The Concept of Mental Dysfunction: A Kantian Critique

Neil Bernard Annett
Registration number 100188147/2

PhD Philosophy

University of East Anglia

School of Philosophy

September 2023

The Concept of Mental Dysfunction: A Kantian Critique

Abstract

My thesis investigates a fundamental presupposition in psychiatric theory and practice: that the symptoms of mental disorders reflect an underlying dysfunction. That is, mental disorder arises from the failure or impairment of one or more mental mechanisms, and is therefore primarily a problem internal to the disordered individual, with external factors being seen as secondary.

I concentrate on what is perhaps the most influential attempt to define the concept of mental disorder in this way, Jerome Wakefield's harmful dysfunction analysis (HDA), and examine it in light of Immanuel Kant's *Critique of Pure Reason*. I first chart Kant's interest in mental disorder and concerns about metaphysics, which I argue are important for the development of the *Critique*. Following this, I draw upon Kant's account of human cognition and its *a priori* conditions of possibility in order to subject the HDA to a critique that questions its status as a purportedly objective analysis of the mental disorder concept.

It transpires that Wakefield's analysis is motivated by – partly ethical – concerns prompted by critics within the so-called 'antipsychiatry' movement. I argue that while his worries are justifiable, his response also has ethical implications as a result of defining mental disorder as a categorical break with nondisorder. I present an approach to understanding one particular disorder, schizophrenia, through a broadly Kantian dialectic of the self, and argue that on this view even this most perplexing of mental disorders can be understood in terms of confusions that are basic to all human experience, and the schizophrenic can come to seem less decisively 'other'.

Contents

Introduction

This thesis is about the problem of defining the concept of mental disorder, particularly as it stands in the context of psychiatry. Although madness and melancholy have been topics of interest to physicians since antiquity, the focus has traditionally been on individual conditions or clusters of symptoms, rather than on the overarching concept that we today call mental disorder. Indeed, it is only in the post-war period that serious attempts to define mental disorder as such have been made. As we shall see, the same period has also witnessed outright denials that mental disorder exists.

My project is to query a particularly influential approach to this problem that might broadly be referred to as the 'dysfunction analysis'. This is an attempt to provide mental disorder with a naturalistic definition, i.e., one that purportedly makes no reference to subjective values. Although these analyses differ in their details, the principle that unites them is that genuine mental disorders are taken to be failures of 'natural functions', whether these be conceived in physiological or psychological terms. Borrowing freely from a long-standing debate within philosophy of science, it is argued that the parts of living organisms possess functions, which is to say that they produce effects that appear to fulfil a purpose – ultimately, that of maintaining the organism as what Carl Hempel called "a going concern" (Hempel 1965, 305).

The details are, as we might expect, complex. The general idea, however, is that we can distinguish between merely accidental effects, such as the rhythmic thumping sound made by the heart, and functions, such as the heart's pumping action that circulates blood around the body. It is claimed, moreover, that the judgements we make in identifying natural functions are about how things stand independent of any observer, and not merely about the effects we value. That the function of the heart is to pump blood reflects an objective fact about reality, rather than our understandable tendency to esteem the pumping of the heart for its contribution to our continued existence.

Per the dysfunction analysis, in order to decide that something is a genuine medical disorder, rather than a natural variation or a reaction to transient external factors that we happen to think needs correction, we need to determine that it is a failure or deficiency of a natural function. Since it is purportedly a matter of scientific fact as to which effects are functions, our attributions of disorder will then be entirely independent of our opinions, and disorders will be legitimate targets for medical correction, since this is the accepted task of the profession. This is not to say that physicians should be barred from treating conditions that are not disorders, thus defined, where it seems appropriate. However, questions as to what, outwith the category of genuine disorders, should or shouldn't be treated, must be argued on other grounds. Physicians may query the need to perform a cosmetic procedure, but not whether a broken arm needs setting.

But why is the problem of defining medical disorder a problem at all? Historically, it has elicited very little interest in contrast to the question of identifying and describing particular maladies. What counts as disorder often seems self-evident. The outward signs of physical disorder can be observed, and are often consistent in their presentation. Today, distinct pathophysiologies that can confirm diagnosis have been discovered for most physical disorders. To formulate a precise definition has not seemed like a pressing requirement at all. The same is clearly not true for mental disorders, however. Although research has revealed numerous anomalies correlated to mental disorders, to date no clear biological signs or pathological processes by which we can identify mental disorders have been discovered. Although there is almost universal agreement that they must ultimately be associated with physical states, this is of little use given our current state of knowledge. To all intents and purposes, observable behaviour and reported thoughts and moods are all that *constitute* mental disorders.

For reasons that I shall explore in chapters 1 and 2, in the post-war period this came to be perceived as a serious problem for psychiatry, a problem framed in terms facts and values. For some, the dysfunction analysis seemed like the solution. Simply put, whether or not physical causes can be pinpointed, the concept of natural function provides mental, as well as physical, disorder with a purely factual definition – or so it is argued. The role of dysfunction has secured a special place in psychiatry as a result of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), a publication of the American Psychiatric Association. Starting with its third edition in 1980 it has contained, in its introduction, a general definition of mental disorder. Although this has gone through changes with each new edition, internal or 'organismic' dysfunction has been and remains a key element of the definition. The DSM, although an American publication, has become the *de facto* global standard clinical manual for psychiatry[1] and it

[1] The authors of DSM-IIIR, a transitional revised text published in 1987, note that the third edition had already been translated into 13 languages, and that the World Health Organisation had adopted many of the diagnostic features of DSM-III in the mental disorders chapter of its own International Classification of Diseases (ICD) – which is technically the worldwide standard (*Diagnostic and*

has earned a degree of public interest and awareness that is highly unusual for a work of this type. As such, its definition of mental disorder is a powerful endorsement of the dysfunction analysis.

All this notwithstanding, the analysis is by no means universally accepted. Inevitably, philosophers have disagreed about the role of dysfunction in the concept of disorder, as well as about the concept of natural function itself. In this thesis, I will focus primarily on one of the foremost analyses of ‘disorder’, Jerome Wakefield’s *harmful dysfunction analysis* (HDA), which has gained a formidable reputation since its first appearance in 1992. I will ask whether the definition yielded by the HDA is really as objective as is claimed. While many others have argued against Wakefield’s point of view, my approach is novel in that I will use Immanuel Kant’s critical philosophy as a framework within which to critique his analysis. I will survey Kant’s distinctive view of concepts and definitions in his *Critique of Pure Reason* (CPR), and look at how the notion of ‘dysfunction’ fares in his epistemological schema.

The core question I will pose in this thesis is, therefore, whether naturalist attempts to analyse the concept of mental disorder can succeed in providing us with a value-free definition. Via Kant’s critical philosophy, I will suggest that they cannot. While I have chosen to focus upon Wakefield’s HDA, I believe the Kantian critique will count against any attempt to provide a purportedly objective concept of mental disorder from which subjective values can either be eliminated or isolated. It must be said that there are very few alternatives to Wakefield’s account; the most prominent of these, that of Christopher Boorse, I will assess in chapter 1. Indeed, between them, Boorse and Wakefield dominate the naturalist camp, with other contributors to the debate refining their ideas rather than proposing substantially novel positions.

The key Kantian point is that the concept of natural function – and thus dysfunction – that is fundamental to naturalist analyses runs up against the epistemological constraints of his transcendental philosophy. For Kant, function cannot be considered an empirical concept, since these must only contain properties available to us through sensible intuition. To that extent, empirical concepts make it possible to *experience* objects, or are, in his terminology, ‘constitutive’ of them. Function as such is not a sensible predicate, and cannot fulfill that role. It is, rather, a subjectively occasioned ‘idea of reason’.

Statistical Manual of Mental Disorders: DSM-III-R 1987, xvii).

Understood properly, according to Kant's critical doctrine, such an 'idea' is a legitimate aid to empirical investigation, but should not be taken to refer to a mind-independent facet of reality. It sets us a goal for inquiry, albeit a goal that we should only expect to approach asymptotically rather than establish with certainty.

This element of Kant's mature thought is important in so far as it limits what he took to be an unfortunate, if almost inevitable, tendency in human thought: that of attributing objective reality to subjective conjectures. This tendency was, in his day, prominent in the speculations of rationalist metaphysics. Metaphysical activity of this sort has been virtually expunged from contemporary philosophy, largely due to Kant's own influence, but, despite the well-known 20th century attempt to eliminate metaphysics from science, some such presuppositions have turned out to be indispensable for scientific investigation and theory – as Kant himself believed they had to be. It was his attempt to reconcile this with his empiricism that distinguishes him from Hume, whose scepticism was an important influence upon the *Critique*. For Kant, ideas are entirely justifiable methodological devices, even though they postulate entities that can never enter empirical experience. In the section of the *Critique* in which Kant discusses the role of reason in science, the Appendix to the Transcendental Dialectic, he insists that they have a 'regulative', rather than constitutive, role, serving as prescriptions to seek certain kinds of empirical data rather than playing any role in supplying that data. The details, as we shall presently see, are subtle, and have – like so much else in the Kantian corpus – divided scholarly opinion.

On the Kantian view, therefore, function and dysfunction would be valid in so far as they are ideas generated by reason in its hypothetical use. If we want to know whether mental disorders are caused by defective or failed 'mechanisms' (whether at a neurobiological or a more abstract psychological level), we need to have a clearer understanding of what it means to be a 'mechanism' in this sense – what it means for something to have a function or purpose. Only then can researchers have a determinate grasp of what empirical investigation should be looking for, or what will count as an underlying cause of mental disorder. As hypothetical, however, such ideas are to be distinguished from the concepts for which they are postulated as causal grounds. Above all, ideas do not license any presupposition regarding the concept and may, indeed, exist alongside ideas that posit different grounds for the same concept. Ideas of reason must win our backing only to the extent that evidence is found to support them. As we shall see, however, the Kantian interpretation of function ill-serves the naturalist project.

While I may be sceptical about naturalised definitions of mental disorder, I will not argue that we should abandon attempts to get clearer on our concept, such as it is. The primary problem, as I see it, is with the insistence on definition. Here, too, I will draw upon the resources of the *Critique*. Kant's opinions regarding the analysis and definition of concepts is distinctive: he argues that definitions, properly speaking, can only be given for a restricted class of concepts, namely, mathematical ones. Because mathematical objects are represented in thought (and only – at best – imperfectly given in experience), their properties can be known completely. Our cognitions of the objects to which our empirical concepts refer, by contrast, can always reveal properties previously unknown to us; conversely, we may find that we were mistaken about properties we had thought necessary to the concept. His view, therefore, is that our empirical concepts are never fixed, but are always subject to potential revision in light of new experience. Bringing clarity to our concepts, therefore, is a matter of what Kant calls explication, rather than definition. I would argue, in addition, that this process of explication must, in the case of mental disorder, include an element of negotiation. This is not, however, a project I will undertake here.

I do not think that there is an insuperable problem with the notion of mental disorder as such. However imperfectly, it does pick out a phenomenon that causes widespread suffering, and grounds our therapeutic responses. At the same time, however, judgements of mental disorder themselves can and do have problematic implications for people so judged and those around them. The highly contested nature of the concept is doubtless partly due to this. The correct response, I believe, is to follow Kant's counsel and take a more cautious and exploratory approach, refining our concept rather than attempting to win backing for a definition intended to settle the matter once and for all – whether in the name of purported scientific facts or not. In the final chapter of this thesis I will discuss some of the problems that the definitional approach may exacerbate, and how a serious engagement with the phenomenology of mental disorder could offer a more constructive response.

My choice of Kant is neither arbitrary nor simply due to his enormous influence upon philosophy in general. It happens that from early in his career Kant repeatedly expressed the concern that the theories of rationalist metaphysics displayed an unnerving similarity to the conjectures of the mentally disordered. This worry, as I will argue, motivated the *Critique*, with its groundbreaking arguments against the legitimacy of the metaphysical tradition. Furthermore, he harboured an interest in mental disorder itself, an interest that emerged in his early work and appears again in his late *Anthropology from a Pragmatic Point of View*, as well as being alluded to in more minor late works. My argument in this thesis, therefore, is not only that Kant's critical philosophy contains a number of insights about cognition that undermine some of the claims shared by the dysfunction analysis, but that these insights are themselves a product of his ruminations upon the theme of madness.

Attempts to define mental disorder are motivated by a number of different goals. Importantly, there is widespread recognition that significant abuses of psychiatric power have taken place since psychiatry itself was instituted by the French physician Philippe Pinel at the tail end of the 18th century. In the absence of a robust definition of mental disorder, it was open for psychiatrists, at various times and in various places, to decide that troublesome relatives (particularly female ones), fleeing slaves, and political dissidents, among others, were mentally disordered and could be subject to forms of coercive management scarcely more humane than the treatment meted out to the inmates of the pre-psychiatry asylums. Psychiatry was open to such abuse because in the absence of agreement as to what actually constituted mental disorder, it was difficult to challenge the purported expertise of the physician. One laudable aim of definition is, therefore, to establish a more rigid, scientific, standard by which to constrain clinical decisions. If this is the goal, however, we should submit candidate definitions to rigorous scrutiny. Dysfunction analyses make strong claims, and it is right to ask how well-founded they are, not least because they way we conceptualise disorder affects the way we think about the people we consider disordered.

A note on terminology

In order to discuss the concept of mental disorder, it will be necessary to employ a number of key terms that require some clarification in their own right. Of first importance is a distinction commonly made in medical discourse between 'disease' and 'illness'. Surveying the issue of terminology, Kenneth Boyd makes the contrast in the following way: "Disease . . . is the pathological process, deviation from a biological norm.

Illness is the patient's experience of ill health, sometimes when no disease can be found." (Boyd 2000, 10) My own use will respect this distinction. Many authors do treat 'disease' and 'illness' as synonyms, however, and I will clarify their meaning where necessary. In recent decades, the term 'disorder' has gained some currency as a general medical term under which terms like 'disease' can be subsumed. In what follows, I will use the term disorder, although the term 'disease' will be encountered in the work of some of the authors whose work I will cite. By the same token, I will use the term 'mental disorder' throughout, but where the historical context justifies it I will refer to 'madness' and 'insanity', terms that were in widespread and uncontroversial use well into the 20th century. Furthermore, since I will be discussing the problem of defining the concept of mental disorder, there will be points in the text where I will paradoxically be forced to use the term referentially even while questioning what, if anything, it actually refers to. In these contexts, I intend my own use of the term to mean something like "the sort of thing we *tend to call* 'mental disorder' ", making minimal assumptions about what that amounts to.

Medical disorders are known, initially, as clusters of co-occurring signs and symptoms (together referred to as a *syndrome*). *Signs*, in medical parlance, are observable abnormalities such as coughing, swelling, rashes, and elevated temperature; *symptoms* are subjectively-experienced phenomena such as headache or nausea; they are not directly observable and physicians may have to rely on patient reports for them to enter the clinical picture. Together, they are normally the original indication that a medical condition is present; for most of human history they have been the *only* available evidence of disease or disorder (Feinstein 1977, 190). It is only relatively recently that scientific methods of isolating the causes of medical conditions have been developed, while the rapid advance of imaging technology has made it increasingly possible to observe the mechanisms by which a condition progresses and affects the organism. Consequently, a physical disorder may be more or less well-defined depending on the current state of scientific progress, and will certainly have been more poorly defined and understood at an earlier point in time.

A disorder is represented in medical nomenclature by a diagnostic category, description of which will cover all currently known information including signs and symptoms, cause, progress, and methods of diagnosis, treatment and prevention. These categories are organised into diagnostic manuals that serve physicians as reference works, such as the *Merck Manual of Diagnosis and Therapy*, the oldest extant work of this type. Manuals of this type do not constitute systems of classification as such, although they are inevitably subject to principles of organisation of one sort or another for ease of use. For statistical and other purposes, however, diagnostic categories have been grouped according to taxonomic schemes of various sorts, with the World Health Organisation's *International Classification of Diseases* (ICD) being the global standard. Although the ICD does cover mental disorders, the *Diagnostic and Statistical Manual of Mental Disorders* is undoubtedly the more exhaustive, and the more discussed, work.

Chapter summaries

In **chapter 1** I will look briefly at the historical developments in medicine, and in psychiatry in particular, that led to the problem of defining 'disorder' becoming an issue of substance. I will survey early attempts to address this problem, before examining in greater detail the appearance of 'function' as a key concept in the philosophy of science. Early applications of this concept to mental disorder are found in the work of Robert Spitzer and Christopher Boorse, whose contributions I explicate. With the key philosophical issues identified, I will outline the most influential and exhaustively-defended definition of mental disorder that employs the concept of dysfunction, Jerome Wakefield's *harmful dysfunction analysis*. This analysis will be the primary subject of my 'Kantian' critique in subsequent chapters.

In **chapter 2** I extend the discussion of the dysfunction analysis, and the HDA in particular, by contrasting it with two of the most important alternative perspectives on mental disorder. The first is the position broadly known by the label 'antipsychiatry', which questions, in various ways, the legitimacy of naturalistic claims that mental disorder refers to an objective, internal, disease entity. This is important, I argue, because what motivates Wakefield's efforts to define disorder in terms of dysfunction is his concern that antipsychiatric criticisms present a serious threat to the profession. The seriousness with which he treats these threats lead him to make unusually strong claims for his own definition of mental disorder.

The second perspective is that of biological psychiatry, which views mental disorders as being essentially dysfunctions of neurobiological processes, to the virtual exclusion of other factors. I explain how biological psychiatry differs from Wakefield's view, and his argument that biological abnormalities can only explain mental disorder if they are given a conceptual foundation such as his own.

I introduce Kant's work in **chapter 3**, surveying the development of his pre-critical concerns about the status of metaphysics in 18th century Germany. As I will show, these concerns were fuelled by the parallel he noted between the speculations of rationalist philosophers in the tradition of Leibniz and Wolff, and the delusions of the mentally disordered. Ultimately, the analogy was sufficiently disturbing that it inspired Kant's great work, the *Critique of Pure Reason*. My argument in this chapter is intended to show that the elements of Kant's critical philosophy that I will employ in the remainder of the thesis are themselves rooted in his own thoughts about the nature of madness. To this end, I will also examine his relatively neglected writings on the 'maladies of the head' in both his pre- and post-critical phases. I will link this to Kant's 'pragmatic' task for his philosophy – its purpose, ultimately, of providing human beings with practical guidance and orientation.

Chapter 4 employs Kant's views on concepts and definitions, as expressed in the first *Critique*, to question Jerome Wakefield's 'harmful dysfunction analysis' (HDA) of mental disorder. I will start by getting clear on some of the finer technical points of Wakefield's position. Although apparently very simple, the HDA is in reality a complex hybrid of conceptual analysis and a form of essentialism inspired by the work of Hilary Putnam. Having done this, I contrast his position with what I argue is Kant's anti-essentialist understanding of the way empirical concepts structure our experience.

My central critique of the HDA takes place here: Wakefield claims to have provided the concept of (mental) disorder with a 'factual scientific' definition. On Kant's view, however, Wakefield has not defined a concept at all. Rather, he has formulated an 'idea of reason' that fails to do what concepts (*a priori* and empirical) must do: make experience of objects possible.

I will go on to set my critique of the HDA in the context of Kant's views on the role of reason in empirical inquiry. The subtext of the dysfunction analysis is that mental disorder *must* reflect malfunction: that nothing else could explain its characteristic signs. Kant's distinctive view of the mind provides us with a very different perspective. His critique discloses an inescapable ambiguity in the fundamental structure of human cognition, which in his view involves a delicate balance between the questing, rational mind and the limits of empirical experience. If this is so, Kant's work lends some weight to the possibility that mental disorder is a matter of degree rather than a qualitative break with sanity.

In **chapter 5**, I complete my project by turning to the work of Louis Sass, who has interpreted some of the central symptoms of schizophrenia through a broadly Kantian dialectic of the self. I present the case that Sass's phenomenological hermeneutic encourages us to see the continuity between sanity and madness, and contrast this with the effect of the harmful dysfunction analysis to make mentally disordered persons seem categorically 'other', with unfortunate consequences.

Chapter 1

Mental Disorder and Mental Dysfunction

Introduction

What is mental disorder? Physicians and philosophers alike have been discussing the phenomenon for thousands of years in one form or another, yet disagreement on this fundamental question continues. Although the authors of one editorial consider that “the statement ‘mental illness is like any other illness’ has become almost axiomatic” (Malla, Joober, and Garcia 2015, 147), even this is problematic, since it is by no means clear what ‘other illnesses’ are like. More accurately – and in keeping with the terminological points I have already outlined – it is not clear what, in the medical context, we mean by ‘disorder’.

In this chapter I will begin by surveying the reasons why giving the concept ‘mental disorder’ a clear definition has come to be considered important. I will then briefly consider the history of attempts to give mental disorder an objective, value-free definition, and why this has been considered desirable. It will emerge from this that the concept of *function* has come to be considered a key element in such attempts. I will go on to assess the definition of mental disorder developed by Robert Spitzer for the groundbreaking third edition of the *Diagnostic and Statistical Manual of Mental Disorders* in 1980, for which the concept of ‘organismic dysfunction’ was crucial. Since this is, and remains, the only institutionally-endorsed definition of the term, the importance of understanding its background and goals is considerable, even though, ultimately, its engagement with the concept of biological function is oblique.

I will then give a brief account of the discussions about functional explanation that developed within philosophy of science in the post-war period. These shaped the way contemporary debates about function and dysfunction in relation to medical disorder have been carried out. With this in place, I will finish by examining two influential and much-debated ‘dysfunction analyses’ of disorder: those of Christopher Boorse and Jerome Wakefield. Each represents one of the two main currents of thought regarding functions in scientific discourse. I will devote particular attention to understanding Wakefield’s harmful dysfunction analysis (HDA), which will be the focus of the Kantian critique I develop in later chapters.

Why define ‘mental disorder’?

Robert Spitzer, perhaps the psychiatrist associated more than any other with the transformation of his profession in the last quarter of the 20th century, once observed that “For hundreds of years the medical profession has managed without a set of criteria for defining which conditions are to be considered medical disorders” (Spitzer and Endicott 1978, 38), and as Robert Kendell noted, “Most doctors never give a moment’s thought to the precise meaning of terms like illness and disease . . . They simply treat the patients who consult them as best they can, diagnose individual diseases whenever they can, and try to relieve their patients’ suffering even if they can’t” (Kendell 1975, 306). Somatic medicine has been able to conduct its business quite well without having to be able to precisely define terms such as ‘health’ and ‘disease’, and many individual conditions seem to fit into our unreflective sense of what a medical disorder is without controversy. As the authors of one paper put it, “In medicine, one can approximate a ‘view from nowhere’ because agreement about what is desirable and undesirable *regardless of circumstances* is frequently so easy” (Jacobs and Cohen 2010, 319). This ‘view from nowhere’ – an idealised, context-free perspective – might only be approximate, but it has facilitated the treatment and study of physical complaints with great success, leading to tremendous advances in medicine over the last few centuries. Simply put, for many physicians, and for many of those who specialise in the study and treatment of mental disorder, analysing terms like ‘disorder’ has not seemed like important work.

Ironically, the great advances medicine has made since the 19th century have, in fact, made definitional problems more pressing. As Jackie Scully argues, “biomedicine’s contemporary power means that it can no longer adopt ambient ideas about disease and disability without running into tricky areas of ambiguity and potentially, ethical difficulties” (Scully 2004, 652). The increasingly profound effects that medical treatments can have on individuals, and the influence of medicine upon society in general, seems to demand a more a more precise demarcation between health and disorder, for a number of reasons. For psychiatry, of course, “ethical difficulties” have always lurked in the background. The effects that mental disorders have on human decision-making and behaviour means that psychiatry possesses a degree of power rarely encountered in other fields of medicine. On the word of a psychiatrist a person may be detained, and perhaps subjected to certain kinds of medical treatment, against their will.

In court, the expert opinion of a psychiatrist may count decisively for or against a defendant. And yet, the difficulty of defining 'disorder' has made psychiatry vulnerable to the charge that it is simply a tool of social control masquerading as a branch of medical science. On this view, psychiatrists are actually making subjective, often morally-grounded, judgements, while representing them as scientific statements about matters of fact. Accusations of this sort were prominently associated with the so-called 'antipsychiatric' movement, to which I shall return in more detail in chapter 2. Although antipsychiatry enjoyed a period of considerable public influence between the 1960s and 1970s, many today dismiss it as little more than a historical anomaly. For all that, these concerns about the concept of mental disorder, and the institutional power of psychiatry, have not gone away.

The kind of rough consensus that may do for somatic disorders therefore doesn't serve psychiatry very well. There is much ambiguity between conditions that we think ought to come under the concept 'mental disorder' and "situations more related to cultural, moral, and religious values . . . ." (Telles-Correia, Saraiva, and Gonçalves 2018, 2). The signs and symptoms of mental disorder are kinds of thoughts, moods, and behaviours that a person can report, or a third party observe; there are no physical lesions or pathogenic agents to which the psychiatrist can appeal.[2] Despite much research into mental disorders, it remains the case that "Unlike many other diseases, there are no approved clinical tests for psychiatric disorders beyond mental and behavioural evaluation" ('Biologically-Inspired Biomarkers for Mental Disorders' 2017, 1). To be sure, biologically-minded psychiatrists remain optimistic about a future in which underlying pathological processes *will* be found for mental disorders. The contrast with physical medicine, however, remains stark. The question, for the time being, is how we can determine that some such thoughts and moods are to be expected – 'normal' – and others pathological, since we do not possess any obviously objective criteria by which the distinction could be made.

In short, while the problem of defining terms like 'disorder' is certainly more complex for psychiatry, it has increasingly been seen as an important task for medicine in general. Conversely, those who want to confirm that psychiatry is a *bona fide* branch of medical

[2] The DSM does include a section on neurocognitive disorders that have recognised pathophysiologies, including conditions such as Alzheimer's and Parkinson's diseases, but these constitute a special case. The vast majority of the conditions listed in both the DSM and the World Health Organisation's *International Classification of Diseases* (ICD) do not have known biological etiologies.

science must find a definition that applies equally well to physical and psychological disorders. In other words, what is wanted is a definition not of *mental* disorder, but a definition of *medical* disorder that encompasses both mental and physical conditions. Consequently, the post-war period has seen a number of physicians and philosophers turning their attention to these issues.

Marc Ereshefsky (Ereshefsky 2009) identifies three broad approaches that have emerged in response to the problem of defining medical terms like 'disorder'. *Naturalists* look for definitions that claim to be objective. Such a definition would be a proposition concerning biological or psychological facts alleged to obtain independently of any observer, and in practice they often appeal to the notion of 'function', to which I shall return in due course. *Normativists* hold the opposite view: a concept like 'disease' reflects subjective judgements about the physical and mental states that we disvalue. Even if there is broad agreement between the members of a society, these judgements may vary widely across different cultures and over periods of time. These approaches question the appropriateness of providing the term 'disorder' with a definition, strictly-speaking. A third perspective combines elements of the first two. On this view, 'disorder' includes a subjective judgement of harm or disability and some kind of objectively identifiable atypicality. Both are required in order for the concept 'disorder' to apply. By contrast, straightforwardly naturalist analyses consider evaluations of harm etc. to be independent: a disorder is such whether or not it is valued or disvalued.

Serious naturalist attempts to define medical disorder started in the late 1950s. An early advocate of this project was not a psychiatrist but a specialist in thoracic conditions, J. G. Scadding. He noted that "A poorly defined word may mislead us in many ways. In a discussion, it may be used in different senses by the discussants. They will then resemble the two women shouting abuse at each other from the windows of their respective houses, about whom the English wit, Sydney Smith, remarked that they would never agree because they were arguing from different premises" (Scadding 1963, 1425). A good joke, and a point well made: in the absence of commonly agreed definitions of key concepts, there is the ever-present risk that clinicians may unwittingly find themselves talking at cross purposes. In the healthcare context, this could well have serious consequences.

Scadding's proposed definition is that "A disease is the sum of the abnormal phenomena displayed by a group of living organisms in association with a specified common characteristic or set of characteristics by which they differ from the norm for their species in such a way as to place them at a *biological disadvantage*" (Scadding 1967, 877; emphasis mine). He further notes that the definition "implies a statistical basis for the concept of abnormality." (Ibid.) The 'biological disadvantage' requirement places a vital constraint on the 'abnormality' criterion. Clearly, not every statistically unusual variation, such as being taller than average, is considered a disease, so this criterion aims to establish a distinction between the merely atypical and the diseased. Scadding did not, however, believe that biological disadvantage could be given a more precise explication, leaving it vague as to how it would be established in practice. In various ways, later attempts at definition would draw upon and develop this aspect of his work.

Robert Kendell (Kendell 1975) recognised the need to get clearer on the notion of biological disadvantage. His response was to frame it in terms of survival and reproduction; although he did not mention evolutionary theory, these are, of course, key elements in the process of natural selection. One problem that he himself pointed out is that his definition would exclude many conditions that physicians generally do consider disorders, but that do *not* significantly impact survival and reproduction, such as psoriasis. In other words, his analysis is contrary to existing intuitions and practice – it is prescriptive rather than descriptive. At the same time as Kendell was writing, others were working on analyses of disorder that would accord a central role to the concept of function. In terms of sheer reach, the most important of these would be that of Robert Spitzer and his colleagues working on the third edition of the DSM.

The DSM and the definition of mental disorder

The *Diagnostic and Statistical Manual of Mental Disorders*, or DSM, a publication of the American Psychiatric Association, has been called the 'bible of psychiatry' (Horwitz 2021). First published in 1952, it is now in its fifth edition, with the latest revisions appearing in 2022. It is, as the name suggests, partly a diagnostic tool, currently containing over 300 categories. Prior to its third edition, published in 1980, the DSM had been psychodynamic in orientation, influenced by the work of Sigmund Freud in Europe and Adolf Meyer in the US (Kawa and Giordano 2012, 3). In keeping with this framework, relatively little emphasis was placed on distinguishing different categories of disorder, and the categorical descriptions that were on offer were "short, general, and infused with theory" (Horwitz 2015, 951).

The pre-eminence of psychodynamic theory in mid-century psychiatry in the US was already under pressure by the time of DSM-II, as advances in pharmacology in the 1950s made drugs such as the anti-psychotic chlorpromazine and anti-depressants like imipramine and iproniazid available. These were among the first pharmacological treatments for which any efficacy could be demonstrated in trials involving psychiatric patients (Zachar 2000, 34f). They could be made widely and cheaply available, making drug therapy feasible for people who could not access or afford psychoanalysis, as well as to people suffering from acute conditions (such as schizophrenia) that psychodynamic theory had, historically, been ambivalent about treating (Shorter 1997, 205). The apparent advantages of pharmacology opened up a genuine alternative to psychodynamic theory, and appeared to support the possibility of a more positivistic scientific approach to understanding and treating disorder. Psychoanalysis was unable to respond effectively to this challenge, since its practices were not based on systematic research and evidence. Indeed, from a scientific perspective, psychoanalytic doctrines could appear remarkably arbitrary (see, for example, Kandel 1999). The pre-eminence of psychoanalysis was in question, but its position in North American psychiatry was deeply entrenched. Unsurprisingly, it did not give way at once.

The decisive shift away from psychodynamic theory and towards a more medical, empirically-informed orientation for the DSM was effected by Robert Spitzer, chair of the task force charged with revising the manual, and his colleagues. Nearly a decade of wrangling eventually yielded the DSM-III, published in 1980. In 1978 Spitzer and his colleague Jean Endicott published a paper that sheds much important light on the work then being undertaken to produce the new edition. Therein, they recall there was much debate over their personal view that a general definition of mental disorder was required. They observed that "without some definition of mental disorder, there would be no explicit guiding principles that would help to determine which conditions should be included in the nomenclature, which excluded, and how conditions included should be defined" (Spitzer and Endicott 1978, 16).

With the stated caveat that a truly precise description was too much to be hoped for, a general definition of mental disorder appeared in DSM-III in 1980. With some minor amendments, it has remained in subsequent editions. In its current formulation, it asserts that

> A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that *reflects a dysfunction in the psychological, biological, or developmental processes*[3] *underlying mental functioning.* Mental disorders are usually associated with significant distress or disability in social, occupational, or other important activities. An expectable or culturally approved response to a common stressor or loss . . . is not a mental disorder. Socially deviant behaviour . . . and conflicts that are primarily between the individual and society are not mental disorders unless the deviance or conflict results from *a dysfunction in the individual*, as described above. (*Diagnostic and Statistical Manual of Mental Disorders: DSM-5* 2013, 20; emphasis mine)

Spitzer and his task force explicitly intended the DSM to be, first and foremost, an *atheoretical* diagnostic manual calculated to serve psychiatrists in clinical practice regardless of their particular beliefs regarding the nature and causation of mental disorder. Consequently, "implicit in DSM-III-R is that all theories of mental disorder presuppose a common pretheoretical concept of mental disorder, as expressed in DSM-III-R's theory-neutral definition" (Wakefield 1992a, 232).[4] The definition, and the individual diagnostic categories that sit under it, advance no views regarding the specific biological or psychological dysfunctions that particular mental disorders supposedly reflect. Rather, they provide criteria which, if met, are considered sufficient to make a diagnosis. In particular Spitzer and Endicott made it clear that "Some dysfunctions may be understandable only on the basis of psychological concepts, such as learning or conflict, and may never be reducible to biochemical or neurophysiological constructs" (Spitzer and Endicott 1978, 27).

This reflects the self-appointed job of Spitzer's task force to reconcile as wide a variety of psychiatric opinion as it could within the constraints of clinical practice. During the

[3] The term 'developmental processes' appears in the definition for the benefit of the category of *neurodevelopmental disorders*, those which emerge in early childhood and include learning disorders and autism. Presumably, however, developmental processes are themselves psychological and/or biological.

[4] Wakefield was discussing the 1987 interim revision, DSM-III-R, but his comments apply equally to the earlier volume.

1970s, the most notable division was between the psychoanalysts and more medically-inclined factions within psychiatry, although this rapidly become a merely historical episode as the place of psychoanalysis in psychiatry declined in the decade following the publication of DSM-III. In the 1970s, new technologies that promised to reveal much more about the brain and its activity were only just emerging. What was 'medical' in psychiatry at that time was not so much neurobiology as an approach to diagnosis. (Spitzer himself had made his name by developing computer-aided, data-driven models of diagnosis employing carefully designed questionnaires – see, e.g., Spitzer and Endicott 1972). Today, psychiatric opinions are more prominently split between those who pursue a rigorously biological approach that views mental disorders as disorders of the brain *simpliciter*, and those who cleave to a weaker medical model that, as Spitzer and Endicott maintain, leaves room for purely psychogenic explanations.

Spitzer's project was unusual in comparison to the other attempts at definition that I will examine below. Although a pretheoretical analysis of disorder might be considered desirable on its own merits, his own efforts in this direction were determined by political pressures within psychiatry, and must be seen in that context. He did not enjoy the freedoms of the professional philosopher to develop an analysis that he found personally satisfying, and the constraints within which he was working led him to take a highly distinctive approach.

In their 1978 paper, Spitzer and Endicott wrote that "the approach taken here is unique in providing not only a definition of medical disorder, but detailed operational criteria" (Spitzer and Endicott 1978, 16). Although operationalism, properly so-called, originated with the physicist Percy W. Bridgman in the 1920s, Kenneth Kendler attests that Spitzer and colleagues were largely unaware of the historical and philosophical background and were instead "in favour of the benefits of practical operationalism – giving psychiatrists rules by which to assemble symptoms and signs into diagnoses as a means of improving reliability" (Kendler 2017, 2055). This is, in outline, still faithful to Bridgman's well-known dictum that "In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations" (Bridgman 1958, 5). An operational definition, then, is one that eschews abstract 'properties' that may not in fact exist (as in the example Bridgman gives of Newton's definition of absolute time) in favour of identifying the concept with the operation used to measure or otherwise determine it. For the DSM, this means providing descriptive criteria by which a diagnostic judgement may be reached. These criteria make no reference to causes (except in the very few cases where known), nor to specific theoretical constructs such as 'repression', which are to be found in the first two editions of the DSM. As Kendler writes, the criteria are primarily lists of signs and symptoms, a minimum number of which must be present for a particular diagnosis to be made. In this way, Spitzer and colleagues were trying to achieve a common standard of psychiatric diagnosis to be made regardless of theoretical persuasion.

Regarding the general definition of mental disorder, Spitzer and Endicott identify two concepts of central importance: negative consequence and "inferred or identified organismic dysfunction" (Spitzer and Endicott 1978, 17). They note, however, that "such terms as 'dysfunction', 'maladaptive', or 'abnormal' . . . themselves beg definition" (Ibid.). Their approach, therefore, is to give operational criteria for the aforementioned concept of negative consequence. In the DSM-III definition, these are "painful symptom (distress) or impairment in one or more important areas of functioning (disability)" (*Diagnostic and Statistical Manual of Mental Disorders: DSM-III* 1980, 6). In fact, the definition as it appears in DSM-III is not fully operationalised, including as it does the requirement that "there is an inference that there is a . . . dysfunction . . . ." (Ibid.). By the time of the interim revision (DSM-III-R) seven years later, this had become the requirement that these negative consequences "not be merely an expectable response to a particular event" (*Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R* 1987, xxii). The principle, therefore, is that the troublesome concept of organismic dysfunction can be reduced to the relatively unambiguous criteria of distress or disability, which must themselves not be a predictable reaction to stressors. These conditions are, in Spitzer and Endicott's view, "sufficient evidence for . . . an organismic dysfunction . . . ." (Spitzer and Endicott 1978, 18).

By eschewing theory, the third edition of the DSM marked a profound change in the way that psychiatric diagnosis and its associated nosology was carried out. While Spitzer emphasised the role of dysfunction in the concept of disorder, and accorded it a kind of prominence it had previously lacked in authors like Scadding and Kendell, he was not alone in so doing. Indeed, where he baulked at subjecting 'dysfunction' to analysis in its own right, others were willing to face the task head-on. Any analysis of dysfunction, however, must rest upon a concept of function as such. Interest in this concept was already well established in philosophy of science. In order to properly understand the philosophical analyses of disorder with which the DSM competes, we must look back at the history of this project.

The concept of function

The word 'function' has a number of different uses, but in the special sciences – particularly in biology – talk of functions occurs in what appears to be a distinctive mode of explanation that doesn't occur in physics. To use a well-worn example, when we ask for a functional explanation of the activity of the heart, we are not asking for the cause of its pumping – what makes it pump. We are asking for what *purpose* or to what *end* the heart pumps.[5] In physics, by contrast, we do not ask what end is served by gravitation, for example, and we do not ascribe a function to it. We simply say that gravity is what causes certain phenomena. The hallmark of functional-explanation, on the face of, is therefore *teleological*.

Although ubiquitous in the life sciences, this way of talking is controversial because the notions of purposes, ends, or goals inevitably suggest conscious evaluation. When we *are* talking about the actions of intentional agents, of course, the teleology is uncontroversial. When it comes to explaining natural phenomena scientifically, however, this possibility must be excluded, since there is no empirical evidence that they are products of conscious design. The problem, then, is to try to understand what exactly we mean when we talk about functions, and whether function-talk is inescapably teleological.

Although talk of functions in this sense goes back to at least the work of Plato and Aristotle, the contemporary debate around the nature of functional explanation is rooted in the work of two philosophers of science, Carl Hempel (Hempel 1965, [1959]) and Ernest Nagel (Nagel 1961a). Both men subscribed to the deductive-nomological (D-N) model of scientific explanation in which the explanation of some phenomenon is comprised by a set of statements (the *explanans*) from which the *explanandum* (the thing to be explained) can be logically deduced. It is usually expected that at least one of the statements in the explanans will state a general law (hence *nomological*, or lawlike), and at least one of them will be a particular factual proposition. For example, the appearance of a rainbow can be explained by particular conditions that obtained at a particular time and

[5] Of course, causal explanations *can* be sought. The point is that this sort of explanation does not answer the sort of question characteristic of the special sciences.

place, and general laws (in this case of refraction) of which these facts are an instance. In their analysis, functional explanation, if it were to be properly scientific, would have to reduce to a D-N explanation. Consequently, they take it that functional-explanations should reduce to a set of statements from which the presence of the function-bearer (the thing whose activity we take to be purposive) can be inferred.

An important difference between them would determine the form of future debate regarding function-talk in the special sciences. For Hempel, functional-explanation does not reduce to D-N explanation, and is therefore not a genuinely scientific form of explanation, even if it is useful in certain fields of investigation. His reason for this conclusion is that if the functional explanation of the heart is that it pumps blood, then one should be able to deductively infer that in an organism (a system) in which blood circulation is taking place a heart is necessarily present. But the presence of the heart is only sufficient, not necessary, since it is logically possible that something other than a heart could fulfil this function (and in fact, artificial pumps *can* fulfil the same function). Therefore the argument is deductively invalid.

Nagel is perfectly aware of this problem, but he maintains that analyses of this sort "are not explorations of merely logical possibilities, but deal with the actual functions of definite components in concretely given living systems." (Nagel 1961a, 404) Using the example of chlorophyll in plants, to which the function of enabling photosynthesis is attributed, Nagel maintains that in point of fact, photosynthesis in plants only occurs if chlorophyll is present. The 'merely logical' possibility of alternatives can be dismissed, and the presence of chlorophyll can be validly deduced from statements regarding a specific activity performed by plants and the *necessity* of chlorophyll for that activity. This latter statement is non-teleological, since it makes no reference to purposes, ends, or goals. As he puts it, "The difference between a teleological explanation and its equivalent nonteleological formulation is thus comparable to the difference between saying that *Y* is an effect of *X*, and saying that *X* is a cause or condition of *Y*. In brief, the difference is one of selective attention, rather than of asserted content" (Nagel 1961, 405). On Nagel's view, then, one can straightforwardly say that *X*'s being the cause of *Y* does explain the presence of *X*, in the simple sense that if it were absent, *Y* would not in fact occur. One need not appeal to purposes or goals at all. Hempel, of course, imposed a stricter logical standard upon his own analysis, and later writers have also not been satisfied with Nagel's willingness to treat what *is* the case as if it were a normative *ought*.

