Do Psychiatric Diagnoses Explain?

A Philosophical Investigation

by Hane Htut Maung

MA (Cantab), MB BChir, MPhil, MRCPsych

Submitted for the Degree of Doctor of Philosophy

December 2016

Abstract

***Do Psychiatric Diagnoses Explain? A Philosophical Investigation*, submitted by Hane Htut Maung for the degree of Doctor of Philosophy, December 2016**

This thesis is a philosophical examination of the explanatory roles of diagnoses in psychiatry. In medicine, diagnoses normally serve as causal explanations of patients' symptoms. Given that psychiatry is a discipline whose practice is shaped by medical traditions, it is often implied that its diagnoses also serve such explanatory functions. This is evident in clinical texts that portray psychiatric diagnoses as referring to diseases that cause symptoms. However, there are problems which cast doubt on whether such portrayals are justified. I address these problems and examine whether psychiatric diagnoses provide explanations of symptoms. The first problem is conceptual. In diagnostic manuals, psychiatric diagnoses are defined by their symptoms. This suggests that invoking them as explanations of the symptoms amounts to circularity. I argue that this can be resolved with an appropriate conceptual framework that captures the complex semantic values of diagnostic terms and their different uses in clinical discourse. I put forward such a framework based on two-dimensional semantics. The second problem is ontological. Empirical research suggests that diagnostic categories in psychiatry do not correspond to invariant causal types, but are associated with variable combinations of diverse causes that interact across biological, psychological, and social levels. Given this heterogeneity, I argue that psychiatric diagnoses fall short of paradigmatic cases of causal explanation, but that some can still provide other sorts of useful causal explanatory information. The original contribution of this thesis is the illumination of the conceptual relations between diagnoses and symptoms. This philosophical work is important, because it can be brought to valuable application in modifying psychiatric practice.

Acknowledgments

I would like to express my gratitude to the many people who greatly helped me in completing this thesis. Most of all, I would like to thank my supervisors, Rachel Cooper and Brian Garvey. Their guidance, support, and criticism were invaluable in motivating and shaping the thesis. Thanks also go to my examiners, Stephen Wilkinson and Alexander Bird, for taking the care to study my thesis and for engaging in a thoroughly stimulating dialogue during my *viva voce* examination.

I greatly benefited from discussions with a number of colleagues in philosophy, to whom I am very grateful. These are Sam Fellowes, Dan Degerman, Moujan Mirdamadi, Faye Tucker, and Tomasz Herok. I would also like to thank my former colleagues in psychiatry, Neil Hunt, Gaetano Dell'Erba, and Dane Rayment, for helping me attain the clinical knowledge that was indispensable for this project. Special thanks go to Elizabeth Fistein, Tony Hope, and Werdie van Staden for encouraging my decision to enter academic philosophy of psychiatry.

I am grateful for the financial support that enabled me to undertake this project. Funding for the project was provided by the Wellcome Trust (grant number 104897/Z/14/Z). I am also grateful for the supportive environment provided by Lancaster University's Department of Politics, Philosophy, and Religion.

Parts of this thesis have been published as peer-reviewed journal articles. A version of Chapter 3 has been published in *Theoretical Medicine and Bioethics*, and versions of Chapter 4 and Chapter 6 have been published in *Studies in History and Philosophy of Biological and Biomedical Sciences*. I am grateful to the journal editors and reviewers who offered useful comments on the manuscripts.

Finally, I would like to thank my wife Shuna Gould and our son Theo for their love, support, and patience that made the completion of this thesis possible.

Declaration

I hereby declare that this thesis is of my own composition and that it contains no material previously submitted for any other degree of qualification. The work in this thesis has been produced by me, except where due acknowledgement is made in the text. I confirm that this thesis does not exceed the prescribed limit of 80,000 words, including the main text and any footnotes but excluding the bibliography.

Hane Maung

Publications from this Thesis

A version of Chapter 3 has been published as: Maung, H. H. (2016b). "The Causal Explanatory Functions of Medical Diagnoses". *Theoretical Medicine and Bioethics*. Published online first 16th September 2016. DOI: 10.1007/s11017-016-9377-5.

A version of Chapter 4 has been published as: Maung, H. H. (2016c). "To What Do Psychiatric Diagnoses Refer? A Two-Dimensional Semantic Analysis of Diagnostic Terms". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 55: 1–10.

A version of Chapter 6 has been published as: Maung, H. H. (2016a). "Diagnosis and Causal Explanation in Psychiatry". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 60: 15–24.

Contents

1. Introduction

1.1 The problem

1.1.1 Overview

Diagnoses are central to the practice of medicine. For clinicians, they provide labels for conditions to aid communication, inform predictions about clinical outcomes, and guide therapeutic interventions. For patients, they legitimise sickness, sanction certain behaviours, and authorise access to therapeutic, social, and financial resources. In addition to these denotative, therapeutic, and social functions, diagnoses in medicine often serve as causal explanations of patients' symptoms. For example, when a patient presents with the symptom of abdominal pain, the diagnosis of acute appendicitis explains why he or she has abdominal pain by indicating the condition that is causing it. This thesis is a philosophical investigation of whether diagnoses in psychiatry similarly serve as explanations of patients' symptoms.

The ways that psychiatric diagnoses are sometimes described in clinical resources suggest that at least some clinicians consider them to have such causal explanatory roles. For example, textbooks targeted at psychiatrists, physicians, and medical students sometimes portray psychiatric diagnoses as referring to the causes or providing explanations of symptoms, as shown by the following passages:

> Depression is more common in older persons than it is in the general population. Various studies have reported prevalence rates ranging from 25 to almost 50 percent, although the percentage of these cases that are *caused by* major depressive disorder is uncertain. (Sadock and Sadock, 2008: p. 215, italics added)

> The diagnosis of antisocial personality disorder is not warranted if the symptoms can be *explained by* schizophrenia, mania or mental retardation. (Sethi, 2008: p. 109, italics added)

> Most auditory hallucinations not associated with falling asleep or waking up are *caused by* schizophrenia or depression. (Collier *et al*., 2013: p. 317, italics added)

> In other words, a positive response to screening for a psychiatric disorder in a relative of a patient increases the likelihood that a psychiatric disorder is the *cause* of the patient's symptoms. (Schneider and Levenson, 2008: p. 8, italics added)

> As an example, think of the differential diagnosis of a patient with episodes of anxiety and breathlessness. These symptoms are often *caused by* panic disorder. (Stevens and Rodin, 2010: p. 74, italics added)

Similar claims can also be found in health information resources about psychiatric disorders that are written by clinicians and targeted at the general public, as shown by the following passages about schizophrenia, generalised anxiety disorder, and major depressive disorder from Patient.info and NHS Choices, two of the leading health information websites in the United Kingdom:

> Schizophrenia is a serious mental health condition that *causes* disordered ideas, beliefs and experiences. (Patient.info, 2013, italics added)

> GAD is a long-term condition that *causes* you to feel anxious about a wide range of situations and issues, rather than one specific event. (NHS Choices, 2014a, italics added)

> Depression affects people in different ways and can *cause* a wide variety of symptoms. They range from lasting feelings of sadness and hopelessness, to losing interest in the things you used to enjoy and feeling very tearful. (NHS Choices, 2014b, italics added)

Furthermore, such claims are even made in clinical research, as shown by the following two passages from scientific papers. In the former the diagnoses of schizophrenia and major depressive disorder are invoked as explanations of anhedonia, while in the latter it is suggested that a diagnosis of major depressive disorder takes explanatory precedence over a diagnosis of chronic fatigue syndrome:

> Fourteen percent of the anhedonia was *explained by* schizophrenia, 13 percent was *explained by* depression, and 73 percent was not explained. (Loas *et al.*, 2000: p. 503, italics added)

> When a well-recognized underlying condition, such as primary depression, could *explain* the subject's symptoms, s/he was classified as having "CFS-*explained*". (Jason *et al.*, 2014: p. 43, italics added)

These passages show that psychiatric diagnoses are sometimes communicated to clinicians, researchers, and the public as if they are causes or explanations of patients' symptoms, much like the diagnoses in other medical specialties. Of course, this is not to

say that such language is universal in clinical texts. For example, the *Shorter Oxford Textbook of Psychiatry* takes great care not to refer to diagnoses as causes of symptoms, instead referring to symptoms as "occurring in" disorders (Cowen *et al.*, 2012). Nonetheless, the passages quoted above indicate that the idea that psychiatric diagnoses refer to conditions which cause symptoms has significant influence in contemporary psychiatric discourse. This is perhaps not surprising when we consider psychiatry's historical and cultural underpinnings as a medical discipline. As noted by Jeffrey Poland (2014: pp. 31–33), psychiatric practice occurs in a context shaped by medical roles and traditions, and so it is understandable that its practitioners apply to it the methods and rhetorical tropes of other medical disciplines.

However, there are worries about psychiatric diagnoses that call into doubt whether they actually do explain their symptoms. One such worry concerns the way that psychiatric diagnoses are defined. According to the most recent editions of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), which is the dominant classification system in psychiatry, psychiatric diagnoses are formally defined in terms of their symptoms, which suggests that they merely describe, rather than explain, these symptoms. Another worry concerns the natures of the causes that underlie psychiatric syndromes. Even if diagnostic terms are taken to refer to whatever causal structures underlie the symptom clusters, it is doubtful whether these causal structures exhibit enough stability for their respective diagnostic categories to have genuine explanatory value.

In this thesis, I address these worries to attain a better understanding of the epistemic functions of diagnoses in psychiatry. I begin by exploring the functions of diagnoses and the nature of diagnostic explanation in medicine more generally. I then examine the semantics of diagnostic terms with appeal to recent theories in the philosophy of language in order to tackle the conceptual problem of whether the

descriptive definitions of psychiatric diagnoses in the most recent editions of the *DSM* necessarily preclude them from referring to the causes of their symptoms. After tackling the conceptual problem, I review the current empirical evidence and theoretical models in psychiatry in order to address the ontological problem of whether the causal profiles associated with the current diagnostic categories in psychiatry are stable and repeatable enough for the diagnoses to genuinely serve as explanations of patients' symptoms in clinical practice. It is now worth laying out these conceptual and ontological problems in more detail.

1.1.2 The conceptual problem

As noted above, the *DSM* is generally considered to be the dominant system for diagnostic classification in psychiatry, with the current edition being *DSM*-5 (2013). The manual offers a standard classification and formal definitions of psychiatric diagnoses that are used by clinicians, researchers, the pharmaceutical industry, insurance companies, and policy makers. Importantly, the definitions of psychiatric diagnoses in the *DSM* are in terms of symptom criteria, as the following excerpts from *DSM*-5 demonstrate:

> The *essential feature* of delusional disorder is the presence of one or more delusions that persist for at least 1 month (Criterion A). (American Psychiatric Association, 2013: p. 92, italics added)

> The *essential feature* of a major depressive episode is a period of at least 2 weeks during which there is either depressed mood or the loss of interest or pleasure in nearly all activities (Criterion A). In children and adolescents, the mood may be irritable rather than sad. The individual must also experience at least four additional symptoms drawn from a list that includes changes in appetite or weight, sleep, and

> psychomotor activity; decreased energy; feelings of worthlessness or guilt; difficulty thinking, concentrating, or making decisions; or recurrent thoughts of death or suicidal ideation or suicide plans or attempts. (American Psychiatric Association, 2013: p. 163, italics added)

> Panic disorder *refers to* recurrent unexpected panic attacks. A panic attack is an abrupt surge of intense fear or intense discomfort that reaches a peak within minutes, and during which time four or more of a list of 13 physical and cognitive symptoms occur. (American Psychiatric Association, 2013: p. 209, italics added)

> The *essential feature* of generalized anxiety disorder is excessive anxiety and worry (apprehensive expectation) about a number of events or activities. (American Psychiatric Association, 2013: p. 222, italics added)

While the *DSM* is widely accepted as the authoritative classification system in psychiatry, it is worth mentioning that there is another diagnostic manual, namely the World Health Organisation's *International Classification of Diseases* (*ICD*), which includes a chapter dedicated to mental and behavioural disorders. The current revision is *ICD*-10 (1992). There are some mostly minor differences between *ICD*-10 and its approximate contemporary *DSM*-IV, some of which reflect their respective European and American origins (Bolton, 2008: p. 1). However, throughout their respective revision processes, the World Health Organisation and the American Psychiatric Association made efforts to bring the *ICD* and the *DSM* in line with each other (First, 2009). As a result, the similarities between them are so significant that at present they can hardly be considered to constitute distinct classifications of psychiatric disorders (Cooper, 2014: p. x). In *ICD*-10, as in *DSM*-5, psychiatric diagnoses are defined in terms of their symptoms, which the

following passages about generalised anxiety disorder and obsessive-compulsive disorder respectively demonstrate:

> The *essential feature* is anxiety, which is generalized and persistent but not restricted to, or even strongly predominating in, any particular environmental circumstances (i.e. it is "free-floating). (World Health Organisation, 1992: p. 140, italics added)

> The *essential feature* of this disorder is recurrent obsessional thoughts or compulsive acts. (World Health Organisation, 1992: p. 142, italics added)

These descriptive definitions in *DSM*-5 and *ICD*-10 suggest that psychiatric diagnoses are constituted by their symptoms. For at least two and a half centuries, it has generally been accepted in philosophy that causes are distinct from their effects (Hume, [1748] 2000; Swain, 1980; Lewis, 1986a). That is, something cannot be its own cause. Therefore, if psychiatric diagnoses merely describe clusters of symptoms as suggested by the *DSM*-5 definitions, then they cannot refer to the causes of these symptoms.

The *DSM* and *ICD* formalised the descriptive approach to defining psychiatric diagnoses, but the worry that psychiatric diagnoses merely have definitional connections with their respective symptoms had been present even before the introduction of the fully descriptive nosology in *DSM*-III (1980). In "The Myth of Mental Illness" (1960), the psychiatrist and leading figure of the antipsychiatry movement Thomas Szasz presents two arguments against the validity of the concept of mental illness. The first argument is that mental illness diagnoses are mere shorthand labels for certain kinds of behaviour, and so cannot also be invoked to refer to causes of these behaviours:

> This is obviously fallacious reasoning, for it makes the abstraction "mental illness" into a *cause*, even though this abstraction was created in the first place to serve only as a shorthand expression for certain types of human behaviour. (Szasz, 1960: p. 114)

The second argument is that mental illnesses are not genuine disorders because, unlike bodily illnesses, they are not characterised by pathophysiological lesions, but by deviations from social and moral norms. Critics of Szasz have tended to target the second argument (Kendell, 1975; Fulford, 1989; Shorter, 2011), with the first argument receiving considerably less attention. However, it is the first argument that is at the core of the conceptual problem regarding psychiatric diagnoses.

Tim Thornton (2007: p. 16) offers an interpretation of Szasz's first argument in terms of necessity and contingency, with reference to David Hume's analysis of causation in *An Enquiry Concerning Human Understanding* ([1748] 2000). According to Hume, causal connections are contingent. We perceive causes and effects as distinct events, but do not perceive any necessary connection between them. Even if the causal chain is broken down further, we only perceive a finer succession of distinct causes and effects, but not any glue between them. Hence, one can conceive one event occurring without the other. For example, while it may be the case that a particular patient's abdominal pain is caused by acute appendicitis, it is conceivable that acute appendicitis could occur without abdominal pain or that abdominal pain could occur without acute appendicitis. However, if a psychiatric diagnosis is defined by its symptoms, then the connection between the diagnosis and the symptoms is not contingent, but necessary. Since causal connections have to be contingent, it follows that the connection between a psychiatric diagnosis and its symptoms is not causal.

Szasz's argument can also be hinged at the level of language, and examined in terms of analyticity and syntheticity. According to Immanuel Kant ([1781] 1998), an analytic proposition is true in virtue of its meaning, as its predicate concept is contained in its subject concept. A classic example is the proposition, "all bachelors are unmarried". This proposition is analytically true, because the concept "unmarried" is contained in the concept "bachelor". By contrast, a synthetic proposition can only be true in virtue of its relation to the state of affairs in the world, because its predicate concept is not contained in its subject concept. For example, the proposition, "all bachelors are unhappy", is synthetic, because the concept "unhappy" is not contained in the concept "bachelor". Applied to diagnoses, the proposition, "this patient with acute appendicitis has abdominal pain", is synthetic, because the concept "abdominal pain" is not contained in the concept "acute appendicitis". However, the proposition, "this patient with panic disorder has recurrent unexpected panic attacks", is analytic, because, according to the *DSM*-5 definition, the concept "recurrent unexpected panic attacks" is contained in the concept "panic disorder". Again, this suggests that the relations between psychiatric diagnoses and their symptoms are not empirical, but definitional.

The above considerations raise serious doubts about whether psychiatric diagnoses can serve the same causal explanatory functions as medical diagnoses. This is articulated by Jennifer Radden in her paper, "Is This Dame Melancholy? Equating today's Depression and Past Melancholia" (2003). Radden notes that while the purely descriptive approach to defining and classifying psychiatric diagnoses in the most recent editions of the *DSM* does permit probabilistic predictions, it renders its diagnostic categories devoid of explanatory value. Because the connection between a psychiatric diagnosis and its symptoms is definitional rather than causal, such a diagnosis does not explain its symptoms, but merely describes them. The diagnosis of panic disorder does not explain

why a patient has recurrent and unexpected panic attacks any more than a man's bachelorhood explains why he is unmarried.

There appear, therefore, to be two kinds of talk going on regarding psychiatric diagnoses. As noted in §1.1.1, some clinical textbooks and public information resources refer to diagnoses as if they are causes of their symptoms, but the *DSM* diagnostic criteria refer to diagnoses as if they are constituted by their symptoms. On closer examination, such instances of such ambiguity are even present within the pages of *DSM*-5. The diagnostic criteria for major depressive disorder, for example, suggest that the diagnosis is constituted by its symptoms. However, the exclusion criteria for brief psychotic disorder refer to major depressive disorder as if it can be an explanation of symptoms:

> The disturbance is not better explained by major depressive or bipolar with psychotic features or another psychotic disorder such as schizophrenia or catatonia, and is not attributable to the physiological effects of a substance (e.g. a drug of abuse, a medication) or another medical condition. (American Psychiatric Association, 2013: p. 94)

The problem is that these two kinds of talk are in tension. If psychiatric diagnoses refer only to clusters of symptoms, then this suggests that they cannot be causal explanations of these symptoms.

This ambiguity might seem to be dissolved by the suggestion in the introduction of *DSM*-5 that the diagnostic criteria are supposed to be summaries, rather than complete definitions:

> The symptoms contained in the respective diagnostic criteria sets do not constitute comprehensive definitions of underlying disorders, which encompass cognitive,

> emotional, behavioral, and physiological processes that are far more complex than can be described in these brief summaries. Rather, they are intended to summarize characteristic syndromes of signs and symptoms that point to an underlying disorder with a characteristic developmental history, biological and environmental risk factors, neuropsychological and physiological correlates, and typical clinical course. (American Psychiatric Association, 2013: p. 19)

This is also the stance assumed by Lawrie Reznek (1991: p. 188), who suggests that psychopathic personality disorder does not refer to a cluster of symptoms, but to whatever causal structure explains these symptoms, just as gold refers not to the surface properties of yellowness and solidity, but to the atomic structure that explains these properties. However, this analogy is not accurate. Whereas yellowness and solidity are contingent properties that are not essential for a substance to qualify as gold (Kripke, [1972] 1980: p. 123), *DSM*-5 explicitly states that the symptom criteria for a given diagnosis are essential to the diagnosis and necessary for the diagnosis to be made, such as the presence of one or more delusions being the "essential feature" of delusional disorder (American Psychiatric Association, 2013: p. 92). Hence, in spite of the above quoted passage, the diagnostic criteria in *DSM*-5 still suggest that certain symptoms constitute part of the meaning of the diagnosis.

And so, the tension still remains between the two kinds of talk regarding psychiatric diagnoses. If a psychiatric diagnosis is supposed to point to an underlying disorder, as suggested by the above quoted passage, then the connection between the diagnosis and its symptoms would be expected to be contingent, just as acute appendicitis and abdominal pain are contingently connected, and gold and yellowness are contingently connected. However, in spite of the above passage, the *DSM*-5 criteria themselves suggest that the connection between a psychiatric diagnosis and certain symptoms is

necessary. This is the conceptual problem regarding the explanatory status of a psychiatric diagnosis, which I address towards the middle of the thesis.

1.1.3 The ontological problem

As described above, the conceptual problem regarding diagnostic explanation in psychiatry results from the diagnoses being defined by their symptoms. However, even if we take psychiatric diagnoses to refer to whatever causal structures underlie the respective symptom clusters, we face a further ontological problem. It is far from clear whether the current diagnostic categories in psychiatry actually do correspond to causal structures that are sufficiently repeatable to be explanatorily valuable with respect to individual clinical cases.

Again, this problem is related to the descriptive ways in which psychiatric diagnoses are defined and classified in the recent editions of the *DSM*. In the early editions, *DSM*-I (1952) and *DSM*-II (1968), syndromes were often defined on the basis of their supposed causes (American Psychiatric Association, 1952: p. 12). By contrast, the following edition, *DSM*-III (1980), presented a largely atheoretical system of diagnostic classification based on descriptions of observable symptoms, rather than on theoretical assumptions about aetiology. This classification system drew heavily from the symptom-based diagnostic criteria for psychiatric disorders developed by Feighner *et al.* (1972) and other related schemes. The next edition, *DSM*-IV (1994), placed less emphasis on the classification being purely atheoretical, but retained the *DSM*-III's descriptive approach to defining disorders. Following calls by the *DSM*-5 taskforce for a more theoretical approach to classification (Kupfer *et al.*, 2002), there was a modest attempt to move towards an aetiologically-informed taxonomy in *DSM*-5 (2013). However, the actual change was slight and was largely restricted to a revised chapter organisation, wherein disorders that are believed to have similar aetiologies were placed adjacent to each other. Despite this

new chapter organisation, the definitions of individual diagnoses in *DSM*-5 have remained descriptive. Correspondingly, the current *ICD*-10 (World Health Organisation, 1992) also employs a descriptive approach to psychiatric classification that is largely neutral with respect to causes.

The descriptive approach to diagnostic classification in psychiatry has been defended on the grounds that standardised operational criteria based on observable symptoms would increase reliability and facilitate communication between health professionals with different theoretical perspectives. Hence, *DSM*-III notes that "the inclusion of etiological theories would be an obstacle to use of the manual by clinicians of varying theoretical orientations" (American Psychiatric Association, 1980: p. 7). On the other hand, the downside of such an aetiologically neutral classification system is that it allows for the possibility of its diagnostic categories failing to correspond to distinctive and invariant causal structures. Such a concern is raised by Kendell and Jablensky:

> [T]he surface phenomena of psychiatric illness (i.e., the clustering of symptoms, signs, course, and outcome) provide no secure basis for deciding whether a diagnostic class or rubric is valid, in the sense of delineating a specific, necessary, and sufficient biological mechanism. (Kendell and Jablensky, 2003: p. 7)

In the recent philosophical literature, this problem has been framed as a debate about whether psychiatric disorders can be considered natural kinds (Zachar, 2000; Cooper, 2005; Beebee and Sabbarton-Leary, 2010; Kendler *et al.*, 2011; Tsou, 2013; Haslam, 2014). The worry is that if a diagnostic category does not capture a stable kind of pathological process, then we have grounds to doubt its epistemic value.

The problem also partly recalls Szasz's (1960) second argument in "The Myth of Mental Illness", previously mentioned in §1.1.2, that mental illnesses are unlike physical

illnesses because they are not characterised by causative pathophysiological lesions. It is perhaps fair to say that our scientific understanding of some psychiatric disorders has progressed since the publication of Szasz's paper and subsequent book (Shorter, 2011). However, Szasz's argument still resonates strongly. Empirical research has revealed an array of causes associated with many of the major psychiatric syndromes, but the story has not been one of characteristic lesions (Bolton, 2012). Rather, it has been one of complexity and heterogeneity at multiple levels of analysis, including the biological, psychological, and social (Poland *et al.*, 1994; Murphy, 2006; Kendler, 2008; Hyman, 2010). As such, it may be that a given diagnostic category in psychiatry does not correspond to a distinctive causal structure, but is associated with a range of possible causal pathways, each involving the complex interactions of diverse factors across different levels. The question, then, is whether this lack of unity undermines the use of the diagnosis as an explanation of a patient's symptoms in clinical psychiatry. I address this ontological problem regarding diagnostic explanation in psychiatry in the latter half of this thesis.

1.2 Significance of the research

1.2.1 Importance of the problem

The question of whether or not diagnoses in psychiatry explain their symptoms is not only of philosophical interest, but has implications for clinical psychiatry. As noted in §1.1.2, Szasz (1960) argues that psychiatric diagnoses are just shorthand labels for certain behaviours and that they are not determined by distinctive causative lesions. He presents his arguments as undermining the legitimacy of psychiatry as a scientific and medical discipline. Although it has been over half a century since Szasz first published "The Myth of Mental Illness", the key points of his arguments continue to resonate in contemporary critiques of psychiatry. Some authors appeal to the ways that psychiatric diagnoses are

defined descriptively through their symptoms to contest particular diagnoses, including post-traumatic stress disorder (Summerfield, 2001) and attention-deficit hyperactivity disorder (Saul, 2014). Other authors appeal to the failure of psychiatric diagnoses to correspond to invariant causal structures, arguing that this warrants changes in research, clinical, and educational practices (Bentall, 2003; Poland, 2014), and even suggesting that it shows psychiatric diagnoses to be little more than political devices used to enable various social arrangements (Ingleby, 1982; Moncrieff, 2010).

In the context of everyday clinical practice, the explanatory role of a diagnosis can influence how a patient perceives and responds to his or her illness. For instance, it may confer a sense of alleviation by providing the patient with the understanding of why he or she is unwell (Chiong, 2004; Kirmayer *et al.*, 2004; Jutel, 2011). Furthermore, the explanatory status of a diagnosis can influence the judgments of clinicians and patients regarding the legitimacy of the disorder. This is evident, as we shall see in Chapter 2, in cases of chronic fatigue syndrome and fibromyalgia, where the lack of medical explanations for the symptoms can frustrate patients and leave clinicians sceptical (Ware, 1992; Nettleton *et al.*, 2004).

These considerations indicate that clinicians and patients consider the ability to explain symptoms a desirable feature of a diagnosis. Hence, the issue of whether or not psychiatric diagnoses genuinely explain their symptoms is not trivial, but has ethical implications for the care of patients. As we saw in §1.1.1, psychiatric diagnoses are often communicated to the public as if they refer to conditions that cause certain symptoms. Moreover, an interview study by Young *et al.* (2008) and recent fieldwork performed by the psychologist Svend Brinkmann (2014) indicate that patients who are diagnosed with attention-deficit hyperactivity disorder do indeed assume that their diagnoses provide explanations of their problems. However, if psychiatric diagnoses do not serve as explanations of symptoms, then it is likely that patients and the wider public are being

misinformed about psychiatric diagnoses. This raises the possibility that patients are misled into believing that their symptoms are being explained, when they are merely being labelled.

