintentional, concepts for which naturalists have also attempted philosophical reduction (e.g., necessity, causality). Again, the failure of an attempt at reduction in these cases need not undermine their reality.

McDowell's views are influenced by a reading of Aristotle's ethical views. He argues, elsewhere, that that both moral values and also secondary qualities form part of the fabric of the world (McDowell 1983). The suggestion is that there may be features of the world for which one needs a particular kind of mind, perhaps formed partly as the result of training, to detect, respond to, and even to conceptualize. Thus, one needs an appropriate moral education to understand, be sensitive, and resonate, to the demands that, say, kindness, makes on one in particular circumstances (McDowell 1979). But

just because these are demands that can be understood only from such a perspective does not undermine their basic reality.

Applied to the concept of disorder, such a conception of philosophical naturalism suggests a different approach to the prima facie normativity of mental illness and disease from reducing it. By contrast with apportioning it either to the irreducibly normative value of harm or to functions anchored, descriptively, in evolutionary theory, a more relaxed conception of nature allows that the normativity may be both more complex but no less part of the natural world. This has an important consequence. A substantial theory of disorder such as Wakefield's forges a connection between it and dysfunction. On a reductionist interpretation, the latter concept has to be understood independently of, and hence shed independent light on, the former. That connection, however, may be informative even without the reductionism. Our grasp of disorder may contribute to our grasp of function and dysfunction and vice versa. And hence Wakefield's analysis of the difference between, for example, sadness and depression may remain suggestive and helpful even in a new philosophical setting.

Acknowledgments

This chapter was written while a fellow of the Institute for Advanced Study, University of Durham. My thanks both to the IAS, Durham, and the University of Central Lancashire for granting me research leave.

References

Blackburn, S. 1984. The individual strikes back. *Synthese* 58(3): 281–301.

Boorse, C. 1975. On the distinction between disease and illness. *Philosophy and Public Affairs* 5(1): 49–68.

Cooper, R. 2002. Disease. *Studies in History and Philosophy of Biological and Biomedical Sciences* 33(2): 263–282.

Cooper, R. 2007. *Psychiatry and Philosophy of Science*. Acumen.

Cummins, R. 1975. Functional analysis. *Journal of Philosophy* 72(20): 741–765.

Dennett, D. 1991. Real patterns. *Journal of Philosophy* 88(1): 27–51.

Fodor, J. A. 1987. *Psychosemantics*. MIT Press.

Fulford, K. W. M. 1989. *Moral Theory and Medical Practice*. Cambridge University Press.

Fulford, K. W. M. 1999. Nine variations and a coda on the theme of an evolutionary definition of dysfunction. *Journal of Abnormal Psychology* 108(3): 412–420.

Fulford, K. W. M. 2000. Teleology without tears. *Philosophy, Psychiatry, and Psychology* 7(1): 77–94.

Fulford, K. W. M. 2004. Ten principles of values-based medicine. In *The Philosophy of Psychiatry: A Companion,* J. Radden (ed.), 205–234. Oxford University Press.

Horwitz, A. V., and J. C. Wakefield. 2007. *The Loss of Sadness*. Oxford University Press.

Kendell, R. E. 1975. The concept of disease and its implications for psychiatry. *British Journal of Psychiatry* 127(4): 305–315.

McDowell, J. 1979. Virtue and reason. *The Monist* 62(3): 331–350.

McDowell, J. 1983. Aesthetic value, objectivity, and the fabric of the world. In *Pleasure Preference and Value,* E. Schaper (ed.), 1–16. Cambridge University Press.

McDowell, J. 1994. *Mind and World*. Harvard University Press.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

Millikan, R. G. 1993. *White Queen Psychology*. MIT Press.

Millikan, R. G. 1998. In defence of proper functions. In *Nature's Purposes: Analyses of Function and Design in Biology,* C. Allen and G. Lauder (eds.), 295–312. MIT Press.

Papineau, D. 1993. *Philosophical Naturalism*. Blackwell.

Quine, W. V. O. 1960. *Word and Object*. MIT Press.

Quine, W. V. O. 1987. Indeterminacy of translation again. *Journal of Philosophy* 84(1): 5–10.

Sober, E. 1984. *The Nature of Selection*. MIT Press.

Szasz, T. 1960. The myth of mental illness. *American Psychologist* 15: 113–118.

Wakefield, J. C. 1992. The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist* 47(3): 373–388.

Wakefield, J. C. 1999. Mental disorder as a black box essentialist concept. *Journal of Abnormal Psychology* 108: 465–472.

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

Wright, C. 1992. *Truth and Objectivity*. Harvard University Press.

Wright, L. 1973. Function. *Philosophical Review* 82(2): 139–168.