In summary, as Peter McLaughlin puts it, "later analyses in the tradition of Hempel tend to be called 'backward looking', 'etiological', or 'teleological'. Those in the tradition of Nagel tend to be called 'forward looking', 'dispositional', or 'causal role'" (McLaughlin 2003, 70). This calls for some clarification, however. Hempel, as we have seen, took it that functional-explanation was meant to literally explain the presence of the function-bearer; to explain why we have hearts. However, he did not think that this mode of explanation was genuinely scientific, in the deductive-nomological sense. Those thinkers following in his footsteps attempt to show that function-talk is valid because the origins or 'etiology' of functional items *can* be established. By contrast, Nagel – somewhat tendentiously – argued that function-talk is legitimately scientific, but the sense in which he took it to explain the presence of functional items is explanatorily shallow. For this reason, although both men assume that functional-explanation answers the question "Why is it there?", the response provided by Nagel's formulation is so insubstantial that some writers characterise him as simply addressing the question "What does it do?" At any rate, this distinction does, I believe, more accurately capture the way that their respective positions have been developed by later thinkers – as reflected in McLaughlin's comment.

One of the most important developments of Hempel's position was set out by Larry Wright in 1973, as can be seen from his statement that

> The function of *X* is *Z* means
>
> (a) X is there because it does *Z*
>
> (b) *Z* is a consequence (or result) of *X*'s being there (Wright 1973, 161).

In common with other analyses that follow in Hempel's footsteps – 'etiological' analyses – Wright includes (in (a) above) a feedback loop that accounts for the presence – more specifically, the *origins* – of functional item *X*. For Hempel, the presence of *X* would have to be a *logical* consequence of its activity; he did not ask as to its actual origins, and, of course, did not think that functional explanations met the logical requirements of the D-N model. Wright abandons the strictures of the D-N model (whose influence declined sharply in the 1960s). Functional explanation is not, for him, a species of deduction. Rather, what a function explains is the causal origin of the function-bearer: the cause of the function-bearer *X* having come into being is its effect (function) *Y*. Wright also rejects the idea, common to Hempel and Nagel, as well as many others, that a function should *benefit* the system within which it occurs.

Wright's position is also marked by his interest in the fact that we also talk about the functions of consciously designed artefacts and their component parts. His analysis proceeds on the principle that the same account of function should explain why we attribute functions to designed artefacts as well as to parts of living organisms (he does not consider the use of the concept in the social sciences). Functional explanation of an artefactual component does seem to account for the presence of that item: it is present because a conscious designer intended it to be there, and to do what it does, because of its contribution to the function of the artefact itself. The effect is indeed, in this sense, the cause of the component's being part of the artefact. In the case of natural functions, Wright simply invokes natural selection: "If an organ has been naturally differentially selected-for by virtue of something it does, we can say that the reason the organ is there is that it does that something" (Wright 1973, 159). Although the intention of a designer and the process of natural selection are feedback mechanisms by which the function-bearers come to be present, the analysis itself makes no reference in *either* case to conscious intent – indeed, is completely agnostic regarding the possible mechanisms by which the feedback condition might be effected.

In this view, the contribution that some effect makes to the system it occurs in is not what confers the status of a function. In principle, an effect may make no contribution at all and still be considered a function as long as it has the right kind of history. In principle, any effect whatsoever deserves to be called a function if it played the relevant role in explaining the origin of the functional item. This means that Wright does not require functional items to confer benefit on the containing system. On the other hand, none of his examples show how a feedback mechanism could operate other than on the principle of benefit.

If Wright's work is the most prominent example of an analysis of function in the vein of Hempel, then the best-known counterpart in the tradition of Nagel is undoubtedly due to Robert Cummins (Cummins 1975). Cummins takes both Hempel and Nagel to task for their analyses, however. In his view, there *is* a sense in which functional explanation addresses the question of why some item is present – but this is *not* the role that function-talk plays in science. Where both Hempel and Nagel believed that an explanation of function-talk in science would have to involve reducing it to the same form that other explanatory statements in science took, Cummins provides "an account of functional explanation which takes seriously the intuition that it is a genuinely distinctive style of explanation" (Cummins 1975, 757). In his view, "what we can and do explain by appeal to what something does is the behaviour of a containing system" (Cummins 1975, 748). That is, functions are those effects of a structure or mechanism that explain a capacity possessed by the system in which they occur. The circulatory system has the capacity of moving blood around the body, so the pumping of the heart is identified as a function because this is the effect that explains the movement of the blood. The rhythmic thumping sound the heart makes is an effect, but it does not contribute to explaining circulation, and is thus not a function.

What this means for Cummins is that talk of functions occurs in a form of *analysis*, since the investigator must identify a capacity of a system and then analyse it into its component parts; those activities of the component parts that explain this system-capacity are functions. However, as he recognises, a system may possess any number of 'capacities' (effects) that we would intuitively deny are functions. His example, unsurprisingly, is that of the "variously-tempoed" sounds made by the circulatory system (Cummins 1975, 763). Within this system, it is the beating of the heart that generates these sounds, so that, on this – perfectly true – analysis, the function of the heart turns out precisely to be making a rhythmic pumping sound. The problem of distinguishing mere effects from functions therefore seems to be relative to the system-capacity that the investigator is interested in. If this is the case, then it would appear that any number of effects that a thing has can legitimately be called functions – it simply depends on the context of interest.

Cummins view here is twofold. Firstly, he does not think that his view will lead to a proliferation of interest-relative functions. It seems implausible that we would analyse the capacity 'circulatory sound-making' in terms of the heart's pumping because this is not a case of explaining something complex in terms of simpler component processes: pumping doesn't seem any more basic than noise-making. Circulation, however, is much more complex than the pumping of the heart, so Cummins argues that this analysis is more likely to provoke interest. His second point is that "It must be admitted . . . that there is no black-white distinction here, but a case of more-or-less. As the role of organization becomes less and less significant, the analytical strategy becomes less and less appropriate, and talk of functions makes less and less sense. This may be philosophically disappointing, but there is no help for it" (Cummins 1975, 764).

Of course, some philosophers have found this aspect of his approach *too* disappointing to accept. Of the stand-off between these two positions, those of Wright and Cummins, it was once predicted that "there is a risk that it will decay into the dull thud of conflicting intuitions" (Bigelow and Pargetter 1987, 196). This fear seems to have been realised: although variations on both the etiological and causal-role interpretations of function have been developed in the intervening decades, the basic terms of the dispute remain essentially as Wright and Cummins determined them in the mid-1970s.

This settling-down of the debate into two entrenched positions is reflected, predictably, in discussions around the concept of dysfunction in the medical context. At this juncture I will turn to two of the most influential dysfunction analyses of medical disorder, analyses that draw directly on the debate that I have surveyed above. The first is due to Christopher Boorse, whose analysis was contemporaneous with that of Spitzer. The second is Jerome Wakefield's harmful dysfunction analysis (HDA), which appeared over a decade later, and will be my primary focus in the remainder of my thesis.

Boorse and the biostatistical model of dysfunction

While Spitzer and his colleagues were trying to formulate a definition of mental disorder that would satisfy a theoretically diverse constituency of practicing psychiatrists, philosopher Christopher Boorse, free from such exigencies, proposed a definition of mental disorder that likewise drew upon the concept of dysfunction. In his paper 'What a Theory of Mental Health Should Be', Boorse argued that "diseases are interferences with natural functions . . . since the functional organisation typical of a species is a biological fact, the concept of disease is value-free" (Boorse 1976a, 63).[6]

Boorse follows in the tradition established by Nagel, employing a concept of function very similar to that developed, contemporaneously, by Cummins. In so far as he intends his analysis to apply to both physical and mental disorder, he notes that "mental-health theory and practice have not sprung up in a vacuum. On the contrary, they originally arose within physiological medicine, a mature and fairly well-articulated body of

[6] Boorse's preferred locution early on is 'disease', while he later substitutes 'pathology' or 'pathological condition'. I will treat both as variations on the term 'disorder' and, in order to minimise confusion, will use this latter term.

thought. From this established discipline they borrowed both the root notion of health and the many unspoken assumptions that surround it . . . ." (Boorse 1976a, 61).

In his view, as in that of Cummins, "Functions are, purely and simply, contributions to goals" (Boorse 1976b, 75). For this reason, although Boorse developed his analysis of function (separately, in the paper just cited) independently of Cummins, in the literature Boorse's analysis of medical disorder is often referred to as employing 'Cummins functions', although the more neutral 'causal-role function' is also frequently used. He takes it that the pre-eminent goal for all biological organisms must be survival and reproduction. Although this is in keeping with evolutionary theory, he does not involve natural selection directly in his analysis, pointing out that function-talk in biology pre-dates modern evolutionary theory. He cites the standard example of William Harvey, who in 1620 established the function of the heart, and points out, against proponents of the etiological view, that "When Harvey, say, claimed that the function of the heart is to circulate the blood, he did not have natural selection in mind. Nor does this mean that pre-evolutionary physiologists must therefore have believed in a divine designer. The fact is that in talking of physiological functions, they did not mean to be making historical claims at all. They were simply describing the organization of a species as they found it" (Boorse 1976b, 74).

Boorse's argument is that the *origin* of a function is irrelevant—what matters *qua* health and disorder is what something does, not why or how it has come to be there. This also means that *current* contribution to a goal is what counts, since 'what something does' in this sense can only be assessed in the present. Boorse's analysis is rather like Cummins in another important respect: for him, function-ascription is interest-relative. He talks about strong and weak function statements, noting that "What converts a function *X* performs into 'the function of *X*' is our background interests in the context in which the function statement is made . . . 'The function of *X*' will be simply that one among all the functions performed by *X* which satisfies whatever relevance conditions are imposed by the context of utterance" (Boorse 1976b, 81f).

Boorse's case is heavily based on the quite diverse ways in which the term 'function' is used in biology. For example, in ecology it is normal to apply the term at a far more general level than the organism, let alone its component parts. As he sees it, "In these contexts talk of functions has a clear and legitimate use without any etiological implications" (Boorse 1976b, 86). On Wright's analysis, by contrast, most of these applications would not be warranted – indeed, would be misleading, since he considers the reference of the concept 'function' to be very specific. For Boorse, however, context sensitivity is not a problem, precisely because his analysis provides a 'context of utterance', i.e., that functions, in the relevant sense, are "species-typical causal contributions to individual survival or reproduction" (Boorse 2014, 687).

The typicality requirement provides us with a sense of normal or standard function even though individuals will vary widely within a species. But, as one commentator paraphrases it, "In many cases, it is impossible to identify variables that are statistically typical for the *entire* species" (Varga 2011, 3). In other words, what might be statistically average for a particular group within the species may not be for another. In order to deal with this Boorse introduces what he calls 'reference classes' within the species. The classes he has specified are biological sex and different age groups (he also entertains the possibility that racial groups might also be appropriate classes (Boorse 1977, 558)). Ultimately, then, he argues that within these classes "Abnormal functioning occurs when some function's efficiency falls more than a certain distance below the population mean . . . this distance can only be conventionally chosen, as in any application of statistical normality to a continuous distribution. The precise line between health and disease is usually academic, since most diseases involve functional deficits that are unusual by any reasonable standard" (Boorse 1977, 559).

On this analysis, medical disorder just is abnormal functioning, or dysfunction. This has subsequently become known as the *biostatistical* model (BST). The statistical-norm requirement is yoked to the central concept of biological function: like Scadding, Boorse recognises that not all norms (and deviations from them) count towards distinguishing the healthy organism from the unhealthy. Nonetheless, the quote above sees him admitting that statistical typicality remains a matter of convention. Ultimately, this means that function and dysfunction involve subjective human judgements about where the boundary between them lies. As we shall see, this is something that his most vigorous competitor, Jerome Wakefield, finds unacceptable.

As with Spitzer, and with most proponents of dysfunction analyses of disorder, Boorse's model is intended to capture both physical and mental dysfunctions. In so doing, he acknowledges that "we may expect normal psychological functions to be somewhat less specific than their physiological counterparts. The outstanding feature of human mentality is its plasticity" (Boorse 1976a, 64). The "somewhat less specific" seems to me to be a breathtaking understatement (and, as we shall see, the assumption is not unique to Boorse). Be that as it may, he suggests that we can at least provide a general characterisation of mental functions, such as that perceptual processing serves to provide information about the external world, and that this is enough to provide us with a concept of mental health. However, he says nothing about how we might actually establish statistical norms for mental functions. Indeed, he seems satisfied to conclude that "given a few plausible assumptions . . . It is quite likely that there is such a thing as mental health" (Boorse 1976a, 68). In other words, his ambitions in his discussion of specifically mental disorder appear to be limited to establishing that mental health and disorder are matters of scientific fact *in principle*. The problem of how to put this into practice is not one he addresses.

Boorse's analysis is one of the earliest attempts to give mental health and mental disorder a value-free definition via the concept of function as it is used in biology. It is one that he has defended and modified over a period of several decades, and remains influential. In 1992, however, an alternative dysfunction analysis appeared that has gone on to become arguably *the* most widely discussed and debated definition of mental disorder: Jerome Wakefield's harmful dysfunction analysis.

Wakefield's harmful dysfunction analysis

Wakefield introduced his analysis in two papers. In one, 'The Concept of Mental Disorder', he surveys a number of different analyses of the concept of mental disorder[7] both naturalist and normativist, before presenting his own harmful dysfunction analysis. In the other, 'Disorder as Harmful Dysfunction', he critiques the DSM definition, but his own analysis of disorder also employs two elements that Spitzer had drawn attention to: organismic dysfunction and harm. The HDA has since become prominent in the debate around the concept of mental disorder, partly due to Wakefield's extensive promotion of its superiority to alternatives and his tenacious defence of it against criticism. In the remainder of my thesis, I will concentrate on the HDA and the definition of mental disorder that originates from it.

[7] On the first page, (373) he notifies the reader, however, that "The focus is on disorder rather than mental . . . ." and asserts that "the general concept of disorder . . . applies to both mental and physical conditions . . . ." (374).

It is important to look at Wakefield's critique of the DSM's definition of mental disorder, since he treats the latter as a template for his own analysis. Following a very careful reading of Spitzer's own account of how the definition was formulated, Wakefield resolves it into the following criteria: "A mental disorder is a mental condition that (a) causes distress or disability and (b) is not a statistically expectable response to external events" (Wakefield 1992a, 238). As noted above, the relatively unambiguous criterion of 'unexpectable response' – with the response being whatever distressing or disabling symptoms prompt interest in the condition in the first place – effectively replaces the problematic concept 'dysfunction'. However, Wakefield shows that the criterion is subject to some significant counterexamples. If unexpectability is not to be a merely subjective judgement, then it must mean 'statistically unlikely'. But this would admit some potentially unexpectable conditions that it seems clearly counterintuitive to call dysfunctions: "selfishness, cowardice, slovenliness, foolhardiness, gullibility, insensitivity, laziness, and sheer lack of talent are a few examples . . . that can cause harm and that can be statistically deviant either in the nature of the response or in the response's intensity . . . ." (Wakefield 1992a, 238). On the other hand, there are some conditions that strike us as disorders despite being expectable. For example, "an expectable response to extreme, sudden pressure on the arm is for the arm to break" (Wakefield 1992a, 239). Broken arms nonetheless seem – almost definitively – to be dysfunctions. If the operationalism of the DSM fails to encompass such an apparently unambiguous case of somatic dysfunction, then it clearly faces a serious difficulty.

A second problem can be traced to the DSM's inclusion of 'disability' as a form of harm, and its relation to 'unexpectable response'. The disability requirement is intended to capture the fact that not only distress but also problems carrying out tasks essential to basic physical maintenance, can be consequences of disorder. Clearly, the authors have in mind some set of universal human abilities that are affected by disorder. As it stands, however, this requirement has the unintended consequence of making almost *any* lack of ability dysfunctional. In order to know what a disability is, we need to know which abilities are relevant. Is not being able to read, for example, a disability in the sense relevant to psychiatry? It certainly disadvantages the individual significantly. If we say it is a disability, we will be obliged to say it is a mental disorder, since in many parts of the world illiteracy is now rare, making it an unexpectable disability. But this is clearly not what we would want to do, so this criterion turns out to be too permissive.

Wakefield's discussion goes into greater detail, but these examples capture the core problems that he perceives in the DSM model. He therefore proposes an alternative analysis which is essentially very similar but "wherein *dysfunction* is a scientific and factual term based in evolutionary biology that refers to the failure of an internal mechanism to perform a natural function for which it was designed, and *harmful* is a value term referring to the consequences that occur to the person . . . ." (Wakefield 1992b, 374). He takes it that 'dysfunction' can be given a value-free analysis of its own, but perhaps the most distinctive aspect of Wakefield's analysis is his argument that the concept of disorder also requires an evaluative element, making it an example of Ereshefsky's 'hybrid' approach. This contrasts with the positions of both Spitzer and Boorse. Spitzer translated harm (or, in his words, negative consequence) into supposedly objective criteria such as distress and disability. For Boorse, on the other hand, harm *per se* doesn't enter into the analysis at all: rather, by defining functions as contributions to survival and reproduction, and then defining health as statistically normal functioning, disorder becomes "a type of internal state which impairs health, i.e. reduces one or more functional abilities below typical efficiency" (Boorse 1977, 555). Contribution to survival (and perhaps reproduction) is something we may well make an evaluative judgement about, but is not itself evaluative, Boorse argues. For Wakefield, however, the subjective judgement of harm is inextricable from the concept of disorder: it is in fact how we distinguish between mere dysfunction and medical disorder, a distinction that Boorse does not make.

The judgment is not an individual one, of course – rather, Wakefield concludes that "*harmful* is a value term referring to the consequences that occur to the person because of the dysfunction and are deemed negative by sociocultural standards" (Wakefield 1992b, 374). One commentator puts it more perspicuously, stating "a condition is harmful in the sense relevant to his account when the cultural value system of the group to which an individual belongs implies that it is harmful, irrespective of whether the affected individual herself considers it harmful" (Dussault 2022, 678). The assumption is that most of the conditions currently contained in nosologies such as the DSM *are* thus disvalued by current cultural standards (in the west, at least), which is primarily why they come to the attention of psychiatrists in the first place. At the same time, this is to acknowledge that (and explain why) disorder judgements can vary across cultures and over time. Nonetheless, the more pressing question for Wakefield is whether all these conditions involve dysfunctions.

On this issue, Wakefield also diverges from Boorse. Although he agrees that it "is a purely scientific concept" (Wakefield 1992a, 236), he adopts the analysis of function developed by Wright. While Boorse's analysis of 'function' is very similar to Cummins', it was developed independently, and his application of the concepts of health and disease is facilitated by the elements of statistical norm and associated reference classes. Wakefield's analysis of disorder – at least in its original formulation – sees Wakefield borrowing Wright's analysis of function wholesale. As a result, we can complete our summary of the HDA by returning briefly to Wright.

As discussed above, the causal-role model employed by Boorse answers the question of what the function of something is. On this view, something counts as a function by virtue of making a causal contribution to a specific goal within the containing system. Wright's analysis, by contrast, answers the question why some feature of a system exists, and it tells us that the determining factor is historical. By his lights, something counts as a function because it is an effect that explains the presence of the function-bearer. No reference is made to what goal-contribution (if any) the effect makes. Wakefield himself writes "the concept of natural function can be analysed as follows: A natural function of an organ or other mechanism is an effect of the organ or mechanism that enters into an explanation of the existence, structure, or activity of the organ or mechanism" (Wakefield 1992b, 382).

We should recall that as far as designed artefacts are concerned, a causal history of this sort is easy enough to make sense of: something is said to have a function if the designer intended it to have that specific effect. This applies equally to artefacts themselves (e.g. a pen or a car) and their constituent parts, whose individual functions combine to serve the overall function of the artefact. Therefore, the explanation for the existence of anything that has a function is that is has that effect. In the case of living organisms, which lack a designer, Wright sees natural selection working on random mutation as a "natural extension" of the same principle (Wright 1973, 163). However, because no mention is made of what contribution a function makes, in both cases he is apparently able to avoid all talk of *benefit*. As a result, it appears that he has provided what Peter McLaughlin describes as "a completely nonevaluative explanation of the origin of the function bearer based on its effects" (McLaughlin 2003, 94). This fills in loose talk about what things are 'meant' to do: they are meant to have the effect or effects that explains why the function-bearer is part of the artefact or organism, whatever that explanation is. Whether or not that effect is beneficial is, in principle, irrelevant.

Wakefield's primary concern is to distinguish disorder from non-disorder, a matter of what he calls *conceptual validity* (Wakefield 1992a, 232) The HDA is not intended to assist with actual diagnosis – unsurprisingly, since physicians concern themselves with diagnosing specific conditions, not with making judgements of disorder as such. As with Spitzer's definition, one of its key roles is provide the standard by which specific conditions are granted the title of disorders. If a candidate for consideration as a mental disorder fails to conform to whatever definition is adopted, then it should be excluded from a work such as the DSM. This would then mean that every diagnostic category available to the clinician has met the relevant condition/s, and is a genuine disorder – to the extent that one accepts the conditions given by the definition, of course.

Validity, in the language of diagnosis, usually refers to the property a category has of tracking objective features of reality, particularly those held responsible for the etiology of a disease (Kendell 1989, 46). This is sometimes referred to as *construct* validity. Cholera, for example, is represented by a robustly valid diagnostic category in so far as it relates a constellation of signs and symptoms to a discrete real-world causal entity, namely the bacterium *Vibrio cholerae*. What Wakefield means by *conceptual* validity is somewhat different, however. It is the extent to which the definition of a concept succeeds in creating homogenous classifications, regardless of whether it actually refers to a mind-independent entity. For example, Wakefield holds that the DSM definition is conceptually invalid to the extent that by using its 'unexpectable response' criterion we would end up including variations on normal ability in the category of disorder.

Spitzer actually made reliability, not validity of any kind, his priority, both for his general definition and for the conditions coming under it (Spitzer, Endicott, and Robins 1975). Reliability "refers to the consistency with which subjects are classified" (Spitzer and Fleiss 1974, 341). It is the property of a category's description that guides different clinicians to the same diagnosis, given the same information. A description of cholera is reliable to the extent that any number of clinicians can, with its aid, determine that a person exhibiting the relevant symptoms is suffering from that disease and not any other, including those that might be very similar in appearance. Spitzer's decision to use the operational approach was influenced by his concern for reliability, since it minimises interpretation (which was very much a feature of the theoretical character of the early DSM) in favour of relatively unambiguous criteria by which category membership can be determined. Nonetheless, a reliable category may not be conceptually valid. The criterion of unexpectable response at the heart of Spitzer's definition of disorder might well be reliable – it is relatively easy to judge what meets that criterion. Unfortunately, as we have just seen, the criterion is also met by things that we would not want to include under the concept of disorder.

Wakefield's concern with conceptual validity is not to establish validity empirically, and for individual disorders, but to establish an *a priori* form of validity for 'disorder' (physical and mental) through conceptual analysis. His is a specifically philosophical task geared towards determining what we mean – and what we *always* meant, consistently, across disparate cultures, for at least two millennia. And what he achieves by this is to avoid – or to minimise – false positives such as unusual variations of normal ability. As he puts it, "The requirement that a disorder must involve a dysfunction places severe constraints on which negative conditions can be considered disorders and thus protects against arbitrary labelling of socially disvalued conditions as disorders" (Wakefield 1992b, 386).

Although Wakefield has written that "A central impetus for my formulating the HDA was my rejection of Boorse's (1976) critique of Larry Wright's (1973) seminal work on the etiological approach to function when applied to the specific case of biological functions" (Wakefield 2021b, 267f), Boorse is hardly mentioned in his earliest work on the HDA. In more recent work he has been much clearer that the issue is principally to do with the question of conceptual validity. Most importantly,

> . . . Boorse's statistical view has no answer to the question of how to set the range of normality of a function, declaring the boundary between normal function and dysfunction to be wholly arbitrary . . . accepting Boorse's . . . position would be disastrous for achieving one of the primary goals that motivated the search for a definition of disorder in the first place: to limit false-positive diagnoses in which social deviance is mislabelled mental disorder and thus to respond to

> antipsychiatric claims that psychiatric diagnosis is misused for social control purposes by creating overly inclusive categories. (Wakefield 2021c, 235)

There is also the problem that, as with Cummins, for Boorse functions are relative to the interests of the investigator, meaning that function as 'contribution to a goal' is relative to a subjective formulation of what the goal is. Essentially, as discussed above in the discussions of Cummins and Boorse, the problem is that goals themselves are not really objective. This is exactly the point about Wright: he is uninterested in goals and contributions to them, and interested purely in the historical story about how some effects, and not others, are decisive in explaining the presence of the thing that has that (and other) effects. So Boorse does give a naturalist account, in a sense: our judgements of disorder are about natural facts. But, since subjective choice is involved in *which* natural facts are of interest, it homes in on what 'a' function of something is rather than what 'the' function is (what Boorse refers to as weak and strong function statements). In the medical context – and, more pertinently, in the psychiatric one – it can be argued that what we are after is 'the' function of a thing, since if that thing can, in a weaker sense, have several effects that can be judged functions, depending on interest, we are left without a decisive way to establish which of them (if any) should be considered a function *independent* of interest, which is precisely what objectivity is often taken to imply.

The enormous importance he attaches to this task will be part of the subject of the following chapter. For the moment, it is enough to acknowledge that this is the primary objective of his analysis. This obviously contrasts with Spitzer and the DSM, because this was driven by slightly different (i.e. professional political) concerns that led him down the path of reliability, and he openly expressed his suspicion that a precise definition of disorder would not be achievable (although that does not mean he didn't believe that the concept itself was an objective one). He presumably did not realise that there was a problem with heterogeneity – once recognised, this would have to be a problem for any analysis with pretensions to naturalism. But the principle guiding him was not, on his own account, that of eliminating heterogeneity. Boorse, meanwhile, must similarly believe that the BST will not lead to heterogeneous classifications, although the problem doesn't loom large in his work on the subject.

Conclusion

I began this chapter with a preliminary discussion regarding the possibility of giving 'mental disorder' a general definition, and what purposes such a definition might fulfil, before sketching the history of the *Diagnostic and Statistical Manual of Mental Disorders*, or DSM, which, since its third edition in 1980, has contained a general definition and which has come to represent the mainstream view in psychiatry. In so doing, I drew attention to a feature of psychiatry unusual among medical disciplines: the notable lack of knowledge surrounding the cause, progress, and treatment of mental disorders. This has resulted in a wide range of competing psychiatric theories and the difficulty of characterising mental disorder in a way that would remain neutral with regard to those theories – something that the authors of the DSM felt it necessary to do for both political and epistemological reasons. I showed how the resulting definition employed an operationalised approach as a strategy for avoiding theoretical commitment. Consequently, I turned to post-war debates about the use of functional-explanations in science, which has determined the form that other 'dysfunction analyses' of disorder have taken since the 1970's. I then discussed one of the most famous, and influential, of these analyses, that of Christopher Boorse. Finally, I introduced Jerome Wakefield's highly influential alternative account, which, like Boorse's, aims to provide an objective analysis of function, but is committed to a narrower, 'etiological' understanding of biological function upon which to base his concept of medical (and thus mental) disorder. I will now turn to two competing, though more radical, narratives which will be the subject of the following chapter: antipsychiatry and biological psychiatry.

Chapter 2

The Purpose of the Concept of Mental Dysfunction

Introduction

There is one factor that motivates Wakefield's project of giving disorder a value-free analysis more than any other. It is to "answer the question of why psychiatry is not just a sophisticated form of social control that wraps itself in the banner of medicine, thus a discipline that uses medical technology and jargon to classify and control people but is not really about disorder in anything like the sense that medicine has traditionally understood it" (Wakefield 2010, 10). The question is often associated with the so-called 'antipsychiatry' movement that emerged in the early 1960s on both sides of the Atlantic, and which enjoyed a relatively brief period of public interest.

It is difficult to give a summary account of antipsychiatry but, as Wakefield suggests, a theme that unites the work of its representatives is scepticism about whether psychiatric conditions are disorders in the same sense as other medical complaints. Although this might not sound controversial in itself, some of its most prominent representatives expressed versions of this scepticism that were considered scandalous by mainstream psychiatrists. In sociologist Nick Crossley's words, "Anti-psychiatrists, in contrast to previous and many subsequent critics, did not question particular treatments or policies, nor did they simply argue for a more humane psychiatry . . . they questioned the very basis of psychiatry itself: its purpose, its foundational conception of mental illness and the very distinction between madness and sanity itself" (Crossley 1998, 878). However, the influence it acquired in the 1960s and 1970s had diminished by the time Wakefield introduced the harmful dysfunction analysis. It may seem surprising, therefore, that it should be a matter of such concern to him and his efforts to define mental disorder. In this chapter, I will begin by looking at why he took the claims of antipsychiatry so seriously, and why this is important for understanding his analysis. This, in turn, will require me to look back at the history and development of these views.

It will transpire that some notable antipsychiatric arguments dovetail with what at first sight would seem to be an entirely opposite point of view, that of biological psychiatry. Biological psychiatrists are firmly of the view that mental disorders can and should be understood in purely biological terms, and that research will eventually uncover their physical causes. In order to take on some of these arguments, Wakefield's analysis must address a philosophical puzzle with a very long history – the mind-body problem. For one of the great difficulties faced by psychiatry is the suggestion that if mental disorders are brain disorders – as most would agree they are – then we have a dilemma. If the neurobiological causes of a given psychiatric condition are known, it would seem actually to belong to a different branch of medicine. And if they are not known, then it is questionable as to whether that condition should be considered a medical problem at all. Neither of these options will be palatable to most psychiatrists, and Wakefield had expended considerable effort to show that the dilemma is in fact a false one. In order to get clear on this, it will be necessary, therefore, to examine in turn the philosophical underpinnings of biological psychiatry, as well as those of the antipsychiatrists. With a complete understanding of what the HDA is trying to accomplish, I will then be in a better position, later in this thesis, to question some of its central claims through the lens of Kant's critical philosophy.

Antipsychiatry

In one of the papers in which Wakefield introduced his harmful dysfunction analysis, he notes "Public concerns about misapplication of the term disorder underlie accusations of sexual, racial, and sexual orientational biases in diagnosis . . . as well as more general accusations that psychodiagnosis is often used to control or stigmatize socially undesirable behaviour that is not really disordered . . . . (Wakefield 1992b, 373). In support of this observation he cites various works by authors including Thomas Szasz, R.D. Laing, Michel Foucault, and Erving Goffman – all of them associated with the antipsychiatry label. Decades later he would go so far as to say that "in light of antipsychiatry, the 'dysfunction' component does present what amounts to an *existential issue* for the standard view of psychiatry as a nonoppressive medical discipline" (Wakefield 2021d, 140; emphasis mine).

In the intervening period he has made similar points with varying degrees of force. It is clear that, whatever other benefits he considers the HDA to possess, this is the key challenge that he intends it to address. What exactly this challenge amounts to is perhaps something he could expect his readership to be familiar with, although in fact the antipsychiatric position has often been reduced to a caricature. Wakefield himself, however, seems to be well-versed in its history. To appreciate exactly what he is attempting with his HDA, and why, it is important to look at some of the antipsychiatric arguments that bear most directly on the concept of mental disorder itself.

Although neuroscience progressed by leaps and bounds during the 20th century, our understanding of the mind, and particularly of its discontents, remains very limited. Nonetheless, psychiatry has always insisted upon its medical, therefore scientific, credentials. In the 1950s, however, a number of practicing psychiatrists, alongside a group of social scientists, began to seriously question its medical status. These thinkers shared moral concerns about the treatment of psychiatric patients that had become increasingly prominent in the late 19th century. It was, to be sure, one of psychiatry's early achievements that it ameliorated the atrocious conditions under which the insane were kept. Its founding father, Philippe Pinel, had famously liberated the inmates of the great Paris asylums from their shackles (Porter 2002, 105), and conditions for the seriously mentally disordered undoubtedly improved in the century that followed, in many parts of the world. But abuses continued and were now justified by appealing to the purportedly scientific authority of the psychiatrist.

Campaigners in the late Victorian era publicised, in particular, the ease with which families could have inconvenient members institutionalised virtually at will, without any credible diagnosis being made, simply on the say-so of a sympathetic psychiatrist. Women were particularly vulnerable to this sort of treatment, often at the hands of their husbands or family members (Fauvel 2013). By questioning the very foundations of psychiatry, however, the critics that emerged in the late 1950s were not merely drawing attention to instances of malpractice. They were, in a variety of ways, querying the concept of mental disorder itself.

This rather disparate band of thinkers came to be grouped under the label of 'antipsychiatry'. The two most prominent were Thomas Szasz in the US, and R. D. Laing in Britain. It should be noted that the label 'antipsychiatry' was rejected by many of those to whom it was applied. Popularised by David Cooper, a South African psychiatrist domiciled in London (Nasser 1995, 744), Szasz would devote much ink to defaming the writers he considered to be 'the antipsychiatrists', with Laing as their supposed leader, while Laing himself disavowed the label. The term tends, therefore, to obscure both the actual content of the work it has been applied to and the very significant differences between its individual protagonists. Certainly, not all of its alleged representatives advocated, in any straightforward sense, the literal dissolution of psychiatry. Nonetheless, the label has endured. Though it may misrepresent, it does refer to an identifiable body of writers and, to that extent, is still useful. With this caveat in mind, I will devote the first half of this chapter to getting clear on the antipsychiatric arguments that are most pertinent to Wakefield's analysis. I will begin with perhaps the most important – for our purposes – of these: the arguments of Thomas Szasz.

Thomas Szasz and *The Myth of Mental Illness*

Szasz was a psychoanalyst practicing at a time – the 1950s – when licensed analysts in the US had also to be qualified psychiatrists (Shorter 1997, 194). From the present perspective, so thoroughly influenced by the principle that medicine is founded upon facts more than theory, this must seem rather strange, though, as noted in chapter 1, psychiatry and psychoanalysis were very thoroughly intertwined in the States until the late 20th century. In 1961, Szasz published his first and most famous work, *The Myth of Mental Illness*. He opens the book with a characteristically confrontational statement: "Psychiatry is conventionally defined as a medical speciality concerned with the diagnosis and treatment of mental diseases. I submit that this definition, which is still widely accepted, places psychiatry in the company of alchemy and astrology and commits it to the category of pseudoscience" (Szasz 1974, 1).

Szasz conflates the terms 'disease' and 'illness' which, as discussed in chapter 1, are often given distinct meanings. As his opening statement indicates, however, what he challenges is the notion of a distinctly psychological category of *diseases*. Despite the misleading title of his book, he does not suggest that mental *illness*, in the sense of a subjective sense of psychological suffering, is a myth. His argument is that if somatic causes were discovered for a (so called) mental disease, it would cease to be 'mental' at all, and would properly be classed as a neurological condition. Conversely, in the absence of known somatic causes, psychiatric conditions cannot rightly be considered medical disorders. Given that no biological etiology had been identified for any mental disorder (then as now), psychiatry had no right to call itself a medical discipline. However, this argument rests rather heavily upon the 'lesion' model of disease: "Since in [the] original meaning of it, illness was identified by altered bodily structure, physicians distinguished diseases from nondiseases according to whether or not they could detect an abnormal change in the structure of a person's body" (Szasz 1974, 11)[8] As some of his critics have pointed out (e.g. Bentall and Pilgrim 1993), this is not the only, or even the dominant, mode of considering medical disorder. For example, the approach adopted (of necessity) by psychiatry, of identifying categories of disorder with clusters of symptoms, is also very widely used in physical medicine.

Nonetheless, Szasz does not deny the existence of mental suffering of a distinctive sort. Szasz's preferred locution for what psychiatrists have called mental disorder is 'problems in living', problems that may be caused by personal, social, or ethical challenges. The cornerstone of Szasz's analysis is summarised by his assertion that "Virtually all behaviour with which the psychoanalyst and psychiatrist deal is learned behaviour" (Szasz 1974, 153). The sort of learning at issue takes place, he argues, in the context of norms encoded in social roles and rules that are, by their very nature, aspects of the sociocultural milieu in which the individual exists. Persons may follow or flout these norms but, in either case, their actions are to a great extent determined by them. For Szasz, mental illness, so-called, is fundamentally a clash between the individual and the contingent social matrix of roles and rules in which they are enmeshed. Its symptoms are a form of idiosyncratic communication aimed at expressing what may be socially unacceptable (because contrary to the prevailing system of norms) sentiments. As Szasz explains, "Indirect communications ensure the speaker that he will be held responsible only for the explicit meaning of his message. The overt message is thus a sort of vehicle for the covert message whose effect is feared" (Szasz 1974, 142). Whatever form a person's 'problems in living' take, it is quite likely that they will transgress the norms that, very generally, regulate personal conduct. By Szasz's lights, whatever is perplexing about the behaviour of the mentally disordered person can be attributed to the difficulty, and perhaps the danger, of openly expressing conventionally disvalued attitudes towards social relations. Since this behaviour is not typically interpreted by others as communicative at all, it provides a relatively safe means of expressing these attitudes.

The Myth of Mental Illness is full of insights and observations that are, in themselves, thought-provoking, and Szasz's arguments are often ingenious. Taken as an analysis of mental disorder *per se*, however, it is highly superficial. His argument relies almost exclusively upon the specific category of hysteria, a diagnosis that was anachronistic even at the time the book was first published; he mentions schizophrenia only in passing,

[8] This view is often credited to Rudolf Virchow (1821-1902), see, e.g. Bentall and Pilgrim 1993, 72.

and other diagnostic categories not at all. He asserts from the start that "I believe that the interpretation of hysteria which I shall present pertains fully – with appropriate modifications – to all so-called mental illnesses . . . ." (Szasz 1974, 10) – but nowhere offers any justification for this view, nor any examples of the sorts of 'appropriate modifications' that might be required. It should also be noted that hysteria is a condition in which the patient reports physical symptoms of various kinds that lack any detectable somatic origin. It is not, therefore, characterised primarily by changes in cognitive or emotional states. Consequently, Szasz frames his analysis entirely in terms of overt behaviour and has very little to say about mental states – the states that are generally held to be of special significance in mental disorder.

Szasz is also not very clear about 'problems in living'. Being at odds with other people and with the values and expectations of one's society is, one way or another, quite commonplace. The 'problems in living' formulation is too permissive: it encompasses far more than the conditions thought of as mental disorders. Szasz provides no further insight that would clarify the matter. In fact, Szasz was, by his own admission largely uninterested in understanding or explaining in any detail how mental disorders arose. For him, it suffices to assert that they are in fact social conflicts, not medical phenomena, and his own approach to therapy, as a psychoanalyst in private practice, was directed very much towards, as he put it "mastery of interpersonal processes" (Szasz 1974, 213), with the analyst serving as a kind of advocate for their client.

Szasz's views owe much to his psychoanalytic training. Although stating his views in uncompromising language, they were to some extent a tacit return to Freudian orthodoxy at a time when an ascendant psychoanalysis was widening its field of operations. Freud and his fellow travellers confined their attention mainly to the neuroses; by the immediate post-war years in the US psychodynamic theory was asserting its role in treating the psychoses as well. Indeed, under the aegis of Adolf Meyer's mental hygiene movement, it had taken on a far-reaching vision of purging civilisation of its discontents pre-emptively (Christiansen 2007). Against this largely-forgotten backdrop, Szasz's early arguments make more sense than they otherwise would to a contemporary reader. He was not opposed to the principle of biomedical research into mental disorder, though the field was small and had revealed very little; he presumably considered it a lost cause. Neither was he opposed to psychoanalysis *per se* – he continued his own practice to the end of his life. The object of his ire was very specifically the form that a psychodynamically-oriented psychiatry had assumed in the United States in the post war years. With the passing decades and the decline in influence of psychoanalysis, much of his early writing has come to appear obscure or misinformed. Provided we understand the context in which he wrote, and the changes that have occurred in the way psychiatry is practiced, we will be able to see how at least some aspects of his position continue to pose a real challenge to the idea that mental disorder is a genuinely scientific concept.

Szasz was a particularly fierce critic of the power of psychiatry – the power to incarcerate people against their will, the power to influence legal judgements, and, more generally, its role in what he termed the 'therapeutic state' and its creeping influence in private affairs. He was essentially a right-wing libertarian of a distinctively American sort, whose commitment to personal liberty was yoked to an extreme aversion to any form of collectivism. Szasz's critique of psychiatric power was given a fillip by a now-notorious experiment carried out by psychologist David Rosenhan in 1973.

Rosenhan recruited eight 'pseudopatients', men and women of various ages and occupations with no history of mental disorder, to see if they could gain admittance to psychiatric hospitals by presenting a minimum of evidence. The pseudopatients were to make appointments at various randomly selected hospitals across the US, reporting just one symptom: for three weeks past they had been hearing a voice repeating the words 'thud', 'hollow', and 'empty'. In all other respects they were to act normally and respond honestly to any questions put to them. In the event, all eight were admitted as in-patients, seven of them with a diagnosis of schizophrenia, with one diagnosed as manic-depressive. Upon admission, the experimental subjects were to report that their symptom had ceased, and were to continue behaving normally. All of them were kept at their institutions for a period of at least several weeks before being released as being 'in remission'. None of the staff at the hospitals concerned voiced any suspicion – although their fellow patients did. Furthermore, Rosenhan followed this up with a second experiment, arranged with another hospital, in which he proposed to send a number of pseudopatients over a period of three months and asked the staff to identify them when they presented themselves. Hospital staff flagged 41 patients out of 193 as potential fakes; in fact, Rosenhan hadn't sent any.