Whether or not psychiatric diagnoses are explanations of symptoms also has potential implications for legal cases. As noted by Wilson and Adshead (2004), there is a strong intuition that some people with psychiatric disorders are sometimes not responsible for their actions. This is reflected by the fact that the presence of psychiatric disorder is sometimes considered a defence in criminal law. For example, under Section 2 of the *Homicide Act 1957* in England and Wales, the offence of murder can be reduced to that of manslaughter on the grounds that the defendant has diminished responsibility due to a psychiatric disorder. Section 37 of the *Mental Health Act 1983* in England and Wales also allows the defendant to be admitted to hospital on a compulsory basis instead of receiving a custodial sentence if he or she is suffering from a psychiatric disorder at the time of sentencing. Hence, a psychiatric diagnosis can influence decisions about the defendant's culpability.

The notion that a defendant is to be excused because his or her action was the product of a disorder has been criticised on philosophical grounds (Radden, 1982; Morse, 1999; Wilson and Adshead, 2004). Nevertheless, it still explicitly stated in criminal legislation. For instance, the amendment of the *Homicide Act 1957* by the *Coroners and Justice Act 2009* in England and Wales states that one of the conditions for the defence of diminished responsibility is that the mental disorder causally explains the defendant's behaviour:

> (1) A person ("D") who kills or is a party to the killing of another is not to be convicted of murder if D was suffering from an abnormality of mental functioning which – (a) arose from a recognised medical condition, (b) substantially impaired

> D's ability to do one or more of the things mentioned in subsection (1A), and (c) provides an explanation for D's acts and omissions in doing or being a party to the killing …. For the purposes of subsection (1)(c), an abnormality of mental functioning provides an explanation for D's conduct if it causes, or is a significant contributory factor in causing, D to carry out that conduct. (*Coroners and Justice Act*, 2009: s. 52 (1))

However, if the defendant's diagnosis is only a description, rather than an explanation, of his or her symptoms, then it is doubtful whether it can function as a legal defence for his or her behaviour, as it does not meet the above stated conditions for diminished responsibility. This is particularly the case if the definition of the diagnosis is partly constituted by the defendant's behaviour. For example, the *DSM*-5 definition of antisocial personality disorder includes the following symptoms:

> 1. Failure to conform to social norms with respect to lawful behaviors … 2. Deceitfulness, as indicated by repeated lying … 3. Impulsivity … 4. Irritability and aggressiveness … 5. Reckless disregard for safety … 6. Consistent irresponsibility … 7. Lack of remorse, as indicated by being indifferent to or rationalizing having hurt, mistreated, or stole from another. (American Psychiatric Association, 2013: p. 659)

As noted by Landy Sparr (2009), such personality disorder diagnoses are sometimes used as legal defences in European countries. However, the fact that the diagnosis is partly defined by harmful or unlawful behaviour suggests that its use as an excuse for such behaviour is circular. Therefore, the question of whether or not certain psychiatric

diagnoses explain their symptoms is of direct relevance to the question of whether or not these diagnoses can serve as legal defences in criminal cases.

1.2.2 Positioning of the research

This thesis is a work in philosophy of science as applied to psychiatry. Somewhat unusually for a work in applied philosophy of science, its focus is not primarily on empirical psychiatric research, but on clinical psychiatric practice. Analytic philosophy of psychiatry is a rapidly expanding discipline and since the turn of the millennium has yielded important insights into a variety of longstanding issues related to psychiatric diagnosis, including the validity of psychiatric classification (Kendell and Jablensky, 2003; Poland, 2014; Tsou, 2015), whether diagnostic categories correspond to natural kinds (Zachar, 2000; Cooper, 2005; Beebee and Sabbarton-Leary, 2010; Haslam, 2014), models of explanation in psychiatric research (Murphy, 2006; Mitchell, 2008; Kendler, 2008; Schaffner, 2008), the concept of mental disorder (Wilkinson, 2000; Bolton, 2008), and how values relate to diagnosis (Sadler, 2005; Fulford *et al*., 2006; Thornton, 2007). However, very little has been written from the angle of philosophy of science about the epistemic roles of psychiatric diagnoses in the clinical context. As such, the question of whether diagnoses serve as explanations of patients' symptoms in psychiatry is yet to be addressed in detail in the literature.

I aim to fill this gap. My original contribution to knowledge in this thesis is the improved philosophical understanding of the explanatory relations between diagnoses and symptoms in clinical psychiatry, as well as in medicine more generally. This consists, more specifically, of the application of philosophical models of explanation to unpack the epistemic roles of diagnoses in clinical practice, the development of a semantic framework informed by recent philosophy of language to analyse the seemingly paradoxical uses of diagnostic terms in psychiatric discourse, the detailed examination of

the implications that causal heterogeneity and complexity have for the explanatory roles of psychiatric diagnoses, and the consideration of how diagnoses could complement other epistemic resources to achieve better causal explanations in clinical psychiatry.

Of course, the themes in this thesis overlap with some key issues that have been discussed at length by previous authors in the philosophy of psychiatry and it is inevitable that my ideas draw heavily from the crucial insights of these authors. However, my investigation differs from this previous work in important ways. For example, I approach the topic of diagnostic categories in psychiatry from a different angle. As noted above, some authors frame this as a metaphysical issue by asking whether or not psychiatric disorders are natural kinds (Zachar, 2000; Cooper, 2005; Beebee and Sabbarton-Leary, 2010; Kendler *et al*., 2011; Haslam, 2014), while others frame it as an issue concerning what constitutes a valid psychiatric classification (Kendell and Jablensky, 2003; Poland, 2014; Tsou, 2015). In contrast, I frame it as an epistemological issue concerning whether or not the diagnostic categories in psychiatry function as explanations in the clinical context. While this covers some of the same material as the above mentioned discussions of natural kinds and psychiatric classification, I hope to show that approaching this material from the angle of explanation in clinical practice can offer novel insights into the problem of how good are our diagnostic categories in psychiatry.

My discussion of explanation in psychiatry also differs from previous discussions of explanation by other philosophers in the field. Two kinds of explanatory question regarding diagnoses in medicine and psychiatry can be distinguished (Qiu, 1989: pp. 199–200; Thagard, 1999: p. 20). The first kind of explanation, which I call disease explanation, belongs to empirical research. This is the explanation of a clinical syndrome in general. Here, the goal here is to develop a general model that brings together the relevant causal factors and mechanisms responsible for the syndrome. The second kind of explanation, which I call diagnostic explanation, occurs in the context of clinical practice. This is

where a patient presents with a set of symptoms and the physician makes a diagnosis that explains these symptoms. Much of the philosophical literature on explanation in psychiatry has focused on disease explanation. The problem is how the construction of a general model of a psychiatric disorder is possible given the problems posed by the high degrees of heterogeneity and complexity at every level of analysis (Murphy, 2006; Kendler, 2008; Mitchell, 2008). However, diagnostic explanation has not been discussed at length in the literature. My investigation stands out from previous discussions of psychiatric explanation in that it focuses on the diagnostic question, where the patient's symptoms constitute the *explanandum* and the diagnosis is the *explanans*. A significant implication of this is that it keeps the thesis relevant to clinical psychiatric practice, perhaps more so than to empirical psychiatric research.

Finally, my investigation stands out from previous work in the field by offering some novel solutions to some of the recognised problems regarding psychiatric diagnoses. For example, while the conceptual problem of whether psychiatric diagnoses can be said to explain symptoms when they are defined by these symptoms has been mentioned in the literature, as yet it has not received detailed treatment. In this thesis, I present a solution to this problem involving a new application of the philosophical theory of two-dimensional semantics that has not been attempted before. Similarly, the problem of causal heterogeneity is a recognised problem, but I present an original response by showing how heterogeneous diagnostic categories can provide other sorts of explanatory information that can be genuinely causal in quite satisfying ways. I also present what I hope to be an original defence of a deflationary approach to the problem of diagnostic classification in psychiatry, based on the recognition that clinical psychiatry has another epistemic resource, the individualised formulation, which can complement the categorical diagnosis to arrive at a more satisfactory explanation of the patient's symptoms.

1.2.3 Scope of this thesis

This thesis, as far as I am aware, is the first detailed philosophical investigation of the explanatory functions of psychiatric diagnoses in the context of clinical practice. While I hope to contribute novel insights into the roles and uses of diagnoses in medicine and psychiatry, I concede that my discussion looks at just one aspect of diagnosis from a fairly narrow disciplinary perspective. Specifically, it examines the diagnosis as an explanatory hypothesis about the patient's clinical presentation through the lens of analytic philosophy of science. As such, it must be made clear from the outset that my discussion is not intended to offer a comprehensive treatment of the many other interesting and important issues concerning diagnosis, to some of which I alluded at the beginning of §1.2.2. Three of these issues warrant special mention here due to the prominent positions they occupy in the literature on diagnosis in the philosophy of psychiatry and the philosophy of medicine.

The first issue of note concerns the concept of disorder. The question here is what demarcates disorder *qua* medical problem from other kinds of problem, such as moral and social problems, or indeed from normal health. Interestingly, this debate is also partly inspired by Szasz's arguments in "The Myth of Mental Illness" (1960). As noted in §1.2.1, Szasz presents his arguments as undermining the status of psychiatry as a medical discipline. Mental illnesses, he argues, are not genuine disorders, but "problems in living". Since Szasz initially presented his arguments, numerous theorists have offered philosophical accounts of disorder, with psychiatric disorders often featuring at the centre of the discussion. In reply to Szasz, the psychiatrist Robert Kendell (1975) defends psychiatry as a medical discipline by suggesting a naturalistic account of disorder based on reduced life expectancy and fertility. A more sophisticated naturalistic account of disorder is offered by the philosopher Christopher Boorse (1977), who argues that disorder is a substandard statistical deviation from normal biological function. Perhaps

one of the most influential philosophical accounts of disorder is Jerome Wakefield's (1992) harmful dysfunction analysis, according to which a condition's disorder status is not determined solely by a factual claim about the presence of biological dysfunction, but also requires an evaluative judgement that the condition is harmful. Other normative accounts of disorder include those based on action failure (Fulford, 1989) and on flourishing (Megone, 1998). In more recent years, philosophers have explored more nuanced and pluralistic discussions of disorder that depart from the accounts based on single criteria (Cooper, 2005; Bolton, 2008). There have also been more focused discussions regarding whether or not particular conditions should be considered disorders, including grief (Wilkinson, 2000), ageing (Schramme, 2013), attention-deficit hyperactivity disorder (Saul, 2014), and obesity (Hoffman, 2016).

The problem of demarcation between disorder and non-disorder has been raised in relation to diagnostic validity. Wakefield (1992), for instance, suggests that a diagnostic category is valid if it discerns genuine cases of disorder from cases of non-disorder. However, whether or not a diagnostic category is valid in this sense is a different issue from whether or not it serves as an explanation of a set of symptoms. Of course, I concede that the two issues are related, as a diagnostic category's lack of explanatory value might provide a reason to suspect that the condition denoted by the category should not be considered a genuine disorder. Nonetheless, disorder status and explanatory function can come apart, and so are not necessarily connected. For example, the category of menopause can be invoked to explain a woman's hot flushes, reduced libido, and cessation of menstruation, but it does not follow from this explanation that the condition denoted by the category is a medical disorder. Therefore, once one has established whether or not a given diagnosis serves as an explanation of a set of symptoms, whether or not the condition it denotes should be considered a disorder remains a further question. Addressing this further question would require commitment

to a particular account of disorder, which is beyond the intended scope of my investigation.

The second issue of note concerns the roles of values in diagnosis. Once again, the debate can be traced back to Szasz (1960), who argues that mental illnesses are not genuine medical disorders, because they are characterised by deviations from social and moral norms. As noted above, some theorists responded to Szasz by suggesting accounts of psychiatric disorder that do not invoke values (Kendell, 1975; Boorse, 1977). Other theorists, including Bill Fulford (1989), John Sadler (2005), and Tim Thornton (2007), acknowledge that psychiatric diagnoses are value-laden, but argue that this value-ladenness does not necessarily undermine their scientific validity. Sadler proposes that values are involved at every level in psychiatric diagnosis, including the diagnostic criteria, the stereotype of the condition, the judgement about its disorder status, and the very enterprise of constructing a classification system. I fully accept that values are involved in diagnosis and that understanding their roles is important. However, a comprehensive analysis of values in psychiatric diagnoses is not necessary for my investigation into whether or not psychiatric diagnoses function as explanations of symptoms. The two issues can, for the most part, be kept apart, although I concede that they may be contingently related. As such, I discuss value-ladenness in this thesis only where it is directly relevant to the question of the explanatory role of a diagnosis.

The third issue of note concerns the relation of diagnosis to evidence-based medicine. This is an important topic, particularly given that evidence-based medicine has been portrayed as being a new and dominant paradigm in clinical medicine (Evidence-Based Medicine Working Group, 1992). Accordingly, in recent years, the philosophy of medicine literature has swelled with highly welcome critical discussions of problems in evidence-based medicine, including the hierarchy of study designs, the epistemic purpose of randomisation, the role of tacit knowledge in clinical judgement, and the evidential

value of mechanistic reasoning (Bluhm, 2005; Thornton, 2007; Worrall, 2007; Bird, 2011; Howick, 2011; Andersen, 2012). The topics of diagnosis and causal explanation are certainly related to the topic of evidence-based medicine. For example, populations for statistical trials are usually defined in part by diagnostic criteria, and there is an active debate regarding the respective roles of statistical evidence for correlations and mechanistic causal explanations in guiding clinical decisions (Clarke *et al.*, 2014: p. 346). However, for the specific purposes of this thesis, the discussion of whether psychiatric diagnoses explain symptoms and the discussion of how diagnoses relate to evidence-based medicine can, to a significant degree, be kept apart. Hence, while I do suggest that diagnoses and the explanations they provide can help to inform predictions and guide therapeutic interventions, I do not intend in this thesis to examine precisely how, or indeed whether, such epistemic resources can complement an evidence-based medicine approach.

In summary, the above mentioned issues can be seen as being orthogonal to, rather than challenging, the analysis I provide in this thesis. Of course, there are areas where the topic of my investigation and these other issues meet, and I would be delighted if it turns out that my discussion helps to shed new light on these issues. However, given the specific focus of my investigation, I do not intend to explore these areas in detail in the current thesis.

1.3 Synopsis

The rest of the thesis proceeds as follows. Chapter 2 considers the variety of epistemic, instrumental, and semiotic roles that diagnoses normally serve for clinicians, patients, and society. The overall aim of the chapter is to show that the explanatory role of a diagnosis is of particular importance, because it provides justificatory support for many of its other roles. I back this up with some evidence from sociological research on medically

unexplained syndromes, which suggests that diagnoses that fail to explain also often fail to reliably inform predictions, effectively guide therapeutic interventions, elicit support from social services, and provide hope for patients. I then explore some of the arguments made by prominent critics of psychiatry, which suggest that these concerns may also apply to psychiatric diagnoses.

Chapter 3 examines in more detail how diagnoses in medicine normally explain patients' symptoms. The general aim is to explicate the nature of the explanatory relation in a paradigmatic example of diagnostic explanation in medicine, which can serve as a point of comparison for my later discussion of diagnosis in psychiatry. This proceeds through consideration of models of explanation in the philosophy of science and their adequacy when applied to the medical context of diagnosis. I begin by considering Carl Hempel's (1965a) covering law account of scientific explanation and showing why it does not adequately capture the way in which a diagnosis explains a patient's symptoms. Rather, the nature of diagnostic explanation in medicine is best captured by a causal model of explanation. I endorse the proposal by Margherita Benzi (2011) that many medical diagnoses, though by no means all, are causal explanations based on particulars. That is to say, they explain by indicating the actual causes of the symptoms in individual cases, rather than by subsuming them under general causal regularities. However, in addition to making a simple causal claim of the form "*C* causes *E*", I argue that the diagnostic explanation also relies on some mechanistic knowledge of how *C* produces *E* to make the causal connection intelligible. Drawing on the work of Kenneth Schaffner (1986) and Jeremy Simon (2008), I suggest that this knowledge of mechanisms is supplied by the theoretical framework in which the clinician operates.

Chapter 4 addresses the conceptual problem described in §1.1.2. The descriptive definitions of psychiatric diagnoses in the *DSM* suggest that they refer to clusters of symptoms. Given that causes are distinct from their effects, this might seem to suggest

that diagnoses in psychiatry cannot serve as causal explanations of patients' symptoms in the ways that many medical diagnoses do as described in Chapter 3. In this chapter, I argue that this is not necessarily so. The argument proceeds through examination of the semantics of diagnostic terms with appeal to theories of reference in the philosophy of language. I begin by considering Jennifer Radden's (2002) distinction between descriptive and causal conceptions of diagnostic terms, and the view suggested by Carl Hempel (1965b) and Paul Thagard (1999) that the historical development of a diagnostic term involves a progressive change from the former to the latter. A problem with this is that it implies radical incommensurability between older and newer conceptions of a diagnostic term (Feyerabend, 1962; Kuhn, 1962; Fleck, [1935] 1981). This is untenable, because it contradicts the intuition that scientific discoveries do not merely involve changes in the meanings of disease terms, but actually do increase our understanding of the respective diseases. I then look at how the causal theory of reference developed by Saul Kripke ([1972] 1980) and Hilary Putnam (1975a) can offer a more reasonable account of the meanings of diagnostic terms that avoids the implication of radical incommensurability. In spite of its strengths, a problem with a pure causal theory of reference is that it relegates the symptoms of psychiatric disorders to mere contingent features of the diagnoses, which contradicts the fact that such symptoms are often necessary conditions for applying the diagnoses according to *DSM*-5. To resolve the problem, I draw on the conceptual framework of two-dimensional semantics, as developed by Robert Stalnaker (1978), David Chalmers (1996), and Frank Jackson (1998). Such a framework permits a semantic pluralism that accommodates the actuality of diagnostic terms being defined through their symptoms, yet being used to refer to the putative causes of these symptoms.

Chapter 5 moves on to the ontological problem described in §1.1.3. Although the solution to the conceptual problem presented in Chapter 4 shows that symptom-based

descriptive definitions do not necessarily preclude psychiatric diagnoses from alluding to the causes of these symptoms, whether categorical diagnoses in psychiatry actually provide satisfactory causal explanations of individual patients' symptoms is also dependent on whether we have enough scientific understanding of these causes and, more fundamentally, on whether the diagnostic categories are respectively associated with distinctive causal profiles that are sufficiently invariant across cases. In this chapter, I review the current findings from empirical research into psychiatric aetiology for some disorders, paying special attention to the example of major depressive disorder. I use this example to illustrate the problems of causal heterogeneity and complexity that are associated with most psychiatric diagnoses. These problems suggest that psychiatric disorders cannot be conceptualised in simple essentialistic terms. In other words, the diagnostic categories in psychiatry do not correspond to distinct and invariant causative pathologies, but are associated with variable ranges of possible causal pathways, each involving complex interactions between diverse biological, psychological, and social factors. I review some recent attempts to conceptualise psychiatric disorders as homeostatic property clusters (Borsboom, 2008; Beebee and Sabbarton-Leary, 2010; Kendler *et al.*, 2011; Tsou, 2013), an idea introduced by the philosopher of biology Richard Boyd (1999) to analyse kinds that are constituted by clusters of unnecessary and insufficient properties that are connected via contingent causal relations. I then present some problems for homeostatic property cluster accounts of psychiatric disorders. Finally, I consider whether the above considerations also apply to common psychiatric diagnoses other than major depressive disorder, including schizophrenia, bipolar disorder, generalised anxiety disorder, the dementias, panic disorder, obsessive-compulsive disorder, and some of the personality disorders.

Chapter 6 examines the implications of the problems discussed in Chapter 5 for the explanatory functions of psychiatric diagnoses. To address the problems of causal

heterogeneity and complexity in psychiatry, theorists have suggested the respective strategies of idealisation (Murphy, 2006) and theoretical pluralism (Kendler, 2008; Mitchell, 2008) in disease explanation. With respect to diagnostic explanation, though, such heterogeneity makes them fall short of the paradigmatic case in medicine, described in Chapter 3, where a diagnosis picks out a specific cause of the patient's symptoms. Nonetheless, I argue that some psychiatric diagnoses, though by no means all, can still supply different sorts of clinically relevant causal information. In particular, I suggest that some psychiatric diagnoses provide negative information to exclude certain medical disorders as causes of the patients' symptoms, some provide probabilistic or disjunctive information about the range of possible causal processes that could be contributing to the patients' symptoms, and some provide causal information about the relations between the symptoms themselves. I also discuss the limitations of these sorts of causal explanatory information and suggest some psychiatric diagnoses to which they do not apply.

Chapter 7 explores the normative and methodological implications for clinical psychiatric practice of the above issues concerning diagnostic explanation. As noted in Chapter 6, categorical diagnoses in psychiatry fall short of the paradigmatic explanatory diagnosis in medicine, although some may provide more modest sorts of causal explanatory information. I consider three strategies for modifying and improving the discourse and practices regarding diagnoses in psychiatry. The first strategy is to amend the ways in which diagnoses are communicated in psychiatric discourse. The problems of causal heterogeneity and complexity suggest that psychiatric diagnoses are often misleadingly essentialised, which Nick Haslam (2014) argues can encourage harmful stigma. I propose that this warrants modification of our language in psychiatry, so that psychiatric diagnoses and whatever explanatory information they might supply are conveyed more accurately to people. The second strategy involves revising diagnostic

classification so that the categories correspond to more distinctive and stable causal structures (Poland *et al.*, 1994; Bentall, 2003; Murphy, 2006; Tsou, 2015). While this is an epistemically respectable project, I argue that there are serious challenges that make it unlikely for a successful aetiologically-based classification to be implemented in the near future. The third strategy, which I endorse, is to supplement the categorical diagnosis with an individualised formulation (World Psychiatric Association, 2003). I show how a categorical diagnosis and an individualised formulation can complement each other to arrive at a more satisfactory causal explanation of the patient's symptoms in the particular case. The upshot is that despite being causally heterogeneous, a psychiatric diagnosis can still serve an important role in the development of a causal explanation. However, again, the quality of the explanation remains limited by our incomplete scientific understanding of the mechanisms through which different causal factors interact, as well as by our ability to match certain causal factors to particular patients.

Chapter 8 is the conclusion of the thesis. Here, I recapitulate my main points and summarise my answer to the main research question. I also tentatively reflect on some of the further questions raised by my investigation that would be interesting to address in future research.

2. The Functions of Diagnoses

2.1 Introduction

The diagnosis is a key concept in contemporary medical practice, and serves a variety of functions for clinicians, for patients, and for society more broadly. This chapter explores these functions and some of the ways in which they are connected. More specifically, I argue that many of these functions receive justificatory support from the explanatory role of the diagnosis. My overall aim is to show why it is desirable for diagnoses to serve as explanations of patients' symptoms. This is significant for the thesis as a whole because it highlights important implications for the status of psychiatry if it turns out that its diagnoses do not explain.

I proceed as follows. In §2.2, I introduce the variety of functions served by diagnoses in clinical practice. Because the intention in this section is to provide a general overview, I draw on examples from across the whole of medicine, including psychiatry and the various specialties of bodily medicine. In §2.3, I argue that the explanatory function of a diagnosis provides justificatory support for many of the other functions. I substantiate this with evidence from the medical and sociological literature regarding diagnoses that fail to explain, namely the so-called medically unexplained syndromes, and the implications of such explanatory failure on the other roles of the diagnoses. I then consider why these concerns might also apply to psychiatric diagnoses.

Before I go on further, I would like to clarify some terminology and distinguish between two commonly used meanings of "diagnosis". Mildred Blaxter (1978) notes that "diagnosis" is an ambiguous term that can refer to either a category or a process. A clinician may use the term to denote the condition from which the patient is suffering, such as "the diagnosis is acute appendicitis", or to indicate how this conclusion can be reached, such as "the diagnosis is clinical and radiological". Similarly, John Sadler (2004:

p. 166) distinguishes "diagnosis-as-denotative-signifier" from "diagnosis-as-epistemic-act", and Kazem Sadegh-Zadeh (2012: p. 110) distinguishes "diagnosis" from the process of "diagnostics". To avoid this ambiguity, I reserve the term "diagnosis" to refer to the categorical conclusion and refer to the process leading to the conclusion as the "diagnostic process".

2.2 The various functions of diagnoses

2.2.1 Hypothesis

The clinical consultation between patient and clinician usually begins with the clinician taking a history from the patient to elicit his or her symptoms and other relevant information, examining the patient to elicit any signs, and reviewing any available investigation results (Stanley and Campos, 2013). In practice, one or more of these steps may be omitted, depending on the particular scenario. For instance, in an emergency scenario involving loss of awareness, the patient is unable to provide a history, and the clinical team have to rely on examination signs and investigation results to make a diagnosis. Conversely, in general practice, many diagnoses are informed by the symptoms and signs, without laboratory or radiological investigations being requested. Nonetheless, these minor differences aside, the diagnostic process normally begins with the gathering of a flexible combination of symptoms, signs, and investigation results, henceforth referred to as patient data.

After the patient data is gathered and consolidated, a diagnosis is inferred from the patient data. Further investigations may then be undertaken to acquire evidence that could support or undermine the diagnosis. Usually, several possible diagnoses are initially stipulated and further assessment is undertaken to help select the correct diagnosis from the list of possibilities, a practice known as differential diagnosis (Longmore *et al.*, 2014: p. 13). For example, after assessing a patient with severe chest pain, a doctor may

stipulate myocardial infarction, pulmonary embolism, and gastro-oesophageal reflux disease as potential diagnoses. After further investigations reveal a positive troponin result that supports the diagnosis of myocardial infarction, a negative D-dimer result that undermines the diagnosis of pulmonary embolism, and no further evidence of gastro-oesophageal reflux disease, the doctor may then conclude that the correct diagnosis from these possibilities is myocardial infarction.

The diagnosis, then, functions as a testable hypothesis about the patient's condition that is informed by the patient data. Indeed, several authors have commented on the similarity between the diagnostic process in medicine and hypothesis formation in science, and consider medical diagnoses to be akin to scientific hypotheses (Rzepiński, 2007; Sadegh-Zadeh, 2012; Aliseda and Leonides, 2013; Stanley and Campos, 2013; Willis *et al.*, 2013). For example, Willis *et al.* suggest that although its conditions are less controlled than those in a laboratory, the diagnostic process is "an example of science in action" (Willis *et al.*, 2013: p. 501). This is both an observation about the scientific knowledge of diseases that is crucial to medicine and about the kinds of method used by clinicians to form diagnoses.

In light of the observed similarities between diagnostic process and hypothesis formation, some theorists have applied the resources of philosophy of science to analyse the inferential practices that take place in the diagnostic process. One popular and plausible view is that the diagnostic process involves abductive reasoning, or inference to the best explanation (Aliseda and Leonides, 2013; Stanley and Campos, 2013). Stanley and Campos defend this view by arguing that neither deduction nor induction are sufficient for the generation of a diagnostic hypothesis. They argue that deduction and induction are too restrictive, because they are limited to the application of general laws or the extrapolation of previously observed patterns to new cases, while the diagnostic process often involves reference to phenomena that are not explicit in the supporting

evidence, which suggests the use of abduction. I also suggest that inference to the best explanation provides a good description of the clinical practice of differential diagnosis, whereby a doctor considers several potential diagnoses before committing to one as the correct diagnosis. Another popular view is that the diagnostic process involves statistical inference (Ledley and Lusted, 1959; Westmeyer, 1975). Willis *et al.* (2013) assume a more pluralist view and suggest that the diagnostic process draws on many kinds of reasoning, including deductive reasoning, inductive reasoning, falsification, inference to the best explanation, and statistical inference.

2.2.2 Explanation

We have just seen that the diagnostic process is akin to scientific hypothesis formation, whereby the diagnosis is inferred from a set of patient data, consisting of symptoms, signs, and investigation results. In turn, it is often the case that the diagnosis explains the patient data. The idea that diagnoses in medicine often and ideally function as explanations of patients' symptoms is generally accepted in the philosophical and medical literature, with many authors endorsing the view that they are causal explanations:

> Discomfort makes the patient think that something is wrong with him, and a why-question arises in his mind. … He complains to the physician of these symptoms. … All these clinical manifestations (symptoms, signs and laboratory data) require an explanation from the physician, and finally a diagnosis is reached. (Qiu, 1989: p. 199)

> When a patient goes to a physician with a set of complaints and symptoms, the physician's first task is to make a diagnosis of a disease that explains the symptoms. (Thagard, 1999: p. 20)

> To solve a clinical diagnostic problem means first to recognize a malfunction and then to set about tracing or identifying its causes. The diagnosis is thus an explanation of disordered function, where possible a causal explanation. (Schwartz and Elstein, 2008: p. 224)

> Once formulated, however, a diagnosis can be synthetically described, from a statistical viewpoint, as a *relation* between a set of findings (signs, symptoms, laboratory test results) and a certain pathological condition attributed to the patient. What kind of relation? According to a common opinion among experts in computational models, medical diagnoses express *explanatory* relations …" (Benzi, 2011: p. 365)

> It is uncontroversial in the medical literature that the ideal diagnosis is a biomedical causal explanation. … Such a diagnosis posits a physiological cause for a set of physical signs and symptoms. (Cournoyea and Kennedy, 2014: pp. 928–929)

And so, there is often a bidirectional epistemic relation between the diagnosis and the patient data. The diagnosis is inferred from the patient data and the patient data is explained by the diagnosis. I analyse in detail the nature of the explanatory relation in Chapter 3, while focusing in the current chapter on why this explanatory function is deemed desirable.

2.2.3 Prediction

In addition to *post hoc* explanation of patient data, a diagnosis serves a predictive function. The clinician is very often able to make reliable predictions about the likely future outcome for a patient based on the diagnosis. These include predictions about the

prognosis, which consists of the clinical course and likelihood of survival, predictions about potential complications, and predictions about responses to treatments. For instance, the diagnosis of acute appendicitis informs the clinician that the patient's condition is likely to deteriorate rapidly without treatment, that a potential complication is peritonitis, and that a good recovery is likely following an appendicectomy. Similarly, the diagnosis of common cold suggests that the patient is most likely to recover completely without treatment within a few days, but also that there is a risk of sinusitis as a potential complication. Therefore, a diagnosis serves the epistemic function of supporting inductive inferences about the future.

2.2.4 Intervention

It is uncontroversial that an important function of a diagnosis is guiding intervention. Indeed, theorists have proposed that the value of the diagnosis must be considered relative to the therapeutic goals of medicine. Caroline Whitbeck argues that the diagnosis is "aimed at obtaining the best medical outcome for the patient" (Whitbeck, 1981: p. 326), while Annemarie Jutel (2011: p. 21) notes that the diagnostic process is very often motivated by the goal to ascertain the correct treatment. Therefore, a diagnosis not only has epistemic significance, but also instrumental utility in guiding treatment, which makes it a key component of practical reasoning in medicine.

As alluded to earlier, this interventional function of a diagnosis is supported by its predictive function. A diagnosis can inform predictions about likely responses to treatments, and so can guide therapeutic decision making. For example, a clinician can predict from the diagnosis of acute appendicitis that the patient is likely to make a good recovery following an appendicectomy, thus supporting the decision to intervene therapeutically with an appendicectomy. Similarly, a clinician can predict from the diagnosis of common cold that the patient is likely to recover without any specific

treatment, thus supporting the decision not to prescribe an antibiotic. This interventional function of a diagnosis is greatly aided by the knowledge provided by evidence-based medicine. Later in §2.3.1, I show how it can also be further supported by the explanatory function of the diagnosis.

2.2.5 Denotation

As well as the above mentioned epistemic functions, the hypothesised diagnosis serves a linguistic function as a denotative label for the condition with which the patient is presenting. It comprises a term that is understood to refer to a state of affairs in the patient, such as "myocardial infarction" referring to ischaemic necrosis of the myocardium due to coronary artery occlusion. Furthermore, as noted by physician and psychoanalyst Michael Balint (1964: p. 25), such a diagnostic term provides a useful shorthand description that organises a variety of disparate clinical features into a unified phenomenon. This is important, because it facilitates the communicative exchanges of clinicians. Hence, diagnostic terms constitute part of a common language with which clinicians can reliably and concisely convey clinical information to each other. I offer a more detailed analysis of the semantics of diagnostic terms in Chapter 4.

2.2.6 Classification

Denotation is closely related to classification. Designating a condition with a specific term implies conceptually distinguishing it from other conditions. The diagnosis of myocardial infarction specifically denotes ischaemic necrosis of the myocardium, which is taken to be conceptually distinct from, for example, inflammation of the pericardium or dissection of the aorta. Moreover, the diagnostic term is not merely taken to denote an individual instance of the condition, but represents a generalised category. Hence, the condition

denoted by the diagnosis is often considered to be a repeatable type, of which individual cases are tokens (Sadler, 2005: pp. 419–420; Sadegh-Zadeh, 2012: p. 172).

Diagnostic terms, then, demarcate and classify diseases into clinically significant categories. Hence, Annemarie Jutel claims that the diagnosis is "one of medicine's most powerful classification tools" (Jutel, 2011: p. 15). This is reflected by the profound and pervasive influences of formal diagnostic classification systems on public policy, health insurance, and pharmaceutical research (Cooper, 2005: p. 1). As noted in Chapter 1, two of the leading formal diagnostic classification systems in current usage are the World Health Organisation's *International Classification of Diseases* (*ICD*), now in its tenth revision (1992), and, in the field of psychiatry, the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), now in its fifth edition (2013). However, even outside the official taxonomies of *ICD*-10 and *DSM*-5, the classificatory functions of diagnoses are deeply embedded in everyday clinical practice. For example, the *Oxford Handbook of Clinical Medicine* (Longmore *et al.*, 2014), which is considered an indispensable resource for medical students and physicians, organises diagnoses into cardiovascular disorders, respiratory disorders, gastrointestinal disorders, endocrine disorders, infectious diseases, malignancies, and so on.

The classificatory function of a diagnosis complements some of the other epistemic functions discussed above. First, it supports the predictive function discussed in §2.2.3. The characterisation of clinical phenomena into a disease that is considered to be categorically distinct from other diseases reflects the assumption that instances of this disease share similarities that are theoretically important and inductively powerful (Cooper, 2012: pp. 61–64). For example, the acceptance of acute appendicitis as a distinct diagnostic category suggests that clinicians recognise that cases of acute appendicitis behave alike in some clinically significant respects. Hence, they can infer that new cases of acute appendicitis will also behave in similar ways. Second, the classificatory function

of the diagnosis also supports the function of the diagnosis in guiding therapeutic intervention. More specifically, it supports the generalisation of a given treatment strategy from past cases to future cases. Because a diagnosis groups together cases under a category based on clinical similarity, an inductive generalisation can be made from the observed treatment responses in studied cases with the diagnosis to all cases with the diagnosis, thus allowing the development of an evidence-based treatment guideline.

2.2.7 Normative

The functions discussed so far have been largely descriptive, that is, they concern the role of the diagnosis in picking out a biological state of affairs that is assumed to be part of the external world, albeit occurring within the body of the patient. However, the diagnosis also has a normative function. Assigning a diagnosis to a patient does not only pick out a state of affairs, but usually implies the evaluative judgement that this state of affairs is abnormal (Bolton, 2008: pp. xiii–xiv).

More specifically, the diagnosis usually implies that the patient has a medical disorder. According to Jerome Wakefield (1992) it is important that a diagnostic category discerns cases of genuine disorder from non-disordered cases, such as variants of normality. While this has some plausibility, I argue that it is not necessarily the case that the condition picked out by a diagnosis has to be considered a disorder, as it is also possible for a diagnosis to indicate non-disorder. An example, previously mentioned in Chapter 1, §1.2.3, is the diagnosis of menopause to account for a woman's complaints of hot flushes, reduced libido, and cessation of menstruation. Nonetheless, even such a diagnosis of non-disorder implies an evaluative judgement about the status of the patient's condition based on the standards of normality and abnormality that are assumed by the medical profession.

The normative function of a diagnosis is often used to offer vindication for some its other functions. With respect to the interventional function discussed in §2.2.4, for example, the normative judgment regarding whether the patient's condition is a disorder informs the decision about whether medical intervention is appropriate at all. As we shall see, this normative function is also closely connected to the semiotic and social functions of a diagnosis.

2.2.8 Semiotic

So far, I have discussed functions of a diagnosis that are useful for the clinician. However, a diagnosis can also serve a useful function for the patient. More specifically, it functions as a "semiotic mediator", or a meaningful label which the patient can use to understand and act upon his or her condition (Brinkmann, 2014). For example, a diagnosis could be taken by the patient as legitimising his or her illness, thus validating his or her personal experience of being unwell as something that deserves to be taken seriously. This draws on the above mentioned normative function of the diagnosis. More broadly, Carl Elliott (1999) proposes that a diagnosis can influence the narrative by which one interprets one's life and shapes one's future. When the effect of the diagnosis on the one's life narrative is significant, such as with a chronic, untreatable, or potentially fatal condition, it can profoundly reorganise one's sense of personal identity and attitude towards what is valuable in life. This reorganisation of the sense of identity can also be collective (Jutel, 2011: p. 11). For example, Roth and Nelson (1997), in their qualitative study of patients diagnosed with human immunodeficiency virus (HIV) infection, found the construction of HIV-positive identities and membership in the HIV-positive community to be prominent themes in the patients' narratives.

Often, the semiotic function served by diagnosis can be helpful for the patient, as it can enable the patient to plan his or her life accordingly. This can be the case even if the

diagnosis is of a serious condition. For instance, polycystic kidney disease is an autosomal dominant inherited disease associated with progressive renal failure and a significantly increased risk of subarachnoid haemorrhage (Longmore *et al.*, 2014: p. 312). A diagnosis of polycystic kidney disease could enable the patient to take measures to control his or her blood pressure, attend regular neuroimaging scans to screen for cerebral aneurysms, consider the possible need for dialysis in the future, and make an informed decision about family planning in light of it being possible that his or her children could inherit the disease. However, it should also be recognised that a diagnosis could also have a harmful effect on a person's life narrative. Rachel Cooper (2012) explores the ways in which people's narratives are influenced by the diagnosis of antisocial personality disorder, which is defined in *DSM*-5 by a number of character traits, including "repeated lying … aggressiveness … disregard for safety of self or others … irresponsibility" (American Psychiatric Association, 2013: p. 659). Given the moral undesirability of these traits, the diagnosis makes it very difficult for the patient to construct a good narrative of his or her life. Drawing on posts on an internet support group for people with antisocial personality disorder, Cooper (2012: pp. 65–66) observes that people tend to respond to the diagnosis in one of three ways. Some challenge the diagnosis. Others consider the diagnosis to legitimise their immoral behaviours and embrace the idea that they are bad people, arguably leaving them worse than they have been before receiving the diagnostic label. Others are left uncertain about what to do with the diagnosis and feel abandoned by mental health professionals. This suggests that while diagnoses can serve many helpful functions, they also have the potential to cause iatrogenic harm.

2.2.9 Social

Finally, a diagnosis has social implications beyond the clinical interaction between the patient and the clinician. Kazem Sadegh-Zadeh (2012: 336–339) characterises the

diagnosis as a performative speech act that generates a social status for the patient, much like a judicial verdict. For instance, he observes that the utterance, "you have acute appendicitis", explicitly appears as a simple description, but expresses the implicit performative, "*I assert that* you have acute appendicitis" (Sadegh-Zadeh, 2012: pp. 55–56). This speech act influences attitudes and behaviours at individual, institutional, and cultural levels.

At the level of individual behaviour, John Sadler (2005: pp. 421–422) notes that the diagnosis endows the clinician with certain privileges. These might include initiating pharmacological treatment, surgical intervention, psychological therapy, and potentially invasive testing. Hence, the social function of the diagnosis is closely related to its interventional function discussed in §2.2.4 and its normative function discussed in §2.2.7.

At an institutional level, a diagnosis entitles the patient to therapeutic, supportive, and financial resources to which he or she had not previously been entitled. For example, the diagnosis of myocardial infarction entitles the patient to a hospital bed, nursing care, laboratory and radiological investigations, medical and surgical interventions, rehabilitation, and outpatient follow-up after discharge into the community. When the illness is more chronic and disabling, a diagnosis can also authorise the patient's access to further supportive and financial resources, including attendance to support groups, carer input, disability benefits, and supported accommodation.

At the level of culture, a diagnosis legitimises sickness and sanctions certain kinds of behaviour (Jutel, 2011: p. 7). The sociologist Talcott Parsons (1951: pp. 436–437) proposes that the patient is thrust into a "sick role", which bestows on him or her certain rights and duties. The patient's rights are to not be considered responsible for his or her illness and to be exempt from some of his or her normal obligations. These are reflected by the intuition that the sick person deserves sympathy and the fact that is sickness is considered a legitimate reason for absence from work. The patient's duties are to try to

get well and to seek appropriate medical care. Again, this draws on the normative function of the diagnosis discussed in §2.2.7.

The sanctioning of certain kinds of behaviour is also relevant in the legal setting. As previously noted in Chapter 1, §1.2.1, psychiatric diagnoses can function as defences and influence sentencing in criminal law. Relevant legislations include Section 2 of the *Homicide Act 1957* and Section 37 of the *Mental Health Act 1983* in England and Wales. A medical diagnosis can also be used to support claims for damages in civil law. An example is the *Mesothelioma Act 2014*, which allows patients diagnosed with mesothelioma to receive damages for past asbestos exposure.

2.3 When diagnoses fail to explain

2.3.1 The importance of the explanatory function

The various functions discussed in §2.2 make the diagnosis a valuable epistemic resource in clinical practice. In this current section, I focus my attention specifically on the explanatory function of the diagnosis which I briefly mentioned in §2.2.2. In particular, I argue that this explanatory function is important because it provides justificatory support for many of the other functions. My claim is not that the explanatory function is necessary for these other functions, but the more modest proposal that these other functions are strengthened by the explanatory function of the diagnosis. I then substantiate this by examining some cases where diagnoses fail to explain.

There is a clear connection between the function of a diagnosis as a hypothesis and its function as an explanation. In general, when we infer hypothesis from a set of data, we want the hypothesis to explain the data. This squares with the idea that the diagnostic process involves abductive reasoning, or inference to the best explanation (Aliseda and Leonides, 2013; Stanley and Campos, 2013). Moreover, as noted by Peter Lipton (2004), explanatory power is a value that is used to judge the quality of the hypothesis. Hence,

explanatory considerations not only motivate and guide the inferential process in diagnostic hypothesis formation, but are appealed to in the evaluation of the hypothesis.

The explanatory function of a diagnosis, in particular its causal explanatory function, supports its predictive and interventional functions. To be clear, this is not to say that causal explanation is necessary for successful prediction or intervention. As noted by Jennifer Radden (2003: p. 46), a diagnostic category that is defined by a cluster of symptoms without allusion to an underlying cause can still permit probabilistic predictions. We may not know what causes this cluster of symptoms, but we could nonetheless make inductive inferences about similar cases based on enumerative induction, which can then inform an evidenced-based treatment guideline. Nonetheless, Radden also argues that a diagnosis that is explanatory is superior to one that is descriptive, because it opens up possibilities for further hypotheses and targeted interventions. Explaining why a patient has a particular cluster of symptoms provides understanding of the underlying causal structure and mechanisms, which can signal potential targets for therapeutic interventions, inform decisions regarding treatment approaches, and allow us to make predictive inferences that go beyond mere enumerative induction.

Holly Andersen (2012: p. 997) argues that this is especially important where the patient's condition is complicated by a comorbid condition. This is because while we may have evidence-based treatment guidelines for individual disorders, it is impractical to expect there to be evidence-based treatment guidelines for all possible combinations of disorders. Andersen gives the example of a patient diagnosed with a particular type of breast carcinoma who also has comorbid type II diabetes mellitus. Here, there may be an evidence-based treatment guideline for breast carcinoma and an evidence-based treatment guideline for treating type II diabetes mellitus, but there may not be trial-based evidence specifically for managing the combination of type II diabetes mellitus and this

particular type of breast carcinoma. Hence, evidence-based treatment guidelines may not be enough to inform the most appropriate treatment in this particular case. Rather, Andersen argues that we can appeal to causal explanatory knowledge, particularly knowledge involving mechanisms, to inform the treatment decision. Consider that there is evidence of one potential breast carcinoma treatment being more effective than another, but also that the clinician has knowledge that this treatment interferes with a chemical pathway that can worsen type II diabetes mellitus symptoms. The clinician can utilise the causal explanatory knowledge of the two diagnoses to assess the potential interactions between the mechanisms involved in the two disorders and their prospective treatments, in order to arrive at a treatment plan that is likely to be optimal for the particular patient.

The denotative and classificatory functions of a diagnosis are also complemented by its explanatory function. As I shall argue in Chapter 3, where a diagnosis serves as an explanation of patient data, it does so partly by denoting a kind of causal structure that is instantiated by the actual patient. For example, the diagnosis of acute appendicitis explains a patient's abdominal pain by denoting a distinctive pathological type, in this case acute inflammation of the appendix, which is causing the abdominal pain. Conversely, causal explanatory considerations partly justify why some conjunctions of clinical phenomena, but not others, are made into diagnostic categories and assigned diagnostic terms. According to Neil Williams (2011b), it is often the case that when seemingly disparate clinical phenomena are clustered together and characterised as a distinctive category, it is because scientists and medical professionals the clinical phenomena to be connected by a unifying causal explanation. Indeed, a diagnostic category can be discarded and replaced by more precise categories if it turns to be too causally heterogeneous to serve as a satisfactory causal explanation, such as dropsy being discarded and replaced by the more precise categories, congestive cardiac failure, cirrhosis

of the liver, and nephrotic syndrome (Peitzman, 2007). Hence, the causal explanatory value of a diagnostic category influences our judgements about the validity of the classification. As we shall see in Chapter 7, this is apparent in many of the recent philosophical critiques of diagnostic classification in psychiatry (Poland *et al.*, 1994; Murphy, 2006; Tsou, 2015).

The explanatory function of a diagnosis is often considered to justify its normative and social functions. As noted by Annemarie Jutel, a diagnosis "explains certain kinds of deviance in terms of disease rather than of moral failing" (Jutel, 2011: p. 229). This is then regarded as a reason to excuse the patient from certain responsibilities and grant him or her certain rights according to the "sick role" (Parsons, 1951). For instance, a child diagnosed with influenza may be temporarily granted absence from school, because his or her failure to concentrate is explained as being due to an unpleasant and unfortunate medical problem, rather than deliberate school refusal. We have also already seen in Chapter 1, §1.2.1, how the presence of a causal explanation is considered to be a legal excusing condition in criminal legislation. The amendment of the *Homicide Act 1957* by the *Coroners and Justice Act 2009* in England and Wales, for example, states that a condition for the defence of diminished responsibility is the presence of a mental disorder that causally explains the defendant's behaviour.

However, it is worth noting that the legitimacy of this sort of reasoning is disputable, because it assumes a dubious dichotomy between the medical and the moral. For example, Jennifer Radden (1982) criticises causal explanation as a legal excusing condition and argues that the mere fact that a disease was causally involved in the production of a criminal action does not justify excusing the action. Rather, she argues that a disease is relevant to the excuse only inasmuch as it can be associated with the traditional excusing conditions, namely ignorance and compulsion. Derek Bolton (2008), in his analysis of the concept of mental disorder, also comments on the dubious

dichotomy between the medical and the social, stating that "there may be no clear basis for distinguishing between mental health problems and social problems, or between mental health problems and 'normal – more or less normal – problems of living'" (Bolton, 2008: p. viii). Hence, to assume that a causal explanation provided by a diagnosis can demarcate the medical from the moral or the social is to commit a conceptual error. Nonetheless, I suggest that it is still possible for the explanatory function of a diagnosis to provide justificatory support for its use as a social tool without assuming the above dichotomy between the medical and the moral. For instance, by explaining that the patient's symptoms are caused by a particular kind of condition, the diagnosis supports the mobilisation of therapeutic, supportive, and financial resources of the sorts and in the amounts deemed by medical professionals and policy makers to be beneficial for this particular kind of condition.

Finally, the semiotic function of the diagnosis draws on its explanatory function. Part of why a diagnosis serves as a meaningful label for the patient is because it is taken to provide an explanation of why he or she has been suffering from his or her symptoms. In a qualitative study of adults diagnosed with attention-deficit hyperactivity disorder, Svend Brinkmann (2014) notes that the participants commonly mediate understanding of their problematic behaviours by invoking their diagnoses as explanations of these behaviours when summarising their stories. While Brinkmann comments on the possible circularity of invoking a syndromic diagnosis like attention-deficit hyperactivity disorder as an explanation of symptoms, his research does at least show that patients consider explanation to be an important function of a diagnosis. Other authors have also written about the way in which a diagnosis helps the patient reorganise his or her narrative and make sense of his or her condition by providing an explanation. For example, Kirmayer *et al.* suggest that "explanations may offer some reassurance and consolidation, promote coping and resilience, and allow the person to plan realistically for the future" (Kirmayer

et al., 2004: p. 664). Similarly, Winston Chiong writes that a diagnosis "can also be an explanation for patients who have had symptoms but do not know their cause", which "may seem to resolve the mystery, such that even patients with intractable, chronic diseases may feel relief when diagnosed" (Chiong, 2004: p. 129). And so, the explanatory function of a diagnosis does not only have epistemic significance and instrumental utility for the clinician, but also has intrinsic value for the patient.

2.3.2 Medically unexplained syndromes

So far, I have given an overview of the various functions served by diagnoses and have argued that the explanatory function of the diagnosis is important because it provides justificatory support for its other functions. However, not all diagnoses function as explanations. There are some diagnoses that are customarily called medically unexplained syndromes, precisely because it is assumed that they fail to explain patients' symptoms. These include chronic fatigue syndrome, fibromyalgia, and irritable bowel syndrome. As we shall see, their explanatory shortcomings are reflected by other epistemic and instrumental limitations. Hence, cases of medically unexplained syndromes provide further support for the idea that causal explanation is a desirable function in part because it strengthens the other functions of the diagnosis.

Medically unexplained syndromes are estimated to account for around a quarter of primary care consultations (Kirmayer *et al.*, 2004). The diagnoses are syndromic, that is, they are not defined in terms of underlying disease processes, but in terms of symptom criteria. For example, the Centers for Disease Control and Prevention define chronic fatigue syndrome as follows:

> A case of the chronic fatigue syndrome is defined by the presence of the following: 1) clinically evaluated, unexplained, persistent or relapsing chronic fatigue that is of

> new or definite onset (has not been lifelong); is not the result of ongoing exertion; is not substantially alleviated by rest; and results in substantial reduction in previous levels of occupational, educational, social, or personal activities; and 2) the concurrence of four or more of the following symptoms, all of which must have persisted or recurred during 6 or more consecutive months of illness and must not have predated the fatigue: self-reported impairment in short-term memory or concentration severe enough to cause substantial reduction in previous levels of occupational, educational, social, or personal activities; sore throat; tender cervical or axillary lymph nodes; muscle pain; multijoint pain without joint swelling or redness; headaches of a new type, pattern, or severity; unrefreshing sleep; and postexertional malaise lasting more than 24 hours. (Fukuda *et al.*, 1994: p. 956)

Cournoyea and Kennedy (2014) argue that such a diagnosis fails to explain, because it merely restates the symptoms without providing any causal information and, importantly, whatever cause there might be for the set of symptoms is currently unknown. Tentative, though plausible, suggestions have been made regarding psychodynamic, cognitive, neuroendocrine, immunological, and cultural factors that may be involved in the conditions, but the precise causal structures of medically unexplained syndromes remain undetermined (Kirmayer *et al.*, 2004: p. 666).

Typically, a diagnosis of a medically unexplained syndrome is only made after investigations have failed to reveal any underlying medical causes for the patient's symptoms and other diagnoses have been eliminated. For example, Cournoyea and Kennedy (2014: p. 929) present the case of Brad, who presents with persistent fatigue, difficulty concentrating, joint pain, and neck soreness. The clinician considers possible explanations for Brad's symptoms, including such autoimmune disorders as systemic lupus erythematosus and rheumatoid arthritis, and such infectious diseases as

cytomegalovirus infection, Epstein-Barr virus infection, and Lyme disease. Only once these possible causes are excluded by tests is Brad given a diagnosis of chronic fatigue syndrome. And so, not only does a medically unexplained syndrome diagnosis fail to serve as a medical explanation of the patient's symptoms, but it specifically implies the absence of medical explanation (Jutel, 2011: pp. 80–81). Such a diagnosis is not so much a positive hypothesis arrived at via inference to the best explanation, but a negative hypothesis, or a diagnosis of exclusion, resulting from a process of eliminative inference.

The absence of a causal explanation is associated with uncertainty and disagreement regarding classification. As noted in §2.2.6, medical disorders are often classified according to the kinds of causal process involved or where they are located. However, if the cause of a disorder is unknown or disputed, then it is left unclear how it should be classified, or indeed if it constitutes a valid category. For example, David and Wessely (1993) contest the assumed classification of chronic fatigue syndrome as an inflammatory disease of the nervous system under the category of benign myalgic encephalomyelitis in *ICD*-10 (World Health Organisation, 1992) and instead argue that it should be classified as a psychiatric disorder under the category of neurasthenia. An unfortunate clinical consequence of this classificatory uncertainty is the unsystematic approach to specialist referral from primary care. Due to the lack of knowledge regarding the underlying causes of the syndromes, patients with medically unexplained syndromes are often repeatedly referred to multiple different specialties (McGorm *et al.*, 2010). This can result in patients feeling like they are being "passed between health specialists" and being unsure about who to approach for help (Nettleton *et al.*, 2005: p. 208).

There may also be other limitations regarding therapeutic intervention, as the absence of a causal explanation leaves the clinician uncertain about how best to treat the patient (Cournoyea and Kennedy, 2014: p. 929). Of course, as noted in §2.3.1, the presence of a causal explanation is neither necessary nor sufficient for there to be an

effective treatment. Diagnoses that do not allude to underlying causes can still permit inductive inferences that inform evidence-based treatment guidelines. Conversely, a diagnosis could provide a causal explanation, but we may currently lack the technological means to therapeutically manipulate this cause. Nonetheless, the presence of a causal explanation can signal targets for therapeutic interventions and justify decisions regarding treatment approaches. The above considerations are illustrated by the example of chronic fatigue syndrome. Here, the absence of a clear explanation for the syndrome is reflected by the fact that there is very little in the way of agreed or successful treatment (Fukuda *et al.*, 1994; Deale and Wessely, 2001). Current treatment strategies are highly miscellaneous and tend to be palliative, rather than being targeted at an underlying disease process. Moreover, while there is empirical evidence supporting the uses of cognitive-behavioural therapy and graded exercise treatment, the expected outcomes are the management of symptoms and the improvement of coping ability, rather than the resolution of whatever disease process might be responsible for the symptoms (Luyten *et al.*, 2008). Such inability to identify the underlying pathology that explains the patient's symptoms can leave the doctor feeling impotent when it comes to treatment (Nettleton *et al.*, 2004: p. 63).