Rosenhan published the results of his experiment in his paper 'On Being Sane in Insane Places' (Rosenhan 1973), in the prestigious journal *Science*. Understandably, it caused a furore. For some it was a conclusive indictment of psychiatry and its institutions; for others (and not only those in the psychiatric profession) the methodology was flawed and the conclusions drawn unwarranted. Robert Spitzer wrote a detailed critique of the experiment which identified a number of important respects in which Rosenhan's claims were not supported by the evidence, as well as arguing that the methods used were weighted against the diagnostic process. Of particular concern to Spitzer was the simple fact that psychiatrists were not in the job of picking out fake patients, but of diagnosing patients who they quite reasonably expected to act in good faith. After all, Spitzer pointed out, the pseudopatients had apparently, and of their own volition, sought help (Spitzer 1975). It was fair to assume that they *needed* help. Arguably, however, this misses the point somewhat. The inescapable fact is that the threshold for being admitted as an in-patient to a psychiatric hospital had been clearly demonstrated to be absurdly low. The experiment confirmed Szasz's suspicions that they were, at that time, far too ready to diagnose serious mental disorders with minimal evidence.

Today, for a number of reasons, the situation has changed. Many psychiatric hospitals in North American and Europe have closed, and the overwhelming emphasis is on treatment within the community. Concerns remain, of course, regarding involuntary confinement, restraint, and treatment of mentally disordered persons. The emphasis has, however, shifted away from the more egregious abuses of earlier times to more subtle problems. The principle at stake, however, remains very much the same. Wakefield himself has (with Allan Horwitz) addressed the case of depression which, as a psychiatric category, he considers to have been grossly expanded to cover what probably ought to be considered ordinary sadness (Wakefield and Horwitz 2007). This looks to be an instance of what Szasz called the 'therapeutic state' (Szasz 1984): the suppression of natural reactions by institutions more concerned with the social, economic, and political status quo than the wellbeing of the individual. Indeed, one author considers Wakefield to actually be a contemporary *representative* of the antipsychiatric view (Whitley 2012), although Wakefield himself would be bound to reject this as a distortion of his opinions.

R. D. Laing and antipsychiatry in the UK

Those identified with antipsychiatry in the US often framed their arguments in terms of the legal and custodial power of psychiatry – that is, in terms of personal liberty, and of the spreading influence of the state. In Britain, by contrast, the emphasis was placed on giving the patient a positive, representative role in their own treatment and recovery, and the concomitant importance of understanding their experience. This was particularly so for R. D. Laing, whose book *The Divided Self* was published in 1960. Like Szasz, Laing was critical of psychiatry as a medical discipline, although he was much more of a realist about mental disorder than Szasz. Interested primarily in schizophrenic experience, Laing insisted (contra most psychiatrists up until that point) that what the schizophrenic subject did and said had its own internally consistent logic, which could be rendered intelligible by a sensitive therapist. In this respect he was, however, preceded by the activities, throughout the 1950s, of the Project for the Study of Schizophrenic Communication, based in the US and led by the British anthropologist Gregory Bateson. Most famously, this group developed the theory of the 'double-bind': a recurring situation in which a person finds that "every move is subsequently demonstrated to have been wrong by the moves which other members of the system make in response" (Bateson 1973a, 241). The 'system' in question is essentially one of communication with significant others, characterised as occurring on various different levels, many of them implicit. Trapped in a situation in which no act or utterance can ever be the 'right' (or even a satisfactory) one, the schizophrenic subject subconsciously evolves coping strategies aimed at surviving what is perceived to be a hopeless situation. Bateson and colleagues argued that, seen in this light, it was unsurprising that the resulting thoughts and behaviours should appear irrational. Indeed, the situations from which they arose were themselves irrational. However, emphasising the communicative aspect, they insisted that what *seemed* absurd or preposterous could be understood or 'made sense of' if the circumstances that were the origin were grasped.

Laing viewed his own approach as 'existential-phenomenological', albeit in a loose sense that made little explicit reference to specific philosophers in the phenomenological tradition or their works. As he put it, the task was "to set all particular experiences within the context of [the subject's] whole being-in-his-world" (Laing 1990, 17). In his view, 'mad' speech and behaviour could only be deciphered if a commitment was made to grasping the particular psychological environment from which it issued. *The Divided*

Self did not offer a theory of schizophrenia – rather, Laing developed a number of concepts that he considered characteristic of the condition and that offered points of access to understanding its more florid symptoms. Of these, the notion of 'ontological insecurity' was perhaps the most important. The ontologically secure individual, no matter how frail they may be according to other evaluative indices, at least possesses "a centrally firm sense of his own and other people's reality and identity" (Laing 1990, 39). Indeed, it is likely an unquestioned presupposition that grounds their experience. By contrast, the ontologically insecure person encounters the world without this anchoring certainty. Their experience is, in a fundamental (if difficult to articulate) way, precarious, attenuated by the near constant fear of their own complete existential dissolution. Tellingly, Laing finds it easier to allude to works of literature and art in an attempt to convey his meaning: Kafka, Beckett, and the painter Francis Bacon are all evoked. This very fact, he suggests, may indicate the presence of such insecurity not very far below the surface of everyday life. We shall see in chapter 5 that the psychiatrist and philosopher Louis Sass would later devote a monograph to a variation on this very theme.

Along with colleagues such as David Cooper and Aaron Esterson, Laing shared with Szasz an interest in the communicative aspects of mental disorder. Unlike Szasz, they were distinctly left-wing in orientation, actively interested in communitarian models of treatment, and deeply invested in the counter-culture of the 1960s and 70s. This, along with a number of experimental 'therapeutic spaces' that did not earn approval from their neighbours and the press (McGeachan 2014), did little to earn them respect in psychiatric circles. Laing, in particular, was treated far more sympathetically in liberal arts circles and by political activists than he was by his fellow psychiatrists. Increasingly eccentric behaviour and pronouncements, including his frank admission that he had administered psychedelic drugs to patients (as well as consuming them himself) effectively disbarred him from occupying a responsible medical position.

The merits of antipsychiatry

The radical and, at times, eccentric nature of some of antipsychiatry's views should not be allowed to obscure the fact that there is still much in the work of these thinkers that deserves our attention. In the decades in which they were most prominent, their criticism of the ways in which the mentally disordered were treated by the institution of psychiatry had considerable merit. Admittedly, the populations of psychiatric hospitals on both sides of the Atlantic were already falling by the 1950s, and state-led programmes of deinstitutionalisation were underway with the Mental Health Act of 1959 in Britain, and the Community Mental Health Centres Act passed in the US in 1963. Widespread reform was also underway in other European countries, Italy in particular, a country in which antipsychiatry also had a strong presence. In short, the Szaszian spectre of psychiatric authoritarianism was already in retreat, for a number of reasons. In more recent times, 'critical psychiatry' has sought to develop some of the themes championed by the antipsychiatrists while dispensing with the more *outré* aspects of its programme (see e.g. Middleton and Moncrieff 2019).

Of more importance from the perspective of my thesis is the development of the idea firstly that the symptoms of mental disorder constituted a comprehensible form of communication, and secondly, the view (closely linked to the first) that at least some forms of mental disorder were causally rooted in social structures and the pressure they exerted on individuals. Neither of these notions were entirely novel. In some respects, antipsychiatry built upon elements of psychoanalytic theory, particularly Freud's *Civilisation and Its Discontents* – unsurprisingly so, given that many of its key figures were psychoanalytically trained.

The concept of dysfunction prominent in psychiatric discourse is virtually absent in the work of the antipsychiatrists, regardless of how much their views in other respects diverge. For all their differences, both Szasz and Laing interpreted mental disorders as more or less rational attempts to cope with external social pressures rather than failures of biological or psychological mechanisms. They represent one variety of the normativist position. An important corollary of this position is that while the expression of the patient's condition may appear irrational or even bizarre, it is nonetheless a form of subconscious communication. As such, the signs and symptoms of mental disorder can, in principle, be given a meaningful interpretation. The dysfunction analysis has little interest in these phenomena, other than as diagnostic data introduced by a damaged or deficient mechanism. For antipsychiatrists like Szasz and Laing, by contrast, symptoms comprise a formally structured, if idiosyncratic, response to the social environment. Laing believed that the specific content of the patients acts and utterances was therapeutically (and, in a sense, morally) important. Although Szasz similarly believed in their meaningfulness, he placed far greater emphasis on facilitating his patients' progress towards a future in which they would exercise ever greater personal autonomy, apparently viewing 'recovery' in terms of competitive advantage.

We have seen that for Szasz in particular, it is entirely possible that at least some so-called mental disorders will turn out to have determinate physical causes. If they do, however, they will not be *mental* disorders at all, but neurological ones. On this view, psychiatry, as a branch of medicine, simply fails to have any distinctive subject matter of its own, since a purely psychological condition isn't – according to the 'lesion' view of disorder – a medical phenomenon. This apparent conundrum, however, did not prevent the rise of a new breed of psychiatric researcher at around the same time that conceptual analyses of disorder were being developed in the 1970s. The introduction of new technologies at this time held out new hope that a distinctively *biological* psychiatry could finally resolve the mystery of mental disorder.

Biological psychiatry

It is a source of frustration to many psychiatrists that, despite the enormous scientific progress that has been made since the late 1800s, there is much that remains mysterious regarding consciousness and the brain. The early psychiatrists, who came to be known as 'alienists', had very little neuroscientific knowledge to draw upon. Freud, whose psychodynamic theories were to dominate psychiatry for half a century, was himself a neurologist by training, and anticipated that brain science would one day transform the understanding of mental disorder. His ideas emerged as a pragmatic response to the limited neurological knowledge available to him and his peers: an alternative, psychological, model was necessary if any immediate progress was to be made.

Biological psychiatry, in the sense in which I discuss it here, is not merely the practice of biomedical research into mental disorder. It is, rather, a philosophical position that, *qua* mental disorder, assigns ontological and epistemic priority to the brain and nervous system. It is an *a priori* belief that mental disorders are essentially brain diseases whose etiologies will inevitably be discovered. As one prominent advocate of the biological approach, Nancy Andreasen, has asserted, "The important question is not 'Will it happen?' It is 'How long will it take?'" (Andreasen 2001, 323). In consequence, biology is held to constitute the correct *foundation* of psychiatry, in line with the rest of medicine. This is expressed in an opinion-piece from 2005 in which Thomas Insel and Remi Quirion talk about "re-defining the foundation of psychiatry as clinical neuroscience" (Insel and Quirion 2005, 2223). To be sure, proponents of this view almost invariably temper it with an insistence upon the continuing importance of psychology, sociocultural sensitivity, and awareness of environmental factors. Nevertheless, it is clear that they can have only a supporting role: the brain and its activity have explanatory priority.

Although the notion that madness has somatic origins is a very old one, the implications for psychiatry have been historically slight. In general, pessimism about the potential of science to contribute practical solutions to the problem of mental disorder has been a recurrent theme from its inception. The early psychiatrists focused their energies on reforming the asylums and applying empirical methods to the signs and symptoms of madness, attempting to construct a 'top-down' representation of the various mental disorders (primarily the psychoses). It was, therefore, something of a provocation when, in 1886, the German psychiatrist Emil Kraepelin argued for a more medical approach, including research into the biological causes of mental illness (Kraepelin 2005). In general, however, the brain proved less amenable to scientific understanding than the other organs of the body. There were advances at the end of the 19th century – above all, the formulation of the 'neuron doctrine' by Cajal and his colleagues (see e.g. Guillery 2004) – but this shed very little light on mental disorder. By the second decade of the 20th century, the tide was turning in favour of the theoretical innovations of Freud and his colleagues.

As noted in the previous chapter, advances in pharmacology and professional concern regarding the status of psychiatry as a robustly medical discipline led to a process of change from the 1960s onwards. The shift was particularly noticeable in the US, where the profession had been dominated by psychoanalysis since before World War Two. In opposition to the highly theoretical, speculative psychoanalytic approach, a number of psychiatrists began to press for a more empirical emphasis, not least through the reorientation of the *Diagnostic and Statistical Manual of Mental Disorders*. What emerged from this process was what has been called the 'biopsychosocial' model of psychiatry, also referred to as the 'weak' medical model. What is 'medical' in this model is an evidence-based approach and an emphasis on diagnosis and nosology, something lacking in the psychoanalytic tradition. Spitzer and his colleagues did not simply renew the emphasis on observation in clinical practice. After all, while psychoanalysis in the US was in the ascendant, this sort of empiricism had to a great extent remained the tradition in Europe and Britain. The key innovation was the methodical collection of empirical data, and its organisation into diagnostic categories designed to improve reliability. From the perspective of the present day, this appears less radical (and perhaps less 'medical') an innovation than it did at a time when psychoanalytic theory held sway.

In the 60s and 70s, advocates for a more medical approach may have believed that biology was ultimately the proper basis for psychiatry, but they recognised that significant advances were likely to lie far in the future. For them, the new data-driven approach to classification and diagnosis offered the brightest short-term prospects for psychiatry. When Gerald Klerman, a colleague of Robert Spitzer, wrote of a 'Kraepelinian revival' in 1978 he argued that "The focus of psychiatric physicians should be particularly on the biological aspects of mental illness" (Klerman 1978, 104). Nonetheless, in his subsequent analysis he paid relatively little attention to biology and concentrated instead upon issues of classification. Samuel Guze, another of Klerman's neo-Kraepelinians, wrote an article titled 'Nature of Psychiatric Illness: Why Psychiatry is a Branch of Medicine' in 1978, but made only the bland observation that "Adherents of the medical model generally consider biologic processes important in the development of psychiatric disorders, while opponents of the medical model generally minimize the role of biologic processes" (Guze 1978, 302f). All of this is to say that agitation for a biomedically-founded psychiatry emerged rather gradually from a 'medical' perspective that was initially more committed to an evidence-based diagnostic approach than to one specifically privileging a biological level of explanation and understanding.

Those pressing for a more biological approach in the 60s and 70s were still constrained by the state of those sciences that, it was hoped, would shed light on the somatic foundations of mental disorder: neuroscience and genetics. In the mid-1970s, when work began on DSM-III, these were still at a relatively early stage of development. Conversely, advances in psychopharmacology made in the postwar period had largely been serendipitous; it became understood that certain brain chemicals were implicated in conditions such as depression and schizophrenia, but it wasn't clear exactly how or why. By the time the new DSM was published in 1980, however, technological advances were sparking a great leap forward, particularly in the field of medical imaging. It was gradually becoming possible to observe the brain *in vivo*, and in increasing detail, using technologies such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) (Turner and Jones 2003). As the decade progressed, a truly biological psychiatry once again seemed possible. By its end, Guze, now in more confident mood, was able to advocate "a conceptual approach to psychiatry that is rooted explicitly in biology" (Guze 1989, 319).

Biology and the quest for validity

The weakly medical, biopsychosocial approach underpinning DSM-III and its successors is pluralistic: it does not assign privileged status to a single level of explanation, whether somatic, psychological, or social. One might more informally say that it hedges its bets. What makes biological psychiatry, as a philosophical position, distinct from this view is its belief that biology should provide the intellectual foundation for psychiatry. There may still be pragmatic reasons for thinking about mental disorder in psychological or environmental terms, but the basic ontology is materialist.

This is reflected particularly clearly in changing attitudes towards the DSM. While its empirical approach was hailed as a win for the more medically-minded psychiatrists, as time wore on there was growing realisation that the doubtful validity of at least some of its categories – the extent to which they actually represented real-world disease entities, as discussed in the previous chapter—was obstructing research. With each new edition of the DSM "the ever increasing fractionation of mental distress into smaller and more numerous categories, without a priori biological validity, makes it harder to find specific biomedical tests that diagnose or predict the disorders" (Kapur, Phillips, and Insel 2012, 1175).

For all its improved reliability, the DSM remains a top-down schema in which signs and symptoms constitute the effects for which a cause is ultimately to be sought. But if the existing category does not classify a natural kind, then a unified explanation is being sought for syndromes that do not in fact have a common cause. Biological psychiatry has, therefore, been increasingly at odds with the biopsychosocial model, culminating in the Research Domain Criteria (RDoC) initiative at the National Institute of Mental Health in the US, the goal of which "is to provide information about the basic biological and cognitive processes that lead to mental health and illness, broadly conceived. The information gained using RDoC may help inform the creation of mental health screening tools, diagnostic systems, and treatments" ('About RDoC' n.d.). This approach is intended to guide basic research into mental disorder and eventually to identify pathological states with clearly defined effects. If successful, the cumulative effect would be the revision of existing DSM categories to reflect psychiatric natural-kind classifications. This – the pursuit of clinical validity – is the holy grail of biological psychiatry, and drives its push for ontological priority.

Various critics have identified biological psychiatry with *reductionism*. It isn't always clear what they mean by the term, nor that the views they critique necessarily deserve to be called reductionistic, except in a very loose sense. A reduction is the description or explanation of an element of scientific knowledge in terms of more basic constituent elements. The canonical example (Nagel 1961b) is of the reduction of the Boyle-Charles law in thermodynamics by way of the kinetic theory of gases. The latter theory belongs to statistical mechanics, a field of physics in which terms such as 'temperature' and 'entropy' – familiar from thermodynamics – have no application. James Clark Maxwell and Ludwig Boltzmann (working independently) were able to account for the Boyle-Charles law in terms of the collision of molecules (kinetic theory), without recourse to concepts particular to thermodynamics. A thermodynamic law had been 'reduced' because it could be stated in terms of more basic entities (molecules).

Reduction has often been undertaken in the normal spirit of scientific enquiry and intellectual curiosity, and does not, in itself, imply a principled stance. *Reductionism*, by contrast, denotes a commitment to the systematic pursuit of reduction. Historically, the aim of reductionism was the complete unification of scientific knowledge. Although a notion with a long history, unification was given a 20th century update through the work of the logical positivists, particularly that of Otto Neurath and Rudolf Carnap. For them, this ideal – the possibility of being able to express all scientific knowledge in terms of a limited set of universal principles – was motivated by concerns about specialisation and the need to facilitate communication and cooperation between domains of science that employed very different conceptual schemes. Although logical positivism has fallen from favour, the essential idea remains powerful. A more modest (though still ambitious) unificationism tends to focus on the explanatory potential of more limited, local reductions. This has been a particularly contentious topic in philosophy of biology, revolving principally around the possibility of reducing genetics to molecular biology.

In psychiatry, the debate has been couched in slightly different terms. Its use has been largely pejorative, and often in a sweeping sense. Gold, for example, suggests that reductionism amounts to the view "that neuroscience . . . and molecular biology will, on their own, eventually provide an exhaustive explanation of mental illness and form the basis for treating it successfully" (Gold 2009, 506), while Brendel associates it with "the exclusion of alternative explanatory concepts" (Brendel 2003, 565). It could be argued that these characterisations are straw men. Arguably, the biological position does have the effect of marginalising 'alternative explanatory concepts', but few of its adherents would argue for their exclusion. Biologically-minded psychiatrists rarely talk about reduction at all (but see Schaffner 2013), let alone about the desirability of reduction as an overarching strategy.

It should be noted, *a fortiori*, that biological psychiatry is not eliminativist – its proponents do not claim that psychology will be replaced by talk of neurobiological mechanisms. Although some eliminative materialists in philosophy of mind have discussed psychiatry in those terms, Kandel, Andreasen, and others have insisted upon the continuing importance of psychology and the language of mentation – albeit in a subordinate role. Kandel – whose early training was in the psychoanalysis – has even written about the potential for bringing psychodynamic theory and neuroscience together by identifying the brain mechanisms in which Freudian concepts such as the ego are actualised (Kandel 1999). Although conciliatory in appearance, however, Kandel's premise remains resolutely biological. Mental disorders, understood in psychological terms (whether in the conceptual language of Freudian psychoanalysis or in some other framework), are nonetheless brain disorders, if Kandel's other statements are to be taken seriously.

Biological psychiatry and the dysfunction analysis

The sense of 'dysfunction' discussed in the previous chapter was abstract: in the case of the DSM itself, its meaning is operationalised, while for writers such as Boorse and Wakefield it is an impairment of a 'psychological mechanism' that has been selected for by evolutionary pressure. These accounts say little or nothing about the material realisation of the function or its supposed deficit. For biological psychiatry, the concept of dysfunction is more concrete: mental disorders are brain diseases for which research – guided by frameworks such as the RDoC – will reveal the physiological basis. As Andreasen puts it, "The brain can . . . become 'broken' in many ways that lead to the disorders known as mental illnesses." (Andreasen 2001, 42) Andreasen's 'broken brain' remains hypothetical: there are as yet no known biomarkers[9] for any type of mental disorder. This effectively means that there are no physical features definitively associated with any mental disorder, such that a laboratory test or similar would confirm or disconfirm diagnosis. But this, it is argued, only reflects the enormous complexity of the brain and nervous system, and the current shortcomings in scientific knowledge and technology. As we have seen, Andreasen, for one, expects that research will eventually reveal mental disorders to be brain disorders.

Wakefield correctly points out that a "more basic problem is that the distinction between normal variation and abnormal physiological functioning itself requires a functional account, so medical disorder is essentially a functional, not anatomical or physiological, concept . . . ." (Wakefield 2003, 988f). Ethan Gorenstein boils this type of observation down into its essentials: "That a particular phenomenon has a cause in no way implies it is a disease" (Gorenstein 1984, 53). If we have not already settled the conceptual question of what a dysfunction is, we are in no position to attach the label to any physical phenomenon whatsoever, regardless of how unusual it is. We need to know that it *ought not* to be as it is, and not in terms of subjective norms regarding what we value or disvalue. If the hoped for discoveries of biological psychiatry are to be as authoritative as its proponents hope, the 'ought' in this case must somehow be dictated by nature. As we have seen, this concern is at the heart of Wakefield's project. He rightly makes the case that projects like the RDoC lack "any serious conceptual component that might effectively connect its ambitious empiricism with the conceptual problems of diagnosis it aims to resolve" (Wakefield 2014, 39f). Although Wakefield hedges his bets somewhat, he seems inclined towards the belief that a mental dysfunction need not entail a biological one. Biologically-minded psychiatrists have indeed shown little interest in the conceptual question, apparently taking it for granted – as so many before them have done – that everybody understands what they mean by such terms as disorder and dysfunction, and that they all mean roughly the same thing.

Given that decades of research have failed to identify any gross neurological abnormalities that are consistently correlated with mental disorder, it is by now inevitable that mental disease entities, if they exist in any meaningful sense, will be

[9] The term 'biomarker' has been defined as "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." (Biomarkers Definitions Workgroup 2001, 91)

diffuse, or *multifactorial.* They will consist of alterations in many biological mechanisms, each perhaps having only a small effect. Gene expression, neurotransmitter levels, the strength and number of synaptic connections, and various other factors may be implicated. There is no longer any serious expectation that anything as definitive as the cellular or macroscopic changes observed in familiar neurological conditions such as Parkinson's disease will be discovered. Although this complexity presents a formidable challenge, Kandel, for one, argues that the biological principle is secure: "All mental processes . . . derive from operations of the brain . . . As a corollary, behavioural disorders that characterize psychiatric illness are disturbances of brain function, even in those cases where the causes of the disturbances are clearly environmental in origin" (Kandel 2005, 39).

The details of this debate are tangential to my central research question: the legitimacy of the dysfunction requirement for mental disorder in general. What I have tried to do here is to show that certain claims have been made for the narrowly biological approach to understanding mental disorder, claims that are *prima facie* persuasive only because they appear to *follow* from an uncontroversial assumption.

Conclusion

In this chapter I have argued that Wakefield's harmful dysfunction analysis is primarily motivated by the allegations made by members of what has been called the antipsychiatry movement. I went on to address the history of this movement and why it matters for the concept of mental disorder. I then turned to the perspective of biological psychiatry in order to understand why and how antipsychiatric arguments remain relevant in the face of considerable scientific and technological advances, and how Wakefield's own position resists the biological argument that scientific investigation alone can reveal the underlying nature of mental disorders.

Having established the philosophical strategy of Wakefield's analysis of the concept of mental dysfunction, I will in the following chapter introduce the philosopher Immanuel Kant, and examine his own views on mental disorder and their significance for his philosophical development.

Chapter 3

Kant, Metaphysics, and Madness

Introduction

It is a relatively obscure fact about the life and work of Immanuel Kant that among his many and varied interests, both within and without the field of philosophy, was a preoccupation with mental disorder. By contrast, he is very well known indeed for his attack on rationalist metaphysics in his most famous work, the *Critique of Pure Reason*. These two themes are, as it is part of my purpose to show in this chapter, in fact related. I want to explicate this association because I will go on, in the remaining chapters of this thesis, to look at what the first *Critique* itself can contribute to a critique of the concept of mental disorder, even though the highly abstract character of that work seems remote from such earthly concerns.

What links these apparently disparate currents in Kant's thought is a feature of metaphysics that disturbed him early in his career. The elaborate or 'subtle' inferences of reason through which metaphysics pursued knowledge of a supersensible realm, it seemed to him, could resemble madness; conversely, the delusions of the insane might pass for metaphysical speculation. This disturbing notion was a factor driving Kant to try to determine the limits of metaphysics, and resulted in a watershed moment in Western philosophy with the publication in 1781 of the *Critique of Pure Reason*.

In this chapter I will first make a survey of the development of Kant's metaphysical worries in the work of the pre-critical period. I will argue that this association informs the view of reason that he formulated in the *Critique of Pure Reason*. Following this, I will examine his attitudes towards mental disorder as expressed in his earliest writing on the subject, the 'Essay on the Maladies of the Head' (1764). Finally, I will look at the development of these ideas as they appear in Kant's last published work, *Anthropology from a Pragmatic Point of View* (1798).

Kant's pre-critical work and the problem of metaphysics

Kant's works span a period of over half a century, between 1749 and 1803, the year before his death. He published regularly throughout this time, with one hiatus between the essay commonly known as the *Inaugural Dissertation* in 1770 and the appearance of the *Critique of Pure Reason* in 1781: the so-called 'silent decade'. What Kant scholars conventionally refer to as the 'pre-critical' period refers to the early phase of his career up to the *Inaugural Dissertation*.

Kant's interests in this period are diverse but, as Alison Laywine comments, "for all the variety, it takes no great insight to see that the early Kant was above all preoccupied by questions on special topics in metaphysics" (Laywine 1993, 11). In mid-18th century Germany, the intellectual landscape was dominated by the so-called Leibnizian-Wolffian metaphysics. Christian Wolff's *German Metaphysics* of 1720 was heavily influenced by Leibniz,[10] with whom he had corresponded, and "enjoyed a position of unequalled authority in German academies through the first half of the 18th century . . . ." (Dyck 2011). In Wolff's taxonomy, the 'special topics' of metaphysics are cosmology (the fundamental nature of the universe as a whole), rational psychology (the nature of the soul) and rational theology, which pertains to the existence and essence of God. It is these weighty investigations that the Kant of the first *Critique* famously showed to be epistemologically illegitimate. However, the overarching project of the pre-critical Kant, as Martin Schönfeld has argued, was to reconcile metaphysics with the emerging sciences as epitomised by Newton, rather than to radically reorient it (Schönfeld 2000). That is, in this early period Kant still hopes that the traditional questions asked by special metaphysics can be answered within a framework that places metaphysics on the same footing as science.

Gradually emerging in these works is a growing disillusionment that culminated in what many commentators interpret as a crisis in Kant's thought, his polemic against the mystic Emmanuel Swedenborg, *Dreams of a Spirit Seer Elucidated by the Dreams of Metaphysics* (1766). Less well-known among Kant's pre-critical works is an article serialised only a few years previously, in 1764, in a journal for the intellectually inclined middle classes of Königsberg, Kant's lifelong home. This piece, the 'Essay on the Maladies of the Head', was occasioned by the appearance near the city of another mystic, an indigent preacher called Jan Komarnicki, accompanied by a young boy whose rustic mien put Kant in mind of Rousseau's 'child of nature'. Briefly feted by the citizens of the East Prussian capital, Komarnicki's religious enthusiasm and unorthodox lifestyle prompted Kant to reflect upon the varied disturbances of the human mind. In these works, the themes of metaphysics, religious enthusiasm, and mental disorder are

[10] The actual extent to which Wolff's philosophy was influenced by Leibniz is a matter of debate and is discussed in (Corr 1975).

interwoven, and in them the overarching concerns of the first *Critique* emerge in embryonic form. Before attending to these works, I will briefly assess Kant's views on metaphysics as they are expressed in some of his prior writings.

Kant's earliest published work was *Thoughts On The True Estimation of Living Forces* of 1746, an intervention into the *vis viva* or living force debate. This controversy concerned the question of two different conceptions of mechanics and the metaphysical – and therefore empirically undecidable – assumptions underlying them. Of interest here is the young Kant's declaration that "Like many other sciences, our metaphysics is indeed only on the threshold of truly sound knowledge, and God knows when one will see that it has been crossed. It is not difficult to discern its weakness in much of what it undertakes. One very often finds that prejudice[11] is the greatest strength behind its proofs" (Kant 2012, 33). Not only this, but he goes on to say that "a metaphysical investigation, especially one so convoluted and complicated, has still countless hideouts on all sides, to which one can escape from enemies who would be incapable of pursuing or pulling one out." (Kant 2012, 88). In this earliest work we immediately encounter substantial concerns about the project he himself is involved in: that it often comes down to dogmatic belief rather than genuine proof, and that it is vulnerable to a kind of intellectual duplicity.

After several fallow years during which he worked as a private tutor in a household outside Königsberg, Kant submitted his doctoral dissertation to the university there. *A New Elucidation of the First Principles of Metaphysical Cognition* (1755) is often summarised as an attempt to reconcile the deterministic worldview of physics with human freedom. As before, the details of this attempt are unimportant. I only want to note the tone in which Kant assesses the contemporary state of metaphysics, asserting for the reader that the metaphysical principles he has developed therein should lead to greater insight and certainty in metaphysics, which "will be found not to be so barren." (Kant 1992a, 45). Taken together, these comments prefigure Kant's characterisation of metaphysics, in the Critique, as a "battlefield of endless controversies" (Aviii)[12] upon which no one is able to secure a decisive advantage.

[11] The German word translated here as 'prejudice' is *Vorurtheil*, which may also be translated as 'bias' or 'preconception'.

[12] All references to the *Critique of Pure Reason* will use the standard A and B edition pagination.

After a period of several years in which he worked mainly on problems in the nascent earth sciences, Kant returned to philosophy in 1762 with the *False Subtlety of the Four Syllogistic Figures Proved*, his first and only work on formal logic. The *False Subtlety* consists primarily of an attack on the syllogistic 'figures' that Aristotle had identified in his *Prior Analytics*. The figures, in brief, are deviations from the regular form of a logical syllogism. Although not invalid, their proof requires additional inferential steps that reduce them to the regular syllogistic form. The 'subtlety' Kant refers to is just this rearrangement of the logical form of an argument that introduces obfuscation and ambiguity into what ought properly to be a field of precision and clarity. In Kant's opinion, "It is easy to to discover what initially led to this subtlety" (Kant 1992e, 100). It is, he suggests, the capricious transposition of the middle term, consequent upon viewing the three-line syllogism "as one would look at a chess-board" (Ibid.). In short, the figures are the result of a familiar human habit, that of discerning connections between concepts without necessarily securing their justification.

The rather polemical tone of this work hardly seems justified by a controversy that had been present in Aristotelian logic from its inception and was, by Kant's time, already of considerable vintage. The pique that Kant seems to have felt has its source not in logic, but in metaphysics. In a passage which lays bare the metaphysical foundations of his argument, he says "when all is said and done, the fate of the human understanding is such that it is either given to brooding over deep matters and falls into bizarre ideas, or it audaciously chases after objects too great for its grasp and builds castles in the air" (Kant 1992e, 100).

What is particularly interesting about the *False Subtlety* is that Kant sets up his discussion of the syllogism in terms of concepts and judgements. As a number of commentators have observed, this discussion resurfaces, many years later, in the section of the *Critique* known as the Discipline of Pure Reason – a section best summarised as establishing "a set of rules for the use of pure reason that, if followed, will mitigate and perhaps even eliminate our tendency to make judgments about supersensible objects" (Chance 2013, 87) I will return to Kant's Critical discussion of concepts and marks in chapter 4.

The *False Subtlety* contains references to epistemology and human cognition, topics that would develop alongside his metaphysical concerns and which would receive their most complete exposition in the *Critique*, particularly in its first edition. In §6 of *False Subtlety*, *Concluding reflection*, Kant makes several remarks on this subject, starting with a description of distinct and complete concepts. A distinct concept, he says, is formed directly through the act of judging, which itself is the recognition that a given property or attribute belongs to a thing; the example he gives is of impenetrability as a property of a body. The concept itself is to be distinguished from the judgement that actualises it. A complete concept, on the other hand, can only result from a syllogism (or from an everyday pattern of inference that can be expressed in the form of a valid syllogism). In other words, it is the product of two distinct concepts conjoined in a way that reveals new knowledge. It is complete, not in the sense that the concept is exhausted by its expansion, but in the sense that the syllogism itself can express nothing further: only one conclusion can be validly drawn from a given syllogism and its particular premises.

In itself this is a merely logical proposal. But in his second remark, he says " the completeness of a concept and its distinctness do not require different fundamental faculties of the soul . . . understanding and reason, that is to say, the faculty of cognising distinctly and the faculty of syllogistic reasoning, are not different fundamental faculties. Both consist in the capacity to judge; but when one judges mediately, one draws an inference" (Kant 1992e, 103). Now Kant is addressing the manner in which the human subject forms knowledge in explicitly functional terms. Although animals can distinguish between at least some objects according to their dispositions, humans, by contrast, can differentiate logically, by *thinking that* A and B are distinct things on the basis of concepts that determine the objects via properties possessed by each of them. Furthermore, Kant tentatively suggests that the "mysterious power" that makes judgment possible is "nothing other than the faculty of inner sense, that is to say, the faculty of making one's own representations the objects of one's thought" (Kant 1992e, 104). We can judge, therefore, because our representational capacity boasts a reflexivity that is absent in animals.

In one sense, this talk of faculties and cognition is quite distinct from the discussion of formal logic with which Kant starts out. As it is presented, however, the relation is clear: formal logic is a science of human reason. Of this science, the *False Subtlety* has already furnished the reader with some important ideas. Humans possess at least two 'faculties', namely understanding and reason, presaging important elements of his *Critique of Pure Reason*, and the function of both is to form judgments, albeit in different ways which lead to different forms of knowledge. Logic is but the way in which these operations can be expressed free from the ambiguity that attends them in everyday life – and even then with restrictive caveats. These ideas were not particularly innovative in themselves. Theories about the nature of human cognition are to be found in philosophical works from antiquity onwards. Their presence in the *False Subtlety* does, however, date Kant's first public speculations on the subject.

Finally, I come to *Attempt to Introduce the Concept of Negative Magnitudes into Philosophy*, published in 1763. In this essay, Kant accuses metaphysicians of freely employing the notion of logical opposition, or contradiction, in their investigations, but neglecting 'real opposition', where "two predicates of a thing are opposed to each other, but not through the law of contradiction" (Kant 1992b, 211). The subsequent discussion is carried out largely in terms of the conflict of forces, such as in the case of the opposition of two objects moving with the same force but in opposite directions and which are consequently motionless. This is not a contradiction and, in contrast with logical opposition, the state of rest is not a 'nothing' or absence. *Negative Magnitudes* is particularly interesting for its discussion of forces and their potential for understanding mental activity, but here I will draw attention only to the preface, in which he promulgates his pre-critical insistence that metaphysics be brought into alignment with science. In a similar tone to that of his earliest work, *Living Forces*, he declares that "it seems easier to linger among obscure abstractions which are difficult to test, than to enter into relations with a science which only admits intelligible and obvious insights" (Kant 1992b, 208). With great irony, however, he expresses his doubt that his rivals will be swayed by what he has to say: "As for the metaphysical intelligentsia who are in possession of a perfect understanding of things, one would have to be very inexperienced to imagine that their wisdom could be increased by any addition, *or their madness diminished by any subtraction*" (Kant 1992b, 210; emphasis mine).

There are, of course, other important works from this period, notably *The Only Possible Argument in Support of a Demonstration of the Existence of God* (1763) and *Inquiry Concerning the Distinctness of the Principles of Natural Theology and Morality* (the 'Prize Essay') of the same year. In these works, too, the theme is metaphysics and Kant's desire to bring it to the status of a true science, although he is less disparaging of contemporary efforts than in the works I have considered above. Within a few years however, Kant would publish (anonymously, though his identity as author was recognised by his colleagues) the book that represents his most complete, and most strident, development of his analogy between metaphysics and madness, *Dreams of a Spirit Seer* (1766).

The book has an unusual backstory. Some years previously Kant had developed an interest in the life and work of Emmanuel Swedenborg, a Swedish scientist and philosopher turned mystic. In middle age, Swedenborg began to experience dreams and waking visions, and underwent a spiritual transformation. Now convinced of his ability to commune with the spirit world, he soon found fame and the patronage of high society, including royalty (for a good summary, see Stroud 2016). Kant acquired and read Swedenborg's voluminous *Arcana Coelestia*, an account of the theology based upon his revelations. His disillusionment led him to write this polemical work, which includes passages that are openly mocking of the Swedish 'seer' and his claims.

Dreams, while motivated by disappointment and, possibly, some personal resentment (Kant had unsuccessfully attempted to enter into personal correspondence with Swedenborg), is more than an *ad hominem* attack. For all its rather un-Kantian lack of politesse, it addresses matters that Kant took very seriously: not only his dissatisfaction with metaphysics, but also the phenomenon of religious enthusiasm and fanaticism, between which he draws disturbing analogies. He compares the rationalist philosophers Wolff and Christian August Crusius to the spirit-seer Swedenborg, noting "a certain affinity between the *dreamers of reason* and the dreamers of *sense*" (Kant 1992c, 329). Both parties claim a kind of knowledge of the 'spirit world', the rationalist psychologists purporting to prove such properties of the soul as immateriality and immortality, mystics such as Swedenborg claiming direct, sensible encounters with the realm of souls or spirits. For Kant, in both cases ideas that arise entirely within the intellect are erroneously taken for the independently real. For the spirit-seer, the ideas assume sensible shape, while for the metaphysician they take the form of presumed knowledge. The latter's 'chimeras' are, however, recognised as arising from his own rational activity rather being given in sensible experience: Kant does not say the metaphysician is subject to hallucinations, as he considers mystics such as Swedenborg to be.

Kant uses the optical metaphor of what he calls the *focus imaginarius* – a device that reappears in the *Critique* in the Appendix to the Transcendental Dialectic (Kant 1998) (A644/B672). In *Dreams*, the *focus imaginarius* is problematic, since in different ways it leads one to accept as externally real what is only internally generated. In the *Critique* Kant expounds a more developed, nuanced view, whereby it may play a legitimate role in rational thought, enabling us to transcend the limitations of our senses – but only on the condition that the resulting 'visions' be treated as useful guiding fictions rather than knowledge, strictly speaking. This, Kant's doctrine of 'regulative' ideas of reason, will play a key role in my critique of Wakefield's HDA in the following chapter.

The comparison is damning, nonetheless, and is related, once again, to Kant's metaphor, in the *Critique*, of the 'battleground' that metaphysical inquiry has become. Wolff, Crusius, and their ilk are "each inhabiting his own world to the exclusion of the others" (Kant 1992c, 329) This point hearkens back to the comments I have picked out in both *Living Forces* and *Negative Magnitudes*: so convoluted and obscure are the metaphysical systems of the rationalists that it is difficult for anyone to decide for or against them. It is Kant's hope that they will one day "wake up and open their eyes upon a viewpoint which no longer precludes an agreement with other minds . . . The philosopher might then be able to live in a common world, just as the exponents of the quantitative sciences have been able to do for some time past" (Kant 1992c, 329). In this pre-critical phase, Kant's worry is that the kind of intersubjective testing of concepts and theories characteristic of science is unavailable to metaphysics because of its method. The works of this period that I have surveyed are tinkering with different aspects of the method with the goal of bringing metaphysics out of this benighted state. The same basic concern for intersubjectivity runs through the the first *Critique* and later works, albeit in a manner transformed by Kant's reconception of metaphysics. It is key to tempering the tendency to project our ideas onto reality that we be able and prepared to submit them to the judgement of others, or to what Kant would later call the common sense or *sensus communis*.

Thus far I have looked at Kant's comparison between madness and certain forms of philosophising. I will now turn to his specific thoughts on mental disorder itself, which he committed to paper in the pre-critical period and would eventually return to towards the end of his life.

Kant on mental disorder

The 'Essay on the Maladies of the Head' was published in 1764 in the *Königsbergsche Gelehrte und Politische Zeitungen* (*Königsberg Scholarly and Political Newspaper)* following the appearance of a Polish religious mystic, Jan Komarnicki, near Kant's home city. Living an austerely simple, nomadic life, Komarnicki and the young boy accompanying him fascinated the worldly citizens of Königsberg. The 'Essay' was prompted by reflections upon life in a state of nature and the effects of civilisation upon the constitution of the mind that Komarnicki's case threw into stark relief. This was only the starting point, however. The bulk of the 'Essay' is taken up by what Kant called an "onomastic of the frailties of the head" (Kant 2007, 66). Kant uses a term that means 'study of names', but to a modern reader it is closer to a nosology – a classification of diseases. Here he describes and labels a series of psychological disorders (in fact, symptoms rather than discrete, bounded pathologies) and, furthermore, links them to distinct "mental capacities" (Kant 2007, 70) which, appear to be the forebears of the faculties of the *Critique*: sensibility, understanding, and reason. In the act of compiling his onomastic, Kant conducts the most overt and detailed investigation into human consciousness so far in his career, an investigation that would play a central role in the *Critique*.