In addition to the above mentioned epistemic and instrumental limitations, the absence of a causal explanation is often taken to undermine the normative, semiotic, and social functions of a diagnosis. As noted by Nettleton *et al.* (2004: p. 48), there is an assumed hierarchy between explanatory and non-explanatory diagnoses, such that medically unexplained syndromes are sometimes considered by patients, clinicians, and social organisations not to be legitimate medical disorders. At the individual level, this can be associated with patient dissatisfaction. A qualitative study by Norma Ware (1992) reports that patients diagnosed with chronic fatigue syndrome often feel betrayed by the lack of explanatory information provided by their diagnoses, tend to consult many other

clinicians hoping that their symptoms might eventually be explained, and become secretive about their conditions due to worries that they won't be perceived as "real". A participant in the study even reported that it would be easier in some ways to have a more serious but more understandable diagnosis like cancer (Ware, 1992: p. 353). Other researchers also report that patients who receive medically unexplained syndrome diagnoses feel let down and frustrated, because their hopes for explanations that would help them make sense of their conditions are left unfulfilled (Kirmayer *et al.*, 2004: p. 668; Nettleton *et al.*, 2004: p. 64). At a wider organisational level, this perceived illegitimacy can be associated with the withholding of services. For example, Joseph Dumit (2006: p. 580) reports that patients with chronic fatigue syndrome in the United States of America are sometimes denied disability benefits on the grounds that such a syndromic diagnosis is not supported by a biological explanation.

And so, the above mentioned problems associated with medically unexplained syndrome diagnoses show that explanation is a desirable function of a diagnosis. Where diagnoses fail to provide explanations for patients' symptoms, there may be uncertainties regarding classification, therapeutic limitations, perceptions of illegitimacy, feelings of dissatisfaction, and dismissive social attitudes regarding these diagnoses. I now examine how some of the above considerations are also of relevance to diagnoses in psychiatry.

2.3.3 Problems with psychiatric diagnoses

In §2.3.2, I presented medically unexplained syndromes as paradigmatic cases of diagnoses that are widely considered not to provide explanations of patients' symptoms and laid out some of the broader implications of their explanatory shortcomings. It may be apparent that some of the sorts of property that are associated with the explanatory shortcomings of medically unexplained syndrome diagnoses are also shared by psychiatric diagnoses, such as syndromic definitions based on symptom clusters, exclusion criteria

that recommend ruling out possible medical causes before the diagnoses are established, and contentions regarding the precise causal structures of the disorders. Given these similarities and the potential associated implications, more detailed investigation is warranted regarding whether or not psychiatric diagnoses actually do provide explanations of patients' symptoms. I do not intend to fully answer this question in the current chapter, as the rest of the thesis is dedicated to this task. Rather, I would like here to highlight some of the critiques of psychiatric diagnoses that are related to their uncertain explanatory statuses, in order to further support the point made in Chapter 1, §1.2.1, that the question of whether or not psychiatric diagnoses genuinely function as explanations of symptoms has significant implications for clinical discourse and practice.

Like medically unexplained syndromes, psychiatric disorders have historically been beset by controversies. Among the most famous of those sceptical of psychiatric disorders are the proponents of the antipsychiatry movement of the 1960s. We have already visited Thomas Szasz (1960) in Chapter 1, who criticises the concept of mental illness. First, he argues that mental illness cannot legitimately be invoked as an explanation of someone's behaviour because it is merely a shorthand label for the behaviour. Second, he argues that mental illness is not determined by a physiological cause, but by moral and social norms. Other antipsychiatrists offer different critiques of psychiatry. For example, Michel Foucault ([1961] 1964) argues that our current ways of thinking about psychiatric disorders as medical problems are the products of contingent historical developments, and so it is possible that these current ways of thinking might not have arisen had history worked out differently, while R. D. Laing (1967), criticises the medical conception of schizophrenia and instead argues that it is a normal and understandable response to an existentially distorted social world.

Szasz's (1961) critique is noteworthy, because it draws connections between the supposed illegitimacy of a psychiatric diagnosis *qua* causal explanation and shortcomings

with respect to its normative and social functions. That is to say, he uses his argument that mental illness diagnoses fail to explain people's behaviours to support the normative claim that mental illnesses are not genuine medical disorders and also to oppose the sanctioned uses of involuntary treatments for mental illnesses. This sort of approach has also been used by proponents of the subsequent critical psychiatry movement. David Ingleby (1982) argues that psychiatric diagnoses are only allowed to instigate the social responses of mobilising clinical resources and sanctioning certain behaviours because they are presented by the psychiatric profession as designating diseases that are responsible for the patients' symptoms, much like diagnoses in other medical specialties. However, the suggestion is that psychiatric diagnoses do not designate genuine diseases that explain the symptoms, and so such social responses are not justified. Hence, Ingleby suggests that if people are made aware that the diagnoses instigate social responses that are not supported by medical explanations, then "questions would immediately arise about the propriety of those responses" (Ingleby, 1982: p. 137). Similarly, Joanna Moncrieff (2010) suggests that the notion that psychiatric diagnoses pick out underlying diseases that cause symptoms is just an assumption and that challenging this assumption could open up the associated social responses to scrutiny.

Writing from an analytic philosophy, rather than a social theory, point of view, Jeffrey Poland (2014) criticises the epistemic shortcomings of the psychiatric diagnoses in the *DSM*:

> The *DSM* categories and associated epistemic practices related to information processing, inferential practice, explanatory practice, and clinical understanding, are ineffective and harmfully biased because, given their atheoretical focus on clinical phenomenology, they do not effectively identify and represent important features, problems, contexts, and processes … (i.e., they do not underwrite sound clinical

> inferences and judgments concerning what is wrong and what is likely to be helpful). (Poland, 2014: p. 48)

Poland's critique suggests that explanatory failure is connected to other shortcomings regarding classification, prediction, and intervention. That is to say, psychiatric diagnoses that do not adequately inform us about the processes underlying patients' problems are poor categories that are unlikely to support reliable inferences or guide effective treatment decisions. He describes such diagnoses as "free riders" that contribute little over and above descriptions of symptoms (Poland, 2014: p. 34).

Moncrieff (2010) also argues that causal explanatory shortcomings are associated with limitations regarding therapeutic interventions in psychiatry. She writes:

> In contrast to most medical conditions like diabetes, tuberculosis and heart disease, no psychiatric condition can be traced to a specific dysfunctional bodily process … There is no evidence that any class of psychiatric drug acts by reversing or partially reversing an underlying physical process that is responsible for producing symptoms … Therefore the idea that the behaviours seen by psychiatrists are indicative of an underlying disease is simply an assumption. (Moncrieff, 2010: p. 373)

Of course, whether or not psychiatric conditions can be traced to specific processes and whether or not psychiatric drugs do act by reversing specific processes are empirical questions that require empirical support. I reserve detailed examination of the empirical data relevant to the former question for Chapter 5. Nonetheless, a more modest point can still be gleaned from the above critique. If it is the case that a psychiatric diagnosis does not provide a causal explanation for a cluster of symptoms, then such a diagnosis

cannot be said to supply a justification for a given treatment for the disorder on the basis of the supposition that the treatment acts by interfering with a particular causal process. Furthermore, there is a sense in which such a treatment would be palliative. Given the lack of knowledge regarding whatever causal process might be underlying the cluster of symptoms, it would seem that only the cluster of symptoms itself, but not any underlying causal process, would be a tangible target for therapeutic intervention.

To sum up, there are controversies regarding the explanatory roles of psychiatric diagnoses. The above critiques show some of the ways in which potential explanatory shortcomings could limit or delegitimise the roles of psychiatric diagnoses in sanctioning certain social responses, predicting clinical outcomes, and guiding therapeutic interventions. Given these controversies and the potential implications for psychiatric practice, it is important to pursue a better understanding of what sorts of explanatory role, if any, are served by diagnoses in psychiatry.

2.4 Conclusion

In this chapter, I have emphasised the role of the diagnosis as a valuable epistemic resource that serves a variety of functions in clinical medicine. It functions as a testable scientific hypothesis, a denotative signifier, a classificatory category, a causal explanatory construct, a predictive indicator, a normative judgement, a therapeutic guide, a semiotic mediator, and a social performative. I have shown that the explanation of symptoms is a desirable function of a diagnosis, in part because it provides justificatory support for many of its other functions. I then supported this with appeal to medically unexplained syndrome diagnoses, where explanatory failures are associated with uncertainties regarding classification, therapeutic limitations, perceptions of illegitimacy, and deeply dissatisfied patients. Finally, I indicated why these could also potentially be concerns for diagnoses in psychiatry, whose explanatory statuses are highly contentious. As we have

seen, a number of critics have argued that potential shortcomings with the explanatory functions of psychiatric diagnoses are connected to serious problems regarding their classificatory, predictive, normative, interventional, and social functions. Therefore, the philosophical question of whether or not psychiatric diagnoses explain has important implications for clinical psychiatric practice and discourse. The rest of this thesis is dedicated to answering this question. In order to answer it, though, we need to understand precisely what it is for a diagnosis to explain a set of symptoms. This will be the focus of Chapter 3.

3. Medical Diagnoses as Causal Explanations*

3.1 Introduction

As we saw in Chapter 2, it is generally accepted that many diagnoses in clinical medicine, though by no means all, serve as explanations of patients' symptoms. This explanatory function is considered desirable, because it can guide interventions, support predictions, and convey understanding to the patient. This indicates a bidirectional epistemic relation between the diagnosis and the patient's symptoms. The diagnosis is inferred from the symptoms and the symptoms are explained by the diagnosis.

In this present chapter, I elucidate the nature of this explanatory relation. My aim is to develop a philosophical account of how it is that a diagnosis serves as an explanation of a patient's symptoms. Of course, as noted in my discussion of medically unexplained syndromes in Chapter 2, not all diagnoses in medicine serve as explanations. Hence, the model of explanation I develop is intended to capture the nature of explanation in those paradigm cases where the diagnoses genuinely do explain the patients' symptoms. Because these paradigm cases come from bodily medicine, I will mostly be dealing with general medical diagnoses in this chapter. However, the relevance for the rest of the thesis is that it will serve as a point of comparison for my later discussion of diagnoses in psychiatry. That is to say, understanding how diagnoses explain symptoms in those uncontroversial medical cases where they do provides a standard with which to assess whether psychiatric diagnoses similarly serve such explanatory functions.

---

* A version of this chapter has been published as: Maung, H. H. (2016b). "The Causal Explanatory Functions of Medical Diagnoses". *Theoretical Medicine and Bioethics*. Published online first 16th September 2016. DOI: 10.1007/s11017-016-9377-5.

The rest of the chapter proceeds as follows. In §3.2, I clarify what the *explanandum* is in diagnostic explanation. This is important, because to "explain symptoms" is an ambiguous expression in need of further specification. I then turn my focus to establishing the nature of the *explanans* in §3.3. In §3.3.1, I look at Carl Hempel's (1965a) deductive-nomological and inductive-statistical models of scientific explanation, and argue that diagnostic explanations are neither explanatory in virtue of their argumentative structures nor in virtue of the general regularities between the diagnoses and the patient data. In §3.3.2, I present Margherita Benzi's (2011) argument that medical diagnoses explain by identifying the actual causes of the patient data in individual cases, rather than by subsuming them under general causal regularities. I then argue in §3.3.3 that although Benzi is correct to stress that diagnostic explanation appeals to actual causation, a more complete account also needs to consider how a successful causal explanation of a patient's symptom presentation not only involves a simple causal claim of the form "*C* causes *E*", but also relies on mechanistic causal knowledge of the form "this mechanism produces this phenomenon" (Darden, 2013: p. 20). In §3.3.4, I suggest that the former is the outcome of the diagnostic search, while the latter is provided by the theoretical framework in which the physician operates. This is supported with appeal to Kenneth Schaffner's (1986) work on theoretical generalisations in medicine and Jeremy Simon's (2008) work on disease ontology.

3.2 The *explanandum*

3.2.1 Contrastive explanation

Before we explore what sort of explanation a diagnosis provides, it is important to clarify precisely what it is that is being explained. It might seem straightforward to say that a diagnosis is invoked to explain why the patient has a certain set of symptoms. However, it is uncontroversial in the philosophical literature that explanations are contrastive. We

do not simply explain "why *P*?", but "why *P* rather than *Q*?" (van Fraassen, 1980: pp. 126–129; Lipton, 2004: pp. 33–37). Peter Lipton refers to *P* and *Q* as the fact and the foil, respectively. For instance, he notes that when we explain why the leaves turn yellow in November, we do not explain this fact *tout court*, but explain "only for example why they turn yellow in November rather than in January, or why they turn yellow in November rather than turn blue" (Lipton, 2004: p. 33). Hence, the information that is required in the explanation depends on which contrastive foil is selected.

Which contrastive foil is selected is guided by our explanatory interests and values. In the context of scientific explanation, these interests and values are not entirely arbitrary, but are shaped by the norms and aims of the field of enquiry. That is to say, certain sorts of contrastive question turn out to be conducive to achieving the goals of certain research programmes. In the context of the clinical consultation, it is supposed that the aim of medicine is to achieve the "the best medical outcome for the patient" (Whitbeck, 1981: p. 324), and that the ideas about what constitute good and bad medical outcomes are shaped by medical theory concerning "the pathological variants of the 'normal' or 'healthy' processes" (Schaffner, 1986: p. 71). Hence, it makes sense that the sort of contrastive question that normally guides diagnosis is why the patient is presenting with medically abnormal symptoms rather than being in an acceptable healthy condition. Tomasz Rzepiński (2007) accordingly characterises a diagnosis as an answer to the following sort of contrastive question:

> "Why X1, X2, …, Xn, when it should be Y1, Y2, …, Yn?" where X1, X2, …, Xn account for a description of improper symptoms, while Y1, Y2, …, Yn account for a description of a properly functioning human body. (Rzepiński, 2007: p. 70)

Here, "improper" is construed to include any bit of patient data that is judged to be abnormal by medical standards and in need of further intervention. Rzepiński gives the examples of a quantitative investigation result such as increased plasma bilirubin concentration, an examination sign such as tenderness on palpation of the right iliac fossa, and a patient's report of certain symptoms being present. The idea of "properly functioning" is construed to include physiological norms based on medical theory and statistical norms regarding quantitative reference ranges.

While Rzepiński's analysis is plausible, I argue that it is incomplete as it stands, because it is too restrictive with respect to what sorts of norm guide judgements about what is proper and improper regarding the patient's clinical presentation. As noted above, Rzepiński suggests that these are informed by physiological and statistical norms based on knowledge from medical science. However, judgements about what is proper and improper regarding the patient's clinical presentation are also informed by a variety of other norms and values, including the patient's evaluation of certain sensations as distressing or disabling, expectations about performance ability relative to the patient's usual baseline performance, and social conceptions of normality and deviance (Fulford, 1989; Wakefield, 1992; Bolton, 2008). These are conspicuously missing from Rzepiński's analysis, but I suggest they could easily be included.

And so, we can construe the diagnostic question as a contrastive question of the sort "why *P* rather than *Q*?", where *P* is the presence of certain symptoms in the patient which are deemed improper according to the above mentioned physiological, statistical, personal, and social norms, and *Q* is the counterfactual state where these symptoms are absent and which is considered more acceptable according to these norms. Of course, this is not to say that there cannot be other sorts of question in the clinical consultation which require different sorts of contrastive foil, such as questions about treatment response and individual differences. Nonetheless, the above construal reasonably

captures the contrastive fact that the physician is seeking to explain with a diagnosis when a patient presents to the clinical encounter with a set of symptoms.

3.2.2 Functional and phenomenal concepts of symptom

In addition to specifying the contrastive structure of the diagnostic question, I argue that we need to be clearer about precisely what feature of a symptom is being explained by the diagnosis. In addition to their observable behavioural manifestations, many symptoms are associated with subjective experiences. Obvious examples, to name a few, include pain, itch, fever, nausea, dizziness, fatigue, depressed mood, and hallucinations. To borrow an expression made famous by Thomas Nagel (1974), there is "something it is like" when one has a symptom.

The subjective quality of experience might appear to present a problem for causal explanation of symptoms. This problem concerns the explanatory gap between physical facts and phenomenal facts (Kripke ([1972] 1980; Nagel, 1974; Jackson, 1982; Chalmers, 1996). The general idea is that the physical facts, which are in terms of structures and dynamics, can yield only further facts about structures and dynamics, but do not encapsulate information about the subjective quality of experience (Chalmers, 1996: p. 107). And so, given all the physical facts about the structures and dynamics of such a system, consciousness remains an extra fact to be considered. Some philosophers take the explanatory gap between the physical and the phenomenal to indicate an underlying metaphysical issue regarding the mind-body problem. For example, philosophers such as David Chalmers (1996), Laurence BonJour (2010), and Martina Fürst (2011) propose that physicalism is false, dualism is true, and consciousness is ontologically fundamental. Note that this is different from the picture suggested by René Descartes ([1641] 1996), where a non-physical *res cogitans* exerts its own influence on physical matter to generate behaviour. Chalmers (1996: pp. 124–125) concedes that the functional properties of the mind

responsible for the production of behaviour can be causally explained in terms of structures and dynamics. Rather, the ontological distinction he proposes is between the physical and the phenomenal, which he suggests are related via correlatory "supervenience laws" (Chalmers, 1996: p. 127). Other authors, such as Thomas Nagel (1974) and Joseph Levine (1983), do not make such metaphysical commitments, but nonetheless concede that the explanatory gap between the physical and the phenomenal is genuine.

In light of this explanatory gap between the physical and the phenomenal, it would appear that if the explanation of a symptom involves the explanation of what it is like to experience that symptom, then information about causes and mechanisms is not wholly adequate for explaining the symptom. Even after one has elucidated the mechanisms responsible for pain, one has not explained the subjective quality of pain. One might be tempted to tackle this by assuming a view that denies the explanatory gap, but I suggest that this is unwarranted. I propose that there is no need to be drawn into the metaphysics of the mind-body problem to defend causal explanations of symptoms. Rather, one just needs to be more discerning with respect to the scope of the *explanandum*. Again, the work of Chalmers is relevant here.

Chalmers (1996: p. 11–22) separates two different concepts of the mental. The psychological, or functional, concept of the mental is that which concerns the causal processes involved in the production of behaviour. The phenomenal concept of the mental is that which concerns the subjective quality of experience, or the "something it is like" of consciousness. These two concepts of the mental tend to co-occur. They are also often conflated in everyday language. For example, pain can be taken to mean a kind of functional state that normally results from actual or potential tissue damage and that normally produces aversive reactions, verbal reports of a part of the body hurting, increased sympathetic nervous system activity, and so on. However, it can also be taken

to mean the kind of phenomenal quality that normally accompanies this functional state. Irrespective of what the metaphysical relation between the functional and the phenomenal concepts of pain might be, there is at least a conceptual distinction between the two. Hence, one can separate trying to explain the functional concept of pain from trying to explain the phenomenal concept of pain.

I argue that for the purposes of causal explanation in medicine, we need only concern ourselves with explaining the functional concept of a symptom. For example, an adequate causal explanation of someone's pain would be an explanation of why he or she is in such a functional state that is associated with aversive reactions, verbal reports of a part of his or her body hurting, increased sympathetic nervous system activity, and so on. It would not require the explanation of why this functional state is accompanied by the patient's subjective experience of pain, or of what it is like for him or her to experience this pain. By restricting the scope of the *explanandum* to the functional concept of a symptom, the adequacy of the *explanans* no longer depends on any attempt to bridge the explanatory gap between the physical and the phenomenal. A causal explanation of symptoms that is in terms of mechanisms can still be adequate in the case that the gap is unbridgeable.

And so, a diagnosis need not elucidate anything profound about phenomenology or the metaphysics of the mind-body problem to be a good explanation of a patient's symptoms. The aim of diagnostic explanation is to explain the functional concept of a symptom and, while we can accept that there is a phenomenal concept associated with this functional concept, there is no need for the diagnosis to explain what this phenomenal concept is like. This is in no way saying that phenomenology is irrelevant to the understanding of disorders. The philosophy of psychiatry has a tradition of phenomenological approaches to psychopathology, which goes at least as far back as Karl Jaspers ([1913] 1997), and which has been continued by contemporary theorists (Fuchs,

2005; Ratcliffe, 2008). I do not claim that medicine and psychiatry are entirely exhausted by empirical science, and I accept that phenomenological research may help us to attain a richer understanding of the aspects of health and illness that are not covered by facts about causes and mechanisms. My claim is merely that the aim of diagnostic explanation is to understand the causes of symptoms from the outside. For this particular purpose, understanding of what phenomenal qualities are like from the inside is not required. A similar attitude is expressed by Dominic Murphy (2006: pp. 16–17) regarding explanation and classification in psychiatry.

3.3 The *explanans*

3.3.1 Covering law models

Having clarified the *explanandum* in the diagnostic context, I now explore the nature of the *explanans* through examination of some prominent philosophical models of explanation. Among the most influential and widely discussed accounts of scientific explanation in the philosophical literature is Carl Hempel's (1965a) covering law account, according to which a phenomenon is explained by subsuming it under a general law or regularity. A covering law explanation has the form of an argument, whereby the *explanandum* is concluded from a set of premises, of which at least one must be a general law that is necessary for the argument. The argument can be either deductive or inductive. The former kind, known as deductive-nomological explanation, has the following form when applied to diagnostic explanation, where S is a set of patient data, D is the diagnosis, and $D \rightarrow S$ is the general law linking the diagnosis with the set of patient data:

$D \rightarrow S$

D

———————

S

For instance, according to the deductive-nomological model, a patient's leg oedema would be explained by deducing it from the diagnosis of heart failure and the general law that links heart failure with leg oedema.

Nonetheless, the deductive-nomological model has a serious limitation in the context of clinical practice. Many regularities in medicine are probabilistic rather than deterministic, and so do not enable sound deductions of the patient data from the diagnoses (Sadegh-Zadeh, 2012: p. 344). In the above mentioned example, the correlation between heart failure and leg oedema is not absolute, and it is possible to have heart failure without leg oedema. This suggests that the premise $D \rightarrow S$ is false and the deduction is not sound. Therefore, the deductive-nomological model is only applicable to a very limited number of cases of diagnostic explanation.

Hempel concedes that the deductive-nomological model cannot account for cases of explanation that do not involve deterministic laws and introduces the latter kind of covering law argument, known as inductive-statistical explanation, to make up for these cases. According to this, to explain a phenomenon is to inductively infer it from a statistical generalisation about previously observed cases. Hempel uses the example of Jones' recovery from a streptococcal infection being explained by his taking penicillin and the statistical generalisation that a high proportion of people who have streptococcal infections recover after taking penicillin. Applied to the example of heart failure, the patient's leg oedema is explained by the fact that he or she has heart failure, along with

the statistical generalisation that a high proportion of people of patients with heart failure have leg oedema:

> Most observed *P*s with *D* had *S*.
>
> *x* is a *P* with *D*.
>
> ═══════ [makes very likely]
>
> *x* has *S*.

The inductive-statistical model accommodates the fact that many relations between diagnosis and symptoms in medicine are probabilistic (Qiu, 1989: p. 203). Therefore, a charitable rendering of a covering law account of diagnostic explanation needs to allow inductive-statistical as well as deductive-nomological explanations.

I accept that some instances of diagnostic explanation may be formulated as covering law arguments of the inductive-statistical kind. There is a certain feature of a diagnosis that permits such a formulation. Covering law explanations appeal to laws or regularities, which in turn depend on the presupposition of repeatable types that instantiate these laws or regularities. In medicine, diagnoses are often treated as such repeatable types (Sadler, 2005: pp. 419–420; Sadegh-Zadeh, 2012: p. 172). They are generalised categories, whose tokens are taken to share certain properties. For example, heart failure is considered to be a type characterised by the following:

> Heart failure is the state of any heart disease in which, despite adequate ventricular filling, the heart's output is decreased or in which the heart is unable to pump blood at a rate adequate for satisfying the requirements of the tissues with function parameters remaining within normal limits. (Denolin *et al.*, 1983: p. 445)

Individual cases of heart failure are tokens of this type that instantiate this feature. This characterisation of diagnoses as repeatable types enables them to support the kinds of regularity and inductive inference that feature in inductive-statistical explanations.

However, it has long been argued that the inductive-statistical model as it stands is too permissive to be a complete account of explanation. There are well-known counterexamples that fulfil the requirements of the inductive-statistical model, yet are not genuinely explanatory. One kind of counterexample concerns explanatory irrelevancies. Peter Achinstein (1983) gives the hypothetical case of Jones, who dies within a day of eating a pound of arsenic. Assume that the actual cause of Jones' death had been an unrelated car accident. If this is the case, then his eating a pound of arsenic is explanatorily irrelevant to his dying. However, according to the inductive-statistical model, Jones' death would still be explained by his eating a pound of arsenic, along with the statistical generalisation that a very large proportion of people who eat a pound of arsenic die within a day. To take another example, a significant proportion of patients diagnosed with left hemispheric stroke present with right-sided paralysis. Now, consider the case of a patient diagnosed with left hemispheric stroke, but who already has right-sided paralysis for a different reason, such as cerebral palsy. In this case, the diagnosis of left hemispheric stroke is explanatorily irrelevant to the patient's right-sided paralysis. Nonetheless, according to the covering law account, the patient's right-sided paralysis would still be explained by his or her diagnosis of left hemispheric stroke, along with the statistical generalisation that a large proportion of patients diagnosed with left hemispheric stroke present with right-sided paralysis.

Another kind of counterexample concerns spurious correlations. Wesley Salmon ([1975] 1998) gives the example of a correlation between a falling barometer reading and a storm. Although there is a significant statistical regularity between these two event types, a falling barometer reading is not a legitimate explanation of a storm. Rather, both

have a common explanation, namely the preceding drop in atmospheric pressure. Applying this to a medical example, there is a statistical regularity between calf pain and pulmonary embolism, such that the probability of a patient having calf pain is higher if he or she also has pulmonary embolism than the probability of his or her having calf pain under any circumstance. However, in this case, the diagnosis of pulmonary embolism does not explain the patient's calf pain. Rather, both the calf pain and the pulmonary embolism, as well as the statistical relation between the two, can be explained by the diagnosis of deep vein thrombosis.

The above counterexamples show that genuine explanatory relations are underdetermined by covering law arguments. In the example of the patient with right-sided paralysis, there are two possible explanations for the patient data, each supported by a different inductive-statistical argument. These are left hemispheric stroke and cerebral palsy, respectively. Here, the correct explanation cannot be determined by the inductive-statistical model on its own. Rather, confronted with two inductive-statistical arguments supporting different diagnoses, the physician has to make a choice, or an inference to the best explanation, based on some other criterion. Hence, the covering law account at best describes only a part of the relation between the actual diagnosis and the clinical data.

What seems to be suggested by the above counterexamples is that a criterion that is required for the relation between the diagnosis and the patient data to be genuinely explanatory is causation. In the case of the patient with cerebral palsy, the reason why left hemispheric stroke does not explain his or her right-sided paralysis is because the right-sided paralysis was caused by another condition, namely cerebral palsy. Also, in the case of the patient with deep vein thrombosis and pulmonary embolism, the reason why the former but not the latter explains his or her chest pain is because it is the former that had caused it. However, inductive-statistical relations are not specifically causal, and so on

their own cannot distinguish between diagnoses that genuinely explain the patient data and those that are merely correlated with the patient data. The upshot, then, is that while the covering law account as described above may capture a part of the relation between a diagnosis and the patient data, it fails to pick out specifically what it is that makes this relation genuinely explanatory.

3.3.2 Causal explanation and actual causation

The above considerations suggest that an adequate model of diagnostic explanation must take causation into account. Over the past half century, the causal model of explanation has attracted a large number of proponents in the philosophy of science, including Wesley Salmon ([1975] 1998), David Lewis (1986b), James Woodward (2003), and Peter Lipton (2004). The basic claim of the causal model is that to explain something is to provide information about its cause. This certainly has intuitive appeal with respect to diagnostic explanation, as it is commonly suggested that the aim of the diagnostic process is to search for the cause of the clinical manifestation (Whitbeck, 1981; Rizzi, 1994; Schwartz and Elstein, 2008). Furthermore, the model's requirement of a causal connection between the *explanandum* and the *explanans* helps to avoid the over-permissiveness of the covering law account. As noted in §3.3.1, physicians seeking explanations of patient data may be confronted with various factors that are correlated with the patient data, some of which may be causally irrelevant or spurious but nonetheless may satisfy the requirements for inductive-statistical explanations. Under the causal model of explanation, though, only those correlations which are genuinely causal would qualify as being explanatory.

Against the causal model of explanation, it might be commented that we do not yet have a fully adequate analysis of causation. However, as argued by Lipton (2004: p. 31), this does not compel us to abjure the model. The notion of causation is indispensable to

philosophy, science, and ordinary life, and we know a lot about the relation even without a full metaphysical account. Hence, a causal model of explanation can appeal to the causal relation as it is without committing to a particular metaphysical account of causation. Accordingly, although I am interested in the role that causation has in explanation, I do not, in this chapter, say much about the large topic of the metaphysics of causation.

Although the causal model of explanation is sometimes described as a reaction to the covering law account, some instances of causal explanation can be formulated as special cases of covering law explanation where the regularities appealed to are causal regularities, or "laws of succession" (Hempel, 1965a: p. 352). Physics and chemistry contain such examples. For example, one could explain why the ice cube in a glass of water melts by appealing more generally to the laws describing how high temperatures influence the hydrogen bonds between H_2O molecules. As noted in §3.3.1, some instances of diagnostic explanation can be formulated as covering law arguments, which suggests that they could be considered cases of covering law explanation that appeal to causal, rather than merely statistical, regularities.

Margherita Benzi (2011) notes that the causal regularities cited in covering law explanations hold between general types. We have already seen in §3.3.1 how a diagnosis, such as heart failure, is treated as a repeatable type. The covering law account also treats the symptom presentation as a repeatable type, such that a causal regularity is taken to hold between the type diagnosis "heart failure" and the type symptom "leg oedema". In diagnostic explanation though, the *explanandum* is not a generality, but a particular fact. That is to say, in the case where the diagnosis of heart failure successfully explains leg oedema, what is being explained is not why leg oedema occurs in general at the total population level, but why this particular patient has leg oedema. To particularise the general regularity to the individual case, the covering law account treats the individual

case as a token of the general type to which the regularity applies. According to this approach, the individual case of leg oedema is explained by the diagnosis of heart failure, because it is a token of the type "leg oedema", and there is a causal regularity between the type "heart failure" and the type "leg oedema".

Indeed, in many cases, the explanation of the individual as if it is a token of a homogeneous type would turn out to yield the correct diagnosis. If a particular type of condition is statistically the commonest cause of a type of symptom in the total population, then it follows that most individual cases of this symptom would be caused by this condition. However, Benzi (2011: pp. 367–368) argues that this does not capture all cases of diagnostic explanation. She draws on the observation by Gorovitz and MacIntyre (1975) that what is crucially important about individual cases in medicine is what is distinctive about them as particulars. Far from being tokens of a homogeneous type, the particular cases of a certain clinical presentation are affected by so many contingencies as to make each case unique. Given this uniqueness, the general causal regularity appealed to in a covering law argument may fail to pick out the actual causal relation in a given case. In other words, the likeliest cause of a clinical presentation in the relevant reference class may not be the actual cause of the clinical presentation in a particular patient.

Consider Benzi's (2011: p. 369) example of a patient presenting to primary care with a new onset of leg oedema, which in this particular case turns out to be caused by acute kidney disease. Also consider that this patient is also known to already have a longstanding history of heart failure. Under the covering law account, the leg oedema could be explained with appeal to a causal regularity between kidney disease and leg oedema. However, in the primary care population, leg oedema is more likely to be caused by heart failure than by kidney disease. Hence, the causal regularity between heart failure and leg oedema would also satisfy the requirements of a covering law explanation, despite

this not being the actual cause of the leg oedema for this particular patient. The upshot is that appealing to general causal regularities cannot discern the actual explanation from the spurious one in the particular case, and so fails to capture what it is that makes the relation between a diagnosis and a set of patient data genuinely explanatory.

Benzi's solution, then, is to propose that the relation between a diagnosis and the patient data is explanatory not in virtue of a general causal regularity, but in virtue of the actual cause of the patient data in the given case. That is to say, a diagnosis explains the patient data if it identifies the actual cause of that patient data. Hence, in the above mentioned example, heart failure may be a more common cause of leg oedema than kidney disease in the general population, but the correct explanation of leg oedema in the given patient is kidney disease, not heart failure, because kidney disease is the actual cause of the leg oedema in that particular case.

The proponent of the covering law account might respond by suggesting that the relevant reference class to which the general causal regularity applies could be narrowed down by including the details of the contingencies emphasised by Gorovitz and MacIntyre (1975) in the description of the reference class. For example, the description of the relevant reference class would not simply be "leg oedema", but something like "leg oedema, male, elderly, smoker, hypertensive, diabetic, proteinuria, raised serum creatinine, family history of kidney disease …", which would strengthen the statistical relation between the reference class to which the patient belongs and the diagnosis of kidney disease. However, there are two problems with this suggestion. First, as argued by Nancy Cartwright (2005) and restated by Stefan Dragulinescu, (2012), a complete description that achieves absolute concordance between the reference class and the correct diagnosis may not be possible. Although we can include certain known risk factors in the description of a reference class, there are also many other contingencies for which we cannot account due to our ignorance of them (Gorovitz and MacIntyre, 1975:

p. 16). To paraphrase Cartwright (2005), there may be no available complete description, but simply individual variation. Second, even if, *à la* Laplace's demon, we were able to specify all of the relevant contingencies and include them in a description, the sheer number of contingencies required to achieve absolute concordance between a reference class and a diagnosis would make the reference class so narrow that we can no longer claim that what we are appealing to in diagnostic explanations are "general causal regularities" rather than instances of singular causation.

And so, an adequate causal account of diagnostic explanation cannot be based on general causal regularities, but needs to appeal to the notion of actual causation in each individual case. As argued by Benzi (2011), the *explanandum*, or the patient's clinical presentation, cannot be characterised as a token of a type, but as a distinctive particular. The *explanans*, or the diagnosis, explains by identifying the actual cause of the clinical presentation in the particular patient. This not only marks an ontological shift from Hempel's (1965a) covering law account due to the commitment to actual causal connections rather than regularities, but also an epistemic shift due to the move away from the claim that explanations are necessarily arguments.

What has been presented here is a descriptive account of what constitutes the explanatory relation between a diagnosis and the patient data, but it does have normative implications for how physicians should reason. It supports the idea, suggested by Dominick Rizzi (1994: p. 316), that while appeal to causal regularities is of relevance to the scientific understanding of what causes a condition in general, it is singular causation that is of relevance to the diagnostic process, where the goal is to ascertain the cause in the individual case. The importance of this is that one of the key functions of a diagnosis is to help determine the correct intervention for the given patient. Settling for the diagnosis of heart failure as an explanation of leg oedema on the grounds that it is normally the cause of leg oedema in general could have disastrous consequences for the

patient whose leg oedema is actually caused by a different condition. Of course, it may be that the precise identification of the actual cause in a given case is not immediately possible due to limitations of resources in the given setting, in which case the best that the physician can practically do may be to treat the patient as a token of a type and infer the most likely cause based on knowledge of causal regularities. I do not dispute that such reasoning may be justified, indeed likely to be successful, given the context. However, with respect to the epistemic status of the resulting relation between the conjectured diagnosis and the patient data, Benzi's (2011) analysis suggests that this relation would only be genuinely explanatory if the inferred likeliest cause does indeed match the actual cause of the patient data in the given case. A diagnosis that cites the wrong cause of the patient data cannot be said to explain the patient data.

3.3.3 Causes and mechanisms

Benzi (2011) is correct to characterise medical diagnoses as causal explanations of symptoms based on particulars. In clinical practice, the diagnostic process is normally aimed at discovering the pathology that is causing a particular patient's symptoms and signs. The diagnosis, which is the outcome of this process, often denotes this cause. For example, the diagnosis of acute appendicitis points to inflammation of the appendix as the cause of a patient's abdominal pain and the diagnosis of myocardial infarction points to ischaemic necrosis of the heart muscle as the cause of a patient's chest pain.

The above suggests that a diagnostic explanation assumes the form of a simple causal claim, "*C* causes *E*", where *C* is the pathology picked out by the diagnosis and *E* is the patient data in need of explanation. This conforms to a variety of causal explanation described by David Lewis, who writes that "an explainer might give information about the causal history of the *explanandum* by saying that a certain particular event is included therein" (Lewis, 1986b: p. 219). Benzi (2011: pp. 369–370) appears to assume this

approach in some passages, such as her counterfactual analysis of a heart problem and a kidney problem as potential causes of a patient's leg oedema.

While I agree with this characterisation of a diagnosis as identifying *C* as the cause of *E*, I argue that its explanatory strength also depends on an understanding of how *C* produces *E*. In other words, knowledge of the causative pathology needs to be supplemented with some knowledge of the mechanisms by which this pathology causes the symptoms. For example, the diagnosis of heart failure may point to the failure of the heart to pump sufficiently to meet the body's metabolic requirements as the cause of the patient's leg oedema, but this is of limited explanatory value unless it is accompanied by knowledge of the mechanisms by which this failure of the heart to pump sufficiently produces the leg oedema. While Benzi does briefly mention mechanisms in her discussion, it is not made clear how they fit into the account of causal explanation presented.

The role of mechanisms in explanation has recently received a lot of attention from philosophers of science. This is, to some degree, inspired by Wesley Salmon's (1984) mechanistic conception of causation, which contrasts with the counterfactual conception of causation advocated by David Lewis (1986a). However, more recent philosophers are in disagreement over how the precise nature of a mechanism should be understood. Some authors take causes to be reducible to mechanisms. For example, Stuart Glennan (1996) argues that causal relations can be explained by mechanisms, while Machamer *et al.* (2000) suggest that the concept of "cause" is vague and can be replaced with more precise mechanistic concepts such as "push", "carry", "burn", and so on. By contrast, James Woodward (2002) suggests that mechanisms are reducible to causes and can be analysed counterfactually. Nonetheless, despite these metaphysical disagreements, it is generally agreed that a mechanistic explanation for a phenomenon should include mention of component parts and their activities organised in such a way that they

produce the phenomenon. This is sufficient for my present analysis of medical explanation, and so the metaphysical debate regarding whether mechanisms can be reduced to causes or *vice versa* can be set aside.

The mechanistic conception of causal explanation has had considerable success in the philosophy of medicine with respect to analyses of disease causation. Examples include the analysis of the relation between smoking and bronchial carcinoma by Russo and Williamson (2007), Mauro Nervi's (2010) analysis of pathological processes, and Lindley Darden's (2013) discussion of the genetic basis of cystic fibrosis. Theorists such as those mentioned above argue that explanations that appeal to mechanisms are desirable in the biomedical sciences, because they provide more detail than simple causal claims, offer justification for believing that a correlation is genuinely causal, inform predictions about outcomes, and identify targets for intervention.

I argue that these also apply to the explanation of patient data in the clinical context. Knowledge of mechanisms makes the causal connection between a diagnosis and the patient data more intelligible. This is perhaps most obvious in the case where a pathological process located in one organ system produces symptoms and signs located in seemingly unrelated organ systems. For example, consider the case of a patient who presents with the recent onset of abdominal obesity, muscle weakness, and fragile skin, who is diagnosed with lung carcinoma. This may correctly identify the cause of the patient data, but it is of limited explanatory value on its own due to the apparent gap between cause and effect. However, the connection is more intelligible if we also know that a small cell lung tumour can secrete adrenocorticotropic hormone, which stimulates the adrenal glands to secrete cortisol, which in turn alters lipid and protein metabolism. Here, the presence of a plausible mechanistic story linking *C* and *E* provides justificatory support for the claim that *C* is the cause of *E*, thus substantiating the value of invoking *C* as a causal explanation of *E*.

Another reason this mechanistic knowledge is important is that it supports the prognostic and therapeutic aims of medicine. Holly Andersen argues that knowledge of mechanisms can "provide grounds for prediction about what would happen to a phenomenon of interest given specific interventions on it" (Andersen, 2010: p. 993). While identifying *C* as the cause of *E* may suggest that treatment ought to intervene on *C* or on somewhere along the causal chain from *C* to *E*, knowing the mechanisms by which *C* produces *E* allows us to isolate particular targets for intervention and, moreover, gives us an indication of how to intervene on these targets. This squares with the notion that the causal information required in an explanation is relative to our explanatory interests, which in clinical medicine are largely to inform prognosis, guide treatment, and prevention. Caroline Whitbeck, for example, argues that the diagnostic process aims for "whatever degree of identification is necessary to achieve the best outcome for the patient and to prevent the spread of disease" (Whitbeck, 1981: p. 322). For this purpose, it may not be enough merely to identify *C* as the cause of *E*, but we may also need to know further details of how *C* produces *E*. Conversely, the prognostic and therapeutic aims of medicine impose negative constraints on how much mechanistic detail is considered relevant in a causal explanation. As Mauro Nervi notes, refining a mechanistic account too much may yield "elementary biochemical events of little or no interest to the researcher" (Nervi, 2010: p. 227). Hence, details that do not aid prediction or intervention in any relevant way may be considered superfluous to the explanation.

The above considerations highlight the importance of mechanistic knowledge in the clinical context of diagnosis. While Benzi (2011) is correct that the contribution of the diagnosis is to identify the actual cause of the patient data, further knowledge of the mechanisms linking this identified cause and the patient data is usually needed for this to be of explanatory value. In the following section, I examine more closely the sources of this mechanistic knowledge.

3.3.4 Mechanisms in a theoretical framework

So far, I have argued that a diagnosis explains by identifying pathology *C* as the cause of the patient data *E*, but the explanatory value of "*C* causes *E*" also depends on understanding the mechanisms by which *C* produces *E*. This raises the question of whence this mechanistic knowledge comes. The account of actual causation presented in §3.3.2 would suggest that it is not explicitly contained in the diagnosis itself, which just identifies and denotes the causative pathology *C*. For example, the diagnosis of heart failure explicitly refers to the failure of the heart to pump sufficiently to meet the body's metabolic requirements, but this description by itself does not provide information about the mechanisms by which leg oedema is produced. Therefore, in such a case, the knowledge of mechanisms must come from sources beyond what is explicitly contained in the diagnosis itself. I suggest that it comes from the broader theoretical framework in which the physician operates.

Jeremy Simon (2008) presents a way of thinking about disease ontology that fits well with this idea. He argues that a model of a disease consists of an explicit description and an implicit addition. The explicit description is the specification of the intrinsic structure of an essential pathological feature. The implicit addition is relational, namely the assumption that this pathological feature is "embedded in an otherwise unspecified living human being, or, more precisely, in an abstract system representing the general physiological features of a living human being" (Simon, 2008: p. 360). For instance, he suggests that cystic fibrosis is defined, in essence, by an abnormal cystic fibrosis transmembrane conductance regulator (CFTR) ion transport system, but there is the implicit assumption that this abnormal CFTR ion transport system occurs within and influencing a broader physiological system. As noted by Simon, "[a] cell cannot have cystic fibrosis by itself" (Simon, 2008: p. 364). Although Simon's account is presented as a metaphysical analysis of the ontological structures of diseases rather than an account of

causal explanation in medicine, it does have an important epistemic implication, namely that knowledge of diseases is embedded within a broader theoretical framework of pathophysiological principles.

A useful way to think about the structure of this theoretical framework is provided by Kenneth Schaffner (1986). Drawing on Thomas Kuhn's (1962) notion that scientific practices take place in the context of a disciplinary matrix, Schaffner suggests that physicians have at their disposal a matrix of theoretical knowledge consisting of a "series of overlapping interlevel temporal models" (Schaffner, 1986: p. 68). He writes:

> Clinicians bring to the examination of individual patients a repository of classificatory or nosological generalizations, as well as a grounding in the basic sciences of biochemistry, histology, physiology, and the pathological variants of the 'normal' or 'healthy' processes. A theory in pathology can be construed as a family of models, each with 'something wrong' with the 'normal' or 'healthy' processes. (Schaffner, 1986: p. 71)

Schaffner suggests that the pathophysiological mechanisms in individual cases can be understood through application of the theoretical knowledge of the processes represented by these models. He argues that this does not involve the subsumption under universal laws as per Hempel's (1965a) covering law account of explanation, but a sort of qualitative comparison which he calls "*analogical extension of biological knowledge*" (Schaffner, 1986: p. 68). The reason for this is the variability between individuals. As noted in §3.3.2, individual patients are not tokens of a homogeneous type, but are unique particulars whose histories are influenced by various contingencies. Given this variability, Schaffner argues that the theoretical representations of mechanisms are idealisations:

> Such a set of overlapping or 'smeared out' models is then juxtaposed, often in a fairly loose way, with an overlapping or 'smeared out' set of patient exemplars. This dual 'smearedness' – one being in the basic biological models and the other in the patient population – typically requires that the clinician work extensively with analogical reasoning and with qualitative and at best *comparative* connecting pathophysiological principles. (Schaffner, 1986: p. 71)

In other words, the pathophysiological mechanisms represented by the theoretical models at best map partially onto the processes going on in individual cases.

However, Schaffner's account is presented as a general account of how theoretical knowledge is applied to cases in the biomedical sciences, not specifically an account of the explanatory functions served by diagnoses. As such, he does not explicitly make clear the particular role that making a diagnosis has in relation to the theoretical knowledge represented by the above mentioned models. We are not told, for instance, whether he conceives a given diagnosis, such as heart failure, as corresponding to a particular model, a particular node or region in a model, or a process involving multiple models.

When viewed in light of my above analysis of the respective contributions of causal claims and mechanistic causal knowledge, though, the relation between a clinical diagnosis and Schaffner's matrix of theoretical knowledge is made clear. The contribution of the diagnosis is the identification of the actual cause *C* of the patient data *E*, such as the diagnosis of heart failure identifying the failure of the heart to pump sufficiently as the cause of the patient's leg oedema. While this description of *C* does not explicitly contain information about the mechanisms by which leg oedema is produced, it is nonetheless implicitly contextualised within a broader matrix of theoretical knowledge consisting of overlapping models of pathophysiological mechanisms. The contribution of this matrix of theoretical knowledge, then, is to provide the background understanding of

the mechanisms that make the link between *C* and *E* intelligible. The upshot, then, is that the diagnosis explicitly identifies a pathology whose causal connection with the patient data is made intelligible in virtue of its being contextualised within a theoretical framework of mechanistic models.

It is worth mentioning three additional points to further clarify the relation between a diagnosis and the theoretical models of pathophysiological mechanisms. First, the mechanisms linking a given diagnosis and the patient data may cross a number of these overlapping models. It is usually the case that a disease has sequelae that affect multiple organ systems and span multiple levels. For example, while cystic fibrosis is, in essence, an abnormality of the CFTR ion transport system at the molecular level, it produces histological abnormalities of the mucosal epithelium, which in turn result in anatomical and physiological abnormalities of the gastrointestinal, respiratory, and reproductive systems (Simon, 2008; Darden, 2013). Understanding these mechanisms, then, often requires us to invoke models at different levels and of different organ systems. In the case of cystic fibrosis, we need to invoke models of ion transport across the cell membrane, mucous stasis in the airways and pancreatic ducts, chronic inflammation, and so forth.

Second, Schaffner describes the theoretical models each as representing "'something wrong' with the 'normal' or 'healthy' processes" (Schaffner, 1986: p. 71), but I suggest that this is not the only way of characterising pathophysiological mechanisms. A recent analysis by Mauro Nervi (2010) suggests that the theoretical understanding of how *C* and *E* are linked can consist of knowledge about mechanism malfunction, knowledge about pathological mechanisms, or a combination of both. The mechanism malfunction conception involves laying out the details of a normal physiological mechanism and depicting the pathology as an impairment of this normal mechanism. This conception aligns with the theoretical knowledge of "pathological variants of the 'normal' or 'healthy'

processes" described by Schaffner (1986: p. 71). For example, the mechanism of a cardiovascular problem can be explicated by laying out the physiological sequence of events that normally occur in a healthy circulatory system and showing how this sequence is interrupted (Nervi, 2010: p. 217). By contrast, the pathological mechanisms conception lays out the details of the pathological sequence of events without explicit reference to normal physiology. Although background knowledge of normal physiology is presupposed, the emphasis is on the progression of pathological processes. For example, the mechanism of diabetes insipidus can be characterised as decreased production of or sensitivity to antidiuretic hormone, lack of permeability of cells of the distal nephron, polyuria, dehydration, hypovolaemic shock, and cardiac arrest (Nervi, 2010: p. 219).

Third, while I think Schaffner (1986) is correct to claim that the theoretical models of pathophysiological mechanisms only partially fit the goings on in actual cases because of the variability across individuals, I argue that the diagnosis itself can still be considered a repeatable type as suggested in §3.3.1. This is because it is often, though by no means always, the case that a diagnosis is explicitly defined by some essential feature that is necessary for a case to qualify as an instance of that diagnosis. As such, every case of that diagnosis must instantiate that feature. A previously mentioned example from Simon is that of cystic fibrosis, which is explicitly defined by the essential feature of an abnormal CFTR ion transport system, such that "regardless of the reason a patient had problems with the CFTR pump system we would consider him to have cystic fibrosis" (Simon, 2008: p. 361) and that a person who does not have an abnormal CFTR does not, by definition, have cystic fibrosis. Similarly, heart failure is defined by the essential feature of the failure of the heart to pump blood at a rate adequate for satisfying the requirements of the tissues, such that only and all patients with heart failure instantiate this feature, despite any variability with respect to their symptoms, signs, and other physiological parameters. Hence, while different cases may deviate from the theoretical models of

pathophysiological mechanisms in varying respects and to different degrees, some diagnoses *qua* generalised categories can be taken to pick out certain repeatable processes embedded within the theoretical framework that are conserved across cases.

To put some of the above considerations into context, let us look at a mechanistic account of how heart failure produces leg oedema from *Davidson's Principles and Practice of Medicine*:

> In patients without valvular disease, the primary abnormality is impairment of ventricular function leading to a fall in cardiac output. This activates neurohumoral mechanisms that in normal physiological circumstances would support cardiac function, but in the setting of impaired ventricular function can lead to a deleterious increase in both afterload and preload. … Stimulation of the renin-angiotensin-aldosterone system leads to vasoconstriction, salt and water retention, and sympathetic nervous system activation. This is mediated by angiotensin II, a potent constrictor of arterioles in both the kidney and the systemic circulation. … Salt and water retention is promoted by the release of aldosterone, endothelin-1 (a potent vasoconstrictor peptide with marked effects on the renal vasculature) and, in severe heart failure, antidiuretic hormone (ADH). … The onset of pulmonary and peripheral oedema is due to high atrial pressures compounded by salt and water retention caused by impaired renal perfusion and secondary hyperaldosteronism. (Newby *et al.*, 2010: p. 544)

The above account demonstrates some of the above mentioned features of how theoretical models of pathophysiological mechanisms relate to a diagnosis. First, it describes mechanisms occurring in different organ systems and at different levels, including haemodynamic mechanisms concerning the regulation of blood pressure and

cardiac output, hormonal mechanisms concerning the stimulation and actions of the renin-angiotensin-aldosterone system, renal mechanisms of salt and water reabsorption, and the hydrostatic mechanisms of oedema formation. This supports the claim that while a diagnosis may explicitly refer to a pathological process in a particular organ system, understanding the mechanisms by which this produces the patient data may require us to invoke models of several other systems.

Second, in keeping with Nervi's (2010) discussion of the different ways mechanisms can be characterised in medicine, this account includes both information about mechanism malfunction and information about pathological mechanisms. Parts of it characterise the leg oedema resulting from heart failure as being due to interruptions of normal physiological mechanisms, including the impairment of ventricular function. Other parts of it detail the progression of pathological processes leading from heart failure to leg oedema, including stimulation of the renin-angiotensin-aldosterone system, salt and water retention, vasoconstriction, and raised atrial pressure.

Third, in keeping with the notion presented in §3.3.3 that knowledge of mechanisms is useful for the therapeutic aims of clinical medicine, the above account of heart failure identifies potential targets for treatment interventions. For example, stimulation of the renin-angiotensin-aldosterone system can be targeted by angiotensin-converting-enzyme inhibitors, sympathetic nervous system activation can be targeted by β-adrenoceptor antagonists, and salt and water retention can be targeted by loop diuretics. And so, while the diagnosis of heart failure tells us what is causing the patient's leg oedema, the importance of the theoretical understanding of the mechanisms by which it produces the leg oedema is that it tells us where and how to intervene.

3.4 Conclusion

This chapter has sought to clarify how diagnoses in clinical medicine provide explanations of patient data. I have argued that the covering law account is inadequate as a general account of diagnostic explanation, even if the general regularities appealed to are causal regularities, and endorsed Benzi's (2011) proposal that diagnostic explanation needs to be conceived of as the explanation of particulars based on the notion of actual causation. That is to say, a diagnosis identifies pathology *C* as the actual cause of the patient data *E* in the particular case. However, this simple causal claim is of limited explanatory value without some understanding of the mechanisms by which *C* produces *E*. Drawing on and bringing together Simon's (2008) work on disease ontology and Schaffner's (1986) work on analogical reasoning from theoretical models, I argued that this mechanistic knowledge is not always explicitly contained in the diagnosis itself, but comes from the broader theoretical framework within which the causal knowledge provided by the diagnosis is implicitly embedded.

4. The Semantics of Diagnostic Terms*

4.1 Introduction

Throughout Chapter 3, I showed that diagnoses in medicine are often invoked to explain patients' symptoms, and that they normally do so by indicating the causative pathologies responsible for producing the symptoms. In this current chapter, I return to the conceptual problem, introduced in Chapter 1, regarding whether diagnoses in psychiatry can also serve such explanatory roles. As previously noted, the language used in some clinical texts suggests that they do. However, this seems to be in tension with the latest editions of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), in which psychiatric diagnoses are defined by the clusters of symptoms themselves. Given that causes are distinct from their effects, it is a minimum requirement of a causal explanation that the *explanans* refers to something other than the *explanandum*. Hence, the fact that psychiatric diagnoses are defined in terms of their symptoms seems to suggest that they cannot legitimately be invoked as causal explanations of their symptoms in the ways that many medical diagnoses can.