Kant would again turn to the matter of mental disorder in his *Anthropology from a Pragmatic point of View* in 1798. Although by then approaching the end of his life, *Anthropology* was the product of a lecture course that Kant had taught since 1772 (Kuehn 2001, 204); therefore, it represents a body of thought that he had been developing since well before the publication of the first *Critique*. A taxonomic schema is once again the heart of the discussion of mental disorder, albeit one significantly different to the 'Essay'. What remains consistent with the earlier work are his more general observations regarding mental disorder.

Neither the 'Essay' nor the *Anthropology* are works of philosophy in the strict sense. Rather, they are 'anthropological' in a sense rather particular to Kant. Historically, that term denotes "a discipline at the crossroads between medicine and philosophy, with changing contours and shifting semantics" (Buchenau 2017, 72 n2), quite different from its current form as an academic discipline. In the context of Kant's development, we can best understand it through the work of his contemporary Ernst Platner, who published his *Anthropology for Physicians and the Worldwise* in 1772. This work drew upon the neurophysiological theories of the physician Albrecht von Haller, who did groundbreaking experimental work on the nervous system in animals, and through it "sought physiological explanations of mental processes while . . . retaining a dualist metaphysics." (Wunderlich 2018, 155)

However, in a letter to his friend Marcus Herz in 1773, Kant contrasted his perspective with that of Platner: "I have read your review of Platner's *Anthropologie* . . . I am giving, for the second time, a lecture course on *Anthropologie* . . . but my plan is quite unique . . . I shall seek to discuss phenomena and their laws rather than the foundations of the possibility of human thinking in general. Hence the subtle and, to my view, eternally futile inquiries as to the manner in which bodily organs are connected with thought I omit entirely" (Kant 1999, 141). Kant's anthropology is, therefore, closer to a form of psychology as it is presently conceived, albeit an anecdotal one. It is, clearly, quite far removed from psychology as an experimental science, which only came into being in the mid-19th century (Wertheimer 2012, 57).[13] What is *pragmatic* is that it is a form of psychology in the service not only of understanding the human character as such, but also the ways in which it could alter its own determination. As he explains, man "has a character, which he himself creates, in so far as he is capable of perfecting himself according to ends that he himself adopts. By means of this the human being, as an animal endowed with the *capacity of reason* (*animal rationabile*), can make out of himself a *rational animal* (*animal rationale*) . . . ." (Kant 1978, 226) Under this rubric, Kant is interested in observation of, and anecdote about, the concrete conditions of human thought, and the practical purpose that this body of knowledge might serve.

Although the 'Essay' precedes Kant's commencement of his lecture course in anthropology, and comes long before the publication of the *Anthropology* itself, it deserves to be considered an early experiment with this instrumental approach to psychology. Michel Foucault argued that "the text published in 1798 [i.e. the *Anthropology*] fits in easily with a number of different writings from the precritical

[13] In a well-known passage from *Metaphysical Foundations of Natural Science* (1786) Kant asserts that "the empirical doctrine of the soul can never become anything more than an historical doctrine of nature . . . but never a science of the soul, nor even, indeed, an experimental psychological doctrine." (Kant 2002a, 186) If we take him at his word, it is clear that he did not anticipate the modern science of psychology, although Patrick Frierson has argued that the true picture is more complex, and that he may not have been categorically denying its possibility (Frierson 2014).

period" (Foucault 2008, 28), noting, above all, the 'Essay' and *Observations on the Feeling of the Beautiful and the Sublime* from the same year, 1764. Even if it predates his anthropology lectures, the 'Essay' reflects their spirit.

Even in outline, Kant's views regarding the nature of madness are both ambiguous and multilayered. Alan Stone captures some of this complexity very well when he writes "The intellectual tradition which Kant exemplifies attempts to isolate the abnormal, madness, at an extreme of the human spectrum . . . [but] in his analysis of the situation of the rest of humanity, Kant is somehow haunted by the analogy to madness . . . ." (Stone 2011, 190). Furthermore, Kant's struggle with the problem portends a more recent dilemma: "The problem is and always has been for psychiatry whether it provides only a theory of madness or a more general theory of human nature as well. Or is it even possible in principle to make such a distinction? Can one explain madness without explaining human nature?" (Stone 2011, 192)

The biopsychosocial model exemplified by the *Diagnostic and Statistical Manual of Mental Disorders* is, in part, an attempt to sidestep this problem by separating the question of diagnosis from matters of theory. Biological psychiatry, by contrast, does its best to exclude questions of 'human nature' from the purview of psychiatry altogether. On the other hand, some of those thinkers grouped under the heading of anti-psychiatry *have* seriously considered the difficulty raised by Stone. I have noted in the previous chapter the problems associated with the anti-psychiatry label, and I do not mean to suggest that Kant's thought on the subject is a precursor to the works of any of those so labelled (although at least one commentator has argued this, see Double 2020, 234). Nonetheless, I will argue that Kant's analysis of mental disorder forms part of an (unfortunately fragmented) attempt to resolve Stone's paradox. Indeed, his interest in the subject is bound up with questions that are foundational to philosophy as such. Few, if any, of the core problems of philosophy are not touched upon by the phenomenon of mental disorder, yet it has been a neglected topic and, in the words of Andrew Quinton, philosophers "ought to have concerned themselves with madness just to the extent that they have taken themselves to be the custodians of the cognitive, of rational belief and valid reasoning" (Quinton 1984, 17).

The previous two chapters have attempted to map out relatively distinct philosophical perspectives within psychiatry. Here I will begin, through several different aspects of

Kant's work, to address the wider question of how mental disorder intersects with the notion of rationality upon which rests the very enterprise of philosophy itself.

The 'Essay on the Maladies of the Head'

As noted above, the 'Essay on the Maladies of the Head' is written in something like the ironic tone that commentators have often noted in Kant's *Dreams of a Spirit-Seer*, published two years later in 1766. It is his intention, he explains, to "imitate the method of the physicians, who believe they have been very helpful to their patient when they give his malady a name, and . . . [to] sketch a small onomastic of the frailties of the head . . . ." (Kant 2007, 206). Clearly, Kant is sceptical about the method he is imitating; the reader might wonder whether his own onomastic, or study of names, is meant to be taken in the same spirit. The passage is illustrative of Krysta Thomason's observation that "his account of mental illness contains classifications that are often fluid and ambiguous. Even though he seems to offer taxonomies, it is unclear how serious he is about them." (Thomason 2021, 189)

The ambiguity is apparent in the first division Kant makes, moving from "frailties of the head which are despised and scoffed at . . . [or] which do not suspend civil community to those in which official care provision takes an interest and for whom it makes arrangements" (Kant 2007, 209). The former 'frailties' are just the varieties of foolishness or eccentricities of temperament to which everyone is prone, and that aren't seriously considered pathological. Kant passes swiftly over the category of 'impotence', or what we now call learning disorders, regarding them as uninteresting. We are left with the category of 'reversal' as *the* class of mental disorders.

Two related features emerge from Kant's onomastic. First, although he treats ordinary foibles separately, Kant clearly sees a continuity rather than a decisive break between the sane and the mad: "in order to recognise these loathsome maladies in their gradual origination, I find it first necessary to elucidate their milder degrees from idiocy to foolishness, because these properties are more widespread in civil relations and lead nonetheless to the former ones" (Kant 2007, 66). Secondly, the reason Kant moves quickly over disorders of impotence is because he sees them as disorders of mental impoverishment: there is simply nothing much to be said about a fundamental deficiency or lack. Insanity, on the other hand, he sees as productive malady, as when he refers to "the frailties of the head, from its *paralysis* in imbecility to its raptures[14] in madness . . . ." (Kant 2007, 66).

Consequently, mental disorder is distinctive for its positive attributes, to which extent Kant again seems to be implying a continuity with sanity, against the dysfunction analysis of mental disorder which sees it as an objectively definable break with normal thought and behaviour. As Monique David-Ménard puts it, for Kant, "madness is an organization of thought. It is made possible by the ambiguity (and hence the possible subversion) of the normal relation between the imaginary and the perceived, whether this pertains to the order of sensation or to the relations between our ideas" (David-Ménard 2000, 86). What is important here is that David-Ménard alights upon the notion of '(re)configuration' implicit in the 'Essay', or what amounts to the implication that what is present in madness is not a deficit or defect, but an altered arrangement or interaction of cognitive mechanisms that are, *in themselves*, not damaged or failing. The 'raptures' Kant mentions are surely complex phenomena that, unlike the 'paralysis' of imbecility, require explanation in terms other than those of mere defect. We may note in passing that the word *derangement* has its roots in mathematics: in combinatorics, the derangement of a set is its reconfiguration such that none of its members occupy the same position as formerly. It is also related, in this combinatoric sense, to the word *disorder*.

It is significant that Kant seems to consider the complexity of developed human society a contributor to the maladies he describes, writing: "Had the brain of the savage sustained some shock, I do not know where the fantastic mania should come from to displace the ordinary sensations that alone occupy him incessantly. Which dementia can well befall him since he never has cause to venture far in his judgment? Insanity, however, is surely wholly and entirely beyond his capacity. If he is ill in the head, he will be either idiotic or mad, and this, too, should happen most rarely . . . ." (Kant 2007, 75).

This passage may have been prompted by Kant's reading of Rousseau (mentioned elsewhere in the *Essay*, albeit in a different context), and the demeanour of the child who

[14] Kant discusses both 'rhapsodic' and 'tumultuous' thinking (both used in his descriptions of madness) in his logic lectures, e.g. in the Vienna logic (Kant 1992d, 287): "A cognition can be rhapsodic, and nevertheless not be tumultuous. For what is opposed to the tumultuous is the methodical, and without method a cognition is tumultuous. But a cognition that is produced methodically, but without system, is rhapsodic. E.g., when we guide ourselves in accordance with the power of comprehension of the subject who is to be instructed."

accompanied Komarnicki, the ‘prophet’ who had intrigued Königsberg society. Kant’s invocation of the ‘savage’ does not seem to illustrate a contrast between the isolated individual and membership of a collective, but one between simple and refined forms of society and culture. In an interesting analysis, Tanehisa Otabe examines the role of the ‘savage’ in Kant’s philosophy, arguing that by the time of the *Critique of Judgement* (1790) it is clear that “Kant agrees with Rousseau that the sciences and the arts, including taste, have aroused various evils that nourish infelicitous inclinations and bring forth decadence or vanity” (Otabe 2018, 49). It is this social artifice to which Kant alludes when he writes, in the ‘Essay’, “The means of leavening for all of these corruptions can properly be found in the civil constitution, which, even if it does not produce them, nevertheless serves to entertain and aggravate them” (Kant 2007, 75). On the one hand, human society and culture may not actually ‘produce’ mental disorder; on the other, without the psychological stresses characteristic of a highly developed society, Kant doubts that it could arise *at all*. As Stone observes in the quote I gave earlier, Kant is not quite able to isolate madness as a qualitative entity – to determine a boundary between it and sanity. This is not to say, however, that Kant is offering a naively idealistic image of a primitive idyll. In the *Critique of Judgement*, he develops this theme, arguing that civilised man cannot simply return to a prelapsarian past, and that regardless of their shortcomings, the refinements of civilised society “have the effect of strengthening the powers of our soul toward morality.” (Otabe 2018, 49) As Kant says in the *Anthropology*, “On the whole, the more civilised human beings are, the more they are actors . . . [but] eventually the virtues, whose illusion they have merely affected for a considerable length of time, will gradually really be aroused . . . .” (Kant 1978, 42).

Kant does speculate about the physical origins of mental disorder, but abstains from drawing firm conclusions, noting simply that “I have also only paid attention to their appearances in the mind without wanting to scout out their roots, which may well lie in the body and indeed may have their main seat more in the intestines than in the brain . . . .” (Kant 2007, 76). The association of the gut with mental states was well established in medicine; the reference Kant makes to the medical journal *Die Artzt* occurs in this passage. More importantly, this comment anticipates the sentiments of Kant’s letter to Herz regarding anthropology and the futility of marrying the metaphysics of the soul with physiology.

Kant's *Anthropology from a Pragmatic Point of View*

As noted above, Kant commenced his anthropology lectures in 1772, some eight years after publication of the 'Essay', and delivered the course, with constant modifications, up until his retirement from teaching in 1796. At this time, and in response to encouragement from friends and colleagues, he wrote up his distinctive 'science of man' into a book for public consumption. *Anthropology From a Pragmatic Point of View* was published in 1798 and, among a wide range of observations regarding human life and behaviour, included what was, in effect, a reworking of the 'Essay' written over 30 years earlier. Indeed, although there are significant differences between the two, it is remarkable how many continuities are also present.[15]

One striking difference is to be encountered at the very beginning of *Anthropology*, in the form of a statement that presents his vision of this philosophical-psychological project: "A doctrine of knowledge of the human being . . . (anthropology), can exist either in a physiological or in a pragmatic point of view. — Physiological knowledge of the human being concerns the investigation of what *nature* makes of the human being; pragmatic, the investigation of what *he* as a free-acting being makes of himself, or can and should make of himself" (Kant 1978, 3). Nonetheless, these themes are implied in the 'Essay' even though Kant was, in the earlier work, a long way from their systematic development. In both works, Kant is very little concerned with the etiology of mental disorders, and more interested in the role they play in human thought *qua* the ends to which thought is applied. By the time of the *Anthropology*, however, Kant had developed this theme through his annual lecture course. For him, the ends of humankind are bound up in the freedom to mould and, potentially, to perfect oneself morally. If there is a 'meaning of life' to be found in Kant, it is this; correspondingly, what we should disvalue in madness is that it impedes the capacity for rational self-determination. If this is compromised, then our very ability to form meaning is diminished. To this extent, Kant's view of madness is explicitly normative. For example: "life as such, which depends on fortunate circumstances, has no intrinsic value of its own at all, and that life has value only as regards the use to which it is put, and the ends to which it is directed" (Kant 1978, 135). By 'fortunate', Kant means that the existence of life is a matter of

[15] In the *Anthropology* he even recycles a number of phrases and observations from the *Essay*, notably 'castles in the air', 'the head is a drum that sounds because it is empty', and the lovers/church steeples example of illusion.

happenstance. Since we cannot look outside ourselves for a source of value, we must create value through choosing ends and therefore endowing life with purpose.

Why did Kant believe that his pragmatic anthropology was the place in which to discuss mental disorder? What specifically *practical* interest could it have for the reader? To a modern reader, the structure as well as the content of the book may seem rather arbitrary in parts. As he points out, however, "the entire use for the cognitive faculty for its own advancement, even in theoretical cognition, surely requires reason" (Kant 1978, 123). In order to advance, we need some empirical knowledge of human thought and behaviour, not only in its idealised specification but also in the manifold ways that it is encountered in the world, including its graver departures from the archetype. This serves the ends of acculturating us to the society of fellow human beings in all their diversity as well as prompting moral reflection upon our behaviour towards them.

It is to this end that Kant's discussion of mental disorder (among other things) is included. We can understand this a little better, perhaps, in the context of Kant's 'cosmopolitanism'. Georg Cavallar identifies a number of 'cosmopolitanisms' in Kant, including epistemological, economic, moral, political, and cultural, arguing that these strands form part of a systematic whole. Of particular relevance here is Kant's "claim that all rational beings . . . should be regarded as ends in themselves and as lawgiving members of 'the universal kingdom of ends'" (Cavallar 2012, 98). Not only is Kant concerned with madness in relation to the philosopher's interest in cognition, but also in relation to an ambitious moral vision.

Kant begins his discussion of mental disorder by making the same distinction between learning impairments and mental disorders proper that he makes in the 'Essay'. In that work, he refers to them as disorders of impotence and of reversal, though those terms are dropped here. He then opines that "Illnesses of the soul with respect to the cognitive faculty can be brought under two main types. One is *melancholia* (hypochondria) and the other is *mental derangement* (mania). With the *former*, the patient is well aware that something is not going right with the course of his thoughts . . . Mental derangement indicates an arbitrary course in the patient's thoughts which has its own (subjective) rule, but which runs contrary to the (objective) rule that is in agreement with laws of experience" (Kant 1978, 96).

This division marks a departure from the taxonomy of the pre-Critical 'Essay', in which Kant does not distinguish between the symptoms of acute mental disorder and milder conditions: all are distributed under the tripartite schema he employs there. This feature of the *Anthropology* foreshadows the distinction between the neuroses and psychoses that emerged during the 20th century. The etymology and relationship of these terms is complex, and emerged from attempts to distinguish between neurological diseases and psychological complaints that appeared to lack somatic causes (Munsche and Whitaker 2012, 224). It is striking, however, that for Kant the distinction (made using alternative terminology) is primarily cast in psychological terms. He does assert that mania is hereditary, situating himself within prevailing attempts at nosology (Kant was known to be familiar, for example, with the work of William Cullen, who coined the term 'neurosis'), but passes over this aspect in a short passage. This is, of course, consistent with his comments regarding 'physiology' and his own distinctive brand of anthropology. The division Kant makes is more in the spirit of Freud (or Jaspers), who used the terms neurosis and psychosis to differentiate between conditions in which a person maintains a connection with reality and those in which the connection breaks down. In a looser sense, it distinguishes between madness and ostensibly less-severe mental disorders.

Mental derangement is here subdivided into amentia, dementia, insania, and vesania; all terms that (in contrast to those employed in the 'Essay') were in widespread use by 18th century physicians (Munsche and Whitaker 2012, 225). These correspond, to some extent, to the cognitive schema developed in Kant's three *Critiques*: imagination, understanding, judgement, and reason. No such division is attempted in the case of melancholia.

Kant describes amentia as 'tumultuous', "the inability to bring one's representations into even the coherence necessary for the possibility of experience" (Kant 1978, 109) The phrasing is unfortunate, suggesting as it does a collapse of cognitive function that would surely reduce the sufferer to a state of utter helplessness. As it happens, Kant goes on to suggest that "talkative women" are its usual victims, because of their "lively power of imagination." (Ibid.) Clearly, amentia is not as catastrophic as might initially be supposed. In fact, the term amentia had been in widespread use for centuries, and was frequently associated with intellectual impairment rather than mental disorder *per se* (Buhrer 2014, 328f). Patrick Frierson, who has studied Kant's taxonomy closely, notes that "it seems a mere deficiency, an *in*ability to order one's representations" (Frierson 2014, 201) but does not pursue the etymological issue, and concludes that "amentia is best described as a disordered understanding that incorporates imaginary distractions into one's stream of thought to such a degree that one can no longer form coherent objective judgments about the world . . . ." (Frierson 2014, 202).

Amentia, in Kant's view, seems to involve some confusion between understanding and imagination wherein the activity of the latter intrudes upon the sufferer's ability to manipulate concepts appropriately: "it is women who, owing to their talkativeness, are most subject to this disease: that is, their lively power of imagination inserts so much into what they are relating that no one grasps what they actually wanted to say" (Kant 1978, 109). Chauvinism notwithstanding, the description – particularly the emphasis on speech – is suggestive of one of the features associated with the schizophrenia spectrum in the DSM-5: "*Disorganized thinking* (*formal thought disorder*) is typically inferred from the individual's speech. The individual may switch from one topic to another (*derailment or loose associations*). Answers to questions may be obliquely related or completely unrelated (*tangentiality*)" (*Diagnostic and Statistical Manual of Mental Disorders: DSM-5* 2013, 88). Given that this is a *formal* thought disorder, affecting the way cognitions are organised, it is easy to see why amentia was historically associated with intellectual deficit. While the contents of psychotic thought and speech are often bizarre, they also reflect a high degree of internal coherence. Disorganised thinking, by contrast, is disruptive of the process of cognition itself.

The second of Kant's categories is *dementia*, "that disturbance of the mind in which everything that the insane person relates is to be sure in conformity with the formal laws of thought . . . but, owing to the falsely inventive power of imagination, self-made representations are regarded as perceptions" (Kant 1978, 109) It is clear that Kant has delusion and hallucination in mind here. Again, a confusion of mental functions is implied, this time between sensibility – our passive capacity for being affected through outer sense – and imagination, such that products of our imagination are incorrectly taken to have the immediacy and givenness of empirical intuition. That this is the case for delusion is not as obvious as it is for hallucination, but Kant impresses the point upon the reader by giving the example of "Those who believe that they are surrounded by enemies everywhere" (Ibid.) In delusion, imaginary supposition imposes itself upon sensibility by giving perceptions a particular kind of significance rather than presenting *as* perceptions, as in hallucination. Dementia is *methodical*, according to Kant – the content of thought, rather than its organisation, is affected.

Also methodical is the third category, *insania*, "a deranged power of judgement in which the mind is held in suspense by means of analogies that are confused with concepts of similar things, and thus the power of imagination, in a play resembling understanding, conjures up the connection of disparate things as universal . . . . (Ibid.). Analogies – the drawing of limited comparisons between properties of things that are in most other respects dissimilar, often for illustrative purposes – are here confused with the function of understanding, so that judgements of partial similarity between particulars are taken to be constitutive of concepts themselves, yielding heterogeneous categories. Thomas Szasz gives the following example (although the cognitive mechanism he attributes it to is different to Kant's): "Since both stags and Indians move swiftly, he [the schizophrenic] equates the two and says that stags are Indians . . . ." (Szasz 1974, 32). A conceptual system whose categories group such diverse entities will clearly be out of step with common understanding. Although still methodical, Kant notes that insania is *fragmentary*, presumably in the sense that the unity generally embodied in concepts is broken up through the inclusion of incongruous elements.

Lastly, Kant describes *vesania*, "the sickness of a deranged reason. - The mental patient flies over the entire guidance of experience and chases after principles that can be completely exempted from its touchstone, imagining that he conceives the inconceivable" (Kant 1978, 110). There are clear echoes here of the sort of metaphysical excess that Kant railed against in some of his pre-Critical works. The patient believes that they comprehend purported *a priori* truths not susceptible of any possible empirical experience or proof: the squaring of the circle, perpetual motion, the mystery of the Trinity.[16] That ambiguity, I argued above, motivated the *Critique of Pure Reason.* For the Critical Kant, reason is the faculty that brings unity to our experiences, and provides us with the regulative ideals that drive scientific investigation, but it is, by its very nature, vulnerable to the kind of overreach described here.

[16] The *a priori* nature of this phenomenon should be emphasised: the patient apparently *grasps* the solution to these mysteries, but is ambivalent about the demand for proof or demonstration. It is of a piece with Kant's account of 'unreason' that the madman occupies a different standpoint wherein the very concept of an intersubjective world has been supplanted. It has consequently been noted in the psychiatric literature that schizotypal beliefs of this kind exist alongside a peculiarly ambivalent attitude to the notions of truth and verification.

As in the 'Essay', Kant pays little attention to the etiology of mental disorder. The sole comment he makes is "The germ of madness develops together with the germ of reproduction, so that this too is hereditary" (Kant 2007, 111). One of the few writers to have looked at Kant's concept of madness as such is Dominic Sisti, who concludes that "Kant's concepts of mental health and illness are those of someone who today would be considered an unabashed naturalist" (Sisti 2012, 5). In Sisti's view, although Kant is a naturalist, his concept of mental disorder is not purely scientific because it involves a teleology of persons. Indeed, it is a "unique kind of naturalism" (Ibid.).

Sisti is right to draw attention to the role Kant accords to moral reason and its principles, but his conclusion that the account is nonnormative seems insupportable. The 'teleological judgement' that Kant discusses in the *Critique of Judgement* is what Kant calls a 'regulative' maxim. Kant introduced the notion of regulative and constitutive principles in the *Critique of Pure Reason*. In Stanley French's words, "A constitutive proposition describes the sensible world. A regulative proposition does not. A regulative proposition *prescribes*. It postulates what we ought to do, or how we ought to think." (French 1967, 624) They are, he says, *heuristics* that guide us in our thought and behaviour, but do not tell us anything about the external world. Clearly, they are value-laden.

Sisti's assessment of Kant as a naturalist ('unabashed' or otherwise), meanwhile, is similarly insensitive to the ambiguity of the Kantian analysis of madness. This is because he rejects 'physiology'. Admittedly, his rejection pertains to explaining the link between body and a metaphysical conception of soul, rather than the body as the *source* of consciousness. At any rate, Kant nowhere comes firmly down on the side of naturalism, and Sisti doesn't even try very hard to show that he does. Robert Butts, in one of the earliest scholarly responses to Kant's thoughts on this subject, reaches a virtually opposite conclusion to Sisti (Butts 1986, 305).

I have argued that Kant's account of madness represents it as fundamentally ambiguous, and also that his reflections on the subject are motivated by what may variously be referred to as 'pragmatic' or 'cosmopolitan' objectives. He does not set about defining the concept of mental disorder, and his discussion is predominantly descriptive rather than analytic. I have noted, however, that the 'Essay' prefigures important elements of the *Critique of Pure Reason*, and that the *Anthropology* is underpinned by the Critical philosophy. It is possible, therefore, to discern philosophically relevant points that, while not explicitly articulated by Kant, provide a more solid foundation for the implication of ambiguity that runs through his work on mental disorder. Of primary importance in this regard is the act of judgement, and a paradox that generated, in the late 20th century, considerable controversy and debate.

A number of commentators (e.g., Frierson 2009, 291, Scholten 2016, 212) have drawn attention to Kant's invocation of 'positive unreason' in relation to the type of mania he calls *vesania*. Here Kant writes "For in this last kind of mental derangement there is not merely disorder and deviation from the rule of the use of reason, but also *positive unreason*; that is, *another* rule, a totally different standpoint into which the soul is transferred . . . ." (Kant 1978, 110). Of this passage, Motohide Saji has observed that "Kant suggests that there is a normative, clear, and definite division between reason and unreason. But Kant also argues that we cannot draw such a division" (Saji 2009, 201). The contradiction is not obvious from this fragment of the text – rather, it must be teased out from comments Kant makes elsewhere in the *Anthropology*, as well as in the first and third *Critiques*. This Saji proceeds to do by appealing to what Kant has to say about reason, rules, and rule-following.

Late in the first *Critique*, Kant says "reason consists just in the fact that we can give an account of all our concepts, opinions and assertions" (A614/B642). In isolation the statement is misleading: Kant does not mean merely that we *can* give such an account of ourselves. Reason is not synonymous with justification. We can only give an account because of the overarching *function* of reason: to bring unity to our diverse and disparate experiences. If and when called to account, we can say why we act in certain ways because we have established relationships between our concepts to form a coherent matrix from which our actions emerge, and within which they have meaning.

This is important for our understanding of Kant's invocation of 'unreason' in the *Anthropology*. In his transcendental psychology, Kant speaks often about 'rules'. In this context, rules are often synonymous with concepts: concepts are rule-like in so far as they provide the criteria for subsuming particulars under general headings. I will return to the significance of rules in short order; it is important only to note that what he means here are not concepts but the '*rule of the use* of reason' – The function of reason is the particular kind of synthesis as described in the first *Critique*. This function, like any function, requires application according to some sort of criterion: a rule. What makes vesania distinctive, Kant says, is that it has its own rule of synthesis. To be sure, the experience of the individual is still unified – this is reason's inalienable function. What does it mean, therefore, for there to be another rule: "positive unreason"? Kant's answer to this question involves contrasting the community with the individual: "from the *Sensorio communi*[17] that is required for the unity of *life* (of the animal), [the soul] finds itself transferred to a faraway place . . . ." (Kant 1978, 110). That is, the rule-of-the-use of reason – the rule according to which logical synthesis is carried out – governs not only the unity of experience of the cognising individual, but also the 'unity of life', or shared universe of meaning, within which individual experience takes its place. This is the mechanism that Kant sees at work in *vesania*: the patient's power of reason certainly yields a systematic account of their actions, but not as part of a common discursive world. As he notes a little later, "it is a subjectively necessary touchstone of the correctness of our judgements generally . . . that we also restrain our understanding by the *understanding of others*, instead of *isolating* ourselves with our own understanding and judging *publicly* with our private representations, so to speak" (Kant 1978, 113).

It is important to distinguish between this and the normativity of prevailing social and cultural values. As Onora O'Neil explains, "Kant does not ground reason in actual consensus, or in the agreement and standards of any historical community; he grounds it in the repudiation of principles that preclude the possibility of open-ended interaction and communication" (O'Neill 1990, 194). In other words, Kant is not proposing a form of social constructivism based upon contingent spatiotemporal conditions. Kant's *Sensus communi*, or in Onora O'Neill's words "the possibility of open-ended interaction", is a basic requirement for agreement or disagreement to arise in the first place. It is, as Wittgenstein put it, "not agreement in opinions but in form of life" (Wittgenstein 1978, PI 241) One afflicted by vesania, however, exists in a (at least partially) closed life-world, one that is peculiarly resistant to the test of intersubjective judgement and correction.

At this point, it would still appear that Kant has successfully drawn a boundary between the derangement of reason (if not other forms of madness) and sanity. But, as Saji argues, Kant also undermines this possibility, and he does it by drawing our attention to a paradox concerning rules. What has come to be known as the rule-following paradox gained prominence through its appearance in Wittgenstein's *Philosophical Investigations*

[17] 'Common sense'; Kant also uses the German *Gemeinsinnes*, which translates the same.

(notably PI 201) and, in particular, via Saul Kripke's analysis of it in *Wittgenstein On Rules and Private Language* (Kripke 1982). In these 20th century works, the emphasis is ostensibly on language, whereas for Kant, the issue has to do with certain kinds of cognitive process. As Paul Boghossian points out, however, "It is hard to see how a convincing meaning scepticism could be confined purely to the linguistic domain, given the intimate relation between thought and language" (Boghossian 1989, 509).

The paradox can be stated thus: at first sight, it would seem that a rule, by virtue of *being* a rule, unambiguously contains the conditions for its application. Upon closer examination we see that to apply a rule is to act a certain way, and that we want to establish under what *circumstances* we act in that way. Without a reason for acting, we could only follow rules arbitrarily. A separate rule might therefore be given to stipulate the appropriate use of the first. But that rule would, in turn, be susceptible to the same problem, and therefore stand in need of yet another rule to guarantee its own correct application—and so on, *ad infinitum*. As Kant puts it in the *Anthropology*, "if there were to be doctrines for the power of judgement, then there would have to be general rules according to which one could decide whether something was an instance of the rule or not, which would generate a further inquiry on into infinity" (Kant 1978, 93). Certainly we do (usually) have reasons for the things we do, but the paradox indicates that in attempting to represent them in terms of rule-following we are fated either to confront an infinite regress or to drawing the perhaps unsatisfying conclusion that our acts cannot be given self-sufficient explanation. In this way, Kant's other manner of characterising reason, as the capacity to "give an account of ourselves", is seen in terms of intelligibility rather than agreement in matters of fact or value.

Kant offers us two rules-of-the-use of reason, one intersubjective, the other merely subjective, and seems to indicate that the latter is the marker of madness. But the indefineability of the criterion of rule-use makes it impossible to represent in words what it is that justifies any particular judgement. To judge that a person is following one rule or another is itself an act that is, as Kant puts it, 'spontaneous'. We can (we hope) offer some evidence – facts – in favour of our judgement. But, as Derek Parfit points out, "Facts give us reasons, we might say, when they count in favour of our having some attitude, or our acting in some way. But 'counts in favour of' means roughly 'gives a reason for'" (Parfit 2011, 31) Facts give us reasons to judge a certain way, but the judgement itself is underdetermined by the facts. Whichever way we look at it, an act of judgement cannot provide its own conditions of use.

As we saw in the passage from the *Anthropology* quoted above, Kant is referring to rules for the 'power of judgement'. In the metaphysical deduction of the *Critique of Pure Reason*, he says that "we can reduce all acts of the understanding to judgements" (A69/B94). So, although Kant's discussion of vesania is cast in terms of reason and unreason, what underlies it is judgement. Indeed, given that Kant conceives judgement as the overarching task of the higher cognitive faculties, not only vesania but potentially many other mental disorders can be understood in its terms. And of this task Kant says that

> the power of judgement is the faculty of subsuming under rules, i.e., of determining whether something stands under a given rule or not . . . if it wanted to show generally how one ought to subsume under these rules, i.e., distinguish whether something stands under them or not, this could not happen except once again through a rule. But just because this is a rule, it would demand another instruction for the power of judgement, and so it becomes clear that although the understanding is certainly capable of being instructed and equipped through rules, the power of judgement is a special talent that cannot be taught but only practiced. (A132/B171f)

Our reasons for judging can never fully determine the judgement we finally make—we must always reach a point at which we stop following the chain of reasons and simply act. Justification therefore always falls short of the regulative ideal that reason itself imposes in the form of the search for an unconditioned condition. This is not, however, to say that our acts are never justified *tout court*. Indeed, as Saji puts it, Kant "suggests that practice precedes rule. That is, it is the bare fact that there is somehow agreement in our practice that makes it possible for us to grasp, after the fact, our practice as a rule-following practice" (Saji 2009, 208).

Indeed, we could go further and argue that if agreement in human life did not have this advantage (no matter how slim the margin), we could not have progressed as we have. On the other hand, progress could hardly be made if our judgements were not open to dispute. In the final analysis, U.S. Justice Potter Stewart's famous concurrence, "I know it when I see it", inadvertently captures the way in which our judgements are ultimately made.[18]

Conclusion

In this chapter I charted Kant's conflicted relationship with metaphysics in his early writing, prior to his 'silent decade' and the appearance of his famous *Critique of Pure Reason*. I showed that in this period he expressed concerns that metaphysical speculation could often seem like the delusions of the insane. I went on to argue that this concern was part of what motivated his project in the *Critique*, that of limiting the pretensions of metaphysics by showing that the proper use of reason was constrained by empirical conditions.

I then discussed Kant's specific work on the subject of mental disorder, from the pre-Critical 'Essay on the Maladies of the Head' to the late *Anthropology from a Pragmatic Point of View*. In these works Kant does not attempt to define mental disorder, but adopts a descriptive approach from which a number of interesting observations can be drawn. In particular, I drew attention to his discussion of judgement and rule-following, and the consequent difficulty of drawing a conceptual boundary between disorder and nondisorder.

[18] Jacobellis v. Ohio, 378 U.S. 184 (1964). The case was an appeal against an obscenity conviction in respect of the film *Les Amants*. Against the original plaintiff's contention that it constituted 'hard-core pornography', Stewart's full response was "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that."

Chapter 4

Mental Dysfunction and Kant's Critique of Pure Reason

Introduction

Kant opens the *Critique of Pure Reason* with a striking statement, and one that sets the tone for everything that follows in this landmark work. He opines that "Human reason has the peculiar fate in one species of its cognitions that it is burdened with questions which it cannot dismiss, since they are given to it as problems by the nature of reason itself, but which it also cannot answer, since they transcend every capacity of human reason" (Avii).

The questions that reason sets us are those of metaphysics, questions about the ultimate constitution of reality that lies entirely beyond sensible experience. Given this, if we are to know anything about the objects or entities that this branch of philosophy posits, it must be through purely mental activity – through the inferences characteristic of our power of reason. By the time of the *Critique*, Kant had lost his youthful faith in this kind of activity – a faith that in any case, as we have seen previously, was tempered by concerns having to do with the status of metaphysics in relation to science on the one hand, and a discomfiting analogy with madness on the other.

The *Critique* addresses a number of problems in philosophy, but its overarching task is bound up with the 'peculiar fate' of reason and "the two goals of establishing that we do have *a priori* knowledge of the most general laws of nature coming from the structure of our own minds and of limiting the validity of such knowledge to the realm of objects that we can actually experience" (Guyer 2010, 5). The settlement that he had finally reached on the subject of metaphysics, therefore, is to radically curtail its ambitions. The big questions about the nature of the soul, the universe, and God, were revealed by Kant to be epistemologically pointless excursions into the realm of speculative fantasy. Peter Strawson captures with considerable clarity what is perhaps Kant's most important insight: "If we wish to use a concept in a certain way, but are unable to specify the kind of experience-situation to which the concept, used in that way, would apply, then we are not really envisaging any legitimate use of that concept at all" (Strawson 2006, 16). This is not to say that mystics such as Swedenborg could not claim to have had experiences of the soul-realm, but his encounters were essentially private and thus not susceptible of ‘specification’, as Strawson puts it. For Kant, concepts can only be legitimately applied to a shared world of experience and the particular situations within it.

In this chapter I will consider Wakefield’s harmful dysfunction analysis in light of Kant’s critique of metaphysics and his assessment of concepts. I will ask, specifically, what kind of concept the HDA delivers, and whether it is one that we can indeed apply to a possible experience. In order to do this, however, it will first be necessary to examine the HDA in further detail. Although apparently very simple, I will show that it is in truth far more complex than is acknowledged, and that once the separate strands that make it up are more clearly identified, a Kantian perspective undermines the objectivity that Wakefield believes it has.

The harmful dysfunction analysis revisited

In one of the two 1992 papers introducing his ‘harmful dysfunction analysis’ of mental disorder, Jerome Wakefield acknowledged the debt he owed to the architect of the path-breaking third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), Robert Spitzer. Spitzer, against much opposition, had insisted on formulating a general definition of mental disorder to preface the manual. Wakefield observed that “Spitzer arrived at the definition through the method of conceptual analysis, which is also used here. In a conceptual analysis, proposed accounts of a concept are tested against relatively uncontroversial and widely shared judgements about what does and does not fall under the concept” (Wakefield 1992a, 233).

Conceptual analysis is often held to be one of philosophy’s defining activities – a paradigmatic example of armchair reasoning. The ‘proposed account’ of a concept draws upon the philosopher’s own intuitions (suitably guided, perhaps, by reflection on common practices); the process may involve refining the definition as and when imagined counterexamples are found until, finally, a definition is reached which appears to capture all our accepted uses of the concept.

Not only the method, but also the presumptive definition Wakefield begins with is taken from the DSM: “There are two fundamental principles that guide DSM-III-R’s definition of mental disorder. The first is that a disorder is a condition that has negative consequences for the person. The second is that a disorder is a dysfunction . . . .” (Wakefield 1992a, 233).

As Wakefield observes, however, the DSM definition does not employ these principles directly. As we saw in chapter 1, Spitzer and his colleagues were wary of trying to define the subordinate concepts 'harm' and 'dysfunction', and chose instead to operationalise them. Wakefield subsequently identified problems of both under- and over-inclusiveness with this approach, and opted for an analysis in traditional terms.

The HDA rather straightforwardly interprets these two principles in terms of a classical analysis, so that they enter the definition as jointly necessary and sufficient conditions (Wakefield 1999a, 377). Rather than try to render the 'harm' element objective, as Spitzer's operationalism attempted, Wakefield openly acknowledges that it will be evaluative and therefore subject to considerable variation, particularly across different cultures. He argues that 'dysfunction', by contrast, is a value-free scientific concept, and, as we have seen, grounds this in Larry Wright's analysis of 'function', whereby

> The function of *X* is *Z means*
>
> (a) *X* is there because it does *Z*,
>
> (b) *Z* is a consequence (or result) of *X*'s being there (Wright 1973, 161).

This 'etiological' analysis takes the effect *Z* to be the historical *cause* of the current existence of *X*, where *X* is a feature or characteristic. Unlike other function-analyses such as that of Robert Cummins, this causal consequence provides us with a notion of what is functionally *normal*. In other words, a function is normative in itself, just by virtue of its role in explaining its own presence, and is therefore (so the argument goes) entirely independent of subjective value judgements. Wright maintains that his analysis applies equally to conscious and natural functions; in biology, natural selection is taken to be the actual process by which this analysis of function is realised (Wright 1973, 162–64).

In common with other scholars who have pursued the project of defining mental disorder, Wakefield wants the HDA to fulfil a number of different tasks, but the most important element of his analysis is this value-free element. Above all, it is this that answers the antipsychiatric worry that judgements of mental disorder are imbued with subjective values, and therefore open to abuse. In its most extreme form, this becomes the allegation that there are no mental disorders at all, if we mean 'mental disorder' to be analogous with somatic diseases identifiable by reference to physical lesions. This is the position held by Thomas Szasz (see chapter 2). As Wakefield writes, "The requirement that a disorder must involve a dysfunction places severe constraints on which negative conditions can be considered disorders and thus protects against arbitrary labelling of socially disvalued conditions as disorders" (Wakefield 1992b, 386). He has made this point repeatedly over the intervening decades, going so far as to describe the issue as an 'existential' one for psychiatry (Wakefield 2021d, 140).

Several years after introducing the HDA, Wakefield extended his account. Starting with a paper published in 1997 (Wakefield 1997a), he introduced a new element: a form of essentialism. Although talk of essences has a history as lengthy as that of conceptual analysis, Wakefield draws upon contemporary discussions in philosophy, citing work by Saul Kripke and Hilary Putnam. The ensuing discussion is clearly indebted to Putnam's classic 1974 essay "The Meaning of 'Meaning'" (Putnam 1997). Working independently from the late 1960s, Kripke and Putnam both developed a theory of reference as a response to the semantic theories of Frege and Russell. Because of the striking similarities between them, they are often referred to jointly as 'Kripke-Putnam Semantics'.

Essentialism and reference

A rough account of the descriptivist theory is that the reference of a name is whatever object answers to a description that the name abbreviates. Kripke (whose interest was initially in proper names) believes that a sound theory of naming must allow that we can use a name to refer to the same individual even if the definite descriptions we apply to them are false. Consider an example from Putnam: "suppose that while Nixon was still president of the United States, someone had said, 'The president would never have become president if his mother had not encouraged him to aim high.' The hypothetical situation envisaged is one in which an entity which is person-identical with the actual president at the time of the speech-act (that is, with Richard Nixon) fails to become president (and hence fails to be denoted by the definite description 'the president)" (Putnam 1992, 58). Nonetheless, it is surely the case that the name 'Nixon' still refers to the counterfactual person-identical Nixon, even though the description 'the president' does not apply to him. Proper names are, in Kripke's parlance, *rigid designators*: they refer to their object in every 'possible world'. Contingent facts about the object can be radically different without altering this relationship.