This chapter explores how theories of reference in the philosophy of language can help to resolve the tension between the uses of psychiatric diagnoses in clinical discourse and their definitions in the *DSM*. The general aim is to show that descriptive definitions of diagnostic terms based on symptoms do not necessarily preclude these terms from referring to the causal profiles underlying these symptoms. In §4.2, I revisit the two kinds of talk regarding psychiatric diagnoses. I consider Jennifer Radden's (2003) distinction

---

* A version of this chapter has been published as: Maung, H. H. (2016c). "To What Do Psychiatric Diagnoses Refer? A Two-Dimensional Semantic Analysis of Diagnostic Terms". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 55: 1–10.

between descriptive and causal conceptions of diagnostic terms, and then examine the idea, endorsed by Carl Hempel (1965b) and Paul Thagard (1999), that the increasing scientific understanding of a disease involves a progressive change from the former to the latter conceptions. I discuss the worry that such a conceptual change implies semantic incommensurability between older and newer conceptions of a diagnostic term.

In §4.3, I consider the causal theory of reference, developed by Saul Kripke ([1972] 1980) and Hilary Putnam (1975a), as a more reasonable account of diagnostic terms that avoids the implication of semantic incommensurability. This also includes a discussion of Neil Williams' (2011a) analysis of Putnam's (1975b) disease kind essentialism and some required modifications to this model. Despite its merits, I argue that something more than the traditional causal theory of reference is required for an adequate analysis of diagnostic terms in psychiatry, on the grounds that the traditional causal theory of reference relegates the *DSM* diagnostic criteria to mere contingent features of the disorders rather than necessary conditions for applying the diagnoses.

In §4.4, I put forward a solution based on the framework of two-dimensional semantics, as developed by Robert Stalnaker (1978), David Chalmers (1996, 2010), and Frank Jackson (1998), which allows the causal analyses of diagnostic terms in psychiatry, while taking seriously their descriptive definitions in the *DSM*. This framework provides one possible way of characterising the semantics of diagnostic terms that is able to accommodate the two different ways in which psychiatric diagnoses are used. However, as I concede in §4.5 and discuss in depth in the following chapters, whether psychiatric diagnoses actually do provide satisfactory causal explanations of individual patients' symptoms is also dependent on empirical facts regarding whether there actually are sufficiently stable causal structures associated with the diagnostic categories. Hence, the claims I make about psychiatric diagnoses in this chapter are to be taken as linguistic claims about how certain expressions, namely diagnostic terms in psychiatry, operate. I

reserve ontological claims about the natures of the kinds denoted by these terms for Chapter 5.

4.2. Descriptive and causal conceptions of diagnostic terms

4.2.1 Two kinds of talk

As we saw in Chapter 1, there are two kinds of talk going on regarding psychiatric diagnoses. Some clinical textbooks and health information resources use psychiatric diagnoses as if they refer to the underlying conditions that cause sets of symptoms. Further to the examples previously provided in §1.1.1, this kind of talk is somewhat prevalent in study guides and textbooks aimed at trainee psychiatrists, especially in the sections on differential diagnoses where psychiatric diagnoses are considered alongside medical diagnoses as referring to potential causes of syndromes:

> There is no doubt that this is a psychotic illness, but what is *the cause*? … [I]f the illness had a more insidious onset (days or weeks) then paranoid schizophrenia would be a likely candidate. (Green, 2009: p. 88, italics added)

> Depression and anxiety *cause* tiredness as do some somatization disorders. … Anaemia, liver failure, coeliac disease, cancer, Parkinson's, alcohol overdose and rare disorders such as myasthenia gravis and motor neurone disease can also cause tiredness. (Wright *et al.*, 2010: p. 152, italics added)

> [T]he symptoms of poor concentration and impaired memory may be *due to* depression, rather than a degenerative brain disorder. (Gulati *et al.*, 2014: p. 139, italics added)

Similarly, *A Guide to Psychiatric Examination* lists schizophrenia, mania, and depression alongside dementia and medical conditions as "common causes of psychoses" (Aquilina and Warner, 2004: p. 79), while *Psychiatry: A Clinical Handbook* lists schizophrenia, schizotypal disorder, schizoaffective disorder, and other psychiatric diagnoses as "causes of psychosis" (Azam *et al.*, 2016: p. 44). These passages suggest that psychiatrists from early in their training are encouraged to think about diagnoses in psychiatry as being analogous to diagnoses in bodily medicine.

However, this contrasts with the formal definitions of psychiatric diagnoses. According to the American Psychiatric Association's *DSM*-5, which is the dominant classification system in psychiatry in use today, psychiatric diagnoses are defined through their symptoms, as demonstrated by the passages previously quoted in §1.1.2. For example, it is stated that the presence of one or more delusions is the "essential feature" of delusional disorder (American Psychiatric Association, 2013: p. 92) and that panic disorder "refers to" unexpected panic attacks (American Psychiatric Association, 2013: p. 209). These definitions suggest that the meanings of psychiatric diagnoses are not determined by the underlying causes of symptoms, but by the clusters of symptoms themselves.

From a sociological perspective, this conceptual unclarity is not entirely surprising. As argued by the feminist philosopher Sally Haslanger (2006), a term can express different concepts to fulfil different ideological functions in different contexts. For instance, Haslanger observes that "parent" is often defined as "immediate progenitor", but used in some contexts to mean "primary caregiver". She respectively terms these the manifest concept and the operative concept, and suggests that the divergences between the two can help reveal the ideological function of a term, as well as open up the manifest concept to normative critique. As I argued in Chapter 2, diagnoses typically serve a variety of epistemic, instrumental, and semiotic functions, including explaining

symptoms, guiding therapeutic interventions, mobilising resources, and legitimising sickness. The divergences between the definitions and uses of psychiatric diagnoses, then, might reflect the expectations for a diagnosis to function both as a label for certain kinds of behaviour and as a scientific explanation of certain distressing symptoms.

However, from an epistemological standpoint, I argue that the conceptual unclarity regarding diagnostic terms is problematic. First, in the case of psychiatric diagnoses, the two kinds of talk are in tension. At least since David Hume's ([1748] 2000) analysis of causation, it has generally been accepted in philosophy that causes are distinct from their effects. For example, David Lewis proposes that causation is to be analysed "in terms of counterfactual dependence between distinct events" (Lewis, 1986a: p. 191). Similarly, Marshall Swain states that for *c* to be called the cause of *e*, "then *c* and *e* must be *distinct*" (Swain, 1980: p. 155). This suggests that while it is possible, indeed common, for someone to be both the immediate progenitor and the primary caregiver of a child, a set of symptoms cannot be its own cause. Therefore, if psychiatric diagnoses are defined by clusters of symptoms as suggested by the *DSM*-5 definitions, then this seems to suggest that they cannot refer to the causes of these symptoms.

Second, although it is certainly the case that terms can express different things in different contexts, the two kinds of talk regarding psychiatric diagnoses often occur within the same context. The term "parent" can be taken to mean "immediate progenitor" or "primary caregiver", depending on whether one is defining the term in a biological context or whether one is a teacher writing parents' evening invitations, but the same psychiatrist who uses a diagnosis to pick out a set of symptoms may also invoke it as a causal explanation of the symptoms within the same clinical encounter. These concerns highlight the need for greater conceptual clarity concerning diagnostic terms in psychiatry.

4.2.2 Ontological descriptivism

A useful approach to characterising the different uses of diagnostic terms in psychiatry is provided by Jennifer Radden in "Is This Dame Melancholy? Equating today's Depression and Past Melancholia" (2003). Radden proposes that there are descriptive and causal conceptions of disorders. A descriptive conception provides a definition of a disorder that consists of a description of its symptoms, without mention of the causal structure underlying these symptoms. As noted earlier, this is the approach used by the most recent editions of the *DSM* to define psychiatric diagnoses. Descriptive conceptions of disorders also occasionally feature in bodily medicine, particularly in what are called syndromic definitions of disorders. For example, as noted in Chapter 2, the Centers for Disease Control and Prevention define chronic fatigue syndrome exclusively through its symptoms (Fukuda *et al.*, 1994). Another example is chronic bronchitis, whose definition includes "cough and sputum expectoration occurring on most days for at least 3 months of the year and for at least 2 consecutive years" (Braman, 2006: p. S104).

In contrast to a descriptive conception, a causal conception does not define a disorder through its symptoms, but in terms of the causal structure that normally produces these symptoms. Causal conceptions of disorders are very commonly used in bodily medicine. For example, as noted in Chapter 3, heart failure is not defined as the conjunction of shortness of breath, leg oedema, and other symptoms, but as the underlying state wherein "the heart's output is decreased or in which the heart is unable to pump blood at a rate adequate for satisfying the requirements of the tissues" (Denolin *et al.*, 1983: p. 445). Similarly, acute appendicitis is not defined as the conjunction of right-sided abdominal pain and other symptoms, but as acute inflammation of the appendix. As noted earlier, some clinical textbooks and health information resources seem to assume causal conceptions of psychiatric disorders.

Radden relates descriptive conceptions of disorders to a view she calls ontological descriptivism, according to which diagnostic terms refer solely to observable clusters of symptoms and not to any underlying causal structures. According to this view, the diagnostic term "major depressive disorder" refers to the conjunction of the patient's low mood, loss of interest, and other associated symptoms. This is not to say that this conjunction of symptoms does not have a cause, but merely that the term "major depressive disorder" does not refer to any causes. Under ontological descriptivism, then, clinical textbooks and health information resources are wrong when they cite psychiatric diagnoses as referring to the causes of certain symptoms. Rather, psychiatric diagnoses refer to the symptoms themselves.

Assuming ontological descriptivism regarding psychiatric diagnoses has important implications. As Radden acknowledges, it suggests that psychiatric diagnoses describe, but do not explain, symptoms:

> Although not without predictive power, an account that is descriptive is not, as such, explanatory. It merely describes. In spite of the commonplace and seemingly irresistible tendency to see explanatory advantage in the assertion that the symptoms of depression are caused by depression, if we accept descriptivism, there is none. (Radden, 2003: p. 46).

This recalls Thomas Szasz's (1960) argument that psychiatrists are wrong to invoke mental illnesses as causal explanations of certain behaviours because they are only shorthand descriptions of these behaviours. Radden also argues that ontological descriptivism precludes cross-historical and cross-cultural equations of certain disorders, such as that of today's depression and past melancholia, and of Western depression and Chinese depression. I return to this point in §4.2.4 in the context of incommensurability.

Further to the above points made by Radden, I argue that assuming ontological descriptivism has implications for how it is that symptoms constitute evidence in support of a diagnostic hypothesis. With a causal conception of a disorder, such as the term "acute appendicitis" being used to refer to acute inflammation of the appendix, the patient's symptoms of right-sided abdominal pain and fever support the diagnosis insofar as they constitute empirical data that can be explained by acute inflammation of the appendix. In this case, the diagnosis of acute appendicitis is an inference to the best explanation about what disease is causing the symptoms. By contrast, with a descriptive conception of a disorder, such as the term "major depressive disorder" being used to refer to the conjunction of low mood, loss of interest, and other associated symptoms, the presence of this symptom cluster in a patient supports the diagnosis insofar as it makes it true by definition. In this case, the diagnosis of major depressive disorder is deductively entailed by the fulfilment of the symptom criteria. This has particular relevance to process of differential diagnosis, where the clinician assesses the available clinical data to select the most appropriate diagnosis from a list of possible diagnoses. With a causal conception this is the case of evaluating which disease best explains the data, while with a descriptive conception it is the case of which definition is fulfilled by the data. Hence, whether we assume a descriptive conception or a causal conception of a disorder has epistemic consequences for clinical practice.

4.2.3 Conceptual change

The notion that there can be descriptive and causal conceptions of disorders is further complicated by a historical dimension. It has been suggested that the historical development of a diagnostic term involves a progressive change from descriptive to causal conceptions. Carl Hempel (1965b) proposes that a scientific discipline proceeds from an early observational stage, when the aim is to describe the phenomena being

studied, to later theoretical stages, when the aim is to explain the phenomena with appeal to general laws and theories. He argues that the classification of psychiatric disorders will follow this trend from descriptive to progressively more theoretical language.

Similarly, Paul Thagard (1999: pp. 118–134) proposes that disease understanding progresses from an early descriptive to later theoretical stages. He describes four stages of disease understanding, which are disease characterisation, cause specification, experimentation, and mechanism elaboration. Disease characterisation involves clustering together a set of associated symptoms and differentiating this cluster from the symptoms of other diseases, such as when the associated symptoms of weakness, swollen limbs, and bleeding gums in sailors were grouped together and characterised as the syndrome of scurvy. Cause specification involves observing factors that correlate with the disease and postulating them as possible aetiologies. In the case of scurvy, damp conditions, salted meat, and nutritional deficiency were observed to correlate with the syndrome, and so were postulated as possible causes. Experimentation involves the gathering of empirical evidence to support a causal hypothesis and disconfirm others, such as the experiments on animals that showed associations between nutrition and symptoms of scurvy. The final stage is the elaboration of the mechanisms linking the aetiology of the disease to its manifestations, such as the discovery that ascorbic acid deficiency leads to defective collagen synthesis, which produces the symptoms of scurvy.

It is also worth noting that diagnostic terms can undergo other sorts of conceptual change in addition to the sort proposed by Hempel and Thagard. For example, there may be a change in the descriptive definition of the disorder, such as when the symptom criteria for schizophrenia were modified between the publications of *DSM*-IV (1994) to *DSM*-5 (2013). There may also be a change from one causal conception to another, such as when Creutzfeldt-Jakob disease went from being considered a disease caused by slow viruses to a disease caused by prions.

The image of disease understanding proposed so far suggests that diagnostic terms undergo conceptual changes throughout their histories. According to Hempel and Thagard, a diagnostic term normally begins as a descriptive concept that refers to a set of associated symptoms, or a syndrome. As the aetiology and mechanisms underlying these symptoms are discovered, it becomes a concept that refers to what normally causes these symptoms. This suggests that psychiatric diagnoses do not currently refer to the causes of their symptoms, but that there is hope that they will in the future, as our theoretical understanding of the disorders increases.

4.2.4 Semantic incommensurability

The move from descriptive to causal conceptions of diseases is largely positive, as it allows greater explanatory power, more accurate prediction, improved prevention, and the development of targeted treatments. However, this conceptual change further complicates the question of what states of affair diagnostic terms denote. As shown by the scurvy example, a diagnostic term can refer to a conjunction of symptoms at one time and refer to the condition that normally causes these symptoms at a later time.

There is disagreement among philosophers over whether this conceptual change is harmless or whether it amounts to a more serious problem of semantic incommensurability. Thagard (1999) acknowledges that conceptual change does occur with changes in disease understanding, but presents this as being largely unproblematic. However, there are philosophers who propose that this kind of conceptual change implies radical incommensurability between the old and new concepts. Notably, Paul Feyerabend (1962) and Thomas Kuhn (1962, 2000) suggest that the gap of meaning between old and new conceptions of a term amounts to linguistic instability, thus precluding meaningful comparison between the term's uses before and after the conceptual change.

Kuhn's discussion of incommensurability first appears in *The Structure of Scientific Revolutions* (1962), where he argues that scientific change follows a cyclical pattern of problem solving in normal science, accumulation of anomalies leading to a crisis, revolutionary replacement of the old paradigm with a new paradigm, a new phase of normal science, and so on. He initially presents incommensurability as quite a general notion which he defends on perceptual, social, and linguistic grounds. Central to this notion of incommensurability is the psychological claim that perception and observation are theory-laden. That is to say, because the observation of data is inevitably influenced by the paradigm within which one is operating, there is no common neutral standard with which one can evaluate hypotheses from different paradigms against each other. In his later thinking, Kuhn (2000) expands more specifically on the linguistic aspect of incommensurability, whereby the above scientific change is marked by untranslatability between the old and new uses of a term (Bird, 2002).

In the same year as Kuhn's *The Structure of Scientific Revolutions*, a version of the incommensurability problem is presented by Feyerabend in "Explanation, Reduction and Empiricism" (1962). According to Feyerabend, the meanings of scientific terms are bestowed by the theories to which they respectively belong. Therefore, when there are theoretical changes, there are also corresponding changes in the meanings of these terms. For example, he notes that the meanings of the terms "temperature" and "entropy" changed when phenomenological thermodynamics was replaced by kinetic theory, and that the meanings of "mass", "length", and "time" changed when classical mechanics was replaced by relativistic mechanics.

Although the problem of semantic incommensurability is perhaps most commonly associated with Feyerabend and Kuhn, it be traced back even further. Its application to medicine goes at least as far back as Ludwik Fleck's *Genesis and Development of a Scientific Fact* ([1935] 1979). Using the example of syphilis, Fleck argues that new concepts of a

disease are not adequate substitutes for the old concepts. Throughout its history, "syphilis" had been defined as a disease that is treated by mercury, a set of characteristic symptoms, and then finally as *Treponema pallidum* infection. These different concepts have different extensions, and so cannot be equated. For example, "a disease that is treated by mercury" excludes treatment-resistant cases of *T. pallidum* infection and "a set of characteristic symptoms" excludes asymptomatic cases.

Radden (2003) reaches a similar formulation in her cross-historical comparison of pre-nineteenth century melancholia and today's depression. Here, the comparison is between two descriptive conceptions of what is often assumed to be the same disorder, as the terms "melancholia" and "depression" are often thought to refer to the same thing. However, Radden argues that they cannot be equated. First, there are differences between the symptom profiles of melancholia and depression, which suggests that the two are not coextensive. For instance, Radden observes that some cases of modern schizophrenia and obsessive-compulsive disorder would also qualify as cases of melancholia. Second, if ontological descriptivism is assumed and the term "depression" refers exclusively to a set of symptoms, then depression and melancholia cannot be equated on causal grounds, because causal factors are not part of the meaning of the term "depression".

Semantic incommensurability challenges the intuition that there is continuity between the past and present concepts of a disease. As previously noted, the term "syphilis" had been defined as a certain set of characteristic symptoms before it was later defined as *T. pallidum* infection. If, after the discovery of *T. pallidum*, it turns out that some of the previous cases diagnosed as syphilis on the basis of their symptoms were not caused by *T. pallidum*, then there is an intuition that such cases were false positives and that it turned out that they were not actually cases of syphilis. We might say that the physicians who identified such cases as syphilis turned out to be wrong. However, the

incommensurability problem suggests that we could not claim that they were wrong, because they were using a different meaning of "syphilis". With respect to their meaning of "syphilis", they were right.

This is untenable, because it seems to suggest that empirical discoveries in medical science do not actually increase our understanding of individual diseases. Joseph LaPorte (2004: p. 114) notes that what appears to be an increase in understanding of a term is actually a case of changing its meaning, so that it refers to a different state of affairs. Before the discovery of *T. pallidum*, "syphilis" used to refer to a set of characteristic symptoms. After the discovery of *T. pallidum*, it referred to *T. pallidum* infection. Let us call these concepts SYPHILIS-1 and SYPHILIS-2, respectively. Rather than resulting in an increase in the understanding of SYPHILIS-1, the discovery of *T. pallidum* resulted in "syphilis" being displaced from SYPHILIS-1 and attached instead onto SYPHILIS-2. Similarly, Howard Sankey (2009: p. 198) notes that if a later concept does not refer to the same phenomenon to which an earlier concept had referred, then the conceptual change does not constitute an increase in knowledge about the phenomenon to which the earlier concept had referred.

The implication of incommensurability would not only make cross-historical comparisons of disorders problematic, but also cross-cultural comparisons. As noted by Radden (2003: p. 44), it is often reported that people with depression in China present with different symptoms from people with depression in the West. In particular, Chinese depression is said to present predominantly with somatic symptoms such as back pain and headache, rather than mood symptoms. Radden even notes that in some cases, there is no apparent commonality between the symptoms of Chinese and Western depression. Again, she argues that if descriptivism is assumed and "depression" is defined exclusively through its symptoms, then Chinese and Western depression cannot be equated because of this lack of commonality between their symptom profiles.

4.3 The causal theory of reference

4.3.1 A solution to incommensurability

In contemporary discussions (LaPorte, 2004; Sankey, 2009), semantic incommensurability is normally presented as being a problem for analyses of conceptual change that presume the descriptive theory of reference. The descriptive theory of reference, as advocated by Gottlob Frege ([1892] 1952) and Bertrand Russell (1905), states that the sense, or intension, of a term consists of a description. The reference, or extension, of the term is what satisfies this description. However, as disease understanding changes, so does the description associated with a diagnostic term. Under the descriptive theory, this change in description amounts to a change in reference, as is suggested by the above case of syphilis where the old and new descriptions are not coextensive.

The problem of semantic incommensurability has attracted different responses. Some have responded by suggesting that the descriptive theory of reference can still be preserved by narrowing down the theory dependence of terms. Alexander Bird (2000, 2004) and Stefan Dragulinescu (2011) make the distinction between thick intensionalism and thin intensionalism. According to thick intensionalism, the term's intension is highly dependent on the contents of theory to which it belongs, such that it includes a wide range of theoretical conditions and descriptions. According to thin intensionalism, the term's intension is much narrower, such that only some theoretical conditions and descriptions are included in it. Dragulinescu (2011: pp. 252–253) argues that under thick intensionalism, where the term's intension includes many theoretical conditions, theoretical change would be likely to result in a change in reference. Moreover, Bird (2000: p. 174) argues that if a scientific realist position is assumed, then thick intensionalism increases the chance of the term having no extension, because it is likely for at least some part of a sophisticated theory not to be true. Thin intensionalism, by contrast, is less likely to be significantly affected by incommensurability, because a narrow

intension that includes fewer theoretical conditions and descriptions confers enough stability of reference across theoretical change. However, while the above strategy may work for comparisons of terms across changes in the underlying theories, it is unclear whether it is applicable to cases such as Radden's (2003) cross-cultural comparison of Chinese depression and Western depression, or the cross-historical comparison between *DSM*-IV schizophrenia and *DSM*-5 schizophrenia, where the differences in the intensions are not due to differences in the underlying theories, but differences in the explicit symptom-based descriptions.

Another highly influential response to semantic incommensurability is the causal theory of reference, developed by Saul Kripke ([1972] 1980) and Hilary Putnam (1975a). This altogether denies that reference is determined by a description, and instead states that it is determined by the nature of the phenomenon being investigated and its causal relation with the speaker. Kripke describes the processes of reference fixing and reference borrowing. Reference fixing involves the initial ostensive dubbing of a paradigmatic sample of the phenomenon by a speaker or group of speakers, such as the disease associated with neurodegeneration and progressive dementia being dubbed "Creutzfeldt-Jakob disease" in the 1920s (Thagard, 1999: p. 123). Reference borrowing involves the transmission of the dubbed term between speakers in the linguistic community via communicative exchanges.

Because reference determination depends on the nature of the phenomenon being investigated and not on the speaker's description associated with a term, the causal theory of reference offers a promising way around the problem of incommensurability. For example, the description associated with "syphilis" has changed over the years due to changes in disease understanding, but the term's reference has not changed, because it is fixed by the initial ostensive dubbing of the sample. Hence, changes in disease understanding do not generally result in changes in the term's meaning, but can result in

better knowledge of the same disease and of what correctly belongs in the extensions of the term.

This is not to say that it necessarily guarantees reference stability, as there are still conceivable scenarios where the referents of terms could change. Gareth Evans (1973) presents the example of "Madagascar", which was originally used by its inhabitants to refer to mainland Africa, before European explorers used it to refer to the island after misunderstanding their interlocutors. Furthermore, Helen Beebee (2013: p. 161) presents a hypothetical example where a paradigm shift results in the entire classificatory framework of chemistry being replaced. Because such categories as element, catalyst, organic, and compound would no longer be recognised, we would no longer be able to point to two chemical samples and truthfully say that they are of the same compound. Nonetheless, while it may not on its own guarantee reference stability, the causal theory of reference at least shows how reference stability is possible, indeed likely, across scientific change of the appropriate sort. I suggest that some cases of changing disease understanding that do not involve the classificatory framework of medicine being overturned can reasonably be considered to fall under this sort of scientific change.

4.3.2 Disease kind essentialism

According to the causal theory of reference, a term's extension is not determined by a description, but by the nature of the phenomenon in the external world. It is commonly supposed in the philosophical literature that the causal theory of reference implies, or at least is related to, a sort of essentialism, whereby a member of a kind has an essential property that is necessary for its identity as a member of the kind (Kripke, [1972] 1980; Ellis, 2001; Bird, 2004; Haukioja, 2015). Bird (2004: pp. 61–64) explicates the relation between the causal theory of reference and essentialism as follows. According to the causal theory of reference, a term such as "water" is a rigid designator that picks out the

same sort of substance in all circumstances. Through empirical discovery, it turns out *a posteriori* that this substance has a certain microstructure consisting of two hydrogen atoms and one oxygen atom, which is denoted with the chemical expression "H_2O". This chemical expression "H_2O" is also a rigid designator that picks out this same kind in all circumstances. From the premises that water is H_2O, and that "water" and "H_2O" are rigid designators, it follows that "water = H_2O" is necessarily true. This supports the idea that water has an essence, namely the microstructure H_2O, as it shows that something cannot be water if it is not H_2O.

It is worth noting here that a distinction can be made between intrinsic and relational essentialism. Intrinsic essentialism states that kind membership is determined by an intrinsic property of the phenomenon, such as its microstructure. For example, as noted above, the essence of water is its microstructure H_2O, such that something must be H_2O for it to be water. By contrast, relational essentialism states that kind membership is determined by a certain relation between the phenomenon and other phenomena, such as its causal history. For example, some philosophers argue that an organism's membership of a biological species depends on its phylogenetic lineage (Millikan, 1999; LaPorte, 2004).

Putnam (1975b) assumes essentialism about disease kinds and, in doing so, supports robustly causal conceptions of diagnostic terms. His position is expounded in detail by Neil Williams (2011a). According to Williams, Putnam proposes that a disease has a relational essence, namely its cause. For example, the essence of polio is poliovirus infection, such that all and only instances of illnesses that involve infection by polioviruses are cases of polio. A case of an illness that resembles polio in its symptom profile but which is not caused by poliovirus infection would not be a case of polio (Putnam, 1975b: p. 329). Conversely, an instance of poliovirus infection with atypical symptoms would still be a case of polio.

Note that one does not require prior knowledge of the nature of the cause of a disease to support Putnam's disease kind essentialism. The cause is discovered *a posteriori*, but this does not change the reference of the diagnostic term, which is fixed by the dubbing of the paradigmatic sample. Even before the discovery of poliovirus, one could still consider the essence of polio to be its hidden causal structure and postulate that all cases of polio share the same kind of causal structure. The subsequent empirical discovery of poliovirus elucidates the nature of this causal structure, and allows speakers to establish which cases have correctly and incorrectly been identified as cases of polio. Hence, "polio = poliovirus infection" is a necessary *a posteriori* fact. It is necessary because all and only cases of poliovirus infection are cases of polio, and it is *a posteriori* because this fact was discovered empirically (Kripke, [1972] 1980).