If this is right, we require an alternative account of how names originally come to refer. Kripke discusses this in terms of a 'baptism' (Kripke 1980, 96). There is, he says, an original act of naming which is communicated to others in the community; although obviously modelled on our practice of naming newborns, he assumes that something equivalent must apply to the names we assign to other things, like planets, mountains, or cities. Particularly in these latter cases, this information may be passed on between generations so that the relation between the name and the thing it refers to may endure in the community across long periods of time. The baptism will typically be made ostensively, i.e., by physically indicating the individual to be named, although description may be used if necessary. In this case, however, the description is *not* a synonym for the thing named, as in the descriptivist theory; it simply secures the referential relationship between name and thing. The significance of the act of naming should not be underestimated. As the authors of one paper put it, "if external matters of fact contribute to reference determination, this is because we ourselves pass the responsibility of reference fixing over to the external world." (Jylkkä, Railo, and Haukioja 2009, 39)

Kripke and Putnam's respective theories are actually quite different in a number of important respects, a fact that the convenience of the 'Kripke-Putnam' label tends to obscure. In particular, Putnam is concerned with the reference of natural-kind terms such as 'water, 'gold', etc. While Kripke does mention natural kinds, his account is rooted in the naming of individuals. Given Wakefield's specific task, it is Putnam's theory that he draws upon specifically. Furthermore, essentialism has strong historical associations with metaphysics. While Kripke does endorse a metaphysical doctrine, Putnam backs away from it. A very few authors, notably Hacking (2007), have pointed out that Putnam explicitly distances himself from metaphysics, and rarely mentions essences or essentialism. For him, 'essences' are scientifically discoverable microstructures. Wakefield, of course, needs to avoid the taint of metaphysics because of his position *vis-à-vis* antipsychiatry. Metaphysical talk, even of a relatively innocent kind, will bring with it the suspicion that values are being smuggled back in, via speculation, to the supposedly objective concept of mental disorder. Nonetheless, he is happy to adopt the language of essences on the condition that it be understood in scientific terms.

Putnam uses his now-famous Twin Earth thought experiment to illustrate his own theory of meaning. Twin Earth is exactly like our own Earth, except that on Twin Earth a liquid closely resembling water is not, at the molecular level, H_2O, as it is on Earth, but has a complex chemical structure represented by the shorthand XYZ. (There is, presumably, *no* H_2O on Twin Earth). Given the traditional understanding of concepts and their associated terms, a person on Earth and their counterpart on Twin Earth will be in the same mental state when they determine, from its surface properties, that a sample of these substances is 'water'. But the extension differs in the sense that the microphysical nature of the substances is different. As he puts it, "it is possible for two speakers to be in exactly the *same* psychological state (in the narrow sense[19]) even though the extension of the term *A* in the idiolect of the one is different from the extension of the terms *A* in the idiolect of the other. Extension is *not* determined by psychological state" (Putnam 1997, 222). Furthermore, Putnam is confident that if an Earthly visitor were to learn that 'water' on Twin Earth is XYX, they would deny that it *was* water, despite its apparent similarity.

While this is a compressed account, the upshot is that concepts, whatever else they may be good for, cannot fix the reference of the corresponding word – at least, for natural-kind terms. As with the example of Richard Nixon, even when our concepts are confused or plain wrong in some respect, there is still a fact of the matter about what, or whom, we are referring to. Like Kripke, Putnam talks about acts of ostensive definition in which terms like 'water' are attached to substances and thereby become rigid designators. The obvious difference is that natural kind terms refer to sorts of things rather than individuals. Because of this, Putnam argues that these terms are like *indexicals* – words like 'this' or 'here', that do not have a meaning as such but are context-dependent. On this reading, "'water' is stuff that bears a certain similarity relation to the water *around here*" (Putnam 1997, 234) We still have concepts of 'water' and other natural kinds, and we will use them to identify stuff that we will call 'water', but the accuracy of our reference is not fixed by them. *Whatever* the content of our concept, 'water' will refer only to that which is similar to the stuff that was originally named 'water'. This 'similarity relation' is not intended to be deciphered in terms of

[19] The 'narrow' sense of a psychological state is one based on the assumption of what he calls 'methodological solipsism', the idea that to attribute a state to a subject entails nothing about the subject's environment. Thus a state such as "*x* is jealous of *y*" (Putnam's own example) is 'broad', since it entails the existence of *y* in *x*'s environment. The idea is that two people in the same narrow psychological state will be in the same state regardless of changes in their environment. It is this that Putnam challenges.

descriptive predicates, but – as the Twin Earth example illustrates – in terms of a scientifically-discoverable microphysical structure that (presumably) causally determines the more superficial surface properties. These latter are what Putnam calls the stereotype: "a standardised description of features of the kind that are typical . . . which in normal situations constitute ways of recognising if a thing belongs to the kind . . . ." (Putnam 1997, 230).

Wakefield seems to adopt all these features of Putnam's theory, although on the face of it this doesn't sit easily with his continuing commitment to conceptual analysis. One of the reasons Putnam was moved to think about meaning and reference was precisely the problem that his Twin Earth experiment seems to highlight. To the extent that concepts have meanings, meaning doesn't seem to determine extension. Analysing them won't fix this, since this will only, at best, clarify the macroscopic properties we include in the concept, which properties turn out to be highly fallible indicators of identity. We need to turn our attention back to Wakefield to work out what he is trying to do.

Inside the black box

Introduced in 1997, several years after the HDA was unveiled, Wakefield calls his approach *black box essentialism*: "in a black-box-essentialist definition, the surface properties of a base set are used to pick out a possibly unknown underlying essence, and the essence is used as the necessary and sufficient criterion for category membership of new instances, whether or not the new instances share the surface properties of the base set" (Wakefield 1997a, 658).

The 'black box' metaphor is intended to capture the idea that we postulate essences for some of our concepts, even in the absence of any corresponding knowledge. The name could give the impression that there is something novel about this, although it is a familiar enough observation. Putnam himself mentions it in "The Meaning of 'Meaning'", but it was noted as far back as the 17th century by Locke in *An Essay on Human Understanding*. He even anticipates Putnam somewhat when he writes "For though in that called *Gold*, one puts into his complex *Idea*, what another leaves out; and *Vice Versa*: yet Men do not usually think, that therefore the Species is changed: Because they secretly in their Minds refer that name, and suppose it annexed to a real immutable Essence of a thing existing, on which those Properties depend" (Locke and Nidditch 1975, 501)

Apart from this, there are some quirks in the way Wakefield describes black-box essentialism, and he substitutes *base set* for Putnam's stereotype. What he means here, is that instead of a series of conjoined macroscopic properties, P_1, P_2, . . . , P_n, we use a number of (perhaps idealised) members of the concept that *display* the typical properties. Thus, "Given some base set of known, agreed, symptomatically recognisable cases of disorder, such as obsessional, depressive, psychotic, and other disorders, 'disorder' might be defined as 'any condition that is either a member of the base set or has the same underlying essential nature as the members of the base set'" (Wakefield 1997a, 658).

In this way he attempts to circumvent the obvious problem of explaining what the stereotype of 'mental disorder' would consist of. This is hard to even conceive, given the diversity of observable 'properties' among all the conditions that we tend to call mental disorders. It is marginally easier to accept that we might use exemplars to establish mental disorder as a (putative) natural kind. The case of mental disorder is also different from Putnam's examples, because it is the genera under which the species concepts 'obsessional disorder', 'depressive disorder', etc. are being collected, whereas 'gold' refers to a substance. Although Wakefield does seem to be using specific cases in his base set, they must *represent* these concepts.

This question remains: why does Wakefield extend his analysis in this way? T. E. Wilkerson makes an illuminating point about semantic externalism that might shed some light on this: "we have *committed* ourselves to applying it [the term 'tiger'] to anything that has that underlying property, that has the relevant sameness relation to our original stereotypical tigers." (Wilkerson 1993, 3; emphasis mine) Although this is the point that Twin Earth is meant to illustrate, it is surprisingly easy for the element of obligation to escape attention. It is useful to see this in the context of Wakefield's HDA: if he can make it plausible that Putnam's semantics applies just as well to mental disorder as to tigers, then it would seem that we have made the same sort of commitment. We will have, as it were, a rational obligation to apply 'mental disorder' only to whatever possesses the essence (if any) possessed by the samples in the base set by way of which the concept was allegedly baptised.

There is a wrinkle, of course. As he writes, "'disorder' is not a purely theoretical, essentialist concept. Many practical concepts require certain effects for membership rather than just an underlying essentialist structure" (Wakefield 1997a, 661). The 'effects' he has in mind here is a reference to the 'harm' element of the HDA. It is *necessary* for a dysfunction to be collectively judged harmful in order to be a disorder, but harm is not *essential* in the relevant sense, since it is precisely the point of essences that they are intrinsic to a thing. So the concept 'mental disorder' is not essentialist in the same way that 'tiger' might be. If we said simply that the base set of disorders had an underlying essence, and that 'mental disorder' should therefore apply to whatever shared that essence, very little would be gained. Without knowing what (if anything) the essential property was, we would have no criteria by which to determine what was or wasn't a mental disorder. Further, given the heterogeneity of what we are accustomed to calling mental disorders, we would have little idea what sort of thing the postulated essence might be or where we should look for it.

Given his analysis, then, the essentialist narrative unfolds in something like the following way. We encounter individuals who exhibit atypical behaviour of a certain sort, note the similarities between these behaviours, and postulate a common underlying nature, something ("we know not what" in Locke's phrase (Locke and Nidditch 1975, 580)) that explains these surface appearances. With this in place, 'mental disorder' is dubbed, and future candidates for membership under this concept must possess that nature, though it be unknown. At this point, however, the concept itself is, like many concepts, not explicitly understood. We were not yet cognisant that, per Wakefield, only *part* of our concept is an essence-bearer. There has been much confusion and debate regarding the nature of the concept, over a long period of time, but the HDA eventually resolves the issue, showing that an intuitive notion of *dysfunction* is what we were actually essentialising. And it transpires that there is already a scientific theory that provides this essence, although indirectly through the concept of 'function': natural selection.

Regarding this, it is interesting to note that Wakefield has taken issue with two prominent philosophers of biological function who might, on the face of it, be his natural allies. Ruth Millikan and Karen Neander have both perceived as intractable the problem of maintaining continuity between the pre-theoretical concept of 'function' and the theory of natural selection. Millikan is highly critical of conceptual analysis, which she sees as a fundamentally confused project. Consequently, she makes it a matter of 'theoretical definition' that natural functions are those shaped through natural selection (Millikan 1989). This is, in effect, a straightforward case of stipulation. Karen Neander, on the other hand, maintains that the biological concept of function has *changed* its meaning post-Darwin (Neander 1991). They both mount detailed arguments to justify their respective positions, a task made easier to the extent that they are primarily concerned with the concept in its scientific use. But the concept 'mental disorder' unavoidably spans science, healthcare, and public interest. If Wakefield were to take a similar route, any appeal to a 'natural' reference-fixing mechanism would be ruled out, and he would instead face the thankless task of selling an engineered definition to an audience with diverging interests.[20] Instead, he is crafting a narrative that seems to show that 'mental disorder' and 'function' (with the latter grounding 'dysfunction') are concepts with a long, unbroken history. If this narrative were to be successful, his whole account of mental disorder will appear to be founded on time-honoured practices that merge seamlessly with scientific discovery. It might even seem unreasonable to hold any other view against it. He insists, therefore, that we share the same basic concept of natural function with thinkers going back at least as far as Aristotle, though our explanatory *theories* may differ (Wakefield 2000).

Although, at first sight, the HDA seems elegantly simple, we have seen that it actually involves a complex series of conceptual and theoretical elements, each contributing to an overall picture wherein the concept 'mental disorder' has emerged naturally from our collective, historical experience with the phenomena, and our associated conceptual practices. In addition to this, the combined elements purport to offer a mechanism by which disorder can be discerned from nondisorder according an objective, scientifically-backed criterion. While I believe that there are, in fact, many flaws in this picture, I will concentrate now on some specific issues prompted by considering the HDA in the light of Immanuel Kant's philosophy, with emphasis on certain sections of his *Critique of Pure Reason*.

Kant, concepts, and analysis

With his *Critique of Pure Reason*, Kant sought a middle way between the two epistemological traditions that dominated philosophy in the 18th century. On one side were the rationalists, like Descartes and Leibniz, who believed that knowledge of the sort possessed by humans could not be supplied by experience alone; this was the tradition in

[20] Christopher Boorse, by contrast, is not interested in demonstrating conceptual continuity across linguistic communites and/or over time. His view is therefore more similar to those of Millikan and Neander.

which he had been schooled. On the other were the empiricists, who insisted that all knowledge whatsoever was, and could only be, derived from experience. In Locke's famous analysis, the human mind was a blank slate waiting for the inscription of sensible intuition. As we have already seen, rationalist metaphysics had come to trouble Kant deeply. In claiming knowledge of supersensible entities, he took these efforts not only to transgress an intellectual boundary, but also to evoke the delusions of madness. The *Critique* was his attempt to determine what kind of knowledge reason could, independently of the senses, claim to yield. In so doing, he is also concerned to mount an anti-sceptical argument that defends the possibility of *a priori* knowledge against the arguments of David Hume, in particular.

Kant identifies three 'faculties' operative in human cognition, sensibility, understanding, and reason.[21] Of the first he says that "Objects are . . . given to us by means of sensibility, and it alone affords us intuitions", which latter is the immediate relation between a cognising subject and an external object. Through this passive faculty we receive representations, or "determinations of the mind in this or that relation of time" (A197/B242). Our intuition of objects is through the senses, and in accordance with the 'forms of sensibility', space and time, which are not, as we might suppose, external phenomena, but subjective structuring features imposed by the human mind. Understanding, by contrast, is the spontaneous[22] faculty of concepts; concepts make judgements possible, thus Kant also describes understanding as the faculty of judgement. The kind of judgement under consideration is primarily of the categorical or subject-predicate form.

Concepts as dealt with in the *Critique of Pure Reason* (henceforth CPR) are primarily the pure *a priori* concepts of the understanding, the categories. These are the most fundamental forms that experience of objects can take, including, crucially, the concept of causality. By invoking the categories, Kant hoped to counter scepticism: contra Hume, causality is not an associative habit of mind brought about *a posteriori* by the 'constant conjunction' of appearances, but a subjectively-contributed condition of the

[21] The latter is construed by Kant alternatively as a particular use of the understanding or an independent faculty. In Kemp Smith's analysis, arguments for both views are present in the Dialectic. The two views are in tension, almost but not fully reconciled (Kemp Smith 1969, 426f).

[22] The nature of this 'spontaneity' is rather obscure, but it is clearly to be contrasted with the passivity of reception – there is some sense in which the contribution of the understanding is an activity. This problem is discussed in Pippin (1987).

very possibility of experience. Only if the content of sensible intuition can be brought under the pure concepts of understanding can experience, strictly speaking, arise. As one of the most frequently quoted lines from the CPR informs us, "Thoughts without content are empty, intuitions without concepts are blind" (A51/B75). And shortly before this, at A50/B74, he says that neither intuitions nor concepts independently "can yield a cognition." Both the faculty of sensibility and that of understanding must work together to produce experience.

We also need *empirical* concepts just in order to move from intuitions of undifferentiated objects to cognition of determinate *kinds* of objects, such as gold, water, tigers, lemons and so on. The categories alone would give us an extremely limited experience of the world. As Houston Smit puts it, "being conscious merely of an object in general (as we are when we represent a thing merely in the categories) does not amount to cognising a thing. To cognise a thing, one must be conscious of a thing in respect of its determinate identity, so as to distinguish it from (some) other things" (Smit 2000, 243)

Kant equates 'experience' with 'empirical cognition' (e.g. B147). It is precisely the unity of intuited object and conceptual determination (fundamentally through the categories, as grounds for empirical concepts) that yields this cognition ("cognition in the proper sense" (A78/B103)), so empirical concepts play a crucial role in making experience, in this sense, possible. When he says that "the understanding can make no other use of these concepts than that of judging by them" (A68/B93), part of what he is referring to is basic object-determination. My experience of the external world is of a multitude of, not merely objects, but of kinds of object: trees, dogs, postboxes and so on. This is explained, for Kant, partly by way of concepts, and therefore by way of judgement. It is not, however, that I consciously (and laboriously) judge that there are trees, dogs, and postboxes in my surroundings just in order to experience them. Kant's point, rather, is that object-determination through concepts can be expressed in the *form* of judgements such as 'this object is a tree'.

Under the rubric of 'judgement' Kant talks also about the *structure* of concepts, which is to say that he explains what object-determination more specifically consists in. Using the example "All bodies are divisible" (A68/B93) he explains that in such a judgement there is a concept "that holds of many, and that among this many also comprehends a given representation, which is then related immediately to the object" (Ibid.). The predicate ‘divisible’ may be applied to (“holds of”) different things, but is here applied to the concept ‘body’, while the concept ‘body’ may be applied to particular objects given to us in sensible intuition; ‘divisibility’ is therefore a ‘mediate representation’, since it is applies to particular objects *through* the concept of a body. To judge that bodies are divisible is for Kant, analytic. It expresses part of the content of the concept ‘body’, or part of what we *mean* by it.

Things are clarified somewhat if we recognise that he cleaves to a “view of concepts that enjoyed wide acceptance in the seventeenth and eighteenth centuries . . . a complex concept can be taken as a conjunction of (more) elementary concepts” (de Jong 1995, 623). Kant calls these predicate-concepts ‘marks’ (*Merkmale*). When we determine that an object is a body, we are intuiting that it possesses various properties or marks that are present in the corresponding concept ‘body’. So when he gives his example ‘All bodies are divisible’, what we are being presented with is a judgement that expresses *part* of the concept ‘body’ (this is also why he sometimes calls these marks ‘partial representations’). To judge that something is a body is, *inter alia*, to judge that this thing possesses the mark of divisibility. The judgement *that* bodies are divisible just lays out the relation of one concept, as a mark, to another concept, the subject. A concept is therefore structured in a certain way, by ‘containing’ a series of marks, which are themselves concepts. Kant’s explanation of the way we acquire empirical concepts makes this more explicit. As he explains in the Jäsche logic lecture notes,

> To make concepts out of representations one must thus be able to compare, to reflect, and to abstract, for these three logical operations of the understanding are the essential and universal conditions for generation of every concept whatsoever. I see, e.g., a spruce, a willow, and a linden. By first comparing these objects with one another I note that they are different from one another in regard to the trunk, the branches, the leaves, etc.; but next I reflect on that which they have in common among themselves, trunk, branches, and leaves themselves, and I abstract from the quantity, the figure, etc., of these; thus I acquire a concept of a tree. (Kant 1992d)

Trunks, branches, and leaves are each concepts in their own right, but are here predicates of, thus contained in, the concept ‘tree’. Concept acquisition is, as the example shows, a case of synthesis or combination. This is not to say that this is a conscious operation, any more than ordinary object-determination is (hence Kant says "one must be able to"). And, in consequence, it will often be the case that we are at best only vaguely aware of the content of our concepts. To get clearer on this is a matter of analysis – in effect, of reverse-engineering. In another celebrated passage, Kant says: "In all judgements in which the relation of a subject to the predicate is thought . . . this relation is possible in two different ways. Either the predicate B belongs to the subject A as something that is (covertly) contained in this concept A; or B lies entirely outside the concept A, though to be sure it stands in connection with it. In the first case I call the judgement analytic, in the second synthetic" (A7/B10).

A synthetic judgement adds something to our cognition of an object, beyond what was already thought through the concept by which the object was made determinate. "This flower is yellow" is synthetic because the concept 'flower' does not contain the property of being yellow: many flowers are not yellow. This judgement adds to our cognition of this particular flower, or 'amplifies' it. In this case, the subject is a particular object, a flower, given in intuition, but synthetic judgements can also hold between concepts, as they do in an analytic judgement such as "All bodies are divisible". The judgement "All bodies are heavy" (B11) relates one concept to another but it is (whether true or not) synthetic, since the property of being heavy is not part of the concept 'body'. It combines, or synthesises, the two concepts.

Kant and harmful dysfunction

With these fundamentals in place, we can start to consider a Kantian response to Wakefield. Per the HDA, what mental disorder essentially is, is mental dysfunction. We may, for our purposes, discount the 'harm' element, given that it is explicitly evaluative and therefore subjective. Wakefield's key claim is that "The HD analysis asserts that disorder, both physical and mental, requires harm, a value criterion, and dysfunction, *a factual criterion* referring to failure of a mechanism to perform a naturally selected function" (Wakefield 2003, 969; emphasis mine).

What is at stake is whether part of the content of the concept is indeed 'factual', which I take to mean that an attribution of mental disorder, if true, is true in virtue of reference to an empirically discoverable entity or 'object' in a broad sense. For Kant, an objective (rather, an 'objectively valid') concept *is* an empirical concept (which itself presupposes the categories). That is to say, it is a concept that makes experience *of* an object possible, in the sense discussed above.

Kant, it should be noted, does not conceive analysis in terms of necessary and sufficient conditions. Given that the usual product of analysis is definition, Kant discusses his own view in one of the later sections of the CPR, the Discipline of Pure Reason, where he considers the attempts of rationalists like Christian Wolff to apply the methods of mathematics to philosophy. Here he says, "To **define** properly means just to exhibit originally the exhaustive concept of a thing within its boundaries. Given such a requirement, an **empirical** concept cannot be defined at all but only **explicated**." (A727/B755) In mathematics, genuine definitions are possible because they are constructed in intuition; their objects are not found in experience. The properties of a triangle, for example, can be derived from the definition, and they can be known in their entirety, i.e., 'exhaustively'. By contrast, concepts derived from empirical experience and that refer to empirical phenomena cannot be given such precise and exhaustive definitions. This is also because, as he says, "One makes use of certain marks only as long as they are sufficient for making distinctions; new observations, however, take some away and add some, and therefore the concept never remains within secure boundaries" (A728/B756).

In other words, on Kant's view empirical concepts are open-ended – they are always revisable in light of fresh experience. This does not mean that they are entirely mutable. Distinguishing between analytic and synthetic marks, Kant says " *The former* are partial concepts of my *actual* concept (marks that I already think therein), while the latter are partial concepts of the *merely possible* complete concept . . . ." (Kant 1992d, 565). Analytic marks are the core of the concept, or its *logical essence*, in so far as they are part of how we apply it, but it is anticipated that an indeterminate number of additional marks may be possessed by the things within the concept's extension. In particular, we can never know, even in principle, whether we have learned enough about the kinds we conceptualise to say that we possess exhaustive knowledge of them, since there is no way experience can tell us that we *have* exhausted all the properties they may possess. This is why he says they cannot be defined. As Patricia Kitcher expresses it, "A definition model implies rigidity in the face of new experience, but the basic [Kantian] theoretical assumption about concepts is that they are moulded by experience" (Kitcher 1990, 212).

This aside, what is notable about the definition given by the HDA is that it does not, as Kant would expect it to, define 'mental disorder' in terms of marks, or observable properties, by which an empirically intuited 'object' could be picked out. Again, it is important to emphasise, with Robert Hanna, that for Kant an empirical concept "contains only phenomenological, 'identificational' sub-concepts . . . ." (Hanna 2006, 147). Neither function nor dysfunction are observable properties, however, though function *bearers* (the traits to which functions can be attributed) may be physical features such as wings or eyes. Nominally, 'mental mechanisms' are function bearers, but are not physical features, so there is a dual sense in which the concept of specifically *mental* dysfunction fails to make experience of an object possible. Wakefield obliquely acknowledges this, noting that "To say that a harm is due to a disorder is to say that the harm is due to the fact that some internal mechanism is not functioning the way it was designed by nature to function. *This attribution is inferential* . . . ." (Wakefield 1992b, 385; emphasis mine).

The inference in question is specifically an abductive inference, or inference to the best explanation[23]: the 'factual' component of mental disorder is a property (dysfunction) postulated to be the most likely explanation of the observable features. This marks a significant difference not only from Kant's understanding of empirical concepts, but also from the semantic externalism of Kripke and Putnam. As we have seen, they were interested in how the reference of a term could be fixed, given that the descriptive content of a natural kind concept consists of surface properties that are highly fallible criteria for concept membership. Reference, in their view, is fixed by a material, scientifically discoverable, microstructure, such as the molecular structure of water, H_2O, something that can be detected independently of the macroscopic appearance of this substance. When it matters scientifically, we most certainly do not *abductively infer* that water is H_2O.

Furthermore, while our Putnamian 'stereotype' of water (for example) is in fact quite a reliable guide to picking out particular samples of water, the same is manifestly not true for mental disorder. As Valérie Aucouturier and Steeves Demazeux point out, "if the concept of mental disorder did pick out some bundle of natural properties like 'water' does, then there would not be any practical or philosophical problem: we would already

[23] Abduction was first discussed as such by C. S. Peirce; there are some minor differences between it and inference to the best explanation, but their interchangability in contemporary use is ubiquitous.

have an agreement on what our object of investigation is" (Aucouturier and Demazeux 2012, 78). It would then be a relatively straightforward task to find out what, if anything, the underlying microstructure was. As regards the surface properties of mental disorder, there is hardly any need to emphasise the enormous diversity of the phenomena that we nonetheless bring under the concept. This is exactly why the issue of definition is such a vexed one, and it is why the Szaszian argument that Wakefield takes so seriously has any bite. Wakefield's analysis cannot escape this difficulty, because inference demands a ground, and the ground in this case is precisely where the problem lies.

It should be said that Putnam's own examples are not entirely uncontroversial. Is water indeed identical with H_2O? As various contributors to the debate about semantic externalism have noted, alternatives that are no less intuitive are available. Consider John Dupré's comment regarding Putnam's Twin Earth experiment, that "it is surely just the absence of experiences like the one Putnam describes that makes it reasonable to attach to molecular structure at least most of the importance that Putnam ascribes to it." (Dupré 1981, 72) His point goes deeper than the commonplace objection that thought experiments strain credulity. If we agree with Putnam, it isn't simply due to the premise of an underlying nature. We will want to ask, *how* similar is water on Twin Earth to our water? Given what we know about the significance of molecular structure, would we be willing to *drink* it? The bare assumption of a 'microstructure' doesn't decide the issue – a wider web of associated beliefs is also required. Conversely, if XYZ was enough like water on Earth for us to treat it as interchangeable, it would simply complicate matters to insist on calling it something else. And, after all, Putnam himself allows that "if H_2O and XYZ had both been plentiful on Earth . . . it would have been correct to say there were two kinds of 'water'" (Putnam 1997, 241).[24]

This is not, at any rate, something that Wakefield considers. Expounding his black-box essentialism, Wakefield asserts that "It turns out that the process that explains the prototypical non-accidental benefits [observed in organismic traits] is natural selection acting to increase inclusive fitness of the organism. Therefore, a function of a biological mechanism is any naturally selected effect of the mechanism" (Wakefield 1999b, 471f).

[24] The comment follows a discussion about jade, which turned out to be two chemically different, though superficially similar, substances: jadeite and nephrite. Since they both do actually occur in relative abundance, Putnam does consider there to be two types of jade. This does not, however, preclude making distinctions between jadeite and nephrite, or H_2O and XYZ, if it is considered necessary for specific reasons.

This is, I take it, supposed to be the 'scientific discovery' that indirectly provides the reference-fixing microproperty for 'dysfunction'. But this is clearly not in the same league as water/H_2O. What the theory of natural selection tells us is that some traits of organisms have been favoured by a certain process, and can be considered adaptations. There is good evidence that some physical traits of various organisms have indeed been shaped in this way. However, the kinds of physical evidence that biologists use to determine what specific traits actually are likely to be adaptations are not available for mental phenomena, for obvious reasons. As far as mental mechanisms are concerned, natural selection tells us no more than that some of our mental capacities may plausibly have been shaped by selection.

Although Wakefield has not committed himself to any particular field of enquiry, evolutionary psychology might seem to be the logical means by which the existence of mental mechanisms and, potentially, their failures, might be discovered and described. He has, in one co-authored paper, expressed enthusiasm for this project (Buss et al. 1998) but, in general, remains silent about how he anticipates empirical research will reveal dysfunctions.[25] He is sanguine nonetheless; in his view "Evolutionary theory, though not yet an explicit diagnostic aid, explains biological design and thus shows that disorder refers to a scientifically identifiable phenomenon" (Wakefield 1999b, 465).

A number of critics have made what Wakefield calls the 'epistemological objection' against the HDA: that there is currently no empirical confirmation of any specific mental dysfunction. In reply he argues that "The epistemological objection is based on the assumption that, to know that there is a dysfunction, one must know the dysfunctional mechanisms and their evolutionary history. This assumption is false. To know that a dysfunction exists, one need only have sufficient indirect evidence – for example, surface evidence that indicates or correlates with the existence of internal dysfunction – to infer that some mechanism is failing to perform as designed" (Wakefield 1997b, 255). The argument as presented here appears to be circular: to know that there is a dysfunction, one only requires evidence to infer that there is a dysfunction. But what counts as evidence? In the same article he gives the example of an automobile that won't start, from which a plausible inference can surely be made that some part is not functioning.

[25] Evolutionary psychology and its forerunner, sociobiology, are controversial subjects in their own right and a meaningful discussion of them exceeds the scope of this thesis. It is worth noting that it faces distinctive methodological constraints that may limit its application to psychiatric disorder.

But Wakefield himself is making an assumption here, and a false one too. What makes the inference plausible is that cars are designed by humans. Their parts have functions by virtue of this fact. Organisms are not designed at all, and if we don't know what mental functions we actually possess, we cannot claim to know what sort of mental states can plausibly be said to represent dysfunctions. In fact, his claim, if we take the above quote literally, is simply wrong: we cannot *know* anything by inference to the best explanation.

Nonetheless, Wakefield has postulated specific mechanisms – for example, "mechanisms designed to constrain food choices" (Wakefield 1997b, 254), "loss-response mechanisms" (Wakefield 1999c, 1008), and "coping mechanisms" (Wakefield 1992b, 381). It is with some justification that Dominic Murphy and Robert Woolfolk accuse him, in lieu of "an established evolutionary science of the mind" of "practicing a highly speculative form of evolutionary faculty psychology in which we infer dysfunction on the basis of behavioural observation" (Murphy and Woolfolk 2000, 245). We may note, of course, that Kant himself postulates mental faculties, as have philosophers and practitioners in a number of more recent disciplines; Jerry Fodor's influential *The Modularity of Mind* is even subtitled *An Essay on Faculty Psychology* (Fodor 1983). There is nothing inherently problematic about so doing. As scholars such as Patricia Kitcher (Kitcher 1990), Andrew Brook (Brook 1994), and others have argued, the critical Kant reveals himself to be a kind of proto-functionalist, in the sense more familiar from contemporary philosophy of mind and cognitive science rather than evolutionary biology. His description of the faculties is an abstract specification of the most basic cognitive operations necessary for experience (as we know it) to be possible. But he does not conceive them as objectively valid, since they are what ground objective validity in the first place.[26] Wakefield is not actually claiming 'factual' evolutionary origins for his particular speculations, so the point made by Murphy and Woolfolk is not entirely warranted, but his willingness to posit 'mechanisms' does muddy the waters somewhat.

To sum up so far, Kant's view of concepts (the categories and empirical concepts) and their role in actually constituting determinate objects of experience helps clarify the lack of 'objective validity' in Wakefield's definition of mental disorder. This is a point that

[26] Bernard Williams makes something like this point when he admonishes some interpreters of Kant for failing to recognise that Kant's "transcendental arguments gave knowledge of how things must be only because the things were not things in themselves" (Williams 1999, 128). Chong-Fuk Lau also addresses the question of trancendental concepts and objective validity in (Lau 2015).

the empiricist would also make, no doubt, although on the basis of a somewhat different view of concepts and their role in cognition. But Kant is neither an empiricist nor a rationalist. It is true that, for him, only empirical concepts make objects possible, by unifying the data given through sensible intuition, and it is only these that can properly be considered factual. There is room, however, in Kant's critical philosophy for concepts of a different sort, which he more often refers to as *ideas*. Considering the HDA in light of this sort of concept offers a deeper Kantian diagnosis of what is at fault in Wakefield's attempt at definition.

The Transcendental Dialectic and concepts of reason

What Kant refers to as ideas are due to a third fundamental faculty of cognition, reason. There is considerable ambiguity as to how Kant conceived reason, since he describes it alternately as understanding operating independently of sensible intuition, and an independent faculty with its own agenda and principles. Regardless of this exegetical point, it is clear enough that while understanding makes experience possible through effecting unity in the otherwise disparate manifold given to sensibility, reason, however we are to characterise it, brings a higher order of unity to that experience – more precisely, it unifies empirical concepts of the understanding by establishing 'conditioning' relationships between them. Such relationships are not to be found in the experience that our concepts (*a priori* and empirical) themselves make possible: reason is a capacity to go beyond what is given to us in experience. It enables us to expand our particular cognitions into an intellectual representation of reality, via concepts, that transcends any experience that finite beings such as ourselves could possibly have.

Kant describes understanding as the faculty of rules (i.e., concepts) that unify the manifold of intuition. Reason, by contrast, he describes as the faculty of principles (A299/B356). Here, too, unity is a central concern, albeit of a different kind. Whereas the categories and our empirical concepts make experience of objects possible by effecting unity in the manifold given to sensibility, reason seeks unity among our manifold of concepts. This is best understood in terms of the explanatory role of concepts. We take ourselves to be in a position to ask, for every thing, why it is the way it is, rather than simply accepting it as being that way. Asking for explanations, or reasons for why a particular thing is as it is, or why something occurs, means to ask for relationships between concepts, because it means first of all establishing what *sort* of thing a given object is – under what concept it comes. What we then want to know is what gives the reason for that concept (that sort of thing) being as it is, or, what are its more general conditions. That is to say, we look to bring our original concept under a more general one. For every explanation forthcoming, we can repeat the demand for conditions so that there is, in principle, either no end to the sequence, or else it must terminate eventually in an unconditioned, something that stands in need of no further explanation. In Kant's view, the task of reason is to seek this kind of unity, with a view to acquiring a completely unified system of knowledge in which higher, more general concepts give the reasons, or conditions, for lower, more specific ones.

In the Introduction to the Dialectic, Kant initially characterises the activity of reason as a 'logical maxim' (A307/B364). is simply a way to organise our knowledge in a hierarchically structured manner. As Kant puts it, "the proper principle of reason in general (in its logical use) is to find the unconditioned for conditioned cognitions of the understanding, with which its unity will be completed." (A307/B364) This is "merely a subjective law for the orderly management of the possessions of our understanding." (A306/B362; Kemp Smith translation) In other words, the maxim prompts us to seek explanatory relationships between items of previously acquired empirical knowledge, with the notional goal ultimately being a completed hierarchy of conditioning relationships terminating in the unconditioned. It does not, however, assert the *existence* of an unconditioned. Most importantly, it "does not prescribe any law to objects, and does not contain the ground of the possibility of cognising and determining them as such . . . ." (A306/B362)

Kant goes on to argue, however, that we inevitably treat this merely logical principle as if it has ontological significance. In an important passage, he explains that "this logical maxim cannot become a principle of **pure reason** unless we assume that when the conditioned is given, then so is the whole series of conditions subordinated one to the other, which is itself unconditioned, also given (i.e., contained in the object and its connection)." (B364/A307-308) This, the 'supreme principle of reason', is an explicitly metaphysical principle, and it is due to what Kant calls 'transcendental illusion', whereby "in our reason (considered subjectively as a human faculty of cognition) there lie fundamental rules and maxims for its use, which look entirely like objective principles, and through them it comes about that the subjective necessity of a certain connection of our concepts on behalf of the understanding is taken for an objective necessity, the determination of things in themselves.” (A297/B353) The supreme principle takes us beyond merely establishing logical connections between concepts, because its goal is to reduce the diversity of concepts to the smallest number of general principles – ultimately, to an unconditioned. Given the logical maxim, if we are missing a cognition that would provide the explanation for something, we would lack anything to establish a logical relation *with*; in order to seek a suitable cognition, however, we would need to think that such a thing was there to be found. There is nothing in the logical maxim, and nothing in empirical cognition, that could supply us with this assumption. The supreme principle, on the other hand, suggests to us that the conditioned objects of our experience actually attest to the real existence of the complete series of their conditions – something that the experience of objects itself cannot contain.

The supreme principle gives rise to more specific ‘ideas’ of reason, notably those that begin to emerge at A323/B380 and which Kant derives logically from the three forms of Aristotelian syllogism: the categorical, hypothetical, and disjunctive. These ideas, which are postively identified at A334/B392, are introduced as the “thinking subject” or soul, “the sum total of all appearances (the world)” and “the being of all beings” (or God), which three correspond to the traditional metaphysical doctrines of (rational) psychology, cosmology, and theology. Given what Kant says in the bulk of the Dialectic, the reader will receive the impression that these three ideas are the *only* ‘concepts’ into which the supreme principle of reason is specified. This serves his purposes in the middle third of the Dialectic, which is taken up with the sections known as the paralogisms, the antinomies, and the ideal of pure reason. These sections subject the three transcendental ideas to critique and constitute the ‘destructive’ part of the Dialectic, which seeks to show that metaphysical speculation about such supposed entities can never yield knowledge, and is thus an empty pursuit. This project is what originally inspired Kant to write the *Critique* and, as I have argued in chapter 3, it emerged from the parallels he drew between rationalist metaphysics and madness. Only later in the Dialectic will it be revealed that the supreme principle of reason generates an open-ended variety of other ideas.

Although by this stage Kant has (to his own satisfaction, at least) shown the inferences of rationalist metaphysics to be fallacious, the reader is left wondering why it is that we should – apparently inevitably – conflate a subjective logical principle with an objective, metaphysically suspect, one. It is this perplexity that Kant explains more fully in the Appendix to the Dialectic of Pure Reason.

The Appendix and the proper direction of reason

It is in the final section of the Dialectic, the Appendix, that Kant explains what positive purpose the apparently disreputable supreme principle of reason serves. It is here that we see that its capacity to generate 'ideas' contributes vitally to scientific inquiry.

At the outset of the Appendix, Kant expresses his view that our faculties must have a "proper direction" (A643/B671), reason as much as sensibility and understanding. He has already explained that, in its logical use, reason performs the important task of establishing connections between our empirical concepts, but it is not clear, for most of the Dialectic, why we should be subject to the apparently inextirpable transcendental illusion and thus prone to taking the logical maxim for a determination of reality itself. After all, he compares it to optical illusion, "that cannot be avoided at all, just as little as we can avoid it that the sea appears higher in the middle than at the shores . . . ." (A297/B354) Anticipating Kant's explanation, Kemp Smith notes that it is not often observed that such illusions are often beneficial: "By their means we acquire the power of compressing a wide extent of landscape into a single visual field, of determining distance, and the like. Their practical usefulness is in almost exact proportion to the freedom with which they depart from the standards of the independently real" (Kemp Smith 1969, 427).[27] Reason's equivalent freedom from the constraints of the empirical may lead us into error just as optical illusions may, but it also affords the capacity to extend our theoretical knowledge beyond what is given to us in experience – indeed, beyond any possible experience. The practical usefulness of *this* is the subject of the Appendix.

Although it is not immediately obvious from the text, the Appendix is very much about the role of reason in science. In fact, Kant does not often use the term 'science', but talks a good deal about *systematic unity* between our concepts. Reason "does not **create** any concepts (of objects) but only **orders** them and gives them that unity which they can have in their greatest possible extension . . . ." (A643/B671) This task is the one prescribed by the logical maxim, as previously discussed. But this principle is limited to

[27] Michela Massimi (2017, 77) makes a remarkably similar point, referencing the Kantian art historian Erwin Panofsky and his discussion of the discovery of perspective drawing: similarly illusory, similarly useful.

bringing unity to the concepts we have already acquired; it "does not direct us to expand our body of cognitions and does not tell us how to do so" (Willaschek 2018, 130). As far as the logical maxim is concerned, the concepts it organises are those that we have, in effect, stumbled upon and, in Kant's view, nothing in our experience of objects itself could make us conscious that there are as yet undiscovered conditions, or explanations, to seek out.

This is where the supreme principle of reason comes into play. As Kant puts it, the "unity of reason always presupposes an idea, namely that of the form of a whole of cognition, which precedes the determinate cognition of the parts . . . ." (A645/B673). It is a matter of something that might initially strike us as self-evident: that there is already a complete, interconnected, system in nature, a totality, of which any particular experience is but a fraction. Such a 'whole' is of course not an object that we could experience as such, but it must be presupposed just in order that we should be motivated to seek out those aspects of reality that will augment our empirical knowledge. If not for this, we might still be subject to animal curiosity and the instinct to conduct *ad hoc* exploration in the pursuit of some end or other, but we could hardly conceive of there being gaps between items in knowledge, much less feel the impulse to conduct principled investigations aimed at filling them. This principle of reason is what "keeps scientific enquiry going and what makes the motor driving it intrinsically reasonable" (Laywine 1998, 280 n1). What Kant is doing is trying to explain what it is about *us*, as rational animals, that justifies this, rather than looking to an external world that cannot provide finite creatures with an experience that would provide independent justification. It is, in short, an aspect of Kant's Copernican turn.

The key Kantian point is, however, that these 'gaps' must not be treated as if they correspond to pre-existing truths determinately located in an independently-existing, rationally-ordered universe. For all that we know – for all that we can ever know – nature may *not* be a systematic unity. There may be regions of discontinuity or of complete unknowability. We must, if we want to avoid straying into ungrounded speculation, be wary of the distinction he makes between the 'constitutive' and 'regulative' use of the ideas of reason. These terms occur earlier in the CPR in the Analytic of Principles, where Kant turns to the problem of how the *a priori* categories can apply to empirical objects, but reappear in the Appendix to the Transcendental Dialectic in a somewhat different guise.[28] As he writes, “the transcendental ideas are never of constitutive use, so that the concepts of certain objects would thereby be given . . . however, they have excellent and indispensably necessary regulative use, namely that of directing the understanding to a certain goal respecting which the lines of direction of all its rules converge at one point, which although it is only an idea (*focus imaginarius*) . . . nonetheless still serves to obtain for these concepts the greatest unity alongside the greatest extension” (A644/B672).