Williams also offers an analysis of Putnam's view on the causal relations between diseases and symptoms. First, Williams notes that Putnam rejects the descriptivist claim that a disease refers to a cluster of symptoms. For instance, Putnam states that "multiple sclerosis" does not mean "the simultaneous presence of such and such symptoms", but "that disease which is normally responsible for some or all of the following symptoms …" (Putnam, 1975b: p. 329). Second, as noted above, Putnam claims that the essence of a disease is its cause, such as poliovirus being the essence of polio. Given that it is generally accepted in philosophy that causes are distinct from their effects, this suggests that although poliovirus is essential for something to count as a case of polio, it is nonetheless distinct from the disease state of polio itself.

This disease kind essentialism, then, assumes a causal chain with three components, which are "the cause of the disease, the disease itself, and the symptoms of the disease (which are caused by the disease)" (Williams, 2011a: p. 167). For example, poliovirus causes polio, which in turn causes infantile paralysis. According to Williams' reading of Putnam, the cause of the disease is a relational essence that is distinct from the disease

itself. This suggests that a disease term refers neither to the cause of the disease nor to the symptoms of the disease, but to an intermediate link in the causal chain. That is, "polio" refers to the disease, poliovirus infection, which is caused by poliovirus and which causes the symptoms of infantile paralysis.

The distinctions between the three steps in this causal chain can be interpreted as paralleling the distinctions between aetiology, pathology, and clinical features that are assumed in clinical textbooks. Clinical features are the symptoms and signs with which the patient typically presents, pathology refers to the internal disease process that causes the clinical features, and aetiology refers to the more remote causal factors which are responsible for the pathology. For example, the entry on polio in *A Synopsis of Children's Diseases* by Rendle-Short and Gray (1967: pp. 386–387) states that the aetiology is the infectious organism poliovirus, the pathology is central nervous system destruction and muscle atrophy, and the clinical features include fever, malaise, and paralysis.

And so, an attraction of the three-step model is that it accommodates different kinds of causal explanatory talk in medicine. As noted above, the cause of the disease, the disease itself, and the symptoms of the disease are considered distinct nodes in a causal chain. This allows diseases to enter into causal explanations of symptoms, such as a case of infantile paralysis being explained by the diagnosis of polio and a case of paresis being explained by the diagnosis of syphilis. Moreover, it accounts for the way in which more general explanations of the diseases themselves appeal to their aetiologies, such as the disease polio being explained by poliovirus.

4.3.3 Some modifications

Although it has its merits, the analysis of disease terms presented here seems overly simplistic. As observed by Williams (2011a), the three-step model complements the germ theory of disease, according to which diseases are caused by pathogens and are classified

on the basis of pathogen species. While this remains relevant for such infectious diseases as polio and syphilis, it is unsuitable for diseases that do not have specific singular causes but result from multiple contributing factors. Hence, the three-step model as it stands does not offer the most charitable rendering of the causal theory as applied to diagnostic terms. I now consider two modifications that help to address this.

The first modification involves relaxing the restrictions on the kinds of phenomenon to which a disease term can refer. Putnam suggests that diseases have relational essences, namely their aetiological agents. However, as noted above, many diseases do not have specific singular causes, and so are not classified on the basis of aetiological agents, but instead on the basis of their pathophysiological mechanisms. In some cases, this can be resolved by replacing relational essentialism with a kind of intrinsic essentialism, such that the essence of a given disease is its pathophysiology. For example, the essence of bronchial carcinoma is uncontrolled cell growth in lung tissue and the essence of appendicitis is inflammation of the appendix. Hence, while different cases of bronchial carcinoma may result from different sets of aetiological factors, every case of bronchial carcinoma involves a particular pathophysiological mechanism, namely uncontrolled cell growth in lung tissue.

However, this may not be quite enough in other cases where the pathophysiological processes are more complex. As noted by Beebee and Sabbarton-Leary (2010: p. 19), there is no guarantee that a linguistic category that follows Kripke's semantics marks out a metaphysically distinctive category. In other words, there is no *a priori* reason to suppose that the paradigmatic sample to which a term is attached makes up an essentialistic kind. In "Arthritis and Nature's Joints" (2011b), Williams examines the category of rheumatoid arthritis. According to the diagnostic criteria for rheumatoid arthritis, a diagnosis can be made if a patient displays at least four of seven anatomical, pathological, and radiological features. Each individual criterion is neither sufficient nor

necessary for a diagnosis of rheumatoid arthritis. Moreover, different cases of rheumatoid arthritis may fulfil different combinations of criteria. In light of this, Williams suggests that rheumatoid arthritis does not have a simple essence. Rather, using a notion coined by the philosopher of biology Richard Boyd (1999), he proposes that rheumatoid arthritis is best conceived as a homeostatic property cluster. Under Boyd's account, members of a given kind do not have to share a single necessary property, but can share clusters of similarities that are causally connected. For example, Boyd suggests that biological species are homeostatic property clusters. Members of a species share a number of common properties, but there is significant variation within the species that no single property is essential for membership within that species. Similarly, Williams proposes that there is a cluster of properties that can be satisfied to varying degrees for something to be a case of rheumatoid arthritis, but it is neither sufficient nor necessary for any particular one of these properties to be satisfied. This potentially allows for more variability between the members of a kind, as different combinations of the properties may be satisfied for kind membership.

Although Williams' suggestion that some diagnostic terms are determined by homeostatic property clusters rather than simple essences is plausible, it is worth noting that such a move may not be necessary in his given example of rheumatoid arthritis. While there is indeed heterogeneity with respect to its clinical features, the pathophysiology of rheumatoid arthritis is somewhat tidier than Williams acknowledges, being characterised by the autoimmune reaction to autoantigens expressed in the joints (Boissier *et al.*, 2012). Therefore, the essentialist could justifiably claim that the essence of rheumatoid arthritis is the erosion of the joint surfaces by autoantibodies. Later in Chapter 5, I examine in more detail recent attempts to conceptualise some psychiatric disorders as homeostatic property clusters.

The second modification involves expanding the three-step causal chain into a more complex causal network. As noted by Thagard (1999), disease causation is usually a complex process with multiple interacting factors. Not only can there be numerous risk and protective factors that influence the development of the disease, but the disease itself can be a causal factor that influences the development of other diseases. This suggests that disease causation cannot be adequately modelled by a simple linear chain. Rather, a more complex causal network is needed to acknowledge the multifactorial aetiologies of some diseases. For example, myocardial infarction can result from the interactions of various causes, such as hypertension, obesity, smoking, and psychological stress, and itself can cause other diseases, such as congestive heart failure, arrhythmia, and cerebral embolism. Again, this is fully compatible with a causal conception of the diagnostic term. The term "myocardial infarction" still refers to the pathology that causes a patient's clinical features. However, rather than just being the intermediate link in a linear causal chain, the model acknowledges that myocardial infarction is embedded within a broader causal network and has complex causal connections with other diseases.

According to Thagard (1999: p. 114), the causal relations in such a network are intended to map onto the actual causal relations in individual cases of the disease. However, not every feature of the network has to be present in every instance of the disease. Thagard states that the causal relations in the model are not deterministic, but statistical. Hence, different instances of myocardial infarction may result from different combinations of aetiological factors. This seems to support the homeostatic property cluster theory, but is also consistent with the sort of intrinsic essentialism where the essence of the disease is its pathophysiology. For instance, it could be claimed that the essence of myocardial infarction is necrosis of the myocardium from prolonged ischaemia, but different cases could differ with respect to what had caused this ischaemic necrosis.

While the above modifications deviate from Putnam's (1975b) disease kind essentialism, I argue that they are compatible with the causal theory of reference and can permit causal conceptions of diagnostic terms. Reference fixing and reference borrowing of diagnostic terms can still proceed as described by Kripke ([1972] 1980), as can the scientific endeavour of discovering the causal structures of the kinds to which the terms refer. However, as well as accounting for disease kinds whose causal structures are determined by specific aetiological agents, the semantic framework can now also accommodate those that turn out to have other sorts of causal structure, including homeostatic property clusters and pathophysiological processes that are embedded within more complex causal networks.

4.3.4 Strengths of the causal theory of reference

To summarise this section, I presented the causal theory of reference as an account of how the reference of a diagnostic term is determined. According to this theory, it is not determined by a description, but by the actual nature of the disease and the causal relations between speakers who use the term. For Putnam (1975b), it is not the symptoms of a disease, but its cause that is essential for the individuation of meaning. The resulting essentialism implies a disease model consisting of three parts, namely the cause of the disease, the disease itself, and the symptoms of the disease. However, I argued that this model is too simplistic and suggested two modifications that allow a more permissive rendering of the causal theory. One modification, after Williams (2011b), is to allow homeostatic property clusters as well as simple essences as determinants of reference. The other modification, after Thagard (1999), is the expansion of the three-step chain into a more complex causal network.

The causal theory of reference supports a robustly causal conception of diagnostic terms, according to which diagnostic terms do not refer to sets of symptoms, but to the

disease processes that cause the symptoms. This is so even before the natures of these disease processes are fully known, as these can be discovered *a posteriori*. As noted in §4.3.1, this can help to avoid the problem of radical incommensurability that affects the descriptive theory of reference. Because reference is not determined by a description, the change in description that results from changing disease understanding do not normally amount to a change in reference. Hence, "syphilis" did not go from referring to a set of characteristic symptoms to referring to *T. pallidum* infection, but has referred to the same disease from the outset. The subsequent discovery of *T. pallidum* simply increased our knowledge of the nature of this disease.

In addition, these causal considerations explain why certain collections of symptoms are characterised by physicians and scientists into distinct syndromes. As noted by Williams (2011b), such actions are those of people who consider the associated symptoms to be connected by a unifying causal structure. In such case as syphilis, it turns out that the symptom cluster is actually the result of a singular kind of pathology. In other cases, it turns out that there are multiple different pathologies, each of which can cause the observed cluster of symptoms. For example, "dropsy" had been used for many centuries to refer to the alleged disease associated with fluid retention. However, it turned out that there are multiple different pathologies that could underlie cases of dropsy, and so the term was discarded and replaced by more specific diagnostic terms, such as "nephrotic syndrome", "congestive heart failure", and "cirrhosis of the liver" (Peitzman, 2007).

If we assume the account presented in this section, a diagnostic term in psychiatry, such as "major depressive disorder", does not refer to a cluster of symptoms, but to the disease process that causes these symptoms. This accounts for the way in which clinical texts use diagnostic terms in psychiatry to refer to the causes of symptoms, as noted in §4.2.1. Furthermore, it provides a possible solution to problem of cross-cultural

incommensurability presented by Radden's comparison of Chinese depression and Western depression in subsection §4.2.4. If "depression" is taken to refer to the putative disease process that produces various symptoms, then this accommodates the possibility of Chinese depression being considered the same disorder as Western depression, based on the assumption that they both involve this same disease process despite their having different symptom profiles.

However, in spite of these attractions, I argue that a pure causal theory has significant shortcomings regarding psychiatric diagnoses. In particular, I argue that by supporting robustly causal conceptions of psychiatric disorders, it downplays the important functions of their symptom-based diagnostic criteria in the *DSM*. In §4.4, I examine this in more detail and propose a two-dimensional semantic framework that preserves the core features of the causal theory while taking the descriptive diagnostic criteria seriously.

4.4 Two-dimensional semantics

4.4.1 Diagnostic criteria in psychiatry

As noted in §4.3, a key premise of the causal theory of reference as put forward by Kripke ([1972] 1980) and Putnam (1975a) is that the reference of a term is not determined by a description of superficial properties. Accordingly, Putnam (1975b) argues that descriptions of symptoms are not necessary to the meanings of diagnostic terms. Rather, descriptions of symptoms constitute stereotypes, which provide conventional ideas of what the disorders look like but are not analytically tied to their associated diagnoses. This does seem to be plausible for some of the medical diagnoses mentioned throughout this chapter, where the symptoms appear to be contingent properties of the diseases. For instance, a painless case of inflammation of the appendix is still a case of appendicitis. However, diagnostic terms in psychiatry, such as "panic

disorder" and "delusional disorder", do seem to allude to their symptoms in ways that suggest more fundamental connections between the symptoms and the disorders. It seems oxymoronic to claim that a person could have panic disorder without having panic attacks, or that a person could have delusional disorder without having delusions.

In response, the causal theorist could appeal to the difference between connotation and denotation. Kripke ([1972] 1980: p. 26) uses the example of the town, Dartmouth. The name may have the connotation of a location at the mouth of the River Dart, but this is not its denotation. According to Kripke, the town could still retain the name "Dartmouth", even if the River Dart changes its course, and so the connection between Dartmouth and its location at the mouth of the River Dart is contingent. Similarly, one might claim that the term "panic disorder" has the connotation of certain symptoms, but denotes the underlying disease that usually causes these symptoms.

However, this analogy is not wholly accurate. What it does not acknowledge is that the symptom-based definitions in *DSM*-5 are not just descriptions of disorders, but necessary conditions for applying the diagnostic terms. While Dartmouth may still retain its name if the River Dart changes course, *DSM*-5 precludes a diagnosis of panic disorder unless panic attacks are present. The symptom criteria in *DSM*-5 set the conditions that something must satisfy for it to qualify as an instance of the diagnosis. Hence, an analysis of the reference of such a diagnostic term as "panic disorder" would need to account for the fact that the presence of the relevant symptom cluster is necessary for the correct application of the diagnostic term. Again, this is unlike the case of a medical diagnosis such as acute appendicitis, where the presence of the stereotypical symptoms is not necessary for the diagnosis to be applied.

Interestingly, the same problem faces the response to Szasz's (1960) argument offered by Tim Thornton (2007: p. 18). Thornton draws on the work of Donald Davidson ([1967] 2001), who notes that events are frequently described in terms of their

effects, whilst actually referring to the causes of them. For example, the event, "firing a gun", can also be described as "the cause of the death of the president". The proposition, "firing a gun caused the death of the president", states a causal connection. However, if one describes "firing a gun" as "the cause of the death of the president", then the proposition becomes "the cause of the death of the president caused the death of the president", which states a necessary connection. Therefore, a causal connection appears to become a necessary connection due to the way in which it is expressed. Thornton proposes that Davidson's analysis can be applied to psychiatric diagnoses. Although psychiatric diagnoses may be described in terms of their effects, they actually refer to their causes.

Again, I argue that this analogy is not accurate, because the connection between "firing a gun" as "the cause of the death of the president" is itself contingent in a way that the connection between "panic disorder" and "recurrent unexpected panic attacks" is not. While the act of firing a gun may have been the cause of the death of the president in this world, it is conceivable in another possible world that it may not have been. For example, the gun may have fired but the president may have survived. However, the same cannot be said about the relation between panic disorder and panic attacks. According to *DSM*-5, panic attacks are essential to the diagnosis of panic disorder, and so it does not seem to make sense to say that a patient can have panic disorder without having panic attacks. Therefore, there is a necessary connection between the diagnosis of panic disorder and having panic attacks which is not present between firing a gun and the death of the president.

The above considerations suggest that a pure causal theory of reference is not adequate for the analysis of diagnostic terms in psychiatry, because it relegates the symptom-based diagnostic criteria to mere contingent features of the disorders. This contradicts the important functions of these symptom criteria as necessary conditions for

applications of the diagnostic terms. In what is to follow, I show how the framework of two-dimensional semantics can overcome this challenge.

4.4.2 An overview of two-dimensional semantics

Before I consider its application to psychiatric diagnoses specifically, I want to lay out the motivations for two-dimensional semantics in more detail. I do so by considering a scientific term that has featured extensively as a standard example in the literature on two-dimensional semantics, namely "water". Not only is this the example which has received perhaps the most detailed analysis by proponents of two-dimensional semantics (Chalmers, 1996; Jackson, 1998), but it is an example that has been used to demonstrate a problem with the causal theory of reference similar to the problem I suggest is presented by certain diagnostic terms. As I note in §4.4.1, the presence of panic attacks seems to be closely connected to the meaning of "panic disorder", such that it is counterintuitive to think of a case of panic disorder without panic attacks. Similarly, it has been argued that notwithstanding the causal theory of reference, the superficial properties of water are still somehow relevant to the meaning of "water". For instance, David Barnett (2000), makes the case that if we were to encounter extremely toxic mushroom-like items on a distant planet, it would be counterintuitive to say that these are "water", even if they turned out to be composed of unfamiliar configurations of H_2O molecules. Hence, although my application of two-dimensional semantics is intended to be a particular solution for the problem of psychiatric diagnoses, my aim in this section is to explicate the framework by presenting a familiar paradigm case where it has been applied and from which parallels can be drawn to the cases of diagnostic terms.

As previously noted, the causal theory of reference states that the reference of a term such as "water" is not determined by a description of water's superficial properties, but by what turns out to be its essence, namely the microstructure H_2O. According to the

causal theory of reference, then, "water" has a single intension, which rigidly designates H_2O. This is taken to apply across all possible worlds, such that "water = H_2O" is necessarily true. However, as noted by Chalmers (1996), there remains an intuition that "water" and "H_2O" differ in some aspect of meaning. The two are not epistemically equivalent. For instance, one could know that the potable liquid found in rivers is water, and not know that it is H_2O. Furthermore, although the potable liquid found in rivers which speakers had dubbed "water" actually did turn out to be H_2O in our world, we can still entertain a hypothetical scenario in which this liquid we call "water" was discovered to be something else.

These intuitions suggest that there is more to the meaning of "water" than having the microstructure H_2O. In the literature (Chalmers, 1996; Jackson, 1998), this is normally explicated in modal terms with a retelling of Putnam's (1975a) Twin Earth thought experiment. Twin Earth is indistinguishable from Earth in almost every way. Like Earth, its rivers contain a colourless, tasteless, and potable liquid, which its inhabitants call "water". The difference between the two worlds is that the stuff we call "water" on Earth was discovered to have the microstructure H_2O, whereas the corresponding stuff on Twin Earth that is called "water" by its inhabitants was discovered to have the microstructure XYZ. According to the causal theory of reference put forward by Kripke and Putnam, this Twin Earth liquid is not water. Because "water" designates rigidly and has been shown by chemistry to pick out H_2O on Earth, "water = H_2O" is a necessary truth that holds across all worlds. Hence, Twin Earthlings are wrong to claim that "water = XYZ". However, if a causal theorist from Twin Earth were to apply the same standards, then he or she would arrive at the opposite conclusion and claim that we Earthlings are wrong to call our liquid "water", because the liquid that Twin Earthlings had dubbed "water" was shown by chemistry to be XYZ. This suggests that if

the world had turned out to be like Twin Earth, then "water" would refer to XYZ instead of H_2O.

The above brings out the tension between the causal theorist's claim that "water" necessarily refers to H_2O and the intuition that it could have referred to something else had the world turned out to be different in the relevant way. Two-dimensional semantics resolves this tension. This is a formal framework developed by Robert Stalnaker (1978), and later championed by David Chalmers (1996, 2010) and Frank Jackson (1998). It proposes that the meaning of "water" is not only dependent on *a posteriori* facts about the world, but also on which possible world we assumed to be the actual world in which the reference is fixed. In the framework of Kripke and Putnam, only Earth is taken to be actual, while all other possible worlds are taken to be counterfactual. Hence, in this scenario, "water" only picks out the substance that was discovered to be H_2O. However, if one assumes Twin Earth to be the actual world that one inhabits, then "water" would pick out the substance that was discovered to be XYZ.

Two-dimensional semantics, then, proposes that a given term, such as "water", is taken to express two intensions. Different authors use different names for these two intensions, but here I follow the terminology of Chalmers (1996). The primary intension of a term is what the term would pick out in a chosen world if that world is imagined to be the actual world in which the reference is fixed. Given that the reference fixing occurs before the discovery of the underlying essential nature of the phenomenon in question, the primary intension roughly approximates to the phenomenon's pre-theoretical mode of presentation. Chalmers (1996: pp. 56–65) initially suggests that the primary intension may be determined by a description of its referent. However, he later concedes that this is not the case for all primary intensions, as there may be some terms that cannot be adequately encapsulated in descriptions (Chalmers, 2002: pp. 143–149). Instead, he suggests that primary intensions can be taken as capturing what the extensions of terms

would be in different epistemic possibilities about how the world could turn out. While these may sometimes be captured by descriptions, this need not always be the case. For instance, the primary intension of "water" (1-WATER) roughly corresponds to the colourless, tasteless, and potable liquid found in rivers. As mentioned above, this liquid that was dubbed "water" turns out to be H_2O in the scenario where Earth is imagined to be the actual world in which the reference is fixed, but turns out to be XYZ in the scenario where Twin Earth is imagined to be actual.

The secondary intension of a term is what the term picks out if we fix our world as actual and then evaluate other worlds as counterfactual relative to it. The secondary intension of "water" (2-WATER) only picks out H_2O, because water was discovered to be H_2O in our world. This identity is taken to be necessary, such that "water" refers to H_2O across all counterfactual worlds. Note that the secondary intension corresponds to the reference of a term as per the causal theory of reference. The reference of 2-WATER is not determined by a cluster of descriptions, but by the microstructure H_2O. Therefore, two-dimensional semantics assimilates the causal theory of reference. Along with it, I argue that it can assimilate the modifications to the causal theory of reference presented in §4.3.3, such that homeostatic property clusters as well as simple essences may be the determinants of secondary intensions.

The above modal story accounts for the way in which there can be two dimensions of a term's meaning, one which is dependent on the pre-theoretical mode of presentation of the phenomenon associated with the term and another which is dependent on what the underlying nature of this phenomenon *a posteriori* turns to be. It also has implications for the notions of necessity and contingency. These implications are useful for understanding the conceptual relations between a term and its associated concepts. For example, the sorts of relation that "water" has with "H_2O" and with "potable liquid found in rivers" depend on whether we assume 1-WATER or 2-WATER. If 1-WATER

is assumed, then "water = potable liquid found in rivers" is necessarily true, because it is defined by this mode of presentation, while "water = H_2O" is contingently true, because the potable liquid that was dubbed "water" in another world could have turned out to be something other than H_2O. On the other hand, if 2-WATER is assumed, then "water = H_2O" is necessarily true, because the liquid that was dubbed "water" in our world *a posteriori* turned out to be H_2O, while "water = potable liquid found in rivers" is contingently true, because H_2O could have had a different form in a counterfactual world.

This can be captured by the idea that there are two sorts of necessity and two sorts of contingency. These correspond to the necessity and contingency when the primary intension of a term is assumed, respectively 1-necessity and 1-contingency, and to the necessity and contingency when the secondary intension of the term is assumed, respectively 2-necessity and 2-contingency (Chalmers, 2010: p. 167). Hence, "water = potable liquid found in rivers", is 1-necessary but 2-contingent, whereas "water = H_2O" is 2-necessary but 1-contingent. The relation of 1-necessity can be thought of as corresponding to the definitional relation between a term and a description, while 2-necessity corresponds to Kripke's ([1972] 1980) *a posteriori* necessity as per the causal theory of reference. As I shall show in the following subsection, this offers a way of analysing the conceptual relations between psychiatric diagnoses, *DSM*-5 symptom criteria, and the pathological processes purported to cause these symptoms.

And so, the two-dimensional semantic framework presented here can be thought of as one way of synthesising the causal theory of reference with descriptivist considerations to capture different aspects of a term's complex semantic value that have useful epistemic roles (Chalmers, 2010: p. 563). In virtue of a term's secondary intension, a causal theorist can accept Chalmers' account of how the reference of that term is determined. However, while the causal theorist might consider this to constitute the

entire meaning of the term, for Chalmers it only constitutes one aspect of its meaning. The term also has a primary intension that corresponds to its mode of presentation. An implication of this is that the two-dimensional semantic theorist can reject the causal theorist's claim that a definition based on superficial properties is not relevant to the reference of a term. A definition is not merely a stereotype, but is a genuine aspect of a term's meaning. As we shall see, this makes two-dimensional semantics better suited than a pure causal theory for the analysis of diagnostic terms in psychiatry.

4.4.3 A two-dimensional semantic account of diagnostic terms

I propose that two-dimensional semantics can make sense of terms whose applications necessitate the presence of certain superficial properties despite the terms being used by speakers to refer to the causes of these properties. This is particularly relevant to psychiatry, where the meanings of diagnostic terms are necessarily tied to descriptions of symptoms despite the terms being invoked to refer to the causes of these symptoms. Although this particular issue is not so obviously a problem for many diagnoses in somatic medicine, the framework can nonetheless accommodate the analysis of medical diagnoses as well.

In §4.4.2, I explicated the principles of the framework with appeal to the classic example of the term "water". I suggest that diagnostic terms are amenable to the same kind of analysis. Like the term "water", a diagnostic term has a pre-theoretical mode of presentation that characterises the primary intension, and an underlying structure that is discovered *a posteriori* and determines the secondary intension. In the case of the term "water", the mode of presentation is roughly the colourless, tasteless, and potable liquid found in lakes and rivers, whereas the *a posteriori* discovered underlying structure is H_2O. In the case of a diagnosis, the mode of presentation is the clinical manifestation and the underlying structure is the disease process that is responsible for the clinical

manifestation. For example, the primary intension of the term "polio" is roughly the transmissible condition presenting with infantile paralysis, whereas the secondary intension is poliovirus infection.

In psychiatry, the clinical manifestations for diagnoses are codified in the *DSM*-5 definitions. A *DSM*-5 definition, then, captures the primary intension of a diagnosis. For example, the primary intension of "delusional disorder" includes "the presence of one or more delusions that persist for at least 1 month" (American Psychiatric Association, 2013: p. 92) and the primary intension of "panic disorder" includes "recurrent unexpected panic attacks" (American Psychiatric Association, 2013: p. 209). As previously mentioned, these definitions serve as necessary criteria for the diagnoses, which squares neatly with Chalmers' (2002: p. 143) suggestion that the role of a description is to provide conditions that give speakers ways to identify the extension of the term. The secondary intensions of diagnoses are whatever turn out to be their respective underlying pathological processes, as per the causal theory of reference.

It is worth noting here that the primary intensions may change across time, as demonstrated by the changes in the criteria for schizophrenia from *DSM*-IV (1994) to *DSM*-5 (2013) mentioned in §4.2.3. This is compatible with Chalmers' account, as he accepts that certain kinds of conceptual change involve changes in an expression's primary intension (Chalmers, 2012: p. 210). However, there is at least the possibility of semantic incommensurability being avoided here by the assumption that the secondary intension is sufficiently invariant. Hence, although *DSM*-IV schizophrenia and *DSM*-5 schizophrenia have different primary intensions, they are assumed to have the same secondary intension in virtue of their being posited to refer to the same causative pathology. This highlights the point made in §4.2.3 that diagnostic terms also go through sorts of conceptual change other than the changes from descriptive to causal conceptions suggested by Hempel (1965b) and Thagard (1999). In particular, they can undergo

changes in the descriptive definitions, which amount to changes in their primary intensions.

Furthermore, Chalmers also suggests that it is possible at a given time for concepts to have different primary intensions but the same secondary intension. For example, "Hesperus" and "Phosphorus" have different primary intensions, as the former picks out the evening star and the latter picks out the morning star, but they have the same secondary intension, as both refer to the planet Venus (Chalmers, 1996: p. 65). I suggest that this can be applied to Radden's cross-cultural discussion of Chinese depression and Western depression. The differences in the symptom profiles of Chinese depression and Western depression can be taken to constitute different primary intensions, but there is the possibility that the two can be equated on the basis of the assumption that they have the same secondary intension, hence potentially avoiding the implication of cross-cultural semantic incommensurability. Of course, whether they can actually be equated depends on the empirical question of whether they turn out to share the same kind of underlying causal structure.