In this passage Kant recycles the device of the *focus imaginarius* from much earlier in his oeuvre, in *Dreams of a Spirit Seer* (Kant 1992c, 332). In that work he uses it to illustrate the folly of rationalists such as Wolff and Crusius, in particular – and, of course, the ‘spirit-seer’ himself, Swedenborg (see chapter 3). These metaphysicians assign self-created objects to external positions within the matrix of sensible impressions; rather than rays of light diverging from empirical objects, Kant rhetorically imagines the metaphysician projecting ideas outwards like rays to converge on a point in space where the purely mental ‘objects’ are conjured into being. Furthermore, he suggests that this “can offer a reasonable explanation of that type of mental disturbance [*storung des Gemuths*] which is called madness . . . the victim of the confusion places mere objects of his own imagination outside himself . . . . (Kant 1992c, 333). In the CPR, by contrast, Kant envisages a more benign application of the metaphor. We do, and must, ‘project’ (A647/B675) certain ideas onto the world. We must apply the supreme principle of reason because, as Susan Neiman puts it, “Without the idea that behind every conditioned stands another conditioned, and so on *ad infinitum*, we would have no reason to question the world as it appears: we could not begin to form the concept of such questioning” (Neiman 1997, 67). Nonetheless, we are not licensed to commit ourselves *a priori* to the independent existence of systematic unity in nature.

While the transcendental ideas of reason are metaphysically imposing ones, the supreme principle also enables us to conceive of any number of theoretical entities that, although they are not to be found in sensible intuition, can play important roles in our attempts to systematise our knowledge. As Kant puts it, “Such concepts of reason are not created by nature, rather we question nature according to these ideas, and we take our cognition to be defective as long as it is not adequate to them” (A645/B673f).

[28] See Banham (2013) for a discussion of the distinction between Kant’s two uses of the term ‘regulative’ .

At this point Kant provides a series of illustrations, beginning with the archaic example of pure earth, water, and air. As Michael Bennett McNulty (McNulty 2015, 7) explains, these concepts were employed by 18th century chemists, who posited these 'pure elements' as the bearers of fundamental causal powers that explained the interactions of the (non-pure) substances of empirical experience. Michelle Grier contends that these ideas are "used to unify a rather particular branch of knowledge (or, correlated with this, a very particular set of phenomena) into a 'whole'" (Grier 2001, 297). Thus, although the three transcendental ideas (and, ultimately, the idea of God alone) provide the most general concepts of unconditioned objects, we can see that reason also generates ideas of more limited scope, tailored to local aims. A few pages later we encounter another example, that of a causal 'power', which Kant applies to the human mind: "the various appearances of one and the same substance show such diversity that one must assume almost as many powers as there are effects, as in the human mind there are sensation, consciousness, imagination, memory, wit . . . a logical maxim bids us to reduce this apparent variety as far as possible by discovering hidden identity . . . ." (A648/B676f).

This is of course the work the concept 'mental disorder' is intended to do: to reduce the variety among different psychiatric conditions by relating them to a 'hidden identity' – in the case of the HDA, the essential conceptual component of dysfunction. Rather than thinking (as we well might, were it not for the prompting of reason) that every condition was unique and stood apart from each other, we *seek out* something that is common to them on the basis of our *idea* of mental dysfunction. But it is only an idea, and should only be treated regulatively, as a goal or, as Kant sometimes puts it, a 'problem'. What we must not do is assume from the get-go that there really is a hidden identity, for then we are treating the idea as if it were constitutive, that is, as if it were the concept of an object that already boasts independent existence.

What is more, it is doubtful that our investigations, whatever fruit they may bear, will ever reveal an idea to in fact have objective validity. Wakefield has pointed out on many occasions that it is a matter of empirical scientific investigation whether dysfunction is present in particular psychiatric disorders (and therefore whether they are disorders, per the HDA). This is part of his response to the 'epistemological objection' discussed above. But, leaving aside the issue of what sort of investigation, or what sort of evidence, would be required, it is improbable that research could ever verify or falsify the HDA. Science might conceivably discover definitive dysfunctions for many of the conditions listed in our nosologies – but this would not confirm that there is a dysfunction for every condition. By the same token, however, failure to discover a dysfunction leaves it open that one exists nonetheless. So it is that, as Kant wryly puts it, "even without our having attempted to find the unanimity among the many powers, or indeed even when all such attempts to discover it have failed, we nevertheless presuppose that such a thing will be found . . . ." (A650/B678).[29]

Certainly, any knowledge we would call scientific must meet reason's demand for systematic unity. Without such unity, we would possess only an aggregate of items of contingent knowledge, without any interconnection. But there is considerable tension between Kant's insistence that the supreme principle and the ideas that it generates are merely regulative, and his repeated assertions that they are transcendental principles – that is, that they too are in some sense necessary for experience. We encounter this in the passage quoted above in which Kant discusses the *focus imaginarius*: the ideas are transcendental, 'indispensable' and 'necessary'. The claim is repeated throughout the Appendix.[30] Since the doctrine of the Transcendental Analytic supposedly showed us that the understanding and its categories alone are what make experience possible, this is a disconcerting development, and one that has, unsurprisingly, divided Kant's commentators.

On one side are those who reject Kant's claim regarding the necessity of reason's supreme principle, in some cases explaining them away as unfortunate vestiges of Kant's thought in earlier stages of the *Critique*'s development that proper editing ought to have excised. This is most clearly expressed by Kemp Smith in his *Commentary*, where the so-called 'patchwork theory' of the *Critique*'s composition is drawn upon to explain what he considers the "extremely self-contradictory" nature of the Appendix (Kemp Smith 1969, 547) More recently, Paul Guyer (another notable translator of the *Critique* into English) argues that systematicity is "only an additional desideratum which reason seeks to find or construct in the empirical knowledge produced by understanding . . . there is no hint that systematicity is a necessary condition for any successful use of the understanding at all." (Guyer 1990, 33)

[29] Compare Wittgenstein: "What a curious attitude scientists have -: 'We still don't know that; but it is knowable and it is only a matter of time until we get to know it!' As if that went without saying" (Wittgenstein 1998, 40).

[30] See, for example, A651/B679f, A663/B691f.

Other commentators insist that the supreme principle and the ideas of reason are indeed necessary for the proper use of the understanding. Michelle Grier (2001, 289) and Henry Allison (2004, 434), for example, both argue that principles of reason are presupposed even for the formation of empirical concepts. Kant briefly argues this point in relation to one of his sub-principles of reason, that of homogeneity or genera, maintaining that "According to that principle, sameness of kind is necessarily presupposed in the manifold of possible experience . . . because without it no empirical concepts and hence no experience would be possible." (A654/B682) The suggestion is that the similarities upon which concepts are based cannot simply be extracted from the manifold of sensibility, but require an *a priori* principle, since we are just as free to consider every appearance as unique. As he points out prior to this at A651/B679, we encounter such diversity in experience that the freely available evidence rather suggests a *lack* of unity in the manifold. Although the understanding is properly the "faculty of concepts" (A160 /B199), it appears to require the principle of systematic unity in order to fulfil its own task.

Much more could be said on the subject of the supreme principle as transcendental, particularly in relation to its role in attributing lawful status to empirical regularities. That every event has a cause is certainly established in the Transcendental Analytic (specifically in the second analogy), but this does not guarantee the repeatability of any observed causal sequence.[31] Henry Allison states the problem thus: "there seems to be no way for the understanding to move from that 'part of the whole of possible experience' with which it is contingently acquainted to the far vaster part with which it is not. But, clearly, if the understanding cannot do this, then it cannot make universally valid claims, which, as the 'faculty of rules', is its proper work." (Allison, 427f) Since we can never be acquainted with future events *per se*, "the introduction of theoretical entities . . . is held by Kant to be the way in which reason introduces unconditioned necessity" (Grier 2001, 299) by positing ideal grounds or 'natures' that determine the behaviour of objects.

The supreme principle guides understanding at a very general level. The theoretical ideas I have briefly discussed above are quite distinct. They are not principles with universal application to experience as such, but are examples of particular specifications

[31] Although disagreeing with this view, Michael Friedman acknowledges that this is the predominant interpretation in Anglophone Kant scholarship (Friedman 1992, 164).

of the general task of achieving systematic unity. Theoretical ideas of reason, such as those of pure substances or of foundational mental powers, are not – even on transcendental readings such as those of Grier and Allison – necessary for our experience of the objects and phenomena of chemistry or psychology. In the same way, if function (and thus dysfunction) is such an idea, it does not ground our experience of mental disorder itself. Far from it – theoretical ideas presuppose experience of objects, since they are posited for the purposes of investigating them. Their task, properly speaking, is only to postulate a specific aim for the inquiry. Without such an aim, our investigations would be directionless. Nevertheless, there is no guarantee that subsequent findings will support the existence of the speculative objects. (Indeed, the 'pure substances' Kant discusses have long since been abandoned by chemistry, underscoring this very point). Above all, we must be wary of dogmatically asserting them to have independent existence.

Whether or not one agrees that the supreme principle of reason, and the ideas that it generates, is truly necessary or indispensable for the use of the understanding, scholars generally agree that they play an essential role in science, even if 'only' at the second-order level that writers such as Guyer and Michael Friedman take it to apply. Whichever interpretation one favours, it does not significantly affect my contention that mental dysfunction is a rational postulate rather than an empirical concept. The overriding point is that a theoretical idea is not drawn from experience but is speculatively generated. As with the propositions of rationalist metaphysics, however, we can easily be led to confuse such ideas with concepts of objects.

Wakefield certainly treats his definition of the concept 'mental disorder' as constitutive, maintaining that "Natural function refers to naturally selected effects, a concept well-anchored in a scientific theory, so dysfunction and disorder also refer to real phenomena" (Wakefield 1999b, 472). His essentialism purports to sidestep the issue of observable identifying properties in favour of an appeal to a scientifically-discoverable underlying nature. This shifts the issue of 'constitution' to whatever microphysical substrate turns out to actually underlie mental disorders (if any such thing can in fact be discovered). But, as discussed above, naturally selected effects are those for which a history of selection can be evidenced, and this poses a particular problem for mental functions.

Kant's discussions of essentialism are, admittedly, fragmentary, but they are unambiguous. In his late metaphysics lectures (the L_2 notes, dated to 1790/91) he is recorded as saying ". . . I observe through experience much that belongs to its existence . . . Now the inner ground of all this is the nature of the thing. We can infer the inner principle only from the properties known to us; *therefore the real essence of things is inscrutable to us* . . . . (Kant 1997, 319).[32] This, of course, mirrors precisely Wakefield's admission that dysfunction can only be inferred from "the properties known to us", but Kant, in keeping with his strict insistence on the role of sensible intuition in cognition, and therefore knowledge *per se*, denies that this essential property can be known. In his Kantian critique of Kripke and Putnam's 'essentialist' theory of reference, Robert Hanna concludes that "*either* one must be completely sceptical about the claims of natural science to know microphysical objects and properties, *or* if one still assumes that *a posteriori* knowledge of them is possible then one must paradoxically claim that our perceptual-empirical mode of access to them is essentially abductive, conceptual, or rational. But then . . . the essentialist must grant that *a posteriori* knowledge of physical microstructures is '*a posteriori*' only in the strictly Pickwickian sense – i.e., it is actually non-empirical or purely rational in character" (Hanna 1998, 511f).

In Hanna's view, H_2O and the atomic number 79, for all that they may be scientifically respectable entities, are regulative ideas, not concepts of objects or phenomena in the robustly empirical sense so important to Kant. They are, it is supposed, *material* entities at least. Wakefield, although he does not directly address the material status of natural functions and their failures, is compelled to obliquely admit their difference to canonical kinds such as gold, and compares them instead to "the modest quasi-essentialist account of artefact categories such as 'chair'; chairs have no material substrate or even physical similarities in common . . . but they are chairs roughly because they share the fact that they . . . [are] a place for someone to sit . . . . (Wakefield 2021a, 179).

It is hard to know what to make of this. Given that objects such as chairs are often used as examples of things that *lack* essences, this 'quasi-essentialist' account seems hardly to be essentialist at all (it is 'essentialist' in Hanna's Pickwickian sense, perhaps). That he uses this strange analogy illustrates just how insubstantial 'dysfunction' is as an essential, reference-fixing property. Yet this is not to say that it is not in some sense plausible, nor

[32] Virtually identical statements are to be found throughout his lectures on logic, as well as in a letter to Karl Reinhold from 1789 (Kant 1999, 299).

that the concept cannot play a useful role. What is at issue is the claim that 'dysfunction' is a factual, value-free concept that picks out an objective facet of the external world. Given his overriding concern regarding the antipsychiatric critique, as discussed in chapter 2, Wakefield is compelled to make this claim.

Viewed through a Kantian lens, I believe one has to conclude that the concept of function (and therefore dysfunction) that underlies Wakefield's concept of mental disorder is an idea of reason. If it is treated as an empirical concept then we have been enticed into error by the transcendental illusion that the subjective demand of reason is objectively significant. And failing to make the distinction, for whatever reason, has consequences. Wakefield is quite right to say "Labelling people as disordered when their distress is due to an oppressive environment is not only incorrect but potentially harmful because it suggests that something is wrong with the person and it directs interventive attention toward the person's internal functioning and away from the person-environment interaction" (Wakefield 1992a, 240) But this cuts both ways. A regulative idea does not sanction us to label people as malfunctioning. This 'concept' of dysfunction does not refer to any possible object of experience. Its presence can only be inferred. To proceed with such confidence, as Wakefield does, on the basis of this inference seems to me to be a remarkably cavalier approach to take. To take oneself, or others, to be internally defective, may have all kinds of unwelcome ramifications, as I will argue in my final chapter.

Reason's conflict with itself

In recent decades, parallels between Kant and Wittgenstein have been made by a number of commentators such as Kurt Mosser, who has written that "taking Wittgenstein's general project as dispelling illusions that an uncritical employment of thought and language generate, I think it is clear that Kant's project is similarly therapeutic" (Mosser 2009, 5). The CPR aims to establish the limits of what we can claim to know in order to prevent our being led astray by – primarily, though not exclusively – our power of reason. In particular, we cannot talk meaningfully about noumenal objects, things that cannot enter sensible experience. But Kant understood that Hume's austere scepticism was problematic in its own way: "scepticism is a resting-place for human reason . . . but it is not a dwelling-place for permanent residence . . . . (A761/B789). As such, his 'Copernican' view is that "we can cognise of things *a priori* only what we ourselves have put into them" (Bxviii). Whatever we can say about mental disorder *a priori*, i.e., independent of sensible experience, is just whatever our concept contains. Wakefield would agree to this extent: he tells us what, in his opinion, our concept contains. But this 'content' would not be acceptable to Kant. In particular, since 'dysfunction' is thought through reason, it can only be a hypothesis.

The account of reason that Kant gives in the CPR is deeply ambiguous, riven by tensions that seem to be built in to the modes by which it gives us the means to acquire knowledge. Willaschek concedes the stark conclusion that ". . . Kant conceives of human reason as being in a tragic position: its very nature makes it ask metaphysical questions about the unconditioned, but the limitation of human cognition (its dependence on sensible intuition) makes answering them impossible" (Willaschek 2018, 270). Sentiments of this sort are to be found throughout the scholarly literature. As his description of transcendental illusion makes clear, there is no foolproof means by which we can avoid being drawn into error, given the 'naturalness' of its appearance. Even if we cultivate the 'discipline' that he prescribes in the second major division of the CPR, the Transcendental Doctrine of Method, there is no question of our ever becoming immunised. As Neiman observes, we cannot help but have a "sense of uneasiness about the notion of a regulative principle. The notion remains elusive, perpetually threatening to become empty or absurd" (Neiman 1997, 203). To make secure use of reason, to employ it only regulatively, is "a balancing act of major proportions" (Neiman 1997, 188). In short, we are presented with an account of our cognitive faculties that shows them to be fundamentally, and inescapably, divided between competing demands that can never be brought into lasting equilibrium. More unsettling still is the way Kant characterises reason as bound up with a form of illusion. To be sure, he accounts for it in terms of a positive as well as a negative effect, but even the manner in which he does this must leave us with a nagging sense of disquiet. It is, of course, our desire for certainty that leads us to trample over our sensible boundaries, and this because sensible experience turns out to yield so very little of it. Telling 'just so' stories is not a remedy, but neither is a retreat to scepticism.

Alan Montefiore talks about two Kants, as "*both* as a philosopher of reason *and* as one of rational self-suspicion" (Montefiore 2000, 95). He talks, too, of "a Reason which was thus itself the source of inescapable paradox and of the self-frustration contained in the knowledge that ultimately the only attainable certainty was to be found in the knowledge of its own ineluctable limits" (Ibid.). The role that paradox plays in Kant's *Critique*, and the fact that Kant, to a considerable extent, accepts these paradoxes, is a novel feature of his philosophy. The metaphysical errors Kant probes in the Dialectic are significant ones indeed – the history of philosophy prior to Kant is full of efforts to grasp the ineffable through the powers of reason, and the decline of this sort of philosophising must be attributed in part to his own efforts. As I discussed in chapter 3, he perceived an alarming parallel between these projects and the manifestations of insanity. It would be too strong a suggestion to say that the very nature of reason, as Kant describes it, explains the possibility of madness. This was not a claim that he made in his own work on the subject, certainly. What I do suggest is that it gives us a model for understanding how it is possible to conceive mental disorder as something other than malfunction, in Wakefield's sense. Robert Fogelin (whose views in this regard are remarkably similar to Montefiore's) notes that "Both Hume and Kant went beyond the claim that reason is simply too weak to give us complete knowledge of the universe we inhabit . . . Both, each in his own way, made the deeper and more disturbing claim that reason, in its purest form, generates illusions that ultimately thwart reason's endeavours" (Fogelin 2003, 70f).

This is a key point. We may well think that reason, however we conceive it, is limited by our finitude. The more remarkable view of Kant (and in his own way Hume) is that reason's activity involves, and must involve, an element of internal conflict or tension. One might want to say, as I think Wakefield would, that reason has been shaped by natural selection to operate within certain parameters, and that 'unreason' gives us grounds to infer that this capacity or 'mechanism' is failing to work as it is meant to, in this purportedly naturally normative sense. We could say, then, that this tension may be present, but that mental disorder is nonetheless a failure of rational function, a dysfunction. But we could also conjecture that reason is by its very nature a capacity to exceed the limitations of sensible cognition, and that there can be no parameters for *this*. To put it another way, reason cannot police itself. That the categories do fix certain boundaries upon what can possibly be given to us in experience could not be a guide to the proper use of reason, since reason is precisely the capacity to go beyond those boundaries. What we would like to know is, *how far* – and that is what Kant, to some extent, tries to tell us.

In chapter 1 I discussed the concept of function and the role it plays in the life sciences. We saw there that function is strongly associated with with teleology, the explanation of phenomena in terms of ends or goals. In the CPR, Kant mentions teleology rather briefly, close to the end of the Appendix, where he returns to the rational idea of God, one of the principle ideas introduced earlier in the Dialectic. Although the 'destructive' part of the Dialectic undermines the metaphysics traditionally associated with these ideas, Kant rehabilitates them somewhat in the Appendix; here the idea of God is effectively identified with the demand for systematic unity, though this be regulative only, and supports no claims about the actual existence of a supreme intelligence. As he writes, "The highest formal unity that alone rests on concepts of reason is the purposive unity of things; and the speculative interest of reason makes it necessary to regard every ordinance in the world as if it had sprouted from the intention of a highest reason. Such a principle, namely, opens up for our reason, as applied to the field of experience, entirely new prospects for connecting up things in the world in accordance with teleological law . . . ." (A686/B714f). In keeping with the theme of the Appendix, the proper role of reason's supreme principle, Kant's point here is a very general one about nature, or external reality, as such. In the third critique, the *Critique of Judgement* (1790; henceforth CoJ), Kant returns to the subject of teleology, and does so in the more specific context of living organisms and their place in the cosmos. Here, the question of purpose (*Zweck*), which approximates to the term 'function' more commonly used in the biological context today, is explored in detail.

In this thesis I have chosen to limit myself, for the most part, to using the resources of the first *Critique* – a task that, in any case, would need to be completed before the themes of the *Critique of Judgement* could be tackled in full. To do justice to the position developed therein, its relation to the earlier work, and the teleological implications of biological function for Wakefield's analysis would have required a chapter in its own right, and would not, I think, have substantially benefited my task. My intention in this chapter was to explore the potential of the CPR not only to challenge the HDA as a definitional account of mental disorder, but also to touch upon Kant's unusual take on the nature of reason. Nonetheless, I will briefly consider the implications of Kant's discussion of teleology and purposiveness for my argument in this chapter.

For Kant, what makes something purposive is that it can only be understood on the assumption that it is the result of intentional design. In his words, "an object or a state of mind or even an action . . . is called purposive merely because its possibility can only be explained and conceived by us in so far as we assume at its ground a causality in accordance with purposes . . . .” (Kant 2000, 105) As we have seen in our earlier discussion of function, man-made artefacts are paradigmatic examples of purposive objects, uncontroversial because they actually are products of intentional design. Complications are encountered when we come to attribute functions to things for which no intention is apparent or can be admitted – a situation that we encounter when we discuss living organisms and their parts in the biological sciences. In the section of the third *Critique* called the Critique of the Teleological Power of Judgement, Kant’s primary interest is what prompts us to attribute purposes to some things and not others. The answer to this question also explains how we come to distinguish living organisms from inert matter – a cognitive capacity that, despite its apparent naturalness, is not easily accounted for.

Kant argues that living organisms are ‘natural ends’ or purposes, and that “a thing exists as a natural end **if it is both cause and effect of itself**” (Kant 2000, 243), which he illustrates with the example of a tree. Leaves are a product of the tree, and dependent upon it for their continued existence, but at the same time they contribute to the maintenance of the other parts of the tree, such as the branches and trunk. The purpose or end, in the case of organisms, is thus their self-maintaining character, something that we do not find in man-made objects. Kant believes, moreover, that it is in this way that cognition of organisms is possible for us.

His argument for this is that forces acting upon matter underdetermine the features we observe in organisms – mere mechanism, in other words, could have produced any number of alternatives, leaving it accidental that just these features are what have come about (see his example of the structure of a bird (Kant 2000, 233)). In this case they would be objects of experience still, but we would cognise them as contingent and, as such, beyond explanation by appeal to efficient-causal laws. We would not cognise them as *organisms*. We do in fact take these objects to be determined in some way, and to admit of explanation, and this is only possible, Kant says, if we contribute the notion of purpose to them ourselves. By so doing, we can make sense of their forms as goal-directed. This subjective contribution therefore enables us to cognise organisms as a distinct class of intelligible entities rather than perplexingly contingent complexes about which there is nothing further to be said. Nonetheless, although the possibility of their cognition gives us grounds for their empirical investigation in terms of standard causation, it does so only by projecting the idea of purposiveness onto them. Our inquiries can never explain whether or how nature itself could make such purposiveness possible: Kant, as John Zammito has forcefully argued, is no naturalist for biological function (Zammito 2006).

Cognition of organisms is achieved, Kant argues, through the reflecting power of judgement. Whereas determinative judgements, or judgements of understanding, bring particulars under empirical concepts, reflective judgements are required where no such concept is available, and one must be acquired. In keeping with the doctrine of the CPR, Kant maintains that purpose as such is an idea of reason.[33] Rather than make experience possible, it guides our actions. The concept of a *natural* purpose is generated by the reflecting power of judgement, "by analogy with the capacity of reason to set itself ends and to direct its activity towards these ends." (Breitenbach 2009, 44) In other words, it is by reflecting upon a rational principle that we arrive at the analogical 'concept' of natural purposes, though it is not a category nor an empirical concept strictly speaking. This distinguishes Kant's view from the design analogy often used to characterise teleological thinking: in his words, "One says far too little about nature and its capacity in organised products if one calls this an analogue of art . . . ." (2000, 246)[34] Here, Kant has recognised a defect with the analogy with artefacts: they are externally caused, whereas a peculiarity of organisms is that they are *self* caused, or "cause and effect of themselves".

What are the implications of this for my critique of the dysfunction analysis? Since I have concentrated on the first *Critique*, I have argued that function, given the doctrine of that work, would have to be an idea of reason. In light of the CoJ this requires qualification. Function, as it appears in the debates about the naturalised reference of the term in the life sciences (which is essentially about what constitutes a component-function, and distinguishes such from mere effects), can still be considered, for Kant, a theoretical idea of reason. Such ideas are invoked for explanatory purposes, not to make experience of objects possible. Conversely, in the third *Critique* Kant is concerned with how we come to cognise organisms as such, rather than how we explain them in terms of

[33] Confusingly he calls it a *concept* of reason in the third Critique (Kant 2000, e.g. 234, 244, 267); I agree with Joan Steigerwald (2006, 717) that he means ideas in the sense of the CPR.

[34] By 'art' Kant is, of course, referring to design rather than fine arts such as painting and sculpture.

their parts (though the latter presupposes the former). Cognition is normally a product of understanding and concepts, but in the unusual case of living things (and of works of art, the other principle subject of the CoJ) no empirical properties suffice to ground our actual way of experiencing them. Hence, he assigns this unusual capacity to the reflecting activity of judgement – and even then, judgement makes analogical use of our own capacity to set ourselves ends, i.e., a rational principle. In short, the emphasis of the CoJ, in relation to function, is somewhat different, though related, to that of the analysis of function I have here critiqued. It is, all the same, in keeping with Kant's treatment of ideas in the CPR. Above all, he maintains that "The concept of a thing as in itself a natural end is not a constitutive concept . . . but it can still be a *regulative* concept for the reflecting power of judgement . . . ." (2000, 247; emphasis mine), employing the same contrast that in the CPR distinguishes concepts of the understanding from ideas of reason.

Breitenbach suggests that, given Kant's position, we may resolve the apparently interminable clash of the causal role and etiological conceptions of function: "If, then, we understand the use of teleological language in the biological sciences as a heuristic means based on analogy, we do not need to decide between the aetiological and the systems theoretic [causal role] approach. Both analyses can be accepted . . . As long as biologists are interested in the investigation of the selection history of a particular organic trait as well as in the causal role that the trait plays in a complex organic system . . . ." (Breitenbach 2009, 46)

This seems acceptable in the biological context of trying to understand living organisms, but will not be found entirely satisfactory in the context of psychiatric research aimed at discovering dysfunctions. Selection history and causal role are not in themselves of particular interest in this context, since what is at stake is not the acquisition of empirical data so much as establishing natural, non-evaluative norms by which dysfunction, and thus disorder, can be judged. The conceptual argument will doubtless continue to rage between those for whom this specific problem is paramount. On the other hand, there is no reason why individual researchers should not follow their intuitions and use their preferred conceptual schema methodologically. It is, after all, only by conducting empirical investigation that either of these analyses can be tested. As I have argued, dys/function for Kant is not an empirical property – not, in other words, something that can be found in our experience of nature. This is not to say that research into mental disorder, whether framed in terms of Wakefield's or Boorse's analysis, or any other, will not be productive and potentially important in various ways. It is only to say that it will not confirm these analyses.

The notions of function and dysfunction are by no means indispensable for the concept of mental disorder itself. The term 'disorder' is – perhaps usefully – ambiguous in meaning. A lack of order might refer to a fact or to a value-judgement (i.e. things out of sequence just are disordered, but whether an arbitrary arrangement of objects betrays a lack of order is a matter of opinion). It is certainly wrong to assume that the problem of disorder must be identified with the problem of functions and their deficits. For example, as we saw in chapter 2, some thinkers among the antipsychiatrists understood mental disorder to be, in various ways, a more or less rational response to an irrational situation – neatly captured by Gregory Bateson's theory of the 'double bind' in relation to schizophrenia. Whatever one might think of such alternatives, they cannot be as easily dismissed as some of their critics suggest. Indeed, new ways to refine the concept are still being developed, among them some that are not so very distant from these earlier accounts. This should not be surprising given just how complex the human mind is and how inadequate our attempts to understand it have been. Here we might think of the work of Denny Borsboom on the 'network theory' of mental disorders (e.g. Borsboom 2017), which suggests that different causal factors, including environmental ones, may independently cause symptoms of various types, which only become 'disorders' when self-sustaining relationships lead to them forming stable 'networks'. Or we might refer to some of the suggestions that have emerged from evolutionary psychiatry (such as those collected in Adriaens and De Block 2011) which (despite being informed by evolutionary theory, favoured by Wakefield as grounding the purported objectivity of his analysis) indicate that disorders can be better understood as variations, rather than failures, of certain fitness-enhancing psychological traits – i.e., not dysfunctions. I am not here advocating any particular alternative, but highlighting that there are many more avenues to explore, and that unless and until compelling empirical evidence convinces us otherwise, we should be wary of narrowing our theoretical options by adopting any one definition of mental disorder.

Conclusion

My task in this chapter was initially to clarify some aspects of Jerome Wakefield's harmful dysfunction analysis of mental disorder – in particular, to get clear on his 'black box essentialism', which borrows primarily from Hilary Putnam's semantic externalism. I was interested, in particular, to understand Wakefield's motive for this development, given that Putnam himself was trying to account for linguistic reference. This expanded upon the discussion of the HDA in chapters 1 and 2. Subsequently, I considered the HDA and its essentialist extension via Kant's view of concepts and their role in cognition. I argued that for Kant, the definition of mental disorder as 'mental dysfunction' (supplemented by an evaluative 'harm' judgement) would not meet the conditions required for an empirical, or objectively valid, concept.

Given Kant's argument in the Transcendental Dialectic of the CPR, I concluded that the concept would be considered a concept or 'idea' of reason that, as 'regulative', should only serve as a heuristic to guide investigation. It does not licence us to say that mental disorder as such essentially refers to a failure or defect of a biological or mental mechanism. I ended by discussing how Kant, in the Dialectic of Pure Reason, shows that our highest cognitive achievements are nonetheless founded on a basic instability or equivocation. Our rationality itself is driven by the illusion that what is only a subjective principle for the connection of our concepts is an objective principle that, in itself, delivers knowledge about a reality that exceeds our sensible experience.

I will now turn, in my concluding chapter, to another example of a fundamental tension in Kant's critical philosophy, and one that bears very directly upon mental disorder – a tension between between the self as knower, and self as known.

Chapter 5

Mental Disorder and the Paradox of the Kantian Self

Introduction

In the *Anthropology*, Kant opines that "The fact that the human being can have the 'I' in his representations raises him infinitely above all other living beings on earth." (Kant 1978, 15) It is not immediately clear in that work why this cognitive accomplishment should be the pre-eminent one, but for anyone who has read the *Critique of Pure Reason* the answer will be familiar: for Kant, the self – in a rather specific sense – is the most basic condition for the possibility of experience and, ultimately, for our highest achievements in the arts and sciences.

The self has a storied history in philosophy, but we shall see that on this topic, as on so many others, Kant's contribution has had a profound influence, and one that is still being debated today. This is particularly relevant because "mental disorder and selves form two sides of the same coin: an analysis of mental disorder is inevitably also an analysis of self" (Dings and de Bruin 2022, 274) In this chapter I will not be concerned so much to critique Wakefield's harmful dysfunction analysis in terms of his own arguments in its favour, as to think about its effects on our attitudes towards mental disorder. I will ask why we might want to resist such an analysis, and explore one alternative approach, that of Louis Sass. Sass has paid special attention to the phenomenon that he calls 'self-disturbance' in schizophrenia, which he has characterised through Kant's distinction between the introspectable empirical self and the self as the pure, knowing subject of 'transcendental apperception'.

I begin by elucidating Kant's view of the self in the *Critique of Pure Reason*, where it emerges as the most fundamental condition for the possibility of experience. Following this, I look at Michel Foucault's *The Order of Things*, which inspires Sass's own application of Kant's work. In this book, Foucault discusses the influence of Kant's dual-aspect self, or what he terms the 'empirico-transcendental doublet', on modern structures of knowing. I will then turn to Sass, with special emphasis on his books *Madness and Modernism* and *The Paradoxes of Delusion*. After laying out Sass's phenomenology of schizophrenia, I will argue for a slightly modified reading of Kant that takes his transcendental philosophy more seriously. Finally, I will argue that interpreting self-disturbance in this revised way allows us to understand schizophrenia – perhaps the most mysterious of mental disorders – in a way that renders it less alien than categorical approaches to mental disorder such as Wakefield's harmful dysfunction analysis. We can, as Karl Jaspers' suggested of the schizophrenic, "bring them closer to ourselves" (Jaspers 1963, 576). I will propose that it is important to maintain that mental disorder – even one as strange as schizophrenia appears – is on a continuum with 'normal' human experience. In particular, this is one way to lessen the burden of stigma, a problem that many decades of public 'psychoeducation' has done little to resolve.

Kant's Deduction and the role of apperception

In the Transcendental Deduction of the *Critique*, Kant develops a response to Hume's famous sceptical attack on the notion of a simple, persisting 'soul', or self.[35] Hume's so-called 'bundle' image of the self portrays it as a rapid succession of changing mental states in constant flux, within which nothing constant and unchanging can be discerned. As he famously put it in his *A Treatise of Human Nature*, "There are some philosophers, who imagine that we are every moment intimately conscious of what we call our Self; that we feel its existence and its continuance in existence; and are certain . . . both of its perfect identity and simplicity" (Hume 1984, 299). Although philosophers today might be more circumspect about this conception of the self, it is a commonplace, possibly universal, folk intuition. Arguably, it is more deeply entrenched in our own time than in any previous era, and it underwrites the individualism that informs almost every aspect of cultural, economic, and political life in the developed world. Hume himself, however, finds no evidence of such a self really existing: "If any impression gives rise to the idea of self, that impression must continue invariably the same, thro' the whole course of our lives, since self is suppos'd to exist after that manner. But there is no impression constant and invariable" (Hume 1984, 300).

To this extent, Kant agrees with Hume, observing that "The consciousness of oneself in accordance with the determinations of our state in internal perception is . . . forever variable; it can provide no standing or abiding self . . . ." (A107). This "consciousness of oneself" Kant terms 'empirical apperception'; like Hume, he can discern no representational content that remains the same throughout the changing states in inner

[35] Kant does not present it in this way: he rarely names the philosophers, including Leibniz and Berkeley, whose views he appears to be responding to. Nonetheless, it is common in the scholarly literature to frame his discussion of the self in terms of Hume. Patricia Kitcher makes a convincing case that Kant *did* have Hume specifically in mind in his discussion of the self (Kitcher 1982, 43).

sense. And yet, besides this, Kant insists that, *contra* Hume, there is a kind of 'self' that persists and remains stable throughout experience. To understand what Kant is getting at here, however, we need to look at how he thinks our disparate representations can become a uniform experience.

In chapter 4 we have already seen how, for Kant, the faculties of sensibility and understanding work in tandem to make experience of the external world possible. Through sensibility we are affected by objects, but it is only through concepts that our sensible intuitions can become determinate experience of those objects. The *a priori* 'pure concepts of the understanding', or categories, provide the fundamental form that any possible experience can take, while our empirical concepts are acquired *a posteriori*, and are the vehicles through which we cognise things as being particular kinds of objects. In the Transcendental Deduction, Kant undertakes to justify our use of the categories, and it is here that he first discusses the self.

We are given, through the faculty of sensibility, a series of representations (essentially, particular determinations of our mental states) ordered in time. In order for these representations to become experience as we know it, however, they must be *connected* somehow. As Kant puts it, "If every individual representation were entirely foreign to the other, as it were isolated and separated from it, then there would never arise anything like cognition, which is a whole of compared and connected representations" (A97). In a mere succession of changing mental states, every representation would be completely novel, without any relation to its fellows. In order for our experience to have the character it does, Kant argues that they must be 'synthesised' to link them together. Although this is a conceptual activity, this unity among our representations also requires that they be related to a single consciousness. As he puts it, "no cognitions can occur in us, no connection and unity among them, without that unity of consciousness that precedes all data of the intuitions, and in relation to which all representations of objects is alone possible. This pure, original, unchanging consciousness I will now name **transcendental apperception**" (A107) This, then, is Kant's response to Hume's sceptical position. We can say that there is indeed a stable, abiding self, because such a consciousness is a requirement – in fact, *the* fundamental requirement – for the possibility of experience: as he goes on to say in the B-edition, this "principle is the supreme one in the whole of human cognition" (B135).

Furthermore, this unity of consciousness is only possible if the mind is *aware* of the way in which it synthesises representations (the manifold). Experience, in other words, rests upon a kind of self-awareness. This point is elucidated, for example, by Wilkerson (Wilkerson 1976, 50), who notes that we can have physical states and remain unaware of them, whereas to have experience is precisely to be aware that it is *my* experience; the experience of a single subject.[36] This is not to say that all my experiences must involve explicit ascription to a pure, unchanging consciousness – this is manifestly not the case. Rather, it must be possible to so ascribe them.[37] As Kant expresses it in the B-edition, "The **I think** must **be able** to accompany all my representations" (B131). Dennis Schulting reminds us that something like this is already to be found in the work of Leibniz and in Christian Wolff, as a "second-order accompanying of one's first-order perceptions of objects" (Schulting 2022, 3). Kant, however, places this kind of self-consciousness at the very centre of his account of how experience is possible.

The doctrine of transcendental apperception has perplexed generations of Kantians. Above all, many of his readers have interpreted it as a metaphysical entity independent of empirical apperception, yielding the conclusion that there are *two* selves. This reading has been disputed by more recent commentators, who provide much textual support for the idea that Kant is talking about two distinct *descriptions* of a single self. In the first place we must remember that the *Critique* was motivated by Kant's antipathy towards rationalist metaphysics. He maintained some faith that a scientific metaphysics was possible, but the 'I think' is not part of it. To understand this, we must attend to his cautionary pronouncements on the matter.

Particularly in its formulation as the 'I think', transcendental apperception could be confused for the Cartesian cogito. In the first edition version of the Deduction, Kant confines himself to discussing the role of transcendental apperception as a condition of the possibility of experience. In the second edition version, Kant takes pains to emphasise a significant point of difference: ". . . I am conscious of myself not **as** I appear to myself, nor as I am in myself, but only *that* I am. This **representation** is a **thinking**,

[36] Although this sense of ownership may be attenuated in some forms of schizophrenia, nonetheless it is only if I have an experience that there can be any confusion or doubt about whether it is 'my' experience. Dan Zahavi, for example, makes precisely this point (Zahavi 2005, 144).

[37] Henry Allison emphasises this, noting also that "Kant is perfectly willing to countenance the possibility of representations (mental contents or states) *in me* that are nothing *to me*, cognitively speaking" (Allison 2004, 164).

not an intuiting" (B157). It follows that "The consciousness of oneself is therefore far from being a cognition of oneself . . . ." (B158). The 'I think' is not an object of knowledge distinct from empirical apperception, and is certainly not a thinking substance, as Descartes infers it to be.

There is, within Kantian scholarship, a broad split between so-called 'one world' and 'two world' interpretations of his distinction between phenomenal and noumenal aspects of reality – between objects as we experience them and as they may be in themselves, independent of the cognitive constraints under which we operate. On the two-world interpretation, the transcendental self is seen to be a noumenal entity, while the empirical self is taken to be phenomenal. Conversely, on the one-world interpretation they are understood to be different perspectives on a single subject.[38] The one-world reading strikes me as the more persuasive, although I do not have room here to justify this conclusion; at any rate, I concur with David Carr when he says that "we can hardly speak of two subjects. We may have two *self-descriptions*, but only one of them [i.e., empirical apperception] can acquire ontological status . . . ." (Carr 1999, 57; emphasis mine). As Carr also points out, however, there is a seeming incompatibility between them. Under one description, I am an object, one that I can cognise and know (albeit constantly changing). Under the other, I am a subject, a 'knower' that is uncognisable, without qualities. Or, as Edmund Husserl expressed it in *The Crisis of European Sciences*, "The paradox of human subjectivity: being a subject for the world and at the same time being an object in the world" (Husserl 1970, 178).

In summary, Kant's view is that there is a single, unitary self that can be given two radically different descriptions, and that one of those descriptions gives us the supreme condition under which we can have any experience at all. He also tells us that we have, and must be able to have, a basic awareness or consciousness of that pure or transcendental self. There is no manifold in this consciousness – it gives us no awareness of properties – but without it the data given to us through the receptive faculty of sensibility would lack the very quality of unity characteristic of experience. Nonetheless, it offers an explanation as to why we have such a persuasive sense of an abiding self even when, as Hume observed, there is nothing in our representations that seems to justify

[38] Michael Oberst credits Karl Ameriks, Richard Aquila, and Béatrice Longueness as defenders of a two-world interpretation, Gerold Prauss, Henry Allison and Lucy Allais as advocates of a one-world reading (Oberst 2015, 53 n1 n2).

such a feeling. On this basis Hume felt justified in denying the self as such, but could not explain the source of the confusion. His error, on Kant's view, is to be looking in inner sense for a representable *object*. With the transcendental self, however, we find that it is not, as we might expect, any kind of representation that furnishes us with the feeling of an abiding self, but a purely formal self-aspect that gives our representations the unity required for them to become experience. Furthermore, Hume is unable to account for the unity of experience. As Kant points out, each representation must be connected in several different ways in order to comprise experience as we know it. Without such a connection, or synthesis, each of our momentary representations would be entirely isolated from every other.

On the face of it, the notion is easy enough to grasp: it is intuitively obvious that experiences must be referred to an experiencer. The problem is that we will inevitably want to know what the experiencer is – what it is like. And this is exactly why Hume (and not only Hume) sought an object. However, Kant's doctrine of transcendental apperception is highly abstract, and certainly does not give us the satisfaction of a continuous self about which any kind of knowledge can be had. The uncomfortable equivocation between the transcendental and empirical selves is also to be observed in the distinction between constitutive and regulative principles discussed in chapter 4.

Kant was not primarily interested in the problem of personal identity *qua* self-knowledge, in its modern form. Apperception enters the picture only because of the cognitive requirement that representations be given unity. But this is precisely what makes the Kantian view compelling. A 'self' of a fixed, persisting sort is a formal requirement of experience and, though it is not an object of experience, it is all too easily misconstrued as such.

Foucault and the 'empirico-transcendental doublet'

In his book *The Order of Things*, published in 1966, Michel Foucault expresses the view that what he calls the 'human sciences' (specifically psychology, sociology, and the history of ideas) are structured in a particular way: they treat as their *object* of study the human being as a cognising *subject*. These sciences, which first emerge in the 19th century, are therefore characterised by a tension between the two very different and, apparently, irreconcilable conceptions of the human being that Kant introduces in the first *Critique*. Gary Gutting summarises it thus: "These sciences deal with man . . . but not in the manner of the empirical sciences of biology, economics, and philology. The latter treat man as part of nature, as an empirical object, presenting his powers of representation as products of the external world. The human sciences . . . are concerned with man as a subject, as a knower whose representations constitute his world and are not just products of it" (Gutting 1989, 208).

That is, the human sciences do not study people merely as acted upon by external forces, but as beings whose interactions are mediated by their own active processes of cognition. Foucault argues that while we are liable to see this as a development of the same conceptual structures that were operational in earlier epochs, as far back as antiquity, the historical reality is different. In earlier periods, this concept of the human being as *simultaneously* knower and known does not exist, and it comes into being quite suddenly, during the latter half of the 18th century. It is only then that the human sciences can and do take shape.

This is part of his notion of the 'episteme', a term that Foucault doesn't provide a succinct definition for, but alludes to as, among other things, "the general space of knowledge" (Foucault 2005, xxv). This abstract space is not, as we might think, neutral, but is always historically determined by complex norms that determine what kinds of enquiry are conceivable, and what kinds of discoveries will pass for knowledge.[39] Furthermore, epistemes can and do change quite abruptly, as dramatically new avenues of thought unexpectedly open up. Perhaps the classic example of this is the change wrought in astronomy by the Copernican hypothesis – a revolution that Kant famously alludes to in describing his transcendental approach to philosophy (Bxiii).

In Foucault's view, this reconfiguration of selfhood is what paves the way for the human sciences, and he traces its source to Kant's Critical philosophy, which calls forth what he refers to as an 'empirico-transcendental doublet' (Foucault 2005, 347). With this, "man appears in his ambiguous position as an object of knowledge and as a subject that knows: enslaved sovereign, observed spectator . . . ." (Foucault 2005, 340) This doublet is, of course, the distinction between empirical and transcendental apperception discussed above. Prior to Kant's 'Copernican revolution', mental representation was taken for

[39] There is an obvious parallel here with Thomas Kuhn's notion of 'paradigms' that structure and constrain thought in periods of 'normal science' (Kuhn 2012). Ian Hacking points out that this idea was already somewhat familiar to French readers through the work of philosophers like Gaston Bachelard (see Hacking 1979, 45).

granted as the foundation of knowing. With Kant, a new realm of enquiry opens up – the subjective conditions that make representation itself possible.

In Foucault's view, this has led to the privileging of human subjectivity as a locus of scholarly attention, what he calls 'anthropology' (and which commentators sometimes call 'humanism' e.g. Paden 1987, although this locution is not to be found in Foucault): "Anthropology constitutes perhaps the fundamental arrangement that has governed and controlled the path of philosophical thought from Kant until our own day" (Foucault 2005, 373). Foucault's complaint is not simply – or even primarily – against Kant. Rather, it is with the post-Kantian trend towards resolving the troubling dualism that he had apparently left dangling. Foucault argues that since the emergence of man as a 'doublet', a persistent oscillation between the transcendental and the empirical is the disconcerting condition from which 'modern thought' seeks liberation through some discourse that would *reconcile* the two. He identifies this project, in particular, with phenomenology (as he makes clear in the Foreword to the English edition (Foucault 2005, xv)).

For Foucault, that project is a sterile one. As Larry Shiner puts it, "The centrepiece of Foucault's critique [of phenomenology] is his rejection of the subject as origin in favour of a body of anonymous rules governing discourse." (Shiner 1982, 312) Although Shiner identifies several arguments that Foucault makes against phenomenology, perhaps the most important is the *epistemological* argument, based on the empirico-transcendental doublet. In his view, "phenomenology's appeal to actual experience is an heroic attempt to bridge this dialectic" (Shiner 1982, 314), albeit one that must always fail. More specifically, Foucault objects to the privileging of the subject whose experience phenomenology attempts to transcribe. For Foucault "the subject itself is shaped through an external process, by its surroundings, by the material practices investing the surface of the body (surveillance, discipline, punishment, confession and so on) and producing thereby the illusion of the self as an origin and as a centre . . . ." (Legrand 2008, 282) It is this 'external process' that Foucault believes should displace 'man', and which, in various forms, he himself studied – most pertinently, in the current context, in his *History of Madness*.

Sass, schizophrenia, and the self

As I have already noted, the very concept of mental disorder implicates the concept of the 'self' – there is no discussion of the former that does not involve the latter, and it is this that makes mental disorder importantly different to physical disorder. This is perhaps nowhere more true than in the case of schizophrenia, the condition that Thomas Szasz ironically termed the 'sacred symbol' of psychiatry (Szasz 1976). The authors of one paper relate that "The notion of the self has a long and distinguished career in understanding and treating schizophrenia and related disorders. It played a central role in the classical theories of schizophrenia offered by such pioneers in the field as Emil Kraepelin (1904), Eugen Bleuler (1950), Karl Jaspers (1963), and Adolf Meyer (1950), and was also given a prominent place in earlier psychoanalytic accounts (e.g. Freud, 1910; Fromm-Reichmann, 1950 ; Schilder, 1976 ; Sullivan, 1940)" (Davidson and Strauss 1992, 132).

To this list can be added the name of psychiatrist and philosopher Louis Sass, who has written extensively about the phenomenology of 'self-disturbance' in schizophrenia, notably in two monographs, *Madness and Modernism* (1992) and *The Paradoxes of Delusion* (1994), as well as in a large volume of subsequent work. Two leading principles emerge from his work. First, that some of the most enigmatic symptoms of schizophrenia can be understood in terms of cognitive contradictions or paradoxes that appear to be coterminous with recurring themes in the art and literature of modernism, and problems in 20th century philosophy. Second, contra the dominant models of schizophrenia (medical and psychodynamic), which conceptualise it as a regressive condition, that some typically schizophrenic phenomena appear to be highly – indeed, excessively – developed forms of consciousness: "Madness, on my reading, is neither the psyche's return to its primordial condition, nor the malfunctioning of reason, nor even some inspired alternative to human reason. It is . . . generated from within rationality itself rather than by the loss of rationality" (Sass 1994a, 12).

In Sass's most comprehensive work, *Madness and Modernism*, he draws numerous parallels between themes in the art and literature of modernism and the symptoms of schizophrenia, and describes "the many paradoxes that lie coiled at the heart of the schizophrenic condition" (Sass 1998a, 324). Concluding the work he identifies, as the most enigmatic of these paradoxes, an equivocation between two apparently irreconcilable conceptions of the *self.* On the one hand "the self can come to seem pre-eminent and all-powerful: rather than drifting somewhere in space . . . one's own consciousness may seem poised at the epicenter of the universe . . . ." (Sass 1998a, 324f). On the other, "many schizophrenic patients tend to lose their sense of active and integrated intentionality. Instead of serving as a kind of anchoring centre, the self may be dispersed outward, where it fragments into parts that float among the things of the world . . . ." (Sass 1998a, 324). Consequently, a disconcerting contradiction can often be observed in which the patient seems to seesaw between conceiving the reality of their environment as somehow dependent upon their own acts of perception, and feeling that their sense of selfhood is fragmenting or dissolving.

Sass describes these two apparently contradictory phenomena as elements that in fact mutually reinforce each other. The first is what he calls *hyperreflexivity*, which he describes as "a relentless kind of introspection" (Sass 1998a, 228). Although it is not a point he makes clear in his earlier work, Sass has more recently stressed that hyperreflexivity "is not, at its core, an intellectual, volitional, or 'reflective' kind of self-consciousness" (Sass and Parnas 2007, 69). It is not a conscious activity of introspection but rather "a tendency . . . for focal attention to be directed toward processes and phenomena that would normally be 'inhabited' or experienced (tacitly) as part of oneself . . . . (Sass 2013, 121) Although this talk of "processes and phenomena" is rather vague, in his earlier work he explicitly links hyperreflexity with Kant's transcendental 'I', the pure subject that gives our representations their unity and therefore has a basic constituting role for experience. As we have seen, Kant associates experience as such with a non-representational 'awareness' of this subject, which is to say that it is not and can never be consciousness of an object. Nonetheless, in self-disturbances of the sort that Sass describes, this self-awareness impinges upon experience, leading to the grandiosity of the schizophrenic who now senses their subjective role in the constitution of reality. This is something we see in some of the writings of Daniel Paul Schreber (1842-1911), the German jurist whose *Memoirs of My Nervous Illness* made him the most celebrated schizophrenic in psychiatric history: "Since God entered into nerve-contact with me exclusively, I became in a way for God the only human being, or simply the human being around whom everything turns, to whom everything that happens must be related and who therefore, from his own point of view, must also relate all things to himself" (Schreber 2000, 233).

The second of Sass's complementary aspects is *diminished self-affection*, "a decline in the (passively or automatically) experienced sense of existing as a living and unified subject of awareness" (Sass and Parnas 2007, 68). Delusions of being controlled from without (the *locus classicus* of which is probably Victor Tausk's report of 'Natalija A.' and her 'influencing machine (Tausk 1992)) or of one's thoughts belonging to another and merely being 'projected' into one's own consciousness are held up as examples of this phenomenon. Sass discusses this in terms that explicate its relationship to hyperreflexity: "nowhere within experience is the self-as-subject, the supposed owner of experience, to be found. It seems, then, that there is no evidence within experience for asserting the important role of one's own self, understood as a unique ego, in grounding the world" (Sass 1994a, 68). In other words, hyperreflexivity (which Sass here likens to the philosophical doctrine of solipsism) undermines itself if one looks for it to be confirmed in experience. Confronted with this failure to affirm the sovereign subject-for-whom, the pendulum swings and the patient is now exposed to profound doubts about the origins of their mental representations, the reality of external objects – even to doubts about their own existence.

In *Madness and Modernism* Sass identifies these same phenomena occurring in modernist arts and letters, primarily in the works of such well-known figures as Franz Kafka, Giorgio de Chirico, Antonin Artaud, Samuel Beckett, and Marcel Duchamp.[40] The paradoxes of the reflexive, he argues throughout the book, do not only appear in the form of exotic psychological aberrations, but pervade Western culture of the past century and more. Sass conducts, for example, a close reading of Kafka's short story "Description of a Struggle", which he closes with the comment that "In Kafka's story, we have the entire progression of a schizophrenic illness: from schizoid self-consciousness and hyperscrutiny through self-alienation and solipsism, and on to the dissolution of both self and world" (Sass 1998a, 323).

Sass's task is to explore connections and to thereby understand schizophrenia in a different way; it is not, as he stresses, to advance causal theories. Still less is it to *define* schizophrenia, or mental disorder as such. His project is different to that of Wakefield and others interested in the concept of mental disorder. Working as he does within the

[40] As I have mentioned in chapter 2, R. D. Laing, in *The Divided Self*, also uses art and literature as devices through which to elucidate the phenomenology of schizophrenia. The first few pages of its chapter 3, 'Ontological Insecurity', in fact read like a brief rehearsal for *Madness and Modernism.*

phenomenological tradition, Sass wants to achieve some kind of understanding of schizophrenic experience, and to note the ways in which that experience overlaps with that of life and culture under the conditions of modernity. Ultimately, his "main goal is simply to reinterpret schizophrenia . . . ; to show, using the affinities with modernism, that much of what has been passed off as primitive or deteriorated is far more complex and interesting – and self-aware – than is usually acknowledged" (Sass 1998a, 9).

In the shorter but more philosophically-inclined *The Paradoxes of Delusion* he discusses the Schreber case in the light of the work of Ludwig Wittgenstein. The pairing is apt for, as Sass explains, "Wittgenstein is famous for likening much of traditional philosophy, with its irrepressible urge towards metaphysical speculation, to a kind of disease – even to a mental illness in need of therapy" (Sass 1994a, 8). As in *Madness and Modernism*, Sass discerns parallels between schizophrenic experience and a form of activity – in this case philosophy – that does not, at first sight, seem to have any relation. Wittgenstein's own awareness of these similarities informs his own idiosyncratic approach to philosophising, a method geared towards revealing that certain intellectual conundrums, rather than being deep problems calling out for resolution, are simply confusions caused by language. Here, too, the phenomenon of self-disturbance is prominent, framed in terms of Wittgenstein's discussion of solipsism and Schreber's claims that "everything that happens is in reference to me" (Schreber 2000, 233). This theme has become a mainstay of Sass's work throughout the intervening decades. In general, what motivates him is the complexity, even the *sophistication*, of the thought processes apparently at work in some of the symptoms of schizophrenia. Are they – can they be – merely products of broken or defective physical or mental 'mechanisms'? Is it not significant that strikingly similar patterns of thinking can be seen at work in human activities that, however odd they may strike us as being, we do not literally take to be the activities of the irrational? Sass of course acknowledges that "the schizophrenic condition can lack orderliness and intelligibility" (Sass 1998a, 8) – but even this, he surmises, may reflect the overworking of rationality rather than its diminishment.

Although Sass works in the phenomenological tradition, he adopts Foucault's view that Kant's philosophy makes possible a new kind of self-awareness. In Foucault's eyes, what looks like a new field of study, 'man', the knowing, world-constituting subject, appears to open up. From the side of philosophy, German idealism and, later, Husserlian phenomenology attempt to investigate the transcendental subject. On the empirical side the new human sciences such as psychology and sociology emerge. As we have seen, Foucault considers these projects to be compromised from the start.

Influenced by Foucault's *The Order of Things*, his suggestion is that the modern episteme, configuring the ways in which human beings think about themselves and the world, has bestowed a distinctive and somewhat disturbing, contradictory character upon our cultural output – and, quite possibly, upon our ways of being mad. As Foucault suggested, the most prominent contributor to this episteme is taken to be Kant, who inaugurates the distinction between the transcendental and empirical selves in the *Critique of Pure Reason*.

Sass and Kant

Sass observes that "Foucault, like many others, singles out as the origin and prime exemplar of modern thought the philosophy of Kant and the neo-Kantians . . . and as the true source of the modernist project of self-reflection" (Sass 1998a, 327). In both his monographs, and in subsequent essays, Sass draws upon this idea that the modern episteme is essentially Kantian in the sense that we have inherited from him a distinctive, and paradoxical, manner of conceiving the self. Foucault himself is concerned, in *The Order of Things*, with specific, and specialised, types of scholarly enquiry that have arisen in the modern period. As Sass writes, "he gives little indication of the implications that these post-Kantian forms of reflection might have at the more immediate level of individual human experience" (Sass 1998a, 330). It is precisely this broader project, of course, that Sass himself is embarked upon.

Sass identifies the contradictory aspects of schizophrenic self-disturbance with Kant's transcendental philosophy and the Copernican turn whereby our own cognitive apparatus contributes to the way we experience the objective world. In his words, ". . . Kant seemed to have shown that it was the world that had to arrange itself in accordance with the conditions of human consciousness rather than the reverse" (Sass 1998a, 328). This corresponds to Sass's hyperreflexity and the privileging of the subjective self as the ground of experience. However, there is also an opposing current in post-Kantian thought: "In addition to being felt as the ultimate subjective centre, the constitutor of the All (or, at least, of all that *we* can know), consciousness beginning with Kant also became a prime object of study . . . ." (Ibid.) In other words, consciousness – including the transcendental *subject*, the ground of conscious experience – becomes a 'thing' that we believe we can in some sense acquire knowledge about. This corresponds to diminished self-affection, since the self is now conceived as something separate, which inevitably raises the question, "Separate from what?" It is in the face of this question that, in schizophrenic experience, the self recedes or appears disperse.

The suggestion running through Sass's book is that the uncanny similarities between modernist culture and mental disorder indicate the extent to which the current episteme has filtered deep into apparently disparate areas of human life in the last two centuries and more. This episteme is dominated by the reflexivity of the relationship between the empirical and the transcendental. For Foucault, this has led to the privileging of human subjectivity as a locus of scholarly attention. Sass suggests something bolder and more disturbing: that the ways in which modern man can make art or be mad have been (re)structured by a landmark shift in thought strongly associated with Kantian thought. It is a suggestion that I agree is compelling. However, Sass only considers these phenomena as an untoward *effect* of Kant's thought[41] – he does not consider the potential of his philosophy as a diagnosis of what it is to be human. What I propose here is that we do just this.

It must be said that Sass's presentation of Kant's critical philosophy is at times hyperbolic. For example, he holds Kant responsible for a "radicalisation" of Cartesianism, arguing that he "draws an absolute distinction between the realm of all possible human experience . . . and that of actual existence or being (the 'noumenal'), thus implying an unbridgeable gap that sunders us eternally from the real . . . ." (Sass 1998a, 92). The point he is making here is that Kant's ideas distance, even alienate, us from the everyday world of external things by reducing them to shadows of a complete, 'really real' domain of objects – a theoretical counterpart to the alienation that he observes in cases of self-disturbance. It is true that many commentators have (less melodramatically, perhaps) interpreted Kant in this way, but there are others, such as Henry Allison, who reject the "erroneous, albeit widely held" view that for Kant "things as they are in themselves are equated with things as they 'really are', whereas things as they appear are things as they are for us . . . ." (Allison 2006, 12).[42] Alternative views such as those espoused by Allison hold that noumena are simply objects considered

[41] Of course, it is not Sass's contention that Kant has single-handedly created the conditions for these ways of doing and being. He takes account, for example, of the rapid social and technological changes wrought by modernity, and his approach overall is speculative rather than assertive.

[42] Allison himself refers to Kant's transcendental idealism as 'therapeutic' in this paper.

“independently of [their] epistemic relation to human sensibility and its conditions” (Allison 2004, 52). Kant’s noumenal realm is not privileged as that of ‘real’ or ‘actual’ existence. Appearances – our cognitions of objects – and things-in-themselves – objects as they may be independent of our acts of cognition – are not different metaphysical entities, but different ways of looking at the same object *qua* knowledge.

Kant identifies a number of rather disturbing paradoxes and tensions in the very structure of consciousness. In chapter 4 I discussed one such tension underlying the faculty of reason, that between constitutive and regulative principles. In this chapter we have encountered the ‘paradox of subjectivity’ that turns out to be fundamental to experience itself. For Kant, these basic tensions or instabilities are necessary features of consciousness. If we take his philosophy seriously we must conclude that these conditions have always been present, and not merely that his account of them has permeated the modern episteme to such a degree that it has made new ways of thinking and experiencing possible. As Carr points out, it is “not quite correct to say that the transcendental subject has a place only in the framework of Kant’s and Husserl’s elaborate theories. On the contrary, both give us the sense, precisely in their theories of the subject, that they are ‘discovering’ rather than merely ‘inventing’ something” (Carr 1999a, 124). What I am suggesting, therefore, is that we try reading Sass in a slightly modified light, whereby we assume that consciousness *is* transcendentally structured as Kant describes. Among other things, this will mean that while we may not, prior to Kant, have conceived the self in terms of the empirico-transcendental doublet, it has nonetheless always been a grounding feature of consciousness.

This adjustment does not require any alteration of Sass’s actual hermeneutic. Indeed, I can agree that the *Critique*, by bringing to light what was previously obscure to us, may well have created new possibilities of the kind documented by Sass. As it happens, in more recent work Sass has discussed the self in terms that resemble the approach I am recommending, though without mentioning Kant. Here he has considered the works of Shakespeare and Rembrandt as reflecting the emergence of a new kind of self-awareness well in advance of Kant and asks (mirroring Carr’s quote above), “Was this a matter of *discovery* on Shakespeare’s and Rembrandt’s part, or should it be described more as an *invention*? The wisest answer, no doubt, is both, though it is the second truth that is the more easily forgotten”(Sass 2022a, 13f). The last clause in that quote may in general be true, but the possibility that Kant has ‘discovered’ “some inherent or foundational reflexivity intrinsic to human consciousness" (Sass 2022b, 53) is something that he has neglected in his own work.

What Kant does with the doctrine of transcendental apperception is to account for a pre-theoretical awareness of self that we demonstrably do have. It is precisely this that gives rise to the metaphysical propositions of rational psychology. When Sass describes self-disturbance as "mind or subjectivity itself as becoming somehow object-like, because, in a Kantian sense, it has paradoxically taken itself as its own object – from within, so to speak" (Sass and Waugh 2018, 19), he is describing the same error that the rational psychologist makes. It is to succumb to the transcendental illusion discussed in chapter 4, the "illusion in the taking of a subjective condition of thinking for the cognition of an object" (A396). There is clearly a difference between the schizophrenic and the metaphysician, but the illusion itself, as Kant tells us, is "natural and unavoidable" (A298/B354). What is omitted in Sass's work, therefore, is that transcendental apperception helps to explain, among other things, philosophical confusions going back, at least, to Descartes. Kant also shows us that, due to its being a formal, contentless condition of experience (recall B158: "The consciousness of oneself is therefore far from being a cognition of oneself . . . ."), we are mistaken in treating the subjective self as an object of knowledge. In Susan Neiman's view, "Metaphysics thus appears to be the result of a perpetual weakness, even neurosis; Kant's assertion that this weakness is a natural part of the human condition scarcely makes it more attractive. On this view, the task of critique, or philosophy, is very much akin to the therapeutic one ascribed to Wittgenstein: to continue to expose the distortions and errors to which reason is inevitably prone and perhaps to remind us of the way to a more satisfactory form of life" (Neiman 1997, 188f).

The parallels between Kant and Wittgenstein have been noted by a number of commentators besides Neiman. Morris S. Engel, for example, argues that Wittgenstein applied Kant's teachings to language. Kant had shown that *a priori* features of cognition could undermine themselves by impelling us to seek the kind of insight that our cognition cannot in fact provide. Conversely, "what Wittgenstein came to see was that philosophical puzzlement arose from our desire to see and introduce into language more consistency and neatness than actually exist in it, or which it can accommodate" (Engel 1970, 508f). Kant, as Engel submits, was conscious of the parallel, as a passage from his *Prolegomena* reveals. There, Kant argues that to reveal in thought the pure concepts of the understanding, the categories, "required no greater reflection or more insight than to cull from a language rules for the actual use of words in general, and so to compile the elements for a grammar (and in fact both investigations are very closely related to one another) . . . ." (Kant 2002b, 115).

We have already seen that Sass's work has been informed by the philosophy of Wittgenstein, and, in particular, that he is impressed by his likening of philosophy to madness, and views his own idiosyncratic approach as a kind of therapy. As I argued in chapter 3, the worry that philosophy could resemble madness motivated Kant throughout his long career, and provided the impetus for the *Critique of Pure Reason*. Subsequently, significant parts of the *Critique* are taken up, not only with exposing metaphysical fallacies, but with an analysis of our cognitive abilities that explains *why* we find this kind of speculation so compulsive. Crucially, it also shows us how we can become more aware of the paradoxical nature of our cognitive framework in order to resist this temptation. Equally crucially, however, he also impresses upon us that there is no question of dissolving the paradoxes we face. David Carr notes that the problem facing Kant was the conflict between an idealism that does away with the material world and a scepticism that does away with the possibility of knowing the material world. "In both Kant and Husserl, the distinction between transcendental and empirical subject is introduced as a response to this situation. It expresses their view that both aspects of subjectivity – being subject for the world and being an object in the world – must be recognized . . . ." (Carr 1999b, 115).

Sass also mentions that Wittgenstein wanted to place philosophers "squarely back in the practical and communal discourse of life." (Sass 1994b, 74) This parallels Kant's 'cosmopolitanism', which informed his understanding of 'anthropology'.

In light of his debt to Wittgenstein, and the presaging of his 'therapeutic' approach in Kant's transcendental philosophy, it appears to me that the general position taken by Sass can be made far more consonant with the findings of the first *Critique*. By making this adjustment, the "endless oscillation" between "what is given in experience and what renders experience possible" that Foucault considered so debilitating for phenomenology and the human sciences (Foucault 2005, 366) can be seen for what it is: part of the 'human condition' rather than an artefact of Kantian theory. The odd situation we find ourselves in is perfectly captured by Hume's famous observations about the self. We are all familiar with the puzzle, though in different ages we may have framed it differently. I believe that it is just this familiarity that gives Sass's phenomenological interpretation of self-disturbance real potential when it comes to how we think about and interact with people suffering with mental disorder.

Sass's hermeneutic and reasons to resist dysfunction

Wakefield's approach, as I suggested in chapter 4, is to fix the HDA as *the* foundational concept of mental disorder – an unusually strong claim. I argued also that, from a Kantian point of view, his concept can only have regulative significance. What I am interested in here is why we should choose not to adopt the HDA as *the* definition of mental disorder.

It is not, primarily, for investigative purposes that Wakefield developed his analysis, as I argued in chapter 2, although he has championed it as a way to give empirical research a conceptual underpinning. Rather, it is to provide a definition of mental disorder that renders it, in principle, objective. It purportedly picks out a fact about reality that would obtain independently of anyone's thoughts about it. The significance of this, for Wakefield, is that the psychiatric profession would have a weighty response to the antipsychiatric allegation that it is a pseudoscience and, even worse, a tool of social control. This is not a regulative, heuristic proposition. In order to guide empirical investigation, it is not necessary to claim that this definition reflects the nature of reality.

I agree that the antipsychiatrists, particularly Szasz and his ilk, raise serious and legitimate ethical and moral concerns. There are many ways to respond to these concerns. The problem with Wakefield's analysis, as I see it, is that it presumptively defines mental disorder in terms that sunder the mentally disordered from the community of the 'normal' – and this, too, has ethical consequences.

As I have already noted, Sass's work addresses very different questions to those of Wakefield. Indeed, it is not incompatible with the harmful dysfunction analysis: Wakefield's response would presumably be that the 'self', however it is conceived, is a naturally-selected 'mechanism' or the product of several related mechanisms. Disturbances of the self would therefore imply malfunctions of such mechanisms, in line with his etiological concept of function as outlined in previous chapters. In at least one publication he has argued something like this, albeit referring to 'personhood' rather than the self (Wakefield 2009). Sass himself is informed about and open to neurobiological research into the physiological correlates of schizophrenia, which he discusses in an appendix to *Madness and Modernism*. His hermeneutic phenomenology does not commit him to any kind of constructivism about this or any other disorder. Neither, of course, does it commit him, as Wakefield is, to the existence of a categorical boundary between disorder and nondisorder. In one co-authored paper Sass explicitly *rejects* the HDA and aligns himself with Peter Zachar's argument that mental disorders constitute an 'imperfect community' (Pérez-Álvarez, Sass, and García-Montes 2008, 222 n7).[43] As Zachar himself describes it, from this perspective "there is no set of properties that all psychiatric disorders share and that distinguish them from nondisorders" (Zachar 2014, 125) – and, by the same token, no single, essential property such as 'dysfunction'.[44]

Why might Sass be disposed to take this view? He believes that the experiences of the mentally disordered – even the supposedly un-understandable experiences of the schizophrenic (Jaspers 1963, 376) – can be rendered meaningful, and that these efforts are vitally important. There is a sense in which this conviction is at odds with the dysfunction analysis, since "when human beings are viewed as malfunctioning cerebral mechanisms, incapable of higher levels of purposefulness and awareness, it will naturally be assumed that they are not only *difficult* to interpret but in some sense *beneath* interpretation, since their behaviour and expression must lack the intentionality and meaningfulness of normal human activity" (Sass 1998a, 18).

These different attitudes to mental disorder reveal themselves in the language of diagnosis. R. D. Laing expressed the opinion that "No one *has* schizophrenia, like having a cold. The patient has not 'got' schizophrenia. He is schizophrenic" (Laing 1990, 34). This is contrary to the orthodox medical view, expressed in a comment in DSM-III that "A common misconception is that a classification of mental disorders classifies individuals, when actually what are being classified are disorders that individuals have" (*Diagnostic and Statistical Manual of Mental Disorders: DSM-III* 1980, 6). Wakefield (2010, 7) concurs with this view, while Sass, as we might anticipate, is more circumspect, suggesting that "The currently preferred phrasing seems unlikely to foster the more complex forms of empathic understanding that may be required for an optimal

[43] There is, I acknowledge, some danger that this could reflect a co-author's view rather than Sass's own, although it is certainly consistent with the general tone of Sass's work.

[44] The term 'imperfect community' is originally due to Nelson Goodman (Goodman 1966).

therapeutic encounter" (Sass 2007, 397). Arguably, neither option is an ideal way to express the conceptual complexities of mental disorder. What the linguistic issue does reveal is that mental disorder is interpreted by the medical model as something separate and, in principle, detachable, from the individual. As Sass observes, this can tend to diminish the importance of the patient's own experience: what matters is the apparently isolable essence of the disorder, i.e., dysfunction.

Sass's approach, at a general level, is not unique. Many other phenomenologists, such as Thomas Fuchs and Matthew Ratcliffe, are also deeply interested in exploring the experiential nature of mental disorder. The relevance, for me, of Sass's contribution is his use – albeit in a somewhat attenuated, Foucauldian form – of a Kantian dialectic. My suggestion has been to revise Sass by taking this dialectic more seriously. In this way, we can come to see that despite the perplexing nature of schizophrenia, it is grounded in a universal feature of consciousness. As I have already shown, in more recent work Sass seems to consider more sympathetically the idea that the conditions of the empirico-transcendental doublet may be a human universal, although Kant himself is absent from these discussions. That is to say that there is at the root of schizophrenic experience something common to us all – that the schizophrenic is still 'one of us'.

His views on Kant notwithstanding, the principle of some sort of continuity between sane and mad does seem to run through Sass's work, though he never to my knowledge states it outright. It is revealing that in one place he does speculate that "rather than indicating a total failure of empathic comprehension, the praecox-feeling which is evoked in the interviewer may in fact indicate *a shared sense* of alienation; it may involve some accurate intuition of the profound self-estrangement which the patient him or herself is really going through" (Sass 1998, 559; emphasis mine). The 'praecox feeling' is a phenomenon first discussed by H. C. Rümke in 1941: a hard to define feeling of unease prompted by "the detachment and alienation of the schizophrenic patient" (Varga 2013, 133). The drift of Sass's thought here seems to be that this feeling may not simply be an expression of helpless perplexity on the part of the physician. Rather, it may be a form of recognition – a sort of intuitive mirroring of the schizophrenic's own dilemma.

Sass's comment contrasts with Rümke's own strained attempt to define the praecox feeling: as he puts it, "Somewhat pathetically one could say: 'the schizophrenic is outside the human community'" (Rümke 1990, 336). It is, I think, to be applauded that

Sass has made such a determined effort to resist sentiments such as these. As he has repeatedly observed, the view that the madman is essentially beyond the pale of rational comprehension has dominated the history of psychiatry. The importance of pushing back against this view of schizophrenia is captured quite well by a comment made by Gregory Bateson, whom we encountered in chapter 2. As he has noted, "Many writers have treated schizophrenia in terms of the most extreme contrast with any other form of human thinking and behaviour . . . so much emphasis on the differences from the normal – rather like the fearful physical segregation of psychotics – does not help in understanding the problems" (Bateson 1973b, 222).

This is a crucial point. As difficult as it may be, it is vital to maintain a place for the schizophrenic – and others suffering from different forms of mental disorder – *within* the 'human community'. To return to the core theme of my thesis, I maintain that the same problem arises in respect of what Sarbin and Mancuso, in a classic paper, called "the mental illness paradigm" (Sarbin and Mancuso 1970, 159), the view that mental disorder is an illness 'like any other', and one that Wakefield appears to support. The paradigm fails to take account of the fact that in the mental realm "what is picked out as disordered or problematic is that there is a disturbance of some kind in the person's thoughts, feelings or behaviours." (Banner 2013, 510) In contrast with physical disorders, a disturbance here is obviously likely to raise concerns about an individual's degree of autonomy and capacity for decision-making.

To be sure, those concerns are almost inevitable regardless of how we conceive mental disorder, and they may well bring in their wake fears that the mentally disordered may be unpredictable, even dangerous. However, if we insist on conceiving mental disorder as an illness 'like any other' – and, in particular, if we treat it as a dysfunction in Wakefield's sense – a far more decisive break with the world of the 'normal' person is effected. If mental disorders are defined by the presence of some kind of internal defect or deficit, the implication is that they are different not in degree but in kind; as the authors of one empirical study have put it, "Viewing those with mental disorders as diseased sets them apart and may lead to our perceiving them as physically distinct. Biochemical aberrations make them almost a different species" (Mehta and Farina 1997, 416).[45]

[45] We must bear in mind that Wakefield does not take 'dysfunction' in psychiatry to necessarily mean biological dysfunction. Nonetheless, moving to an abstract psychological level makes no obvious

This has important ramifications for the issue of stigma. It is well-known that people with mental disorders suffer from various forms of stigma and their attendant disadvantages (for a review see Rüsch, Angermeyer, and Corrigan 2005). One of the justifications often offered for adopting the mental illness paradigm is precisely in order to reduce stigma by changing the general public's *concept* of disorder. Robert Kendell, for example, argues that many lay people, and even some medical professionals "are apt to assume that developing a 'mental illness' is evidence of a certain lack of moral fibre and that, if they really tried, people with illnesses of this kind ought to be able to control their anxieties, their despondency and their strange preoccupations and 'snap out of it'" (Kendell 2001, 492). In other words, if mental disorders are no different to physical ones, no amount of forbearance will make a difference, just as it will not affect the outcome of a broken limb or bacterial infection.

The sort of attitude Kendell refers to certainly exists (for a surprisingly rare philosophical discussion see Scheurich 2002), although it is arguably not the most common response to mental disorder.[46] More significant are perceptions of danger, and awkwardness in social encounters, i.e., the view that the mentally disordered are difficult to interact with (e.g. Hayward and Bright 1997, 350f). It is not obvious why the mental illness paradigm should make any difference to these latter concerns, and may indeed exacerbate them. In a wide-ranging review article, John Read and colleagues suggest that "an illness model may lead people to believe that the ill have no control over their behaviour and may thereby increase the already widespread fear of the unpredictable and dangerous 'schizophrenic'" (Read et al. 2006, 305). Having reviewed research conducted in 16 countries between 1963 and 2005, they found a slight shift in public attitudes from psychosocial explanations for mental disorder to a more biogenetic understanding. This correlates with efforts to educate the public to adopt the mental illness paradigm. However, they measured very little change in stigmatising beliefs. Somewhat similarly, the authors of another study observe that in the US, "intensive efforts through the 1990s to 2006, mounted on the promise of neuroscience, have been

difference – what is really at stake here is the assumption that 'dysfunction' (for Wakefield) and 'biochemical abberations' (as Mehta and Farina put it) are deviations from objective standards of human function. I take it that this is what creates the impression that mentally disordered persons are categorically different from the nondisordered.

[46] That is not to say that it is absent; a study by Crisp et al., for example, reports that this attitude was present among their respondents, although it was far more strongly associated with addiction and eating disorders than other conditions (Crisp et al. 2000).

rewarded with significant and widespread increases in public acceptance of neurobiological theories and public support for treatment, including psychiatry, but no reduction in public stigma" (Pescosolido et al. 2010, 1325).

The antipsychiatrists discussed earlier in this thesis were motivated by worries about stigma, as well as the extent of medico-legal power and – particularly in the case of conditions such as schizophrenia – limited and often damaging forms of treatment. In the UK, R. D. Laing's response to these problems was of a broadly phenomenological bent: he recognised, perhaps more acutely than Jaspers before him, how important it was to grasp the experience of the disordered individual if a more equitable and humane understanding of madness was to be possible. Today, besides the phenomenological approach of Sass and others that I have considered above, attempts have also been made to directly incorporate the experience of the disordered into practical solutions to these challenges. While they have taken a number of forms, I will close by briefly considering two related strategies: co-production and expertise by experience.

Owing originally to the work of political scientist Elinor Ostrom in the 1970s, the term co-production captured the often unrecognised fact that the efficiency of public services relied, in part, on the active participation of service users themselves. In the context of mental disorder, this primarily means engaging service users as active participants in their own care – that is, as agents capable of reflecting upon their lived experience to identify specific problems and possible solutions. This approach thus reconceives health care professionals as facilitators helping their patients achieve those solutions, rather than imposing impersonal clinical 'fixes' upon them (Realpe and Wallace 2010), marking a significant departure from the traditional paternalistic mode of health care. Conceptualising mental disorder as organismic dysfunction presents it in terms of deficits that appear to undermine, in particular, the subject's control and agency. The result is both potentially stigmatising in its own right (despite the narrative that presents it as a solution to stigma) and disempowering for the patient. By recognising patients as partners in, rather than passive receivers of, treatment, co-production may be "more than a method or tool of better decision-making, rather it reflects a political agenda to rebalance inequalities and promote democracy."(Bevir, Needham, and Waring 2019, 197) This is all the more important in so far as factors such as sex and race play a significant role in psychiatric judgements, with women and people of colour disproportionately likely to be given mental disorder diagnoses, raising the suspicion that what counts as normal is a product of narrow – particularly white and male – values and expectations (Rachel Cooper addresses this problem in chapter 8 of her book *Psychiatry and Philosophy of Science* (Cooper 2007)). Co-production has the potential to ameliorate problems such as these, making it easier for patients to challenge biases. It is no panacea, to be sure, and some commentators have expressed serious doubts about its prospects (see, e.g., Kalathil and Rose 2019). Nonetheless, it is surely a step, however faltering, in the right direction. Furthermore, it has been recognised that there is wider scope for co-production beyond the context of individualised treatment. In this context the related concept of 'expertise by experience' has become prominent.

This concept recognises that knowledge of the sort that might be considered 'expert' can be acquired experientally as well as discursively through reasoning and reflection. This can be contrasted with expert knowledge more traditionally conceived, which (as Dings and Tekin (2023, 1420) emphasise by way of analogy with Frank Jackson's famous 'Mary the colour scientist' thought experiment) cannot capture what it is like to experience a phenomenon.[47] An expert by experience *qua* mental disorder is therefore someone who has experienced disorder and, usually, is familiar with concomitant challenges such as negotiating health and social care systems. As Dings and Tekin phrase it, such knowledge ideally "encompasses the full phenomenon in the way it is experienced by the person." (Ibid.) Involving experts by experience may therefore address a significant epistemic gap. To date, the involvement of such experts has been particularly notable in the education of health care professionals, although few experts by experience have been absorbed into academia (Happell et al. 2021). Clearly there is much more potential to extend co-production by enlisting experts by experience in fields such as policy design and research. Anne-Marie Gagné-Julien has made just such a case regarding the specific issue of defining mental disorder (Gagné-Julien 2021).

Gagné-Julien maintains that a definition of some kind remains an important goal, while at the same time adopting a form of constructivism that recognises the inescapable role of values in mental disorder judgments. Crucially, she notes that "the influence of oppressive values in psychiatry is usually implicit: values enter psychiatric classification through the back door. It then becomes difficult to address those prejudicial values if psychiatry doesn't have official structures or mechanisms to make these values explicit."

[47] The authors do, of course, recognise that philosophical opinion regarding this implication of Jackson's experiment is divided.

(Gagné-Julien, 9413) A more promising approach, therefore, would be to acknowledge that values thoroughly penetrate scientific concepts and theories, including that of dysfunction, but to avoid relativism through critical intersubjective dialogue. Gagné-Julien argues that if a definition of mental disorder must be developed, it should be through this kind of dialogue, and that it is essential that experts by experience be numbered among the various experts who would contribute – not least because people with experience of mental disorder are better placed to identify 'prejudicial values' as a result of having been negatively affected by them.

Gagné-Julien's proposal brings me back, at the close of this thesis, to the problem of definition that I started with. It offers one example of an alternative to naturalist analyses of the mental disorder concept that doesn't entail abandoning ourselves to relativism. It is an alternative that recognises how scientific claims, rather than eliminating values, can mask their continued and potentially malign influence, and an alternative that takes seriously the power of lived experience in bringing those values to light. Not only is there reason, as I hope to have shown, to think that the dysfunction analysis fails to deliver a naturalistic definition, there is also good reason to resist the very notion of such attempts.

Conclusion

The harmful dysfunction analysis of the concept of mental disorder defines mental disorder as a categorical break with what is 'healthy' or 'normal'. This reflects the view of its author, Jerome Wakefield, that "one of the primary goals that motivated the search for a definition of disorder in the first place [is]: to limit false-positive diagnoses in which social deviance is mislabeled mental disorder and thus to respond to antipsychiatric claims that psychiatric diagnosis is misused for social control purposes by creating overly inclusive categories" (Wakefield 2021c, 235). This is, in itself, a laudable aim. However, it also has the effect of cutting the psychiatric patient off from the community of the nominally sane, increasing the psychological distance between 'them' and 'us'.

In this chapter I have looked at how Louis Sass has approached the most mysterious of all mental disorders, schizophrenia, through a phenomenological attempt to make sense of the patient's experience. He has looked at the phenomenon of self-disturbance, a central symptom of this disorder, through a broadly Kantian lens which acknowledges a tension between the subjective, knowing self and the self as empirically introspectable.

Against the prevailing view that schizophrenia is fundamentally irrational and thus beyond comprehension, he has shown that "Many strange and paradoxical characteristics of schizophrenia can arise from *within* aspects of experience traditionally equated with maturity and health: from disengagement, self-consciousness, and capacity for reflective distance . . . ." (Sass 1994, 79). I have argued that if we take this perspective we can come to see the schizophrenic less in terms of what is strange and disconcerting than as someone whose difficulties are related to the universal human condition of being both subject and object – a transcendental and an empirical self.

Conclusion

In this thesis I set out to ask whether the dysfunction analysis of medical disorder is justified in its claim to provide a value-free naturalistic definition for *mental* disorder. Specifically, I used elements of Immanuel Kant's *Critique of Pure Reason* to question whether Jerome Wakefield's harmful dysfunction analysis (HDA) provides a legitimate conceptual definition.

My choice of Kant's philosophy was inspired by his own interest, throughout his long career, in mental disorder. This relatively obscure aspect of his work, I argued, had a direct influence on the critical philosophy. The CPR was fundamentally an attempt to determine the legitimate boundaries of knowledge, and to demonstrate that these boundaries are strictly limited by the nature of our cognitive architecture. This project was, in my view, motivated by his thoughts and observations about mental disorder, and his perception that the delusions of insanity were importantly similar to the speculative excesses of his own philosophical forebears and contemporaries. In this, as I have noted, he presages the later Wittgenstein's diagnosis of the 'philosophical disease'.

The HDA fails to define what Kant would call an empirical concept, but rather yields an 'idea' of reason that, while of heuristic or 'regulative' utility, does not make 'disorder' an object of possible experience. Like the empiricists, Kant insists that all knowledge must derive from sense experience. Unlike them, he argues that experience also demands concepts – both the *a priori* categories and empirical concepts. Without concepts, objects must remain indeterminate, and cannot provide any basis for knowledge.

Dysfunction – and, therefore, disorder – lacks any sensible predicates by which it could be conceptually determined. It cannot be an object of experience. This is not to say that we cannot make dysfunction, or disorder, attributions. What it means is that these cannot, as the dysfunction analysis claims, be objective. We need not dismiss them as purely evaluative, but the evaluative element is – contra its advocates – inescapably bound up with the very notion of 'dysfunction'.

I concluded by looking at Kant's response to the problem of the self. I examined Louis Sass's application of Kant dialectic of the self in his work on schizophrenia. I ended by considering the effects on patients of thinking in terms of dysfunction, particularly in terms of stigma. I suggested that Sass's broadly Kantian hermeneutic provides a way to mitigate perceptions of the mentally disordered as categorically different to the 'sane'.

Bibliography

'About RDoC'. n.d. National Institute of Mental Health. Accessed 20 September 2023. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/about-rdoc.

Adriaens, Pieter R., and Andreas De Block, eds. 2011. *Maladapting Minds: Philosophy, Psychiatry, and Evolutionary Theory*. Oxford: Oxford University Press.

Allison, Henry E. 2004. *Kant's Transcendental Idealism*. New Haven: Yale University Press.

———. 2006. 'Transcendental Realism, Empirical Realism and Transcendental Idealism'. *Kantian Review* 11: 1–28.

Andreasen, Nancy C. 2001. *Brave New Brain: Conquering Mental Illness in the Era of the Genome*. Oxford: Oxford University Press.

Aucouturier, Valérie, and Steeves Demazeux. 2012. 'The Concept of "Mental Disorder"'. In *Health, Illness & Disease: Philosophical Essays*, edited by Rachel Cooper and Havi Carel, 75–89. Stocksfield: Acumen Publishing.

Banham, Gary. 2013. 'Regulative Principles and Regulative Ideas'. In *Kant Und Die Philosophie in Weltbürgerlicher Absicht: Akten Des XI. Kant-Kongresses 2010*, edited by Stefano Bacin, Alfredo Ferrarin, Claudio La Rocca, and Margit Ruffing, 15–24. Berlin: De Gruyter.

Banner, Natalie F. 2013. 'Mental Disorders Are Not Brain Disorders'. *Journal of Evaluation in Clinical Practice* 19 (3): 509–13.

Bateson, Gregory. 1973a. 'The Group Dynamics of Schizophrenia'. In *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, 228–43. Paladin.

———. 1973b. 'Toward a Theory of Schizophrenia'. In *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, 201–27. Paladin.

Bentall, Richard, and David Pilgrim. 1993. 'Thomas Szasz, Crazy Talk and the Myth of Mental Illness'. *British Journal of Medical Psychology* 66 (1): 69–76.

Bevir, Mark, Catherine Needham, and Justin Waring. 2019. 'Inside Co-Production: Ruling, Resistance, and Practice'. *Social Policy & Administration* 53 (2): 197–202.

Bigelow, John, and Robert Pargetter. 1987. 'Functions'. *The Journal of Philosophy* 84 (4): 181–96.

'Biologically-Inspired Biomarkers for Mental Disorders'. 2017. *EBioMedicine* 17: 1–2. https://doi.org/10.1016/j.ebiom.2017.03.015.

Biomarkers Definitions Workgroup. 2001. 'Biomarkers and Surrogate Endpoints: Preferred Definitions and Conceptual Framework'. *Clinical Pharmacology & Therapeutics* 69 (3): 89–95.

Boghossian, Paul A. 1989. 'The Rule-Following Considerations'. *Mind* 98 (392): 507–49.

Boorse, Christopher. 1976a. 'What a Theory of Mental Health Should Be'. *Journal for the Theory of Social Behaviour* 6 (1): 61–84.

———. 1976b. 'Wright on Functions'. *The Philosophical Review* 85 (1): 70–86.

———. 1977. 'Health as a Theoretical Concept'. *Philosophy of Science* 44 (4): 542–73.

———. 2014. 'A Second Rebuttal on Health'. *Journal of Medicine and Philosophy* 39 (6): 683–724.

Borsboom, Denny. 2017. 'A Network Theory of Mental Disorders'. *World Psychiatry* 16 (1): 5–13.

Boyd, Kenneth M. 2000. 'Disease, Illness, Sickness, Health, Healing and Wholeness: Exploring Some Elusive Concepts'. *Journal of Medical Ethics* 26 (9): 9–17.

Brendel, David H. 2003. 'Reductionism, Eclecticism, and Pragmatism in Psychiatry: The Dialectic of Clinical Explanation'. *The Journal of Medicine and Philosophy* 28 (5): 563–80.

Breitenbach, Angela. 2009. 'Teleology in Biology: A Kantian Perspective'. *Kant Yearbook* 1: 31–56.

Bridgman, Percy W. 1958. *The Logic of Modern Physics*. New York: MacMillan.

Brook, Andrew. 1994. *Kant and the Mind*. Cambridge: Cambridge University Press.

Buchenau, Stephanie. 2017. 'Herder: Physiology and Philosophical Anthropology'. In *Herder: Philosophy and Anthropology*, edited by Anik Waldow and Nigel DeSouza, 72–93. Oxford: Oxford University Press.

Buhrer, Eliza. 2014. '"But What Is to Be Said of a Fool?": Intellectual Disability in Medieval Thought and Culture'. In *Mental Health, Spirituality, and Religion in the Middle Ages and Early Modern Age*, edited by Albrecht Classen, 314–43. Berlin: De Gruyter.

Buss, David M., Todd K. Shackelford, April L. Bleske, and Jerome C. Wakefield. 1998. 'Adaptations, Exaptations, and Spandrels'. *American Psychologist* 53 (5): 533–48.

Butts, Robert E. 1986. *Kant and the Double Government Methodology: Supersensibility and Method in Kant's Philosophy of Science*. Dordrecht: D. Reidel Publishing Company.

Carr, David. 1999a. *The Paradox of Subjectivity: The Self in the Transcendental Tradition*. Oxford: Oxford University Press.

———. 1999b. *The Paradox of Subjectivity: The Self in the Transcendental Tradition*. Oxford: Oxford University Press.

Cavallar, Georg. 2012. 'Cosmopolitanisms in Kant's Philosophy'. *Ethics & Global Politics* 5 (2): 95–118.

Chance, Brian A. 2013. 'Kant and the Discipline of Reason'. *European Journal of Philosophy* 23 (1): 87–110.

Christiansen, Charles. 2007. 'Adolf Meyer Revisited: Connections Between Lifestyles, Resilience and Illness'. *Journal of Occupational Science* 14 (2): 63–76.

Cooper, Rachel. 2007. *Psychiatry and Philosophy of Science*. Abingdon: Routledge.

Corr, Charles A. 1975. 'Christian Wolff and Leibniz'. *Journal of the History of Ideas* 36 (2): 241–62.

Crisp, Arthur H., Michael G. Gelder, Susannah Rix, Howard I. Meltzer, and Olwen J. Rowlands. 2000. 'Stigmatisation of People with Mental Illness'. *British Journal of Psychiatry* 177 (1): 4–7.

Crossley, Nick. 1998. 'R. D. Laing and the British Anti-Psychiatry Movement: A Socio-Historical Analysis'. *Social Science & Medicine* 47 (7): 877–89.

Cummins, Robert. 1975. 'Functional Analysis'. *The Journal of Philosophy* 72 (20): 741–65.

David-Ménard, Monique. 2000. 'Kant's "An Essay on the Maladies of the Mind" and Observations on the Feeling of the Beautiful and the Sublime'. Translated by Alison Ross. *Hypatia* 15 (4): 82–89.

Davidson, Larry, and John S. Strauss. 1992. 'Sense of Self in Recovery from Severe Mental Illness'. *The British Journal of Medical Psychology* 65 (2): 131–45.

Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 2013. Arlington VA: American Psychiatric Association.

Diagnostic and Statistical Manual of Mental Disorders: DSM-III. 1980. Arlington VA: American Psychiatric Association.

Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R. 1987. Arlington VA: American Psychiatric Association.

Dings, Roy, and Leon C. de Bruin. 2022. 'What's Special About "Not Feeling Like Oneself?" A Deflationary Account of Self(-Illness) Ambiguity'. *Philosophical Explorations* 25 (3): 269–89.

Dings, Roy, and Şerife Tekin. 2023. 'A Philosophical Exploration of Experience-Based Expertise in Mental Health Care'. *Philosophical Psychology* 36 (7): 1415–34.

Double, Duncan B. 2020. 'Critical Psychiatry: An Embarassing Hangover from the 1970s?' *British Journal of Psychiatry* 44 (6): 233–36.

Dupré, John. 1981. 'Natural Kinds and Biological Taxa'. *The Philosophical Review* 90 (1): 66–90.

Dussault, Antoine C. 2022. 'A Clarification on the Boorse-Wakefield Debate About Health: Is the Theoretical/Therapeutic Distinction Dispensable?' *Analysis* 82 (4): 673–81.

Dyck, Corey W. 2011. 'A Wolff in Kant's Clothing: Christian Wolff's Influence on Kant's Accounts of Consciousness, Self-Consciousness, and Psychology'. *Philosophy Compass* 6 (1): 44–53.

Engel, S. Morris. 1970. 'Wittgenstein and Kant'. *Philosophy and Phenomenological Research* 30 (4): 483–513.

Ereshefsky, Marc. 2009. 'Defining "Health" and "Disease"'. *Studies in History of Biological and Biomedical Sciences*, no. 40: 221–27.

Fauvel, Aude. 2013. 'Crazy Brains and the Weaker Sex: The British Case (1860-1900)'. *Clio*, no. 37: 1–25.

Feinstein, Alvan. 1977. 'A Critical Overview of Diagnosis in Psychiatry'. In *Psychiatric Diagnosis*, edited by Vivian Rakoff, Harvey Stancer, and Henry Kedward, 189–206. New York: Brunner Mazell.

Fodor, Jerry A. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fogelin, Robert. 2003. *Walking the Tightrope of Reason: The Precarious Life of a Rational Animal*. Oxford: Oxford University Press.

Foucault, Michel. 2005. *The Order of Things: An Archaeology of the Human Sciences*. London: Routledge.

———. 2008. *Introduction to Kant's Anthropology*. Translated by Roberto Nigro and Kate Briggs. Los Angeles: Semiotext(e).

French, Stanley G. 1967. 'Kant's Constitutive-Regulative Distinction'. *The Monist* 51 (4): 623–39.

Friedman, Michael. 1992. 'Causal Laws and Foundations of Natural Science'. In *The Cambridge Companion to Kant*, edited by Paul Guyer, 161–99. Cambridge: Cambridge University Press.

Frierson, Patrick. 2009. 'Kant on Mental Disorder. Part 2: Philosophical Implications of Kant's Account'. *History of Psychiatry* 20 (3): 290–310.

———. 2014. *Kant's Empirical Psychology*. Cambridge: Cambridge University Press.

Gagné-Julien, Anne-Marie. 2021. 'Towards a Socially Constructed and Objective Concept of Mental Disorder'. *Synthese* 198 (10): 9401–26.

Gold, Ian. 2009. 'Reduction in Psychiatry'. *The Canadian Journal of Psychiatry* 54 (8): 506–12.

Goodman, Nelson. 1966. *The Structure of Appearance*. Indianapolis: Bobbs Merrill.

Gorenstein, Ethan E. 1984. 'Debating Mental Illness: Implications for Science, Medicine, and Social Policy'. *American Psychologist* 39 (1): 50–56.

Grier, Michelle. 2001. *Kant's Doctrine of Transcendental Illusion*. Cambridge: Cambridge University Press.

Guillery, R. W. 2004. 'Observations of Synaptic Structures: Origins of the Neuron Doctrine and Its Current Status'. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1458): 1281–1307.

Gutting, Gary. 1989. *Michel Foucault's Archaeology of Scientific Reason*. Cambridge: Cambridge University Press.

Guyer, Paul. 1990. 'Reason and Reflective Judgement: Kant on the Significance of Systematicity'. *Noûs* 24 (1): 17–43.

———. 2010. 'Introduction'. In *The Cambridge Companion to Kant's Critique of Pure Reason*, edited by Paul Guyer, 1–18. Cambridge: Cambridge University Press.

Guze, Samuel. B. 1978. 'Nature of Psychiatric Illness: Why Psychiatry Is a Branch of Medicine'. *Comprehensive Psychiatry* 19 (4): 295–307.

———. 1989. 'Biological Psychiatry: Is There Any Other Kind?' *Psychological Medicine* 19 (2): 315–23.

Hacking, Ian. 1979. 'Michel Foucault's Immature Science'. *Noûs* 13 (1): 39–51.

———. 2007. 'Putnam's Theory of Natural Kinds and Their Names Is Not the Same as Kripke's'. *Principia* 11 (1): 1–24.

Hanna, Robert. 1998. 'A Kantian Critique of Scientific Essentialism'. *Philosophy and Phenomenological Research* 58 (3): 497–528.

———. 2006. *Kant, Science, and Human Nature*. Oxford: Clarendon Press.

Happell, Brenda, Aine O'Donovan, Julie Sharrock, Terri Warner, and Sarah Gordon. 2021. 'They Are a Different Breed Aren't They? Exploring How Experts by Experience Influence Students Through Mental Health Education'. *International Journal of Mental Health Nursing* 30 (S1): 1354–65.

Hayward, Peter, and Jennifer A. Bright. 1997. 'Stigma and Mental Illness: A Review and Critique'. *Journal of Mental Health* 6 (4): 345–54.

Hempel, Carl G. 1965. 'The Logic of Functional Analysis'. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 297–330. New York: The Free Press.

Horwitz, Allan V. 2015. 'DSM-I and DSM-II'. In *The Encyclopedia of Clinical Psychology*, edited by Robin L. Cautin and Scott O. Lilienfeld, 951–56. John Wiley & Sons.

———. 2021. *DSM: A History of Psychiatry's Bible*. Baltimore: Johns Hopkins University Press.

Hume, David. 1984. *A Treatise of Human Nature*. Edited by Ernest C. Mossner. London: Penguin.

Husserl, Edmund. 1970. *The Crisis of European Sciences and Transcendental Phenomenology*. Translated by David Carr. Evanston, IL: Northwestern University Press.

Insel, Thomas R., and Remi Quirion. 2005. 'Psychiatry as a Clinical Neuroscience Discipline'. *Journal of the American Medical Association* 294 (17): 2221–24.

Jacobs, David H., and David Cohen. 2010. 'Does "Psychological Dysfunction" Mean Anything? A Critical Essay on Pathology Versus Agency'. *Journal of Humanistic Psychology* 50 (3): 312–34.

Jaspers, Karl. 1963. *General Psychopathology*. Translated by J. Hoenig and Marian W. Hamilton. Manchester: Manchester University Press.

Jong, Willem R. de. 1995. 'How Is Metaphysics as a Science Possible? Kant on the Distinction between Philosophical and Mathematical Method'. *The Review of Metaphysics* 49 (2): 235–74.

Jylkkä, Jussi, Henry Railo, and Jussi Haukioja. 2009. 'Psychological Essentialism and Semantic Externalism: Evidence for Externalism in Lay Speakers' Language Use'. *Philosophical Psychology* 22 (1): 37–60.

Kalathil, Jayasree, and Diana Rose. 2019. 'Power, Privilege and Knowledge: The Untenable Promise of Co-Production in Mental "Health"'. *Frontiers in Sociology* 4: 57.

Kandel, Eric. 1999. 'Biology and the Future of Psychoanalysis: A New Intellectual Framework for Psychiatry Revisited'. *American Journal of Psychiatry* 156 (4): 505–24.

———. 2005. *Psychiatry, Psychoanalysis, and the New Biology of Mind*. Washington, D.C.: American Psychiatric Publishing.

Kant, Immanuel. 1978. *Anthropology from a Pragmatic Point of View*. Edited by Hans H. Rudnick. Translated by Victor Lyle Dowdell. Carbondale, IL: Southern Illinois University Press.

———. 1992a. 'A New Elucidation of the First Principles of the Metaphysical Cognition'. In *Theoretical Philosophy, 1755-1770*, translated by David Wolford and Ralf Meerbote, 1-46. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1992b. ‘Attempt to Introduce the Concept of Negative Magnitudes into Philosophy’. In *Theoretical Philosophy, 1755-1770,* 203-242. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1992c. ‘Dreams of a Spirit-Seer Elucidated by Dreams of Metaphysics’. In *Theoretical Philosophy, 1755-1770*, 301–59. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1992d. *Lectures on Logic*. Edited and translated by Young J. Michael. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1992e. ‘The False Subtlety of the Four Syllogistic Figures’. In *Theoretical Philosophy, 1755-1770*, 85-101. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1997. *Lectures on Metaphysics*. Edited and translated by Karl Ameriks and Steve Naragon. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1998. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 1999. *Correspondence*. Translated by Arnulf Zweig. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 2000. *Critique of Judgment*. Translated by Eric Matthews and Paul Guyer. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 2002a. ‘Metaphysical Foundations of Natural Science’. In *Theoretical Philosophy After 1781*, edited by Henry E. Allison and Peter Heath, translated by Gary Hatfield, Michael Friedman, Henry E. Allison, and Peter Heath, 171–270. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 2002b. ‘Prolegomena to Any Future Metaphysics’. In *Theoretical Philosophy After 1781*, edited by Henry E. Allison and Peter Heath, translated by Gary Hatfield, Michael Friedman, Henry E. Allison, and Peter Heath, 29–170. The

Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 2007. 'Essay on the Maladies of the Head'. In *Anthropology, History, and Education*, edited by Günter Zöller and Robert B. Louden, translated by Mary Gregor, 65–77. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

———. 2012. 'Thoughts On The True Estimation of Living Forces'. In *Natural Science*, edited by Eric Watkins, translated by Lewis White, 11-155. The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press.

Kapur, S., A. G. Phillips, and Thomas R. Insel. 2012. 'Why Has It Taken So Long For Biological Psychiatry To Develop Clinical Tests And What To Do About It?' *Molecular Psychiatry* 17 (12): 1174–79.

Kawa, Shadia, and James Giordano. 2012. 'A Brief Historicity of the Diagnostic and Statistical Manual of Mental Disorders: Issues and Implications for the Future of Psychiatric Canon and Practice'. *Philosophy, Ethics, and Humanities in Medicine* 7: 2.

Kemp Smith, Norman. 1969a. *A Commentary to Kant's Critique of Pure Reason*. Bath: Cedric Chivers.

———. 1969b. *A Commentary to Kant's Critique of Pure Reason*. Bath: Cedric Chivers.

Kendell, Robert E. 1975. 'The Concept of Disease and Its Implications for Psychiatry'. *British Journal of Psychiatry*, no. 127: 305–15.

———. 1989. 'Clinical Validity'. *Psychological Medicine* 19 (1): 45–55.

———. 2001. 'The Distinction Between Mental and Physical Illness'. *British Journal of Psychiatry* 178 (6): 490–93.

Kendler, K. S. 2017. 'DSM Disorders and Their Criteria: How Should They Inter-Relate?' *Psychological Medicine* 47 (12): 2054–60.

Kitcher, Patricia. 1982. 'Kant on Self-Identity'. *The Philosophical Review* 91 (1): 41–72.

———. 1990. *Kant's Transcendental Psychology*. Oxford: Oxford University Press.

Klerman, Gerald L. 1978. 'The Evolution of a Scientific Nosology'. In *Schizophrenia: Science and Practice*, edited by John C. Sherstow, 99–121. Cambridge, MA: Harvard University Press.

Kraepelin, Emil. 2005. 'The Directions of Psychiatric Research'. *History of Psychiatry* 16 (3): 350–64.

Kripke, Saul. 1980. *Naming and Necessity*. Oxford: Blackwell.

———. 1982. *Wittgenstein On Rules and Private Language*. Cambridge, MA: Harvard University Press.

Kuehn, Manfred. 2001. *Kant: A Biography*. Cambridge: Cambridge University Press.

Kuhn, Thomas S. 2012. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Laing, R. D. 1990. *The Divided Self: An Existential Study in Sanity and Madness*. Penguin Psychology. London: Penguin, 1990.

Lau, Chong-Fuk. 2015. 'Transcendental Concepts, Transcendental Truths and Objective Validity'. *Kantian Review* 20 (3): 445–66.

Laywine, Alison. 1993. *Kant's Early Metaphysics and the Origins of the Critical Philosophy*. Atascadero, CA: Ridgeview Publishing Company.

———. 1998. 'Problems and Postulates: Kant on Reason and Understanding'. *Journal of the History of Philosophy* 36 (2): 279–309.

Legrand, Stéphane. 2008. '"As Close as Possible to the Unliveable": Michel Foucault and Phenomenology'. *Sophia* 47 (3): 281–91.

Locke, John, and P. H. Nidditch. 1975. *An Essay Concerning Human Understanding*. The Clarendon Edition of the Works of John Locke. Oxford: Clarendon Press.

Malla, Ashok, Ridha Joober, and Amparo Garcia. 2015. '"Mental Illness Is Like Any Other Illness": A Critical Examination of the Statement and Its Impact on Patient Care and Society'. *Journal of Psychiatry & Neuroscience* 40 (3): 147–50.

Massimi, Michela. 2017. 'What Is This Thing Called Scientific Knowledge? Kant On Imaginary Standpoints and the Regulative Role of Reason'. *Kant Yearbook* 9 (1): 63–84.

McGeachan, Cheryl. 2014. '"The World Is Full of Big Bad Wolves": Investigating the Experimental Therapeutic Spaces of R. D. Laing and Aaron Esterson'. *History of Psychiatry* 25 (3): 283–98.

McLaughlin, Peter. 2003. *What Functions Explain: Functional Explanation and Self-Reproducing Systems*. Cambridge: Cambridge University Press.

McNulty, Michael Bennett. 2015. 'Rehabilitating the Regulative Use of Reason: Kant on Empirical and Chemical Laws'. *Studies in History and Philosophy of Science* 54: 1–10.

Mehta, Sheila, and Amerigo Farina. 1997. 'Is Being "Sick" Really Better? Effect of the Disease View of Mental Disorder on Stigma'. *Journal of Social and Clinical Psychology* 16 (4): 405–19.

Middleton, Hugh, and Joanna Moncrieff. 2019. 'Critical Psychiatry: A Brief Overview'. *BJPsych Advances* 25 (1): 47–54.

Millikan, Ruth Garrett. 1989. 'In Defence of Proper Functions'. *Philosophy of Science* 56 (2): 288–302.

Montefiore, Alan. 2000. 'Reason and Its Own Self-Undoing?' In *Philosophy and the Human Paradox: Essays on Reason, Truth and Identity*, edited by Danielle Sands, 85–98. New York: Routledge.

Mosser, Kurt. 2009. 'Kant and Wittgenstein: Common Sense, Therapy, and the Critical Philosophy'. *Philosophia* 37 (1): 1–20.

Munsche, Heather, and Harry A. Whitaker. 2012. 'Eighteenth Century Classification of Mental Illness: Linnaeus, de Sauvages, Vogel, and Cullen'. *Cognitive and Behavioural Neurology* 25 (4): 224–39.

Murphy, Dominic, and Robert L. Woolfolk. 2000. 'The Harmful Dysfunction Analysis of Mental Disorder'. *Philosophy, Psychiatry, & Psychology* 7 (4): 241–52.

Nagel, Ernest. 1961a. 'Mechanistic Explanation and Organismic Biology'. In *The Structure of Science: Problems in the Logic of Scientific Explanation*, 398–446. New York: Harcourt, Brace & World.

———. 1961b. 'The Reduction of Theories'. In *The Structure of Science: Problems in the Logic of Scientific Explanation*, 336–97. New York: Harcourt, Brace & World.

Nasser, Mervat. 1995. 'The Rise and Fall of Anti-Psychiatry'. *Psychiatric Bulletin* 19 (12): 743–46.

Neander, Karen. 1991. 'Functions as Selected Effects'. *Philosophy of Science* 58 (2): 168–84.

Neiman, Susan. 1997. *The Unity of Reason: Rereading Kant*. Oxford: Oxford University Press.

Oberst, Michael. 2015. 'Two Worlds and Two Aspects: On Kant's Distinction between Things in Themselves and Appearances'. *Kantian Review* 20 (1): 53–75.

O'Neill, Onora. 1990. 'Enlightenment as Autonomy'. In *The Enlightenment and Its Shadows*, edited by Peter Hulme and Ludmilla Jordanova, 186–99. London: Routledge.

Otabe, Tanehisa. 2018. 'An Iroquois in Paris and a Crusoe on a Desert Island: Kant's Aesthetics and the Process of Civilization'. *Culture and Dialogue* 6 (1): 35–50.

Paden, Roger. 1987. 'Foucault's Anti-Humanism'. *Human Studies* 10 (1): 123–41.

Parfit, Derek. 2011. *On What Matters*. Edited by Samuel Scheffler. Vol. 1. Oxford: Oxford University Press.

Pérez-Álvarez, Marino, Louis A. Sass, and José García-Montes. 2008. 'More Aristotle, Less DSM: The Ontology of Mental Disorders in Constructivist Perspective'. *Philosophy, Psychiatry, & Psychology* 15 (3): 211–25.

Pescosolido, Bernice A., Martin Jack K., Scott J. Long, Tait R. Medina, Jo C. Phelan, and Bruce G. Link. 2010. '"A Disease Like Any Other"? A Decade of Change in Public Reactions to Schizophrenia, Depression, and Alcohol Dependence'. *American Journal of Psychiatry* 167 (11): 1321–30.

Pippin, Robert B. 1987. 'Kant on the Spontaneity of Mind'. *Canadian Journal of Philosophy* 17 (2): 449–75.

Porter, Roy. 2002. *Madness: A Brief History*. Oxford: Oxford University Press.

Putnam, Hilary. 1992. 'Is Water Necessarily H_2O?' In *Realism With a Human Face*, 54–79. Cambridge, MA: Harvard University Press.

———. 1997. 'The Meaning of "Meaning"'. In *Mind, Language and Reality: Philosophical Papers Volume 2*, 215–71. Cambridge: Cambridge University Press.

Quinton, A. 1984. 'Madness'. *Royal Institute of Philosophy Supplements* 18: 17–41.

Read, John, Nick Haslam, Liz Sayce, and Emma Davies. 2006. 'Prejudice and Schizophrenia: A Review of the "Mental Illness Is an Illness Like Any Other" Approach'. *Acta Psychiatrica Scandinavia* 114 (5): 303–18.

Realpe, Alba, and Louise M. Wallace. 2010. 'What Is Co-Production?' London: The Health Foundation.

Rosenhan, David L. 1973. 'On Being Sane in Insane Places'. *Science* 179 (4070): 250–58.

Rümke, H. C. 1990. 'The Nuclear Symptom of Schizophrenia and the Praecoxfeeling'. Translated by J. Neeleman. *History of Psychiatry* 1 (3): 331–41.

Rüsch, Nicolas, Matthias C. Angermeyer, and Patrick W. Corrigan. 2005. 'Mental Illness Stigma: Concepts, Consequences, and Initiatives to Reduce Stigma'. *European Psychiatry* 20 (8): 529–39.

Saji, Motohide. 2009. 'On the Division Between Reason and Unreason in Kant'. *Human Studies* 32 (2): 201–23.

Sarbin, Theodore R., and James C. Mancuso. 1970. 'Failure of a Moral Enterprise: Attitudes of the Public Toward Mental Illness'. *Journal of Consulting and Clinical Psychology* 35 (2): 159–73.

Sass, Louis A. 1994a. *The Paradoxes of Delusion*. Ithaca, NY: Cornell University Press.

———. 1994b. *The Paradoxes of Delusion*. Ithaca, NY: Cornell University Press.

———. 1998a. *Madness and Modernism: Insanity in the Light of Modern Art, Literature, and Thought*. Cambridge, MA: Harvard University Press.

———. 1998b. 'Schizophrenia, Self-Consciousness, and the Modern Mind'. *Journal of Consciousness Studies* 5 (5–6): 543–65.

———. 2007. '"Schizophrenic Person" or "Person with Schizophrenia"?: An Essay on Illness and Self'. *Theory & Psychology* 17 (3): 395–420.

———. 2013. 'Self-Disturbance and Schizophrenia: Structure, Specificity, Pathogenesis'. *Recherches En Psychanalyse* 16 (2): 119–32.

———. 2022a. '"A Flaw in the Great Diamond of the World": Reflections on Subjectivity and the Enterprise of Psychology (A Diptych)'. *The Humanistic Psychologist* 501 (1): 3–32.

———. 2022b. ‘Intersecting Perspectives: Hermeneutic Phenomenology, Psychoanalysis, and Historical Ontology’. *The Humanistic Psychologist* 50 (1): 46–54.

Sass, Louis A., and Josef Parnas. 2007. ‘Explaining Schizophrenia: The Relevance of Phenomenology’. In *Reconceiving Schizophrenia*, edited by William Fulford, Man Cheung Chung, and George Graham, 63–95. Oxford: Oxford University Press.

Sass, Louis A., and Patricia Waugh. 2018. ‘Modernism and Madness: Louis Sass and Patricia Waugh in Conversation’. Edited by John Foxwell. Hearing the Voice.

Scadding, J. G. 1963. ‘Meaning of Diagnostic Terms in Broncho-Pulmonary Disease’. *British Medical Journal* 2 (5370): 1425–30.

———. 1967. ‘Diagnosis: The Clinician and the Computer’. *The Lancet* 290 (7521): 877–82.

Schaffner, Kenneth F. 2013. ‘Reduction and Reductionism in Psychiatry’. In *The Oxford Handbook of Philosophy and Psychiatry*, edited by Kenneth Fulford, Martin Davies, Richard Gipps, George Graham, John Sadler, Giovanni Stanghellini, and Tim Thornton. Oxford: Oxford University Press.

Scheurich, Neil. 2002. ‘Moral Attitudes and Mental Disorders’. *Hastings Centre Report* 32 (2): 14–21.

Scholten, Matthé. 2016. ‘Schizophrenia and Moral Responsibility: A Kantian Essay’. *Philosophia* 44 (1): 205–25.

Schönfeld, Martin. 2000. *The Philosophy of the Young Kant: The Precritical Project*. Oxford: Oxford University Press.

Schreber, Daniel Paul. 2000. *Memoirs of My Nervous Illness*. Translated by Ida MacAlpine and Richard A. Hunter. New York: New York Review Books.

Schulting, Dennis. 2022. *Apperception and Self-Consciousness in Kant and German Idealism*. London: Bloomsbury.

Scully, Jackie Leach. 2004. ‘What Is a Disease?’ *European Molecular Biology Organisation Reports* 5 (7): 650–53.

Shiner, Larry. 1982. ‘Foucault, Phenomenology and the Question of Origins’. *Philosophy Today* 26 (4): 312–21.

Shorter, Edward. 1997. *A History of Psychiatry: From the Era of the Asylum to the Age of Prozac*. New York: John Wiley & Sons.

Sisti, Dominic A. 2012. 'Was Kant a Normativist or Naturalist for Mental Illness?' *Journal of Ethics in Mental Health* 7.

Smit, Houston. 2000. 'Kant on Marks and the Immediacy of Intuition'. *The Philosophical Review* 109 (2): 235–66.

Spitzer, Robert L. 1975. 'On Pseudoscience in Science, Logic in Remission, and Psychiatric Diagnosis: A Critique of Rosenhan's "On Being Sane in Insane Places"'. *Journal of Abnormal Psychology* 84 (5): 442–52.

Spitzer, Robert L., and Jean Endicott. 1972. 'Current and Past Psychopathology Scales (CAPSS)'. *Archives of General Psychiatry* 27 (5): 678–87.

———. 1978. 'Medical and Mental Disorder: Proposed Definition and Criteria'. In *Critical Issues in Psychiatric Diagnosis*, 15–38. New York: Raven Press.

Spitzer, Robert L., Jean Endicott, and Eli Robins. 1975. 'Clinical Criteria for Psychiatric Diagnosis and DSM-III'. *American Journal of Psychiatry* 132 (11): 1187–92.

Spitzer, Robert L., and Joseph L. Fleiss. 1974. 'A Re-Analysis of the Reliability of Psychiatric Diagnosis'. *British Journal of Psychiatry* 123 (587): 341–47.

Steigerwald, Joan. 2006. 'Kant's Concept of Natural Purpose and the Reflecting Power of Judgement'. *Studies in History and Philosophy of Biological and Biomedical Sciences* 37 (4): 712–34.

Stone, Alan. 2011. 'Psychiatry and Morality'. In *The Tanner Lectures on Human Values*, edited by Sterling M. McMurrin, IV:185–226. Cambridge: Cambridge University Press.

Strawson, P. F. 2006. *The Bounds of Sense*. London: Routledge.

Stroud, Scott R. 2016. 'Style and Spirit in Dreams of a Spirit-Seer: Swedenborg and the Origin of Kant's Critical Rhetoric'. *The New Centennial Review* 16 (3).

Szasz, Thomas. 1974. *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*. New York: Harper & Row.

———. 1976. 'Schizophrenia: The Sacred Symbol of Psychiatry'. *The British Journal of Psychiatry* 129 (4): 308–16.

———. 1984. *The Therapeutic State: Psychiatry in the Mirror of Current Events*. Buffalo, NY: Prometheus Books.

Tausk, Victor. 1992. 'On the Origin of the "Influencing Machine" in Schizophrenia'. Translated by Dorian Feigenbaum. *Journal of Psychotherapy Practice and Research* 1 (2): 184–206.

Telles-Correia, Diogo, Sérgio Saraiva, and Jorge Gonçalves. 2018. 'Mental Disorder-The Need for an Accurate Definition'. *Frontiers in Psychiatry* 9 (64).

Thomason, Krista K. 2021. 'The Philosopher's Medicine of the Mind: Kant's Account of Mental Illness and the Normativity of Thinking'. In *Kant on Morality, Humanity, and Legality: Practical Dimensions of Normativity*, edited by Ansgar Lyssy and Christopher Yeomans, 189–206. Cham: Palgrave Macmillan.

Turner, Robert, and Terry Jones. 2003. 'Techniques for Imaging Neuroscience'. *British Medical Bulletin* 65 (1): 3–20.

Varga, Somogy. 2011. 'Defining Mental Disorder. Exploring the "Natural Function" Approach'. *Philosophy, Ethics, and Humanities in Medicine* 6.

———. 2013. 'Vulnerability to Psychosis, I-Thou Intersubjectivity and the Praecox-Feeling'. *Phenomenology and the Cognitive Sciences* 12 (1): 131–43.

Wakefield, Jerome C. 1992a. 'Disorder as Harmful Dysfunction: A Conceptual Critique of DSM-III-R's Definition of Mental Disorder'. *Psychological Review* 99 (2): 232–47.

———. 1992b. 'The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values'. *American Psychologist* 47 (3): 373–88.

———. 1997a. 'Diagnosing DSM-IV - Part II: Eysenck (1986) and the Essentialist Fallacy'. *Behaviour Research and Therapy* 35 (7): 651–65.

———. 1997b. 'Normal Inability Versus Pathological Disability: Why Ossorio's Definition of Mental Disorder Is Not Sufficient'. *Clinical Psychology: Science and Practice* 4 (3): 249–58.

———. 1999a. 'Evolutionary Versus Prototype Analyses of the Concept of Disorder'. *Journal of Abnormal Psychology* 108 (3): 374–99.

———. 1999b. 'Mental Disorder as a Black Box Essentialist Concept'. *Journal of Abnormal Psychology* 108 (3): 465–72.

———. 1999c. ‘The Concept of Mental Disorder as a Foundation for the DSM’s Theory-Neutral Nosology: Response to Follette and Houts, Part 2’. *Behaviour Research and Therapy*, no. 37: 1001–27.

———. 2000. ‘Aristotle as Sociobiologist: The “Function of a Human Being” Argument, Black Box Essentialism, and the Concept of Mental Disorder’. *Philosophy, Psychiatry, & Psychology* 7 (1): 17–44.

———. 2003. ‘Dysfunction as a Factual Component of Disorder’. *Behaviour Research and Therapy* 41 (8): 969–90.

———. 2009. ‘Mental Disorder and Moral Responsibility: Disorders of Personhood as Harmful Dysfunctions, with Special Reference to Alcoholism’. *Philosophy, Psychiatry, & Psychology* 16 (1): 91–99.

———. 2010. ‘False Positives in Psychiatric Diagnosis: Implications for Human Freedom’. *Theoretical Medicine and Bioethics* 31 (1): 5–17.

———. 2014. ‘Wittgenstein’s Nightmare: Why the RDoC Grid Needs a Conceptual Dimension’. *World Psychiatry* 13 (1): 38–40.

———. 2021a. ‘Can a Nonessentialist Neo-Empiricist Analysis of Mental Disorder Replace the Harmful Dysfunction Analysis? Reply to Peter Zachar’. In *Defining Mental Disorder: Jerome Wakefield and His Critics*, edited by Luc Faucher and Denis Forest, 177–96. Cambridge, MA: MIT Press.

———. 2021b. ‘Can Causal Role Functions Yield Objective Judgments of Medical Dysfunction and Replace the Harmful Dysfunction Analysis’s Evolutionary Component? Reply to Dominic Murphy’. In *Defining Mental Disorder: Jerome Wakefield and His Critics*, edited by Luc Faucher and Denis Forest, 267–315. Cambridge, MA: MIT Press.

———. 2021c. ‘Is the Harmful Dysfunction Analysis Descriptive or Stipulative, and Is the HDA or BST the Better Naturalist Account of Dysfunction? Reply to Maël Lemoine’. In *Defining Mental Disorder: Jerome Wakefield and His Critics*, edited by Luc Faucher and Denis Forest, 213–49. Cambridge, MA: MIT Press.

———. 2021d. ‘Quinian Qualms, or Does Psychiatry Really Need the Harmful Dysfunction Analysis? Reply to Harold Kincaid’. In *Defining Mental Disorder: Jerome Wakefield and His Critics*, edited by Luc Faucher and Denis Forest, 133–53. Cambridge, MA: MIT Press.

Wakefield, Jerome C., and Allan V. Horwitz. 2007. *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*. New York: Oxford University Press.

Wertheimer, Michael. 2012. *A Brief History of Psychology*. 5th ed. New York: Psychology Press.

Whitley, Rob. 2012. 'The Antipsychiatry Movement: Dead, Diminishing, or Developing?' *Psychiatric Services* 63 (10): 1039–41.

Wilkerson, T. E. 1976. *Kant's Critique of Pure Reason: A Commentary for Students*. Oxford: Oxford University Press.

———. 1993. 'Essences and the Names of Natural Kinds'. *The Philosophical Quarterly* 43 (170): 1–19.

Willaschek, Marcus. 2018. *Kant on the Sources of Metaphysics: The Dialectic of Pure Reason*. Cambridge: Cambridge University Press.

Williams, Bernard. 1999. 'Knowledge and Meaning in the Philosophy of Mind'. In *Problems of the Self: Philosophical Papers 1956-1972*, 127–35. Cambridge: Cambridge University Press.

Wittgenstein, Ludwig. 1978. *Philosophical Investigations*. Translated by G.E.M. Anscombe. Oxford: Blackwell.

———. 1998. *Culture and Value*. Edited by G. H. von Wright. Translated by Peter Winch. Oxford: Blackwell.

Wright, Larry. 1973. 'Functions'. *The Philosophical Review* 82 (2): 139–68.

Wunderlich, Falk. 2018. 'Platner on Kant: From Scepticism to Dogmatic Critique'. In *Kant and His German Contemporaries*, edited by Falk Wunderlich and Corey W. Dyck, 155–72. Cambridge: Cambridge University Press.

Zachar, Peter. 2000. *Psychological Concepts and Biological Psychiatry: A Philosophical Analysis*. Vol. 28. Advances in Consciousness Research. Amsterdam: John Benjamins Publishing.

———. 2014. *A Metaphysics of Psychopathology*. Cambridge, MA: MIT Press.

Zahavi, Dan. 2005. *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge, MA: MIT Press.

Zammito, John. 2006. 'Teleology Then and Now: The Question of Kant's Relevance for Contemporary Controversies Over Function in Biology'. *Studies in History and Philosophy of Biological and Biomedical Sciences* 37 (4): 748–70.