Analysing diagnostic terms in psychiatry as having primary intensions and secondary intensions allows us to take their descriptive definitions in *DSM*-5 seriously as necessary criteria making the diagnoses, yet still talk about the diagnoses as referring to the causes of the symptoms that make up these definitions. Let us consider, for example, the connection between "panic disorder" and the *DSM*-5 description "recurrent unexpected panic attacks". If the primary intension of "panic disorder" is assumed, then the connection is necessary, because the primary intension is defined through this *DSM*-5 description. This 1-necessity reflects the way in which the *DSM*-5 symptom criteria are explicitly required for the diagnosis to be made. A diagnosis of panic disorder, for instance, cannot be made unless panic attacks are present. Therefore, unlike a pure causal theory of reference, two-dimensional semantics does not relegate the *DSM*-5 symptom

criteria to contingent features of the disorder, but acknowledges they are necessarily tied to the diagnosis in virtue of the diagnostic term's primary intension.

If the secondary intension of "panic disorder" is assumed, then the term refers to whatever turns out to be the pathology that normally causes recurrent unexpected panic attacks. The connection between the secondary intension of "panic disorder" and "recurrent unexpected panic attacks" is contingent, because it is counterfactually conceivable that the pathology picked out by the secondary intension could be present without it being accompanied by recurrent unexpected panic attacks, just as it is conceivable for inflammation of the appendix to be present without it being accompanied by abdominal pain. This 2-contingency reflects the way in which the diagnostic term is used in medical textbooks and health information resources to refer to what is causing a set of symptoms, such as panic disorder being invoked as the cause of a patient's panic attacks.

A two-dimensional semantic analysis, then, provides one possible way of resolving the tension between the *DSM*-5 definitions of psychiatric diagnoses as symptom clusters and their uses in other clinical texts as terms that refer to the causes of these symptom clusters. Under this framework, a diagnostic term does not have a single intension, but a complex semantic value involving a primary intension and a secondary intension. These two intensions have different epistemic roles that capture the two kinds of talk mentioned above. In virtue of their primary intensions, diagnostic term are defined through their symptoms, thus capturing their symptom-based definitions in *DSM*-5. In virtue of their secondary intensions, they refer to the pathologies that normally cause these symptoms, thus capturing their uses as explanations of patients' symptoms in textbooks and health information resources. This suggests, *pace* Szasz (1960), that although a psychiatric diagnosis is defined through its symptoms, this does not necessarily preclude it from being invoked as a causal explanation of these symptoms.

4.4.4 Other implications of two-dimensional semantics

As well as resolving the tension between the two kinds of talk regarding psychiatric diagnoses, two-dimensional semantics has further strengths as a framework for the analysis of diagnostic terms more generally. First, its semantic pluralism helps to characterise the different kinds of information conveyed by diagnostic terms in the communicative exchanges of clinicians. A diagnosis normally provides both information about a patient's likely clinical presentation and information about the underlying disease process, which respectively correspond to the primary intension and secondary intension of the diagnostic term. For example, "chronic bronchitis" not only informs the clinician that the patient is likely to be presenting with cough and sputum expectoration, but also that the underlying disease process is inflammation of the bronchi.

Second, two-dimensional semantics not only provides a way of interpreting changes in disease understanding that does not imply radical incommensurability, but also has the added advantage of taking seriously the different epistemic possibilities entertained by scientists in the early stages of disease understanding. Before the nature of the underlying pathology is understood, speakers rely on the primary intension of the disease term. For example, before poliovirus was discovered by Karl Landsteiner and Erwin Popper in 1909, doctors applied the term "polio" to cases of infantile paralysis, while aiming to elucidate the underlying causal structure. This primary intension analysis allows for the intuition that the condition presenting with infantile paralysis that was dubbed "polio" could have turned out to be caused by something else had the world been different in the relevant way. In actuality, polio turned out to be caused by poliovirus, which indicates that "polio = poliovirus infection" is 2-necessary. However, before poliovirus was discovered, Landsteiner and Popper had initially tried to look for a responsible bacterial agent for polio (Skern, 2010: p. 1372). This suggests that they had entertained the epistemic possibility that polio could have turned out not to be poliovirus

infection. If this epistemic possible world had turned out to be our actual world, then the condition presenting with infantile paralysis that was dubbed "polio" would have turned out to be a sort of bacterial infection, just as the liquid that was dubbed "water" would have turned out to be XYZ if Twin Earth had turned out to be our actual world. This scenario can be captured with the analysis that "polio = poliovirus infection" is 1-contingent.

Once the nature of polio's underlying pathology was discovered to be poliovirus infection, speakers could then utilise the secondary intension, which is determined by rigidifying this evaluation so that "polio" refers only to poliovirus infection across all worlds. Because the secondary intension fixes the reference of "polio" across all worlds, it establishes which cases of infantile paralysis are cases of polio and which ones are not. Hence, cases of infantile paralysis in the past that were not caused by poliovirus infection were not actually cases of polio. It is this secondary intension that is central to the aims of further scientific research into prevention and treatment. When Jonas Salk and Albert Sabin were developing vaccines for polio, they were developing vaccines specifically to prevent poliovirus infection.

The above suggests that the move from descriptive to causal conceptions of a diagnostic term involves the change in emphasis from the term's primary intension to its secondary intension. This does not involve the semantic incommensurability permitted by the descriptive theory of reference, because the determination of a diagnostic term's reference follows the same processes of reference fixing and borrowing as proposed by the causal theory of reference. The secondary intension of the term rigidly designates what the causal structure of a disease turns out to be, which maintains reference stability. Hence, "polio" denotes only genuine cases of poliovirus infection. Nonetheless, the primary intension accounts for the epistemically possible scenarios where the condition

that was dubbed "polio" had turned out to be caused by other kinds of agent instead of poliovirus.

4.4.5 Objections and replies

I now address some challenges to generalised two-dimensional semantics. Some of these are objections found in the literature, while others are potential challenges that could be raised. Although my own application of two-dimensional semantics is restricted to analysing diagnostic terms, it is nonetheless worthwhile answering these challenges to the framework as a more general theory.

The first challenge is an objection by Diego Marconi (2004), who argues that generalised two-dimensional semantics is implausible because it suggests that all ordinary expressions are ambiguous. If a term has both a primary intension and a secondary intension, then it could refer to either one of two different things. Of course, Marconi concedes that there are some ordinary terms which express different things in different contexts. An example already considered in §4.2.1 is the term "parent", which Haslanger (2006) observes could be interpreted as the immediate progenitor or the primary caregiver of a child. However, according to Marconi, not all ordinary expressions are obviously ambiguous in this way, and so two-dimensional semantics cannot provide a general framework to analyse ordinary expressions.

In reply, I argue that two-dimensional semantics does not entail that terms are ambiguous. Rather, as noted by Chalmers (2010: p. 563), two-dimensional semantics states that a term has a complex semantic value involving a primary intension and a secondary intension, and it has this complex semantic value in all contexts. Furthermore, for a given term, the primary intension and the secondary intension may be coextensive in the actual context of utterance. Consider, for example, the term "water". At first glance, the suggestion that 1-water roughly picks out the potable liquid found in lakes and

rivers while 2-water picks out H_2O might seem to suggest that the term "water" is ambiguous. However, on Earth, it is an empirical fact that the potable liquid found in lakes and rivers is H_2O. Hence, on Earth, the term "water" refers only to the substance that has the molecular structure H_2O, regardless of whether 1-water or 2-water is assumed. The primary intension and secondary intension of "water" only come apart when different modal possibilities are considered, such as scenarios involving Twin Earth or other distant worlds. This suggests that "water", in our ordinary usage of the term in the actual world, is not ambiguous, because the primary intension and secondary intension refer to the same thing on Earth, even though they are associated with different modal relations when other possible worlds are considered. I argue that the same sort of analysis could also be applied to diagnostic terms.

The second challenge is an objection by Scott Soames (2005), who argues that two-dimensional semantics vindicates internalism about meaning. Internalism is the view that meaning is individuated by the internal psychological state of a speaker. This is contrasted with externalism, which states that it is at least partly individuated by the speaker's external environment. Indeed, some proponents of two-dimensional semantics, including Chalmers (1996) and Jackson (1998), suggest that primary intensions are determined by the internal states of speakers. Soames objects to this on the basis that it contravenes the important externalist consequences of the causal theory of reference developed by Kripke ([1972] 1980) and Putnam (1975a).

In reply, I argue that Soames' objection is not applicable to all varieties of two-dimensional semantics, as not all proponents of two-dimensional semantics favour an internalist account of primary intensions. For example, Stalnaker's (1978) interpretation of two-dimensional semantics assumes externalism about intensions. Even Chalmers (1996: pp. 58–59) concedes the possibility that primary intensions might be determined by appropriate causal relations between the referents and the speakers, as per the causal

theory of reference. I propose that my own restricted application of two-dimensional semantics to psychiatric diagnoses is compatible with externalism, as the descriptions that express the primary intensions of the terms are not determined by internal states of the speakers, but are codified in an external resource, namely *DSM*-5.

A third potential challenge is that it is not clear why there should be only two dimensions of meaning. There may be other ways to break down the meaning of a term, and it is plausible that there are other aspects of a term's semantic content that are not captured by a primary intension and a secondary intension. Therefore, the worry is that two-dimensional semantics is too narrow a framework to completely capture the full meanings of terms.

In response to this, I emphasise that the two-dimensional semantic framework I have presented is not to be taken as providing an exhaustive account of the meanings of terms. Rather, as noted by Chalmers (2010: p. 556), two-dimensional semantics is compatible with semantic pluralism, which allows a term to be associated with a number of different semantic relations. The primary intension and secondary intension of the term do not exhaust the meaningful content of the term, but are ways of capturing two of the aspects of a term's complex semantic value. These are not two arbitrary aspects, but two aspects whose semantic relations have useful modal and epistemic roles. Indeed, there may be other aspects of its meaning that are not captured in terms of a primary intension and a secondary intension, but which might be captured by another sort of analysis. However, this can be taken as complementing rather than challenging the two-dimensional semantic analysis presented here. Different sorts of analysis provide ways of capturing different aspects of meaning that are useful for different purposes.

A fourth potential challenge is the worry that the two-dimensional semantic framework I present does not offer an account of the social processes that also influence the semantic practices surrounding psychiatric diagnoses. One such account of these

processes in the literature on philosophy of psychiatry is Ian Hacking's (1999) theory of dynamic nominalism. Using the example of childhood autism, Hacking (1999: pp. 114–115) proposes that psychiatric disorders are interactive kinds. That is to say, categorising disorders results in looping effects that alter the natures of the disorders in question. He argues that since the diagnostic term "childhood autism" was coined, the ideas about the disorder that became prevalent in society have influenced the sorts of behaviour with which new cases present. This suggests that there is an aspect of the meaning of the term "childhood autism" that changes in response to social processes.

Again, in response, I propose that this complements rather than challenges the two-dimensional semantic framework I have presented. In fact, Hacking (1999: pp. 119–124) himself is sympathetic towards the use of the causal theory of reference endorsed by Kripke ([1972] 1980) and Putnam (1975a) as a tool to analyse the semantics of diagnostic terms. For example, he considers the term "childhood autism" being used to designate the putative pathology *P* (Hacking, 1999: pp. 119–124). This suggests that childhood autism is an interactive kind with respect to its prototypical symptoms, but is presumed to be an indifferent kind with respect to *P*. I argue that this is consistent with the analysis that the changes that result from looping effects are with respect to the primary intension of "childhood autism", whereas the secondary intension is posited as remaining stable in virtue of *P*. Of course, it may turn out that *P* is associated with a range of pathologies rather than a single definite pathology, but this might be accommodated with the analysis that the secondary intension of "childhood autism" is disjunctive. Nevertheless, Hacking's important observations highlight that there are social dynamics working at the level of classification that are not specifically expounded by the theories of reference discussed in this chapter.

4.5 Conclusion

This chapter has explored how philosophical theories of reference apply to diagnostic terms, with the aim of resolving the conceptual problem regarding the tension between the descriptive definitions of psychiatric diagnoses in *DSM*-5 and their causal conceptions in other clinical resources. After looking at descriptive and causal theories of reference, I sketched how a two-dimensional semantic framework that assimilates the causal theory of reference with descriptive considerations accommodates the two seemingly contradictory ways in which diagnostic terms are used in psychiatry. The framework I have presented suggests that invoking psychiatric diagnoses as causes of patients' symptoms is not necessarily precluded by the fact that they are defined through symptoms. This partly addresses Szasz's (1960) argument that a mental illness cannot explain behaviour because it is just a shorthand label for this behaviour. However, an important concession must be made, which I now consider.

While the two-dimensional semantic framework I have presented allows a diagnostic term to refer to the causal profile that normally produces a set of symptoms despite being defined through these symptoms, whether or not the diagnosis actually provides a satisfactory explanation of a patient's symptoms also depends on the empirical fact regarding the nature of this causal profile associated with the diagnostic category. For some disorders, there are doubts about whether the underlying causal profiles will turn out to be sufficiently stable and repeatable for their respective diagnostic categories to be considered epistemically useful. In other words, it may turn out that the symptoms associated with a given diagnostic category can be produced in many different ways and that there is no unifying set of mechanisms that is shared by every instance of the diagnosis. Such a diagnosis would be like the case of dropsy mentioned in §4.3.4, where the secondary intension refers to a disjunction of several different pathologies. In a more extreme scenario, it may turn out that a given diagnosis may not be associated with any

discernible regular causes at all, such that every case turns out to have a different causal profile. This is also of relevance to the cross-historical and cross-cultural comparisons of disorders discussed by Radden (2003). I noted in §4.4.3 that a two-dimensional semantic framework accommodates the conceptual possibility of equating Chinese depression and Western depression based on the assumption that their secondary intensions are the same. However, whether it is actually correct to equate Chinese depression and Western depression is ultimately dependent on whether their secondary intensions do indeed turn out to refer to the same kind of causative pathology. This is something that must be ascertained empirically.

In summary, the framework of two-dimensional semantics shows that it is possible for diagnostic terms to be defined descriptively through their symptoms, yet refer to the causal processes that produce these symptoms. However, in order to answer the question of whether or not psychiatric diagnoses provide causal explanations of patients' symptoms, we need to examine the empirical facts regarding the causal profiles associated with the diagnostic categories. This will be the focus of Chapter 5.

5. The Causal Profiles of Psychiatric Disorders

5.1 Introduction

Having addressed the conceptual problem surrounding the definitional connections between psychiatric diagnoses and their symptom criteria, I now turn to the ontological problem regarding the natures of the causal profiles associated with the diagnostic categories in psychiatry. The two-dimensional semantic framework I put forward in Chapter 4 suggests that psychiatric diagnoses can still be taken to refer to the causal structures that produce sets of symptoms, even though they are formally defined through these sets of symptoms. However, we also need to examine what we know from empirical research about the causal profiles of psychiatric diagnoses to assess whether they are stable enough to support causal explanations in individual cases. And so, this chapter reviews the current empirical evidence pertaining to the causal profiles of some psychiatric disorders and the implications of this evidence for theoretical conceptualisations of the disorders.

As noted in Chapter 3, although diagnoses *qua* categories are generalisations, the causal profiles respectively associated with many, though certainly not all, of these diagnostic categories in bodily medicine are invariant in the appropriate respects for them to indicate, with reasonable specificity, the actual causal processes producing the symptoms in individual cases. Despite the various constitutional and biographical differences between individuals, every case of cystic fibrosis involves an abnormal cystic fibrosis transmembrane conductance regulator (CFTR) system (Simon, 2006: pp. 360–362) and every case of heart failure involves the inability to pump blood at a rate adequate for satisfying the requirements of the tissues (Denolin *et al.*, 1983: p. 445). Historically, this reflects a form of essentialistic thinking regarding such diagnoses in medicine, which has been referred to as the "disease entity" model (Hucklenbroich,

2014). For instance, the essential feature of cystic fibrosis is the abnormal CFTR ion transport system, such that a person whose phenotype does not involve an abnormal CFTR ion transport system does not, by definition, have cystic fibrosis. This essentialistic model has perhaps had the most success with respect to infectious diseases and genetic disorders, which have distinctive pathologies and singular aetiologies. However, it could also be applied to some diseases with multifactorial aetiologies, such as heart failure and myocardial infarction, which are still constituted by distinctive pathological processes, although these processes themselves may result from multiple contributory aetiological factors. Every case of myocardial infarction, for example, involves a distinctive pathological process, namely ischaemic necrosis of the myocardium, but this process might itself result from different combinations of aetiological factors in different cases.

As we shall see in this chapter, while essentialism may have had some success with respect to a number of diagnoses in bodily medicine, it is not an appropriate model for many diagnoses in psychiatry. I proceed as follows. In §5.2, I review the findings from scientific research into the causal profile of major depressive disorder, which is one of the commonest psychiatric conditions. I have chosen major depressive disorder as a paradigmatic example, not only because it is a common disorder, but also because it exemplifies some of the salient features that are found to different degrees in the causal profiles of many psychiatric disorders, such as high degrees of heterogeneity and complex interactions of diverse variables across multiple levels of organisation. In §5.3, I look at how theoretical models of psychiatric disorders might accommodate these problematic features. After arguing that simple essentialism is inadequate, I critically examine recent attempts in the philosophy of psychiatry to conceptualise psychiatric disorders as homeostatic property clusters. Finally, in §5.4 I explore to what extent the considerations raised also apply to psychiatric disorders other than major depressive disorder.

5.2 Major depressive disorder

5.2.1 Symptom criteria

In the fifth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*-5), major depressive disorder is described as a syndrome characterised by the following nine symptoms:

> 1. Depressed mood most of the day, nearly every day … 2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day … 3. Significant weight loss when not dieting or weight gain … or decrease or increase in appetite nearly every day … 4. Insomnia or hypersomnia … 5. Psychomotor agitation or retardation … 6. Fatigue or loss of energy … 7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) … 8. Diminished ability to think or concentrate, or indecisiveness … 9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide. (American Psychiatric Association, 2013: pp. 160–161).

A diagnosis of major depressive disorder requires a minimum of five out of the above nine symptoms, at least one of which must be depressed mood or diminished interest. These must be present for at least two weeks, result in clinically significant distress or impairment, and must not be attributable to the physiological effects of a substance or another medical condition.

This diagnostic process based on the fulfilment of a minimum number of criteria from a longer list allows for many different combinations of criteria to qualify for a diagnosis of major depressive disorder. Zimmerman *et al.* (2015) note that there are theoretically 227 different ways in which one can be diagnosed with major depressive

disorder based on the above criteria. Fried and Nesse (2015) also note that three of the diagnostic criteria for major depressive disorder are disjunctive. These are "weight loss when not dieting or weight gain", "insomnia or hypersomnia", and "psychomotor agitation or retardation". This not only increases the number of possible symptom combinations to over 1,000, but it means that different patients with major depressive disorder can have no symptoms in common, given that no single symptom is necessary or sufficient for a diagnosis of major depressive disorder. Moreover, because the disjunctive diagnostic criteria encompass opposite features, two patients with major depressive disorder could present with contrasting symptoms. For example, one patient may be diagnosed with major depressive disorder on the basis of depressed mood, weight loss, insomnia, and psychomotor agitation, while another may be diagnosed on the basis of diminished interest, weight gain, hypersomnia, and psychomotor retardation. Again, this suggests that the symptoms associated with major depressive disorder are highly heterogeneous, with the diagnosis being met by widely varying clinical presentations.

5.2.2 Genetics

Heterogeneity regarding the symptoms of a disorder is not by itself problematic, as many disorders in medicine are known to present in different ways. One example is syphilis, which is notorious for its protean manifestations. These can include ulceration, rash, malaise, weight loss, gastric dysmotility, hepatitis, meningitis, cardiovascular disease, and general paresis. Here, the many different manifestations are unified by a singular cause that is stable across cases, namely *Treponema pallidum* infection. This allows syphilis to be amenable to an essentialistic analysis, despite the heterogeneity at the level of its symptoms.

Following the discovery that *T. pallidum* infection is the defining cause of general paresis of the insane, it was hoped that other psychiatric disorders might also be

constituted by singular causative pathologies (Bolton, 2012: p. 9). One field where the search for singular causes has taken place is that of psychiatric genetics. This has had some success with respect to a small number of disorders. For example, the defining feature of Huntington's chorea was discovered to be the expansion of the CAG repeat on chromosome four. There are also uncommon forms of early-onset Alzheimer's disease involving the genes presenilin one, presenilin two, and amyloid precursor protein (Cowen *et al.*, 2012: p. 328). However, such instances are rare in psychiatry. For the majority of major psychiatric disorders, genetic research has failed to find genes of even moderate effect size (Kendler, 2006). In the case of major depressive disorder, data from family, twin, and adoption studies indicates a heritability of approximately thirty-seven percent (Sullivan *et al.*, 2000). While this indicates that there are genetic factors that increase vulnerability to major depressive disorder, environmental factors remain aetiologically more important.

Regarding the heritable component of vulnerability to major depressive disorder, linkage and association studies have been used to search for specific genes. One genetic variation that has received attention is a polymorphism in the promotor region of the serotonin transporter gene (5-HTTLPR). A study by Caspi *et al.* (2003) suggested that people with one or two copies of the short allele of 5-HTTLPR have a higher risk of developing major depressive disorder in response to stressful life events than people homozygous for the long allele. However, a subsequent meta-analysis found no significant association between the 5-HTTLPR polymorphism and the occurrence of major depressive disorder (Risch *et al.*, 2009). In general, data from association studies indicates that the heritable component of vulnerability to major depressive disorder is not attributable to one or a small number of genes, but to the combined effect of a vast number of genes, each with a small effect size (Shyn and Hamilton, 2010).

The above considerations indicate that a singular defining feature of major depressive disorder is not to be found at the level of genetics. However, it could be argued that heterogeneity with respect to distal causes such as genes is not necessarily a problem for disease explanation, because many medical disorders that have been successfully modelled are known to have multiple risk factors. For example, distal causes for myocardial infarction include genetic vulnerabilities, hypertension, obesity, smoking, and psychological stress, which vary significantly across cases. Nonetheless, these all converge onto a singular proximal cause, ischaemic necrosis of the myocardium, which is the determining property of every case of myocardial infarction. This suggests that we also need to look at whether the proximal causes associated with major depressive disorder are heterogeneous.

5.2.3 Neurochemistry

Regarding the proximal causes associated with major depressive disorder, a lot of attention has been paid to the investigation of neurobiological processes in the brain. Perhaps the most popular neurochemical hypothesis throughout the latter half of the twentieth century has been the monoamine hypothesis. Monoamines are a class of neurotransmitters that include noradrenaline, dopamine, and serotonin. It was observed by Edward Freis (1954) that patients who were treated for hypertension with reserpine, a monoamine antagonist, suffered the side effect of depressed mood. This then led to theorists, such as Joseph Schildkraut (1965) and Alex Coppen (1967), to hypothesise that the underlying pathology of major depressive disorder is underactive monoamine neurotransmission, particularly serotonin neurotransmission. To this day, the recommended pharmacological treatments for major depressive disorder are drugs that elevate serotonin neurotransmission.

However, since the monoamine hypothesis was first conjectured, the neurobiological features of major depressive disorder have been discovered to be far more complicated than initially anticipated. First, some studies have found no indication of reduced levels of monoamine metabolites in the cerebrospinal fluid or urine of patients with major depressive disorder compared to controls (Shaw *et al.*, 1973; Coppen *et al.*, 1979). Second, while reducing levels of serotonin by depleting its precursor tryptophan reduces antidepressant efficacy in a proportion of cases, it neither induces depressive symptoms in healthy volunteers, nor worsen symptoms in unmedicated patients with major depressive disorder (Delgado, 2011). Third, the drug tianeptine has been shown to be effective for the treatment of major depressive disorder, despite it actually reducing monoamine transmission (Wilde and Benfield, 1995).

These findings suggest that underactive monoamine neurotransmission is neither necessary nor sufficient for the occurrence of major depressive disorder. Based on the figures for antidepressant response, Belmaker and Agam (2008) estimate that the mechanism of major depressive disorder may not related to monoamines in up to two thirds of cases. In fact, based on data from rodent studies and the poor responses of a number of patients with major depressive disorder to conventional antidepressants, it has been hypothesised that monoamine neurotransmission may actually be elevated in some cases of major depressive disorder, rather than reduced (Fitzgerald, 2013). Therefore, although underactive monoamine neurotransmission is a factor associated with some cases of major depressive disorder, it cannot be taken as constituting the defining feature of the diagnosis.

More recent neurochemical hypotheses have acknowledged the role that stress has in the aetiology of major depressive disorder (Massart *et al.*, 2012; Palazidou, 2012). One of these hypotheses is that major depressive disorder involves alterations in the hypothalamic-pituitary-adrenal (HPA) axis (Arborelius *et al.*, 1999). The HPA axis is a

neuroendocrine system that is activated in response to stress. The hypothalamus increases its secretion of corticotropin-releasing hormone (CRH), which stimulates the anterior pituitary gland to secrete adrenocorticotropin (ACTH), which in turn stimulates the adrenal cortex to secrete cortisol into the systemic circulation. Negative feedback occurs at each step, such that cortisol inhibits further ACTH and CRH secretion, and ACTH inhibits further CRH secretion. Activation of the HPA axis results in various adaptive physiological changes, including the mobilisation of glucose and amino acids, and the inhibition of inflammation. It also results in neuroplastic changes in the hippocampus and prefrontal cortex (Massart *et al.*, 2012).

It has been suggested that major depressive disorder involves HPA axis overactivity and impaired negative feedback due to a combination of severe stress and genetic vulnerability (Massart *et al.*, 2012; Palazidou, 2012). Evidence supporting this hypothesis includes the increased levels of CRH found in the cerebrospinal fluid of suicide victims who had major depressive disorder (Nemeroff *et al.*, 1988), and the increased levels of salivary and plasma cortisol in depressed patients (Goodyer *et al.*, 1996). The neurotoxic effects of elevated cortisol have also been suggested as an explanation for the reductions in hippocampal volume found in patients with major depressive disorder (MacQueen *et al.*, 2003).

However, HPA axis dysregulation is far from a universal finding in cases of major depressive disorder. A study by Strickland *et al.* (2002) not only failed to find increased levels of salivary cortisol in depressed patients, but also found increased rather than decreased serotonin responsivity. These findings run counter to both the HPA axis hypothesis and the monoamine hypothesis. Belmaker and Agam (2008) note that although HPA axis dysregulation occurs in some cases of major depressive disorder, most people treated for major depressive disorder have no evidence of HPA axis dysregulation, just as most patients have no evidence of impaired monoamine

neurotransmission. Another author estimates the prevalence of HPA axis dysregulation in major depressive disorder as around fifty percent (Palazidou, 2012).

As with impaired monoamine neurotransmission, then, HPA axis dysregulation appears to be an important factor associated with some cases of major depressive disorder. A particular strength is that it offers a promising account of how stress might produce changes at the neurobiological level in such cases. However, it is neither necessary nor sufficient for the development of major depressive disorder, and so cannot be considered to be a defining feature of the diagnosis. These findings support the notion that major depressive disorder is heterogeneous with respect to its neurochemistry.

5.2.4 Brain circuitry

A recent trend in neuroscientific research into major depressive disorder has been to look for mechanisms at the level of brain circuitry. Using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) techniques, some people with major depressive disorder have been found to exhibit changes in the activation of the subgenual anterior cingulate cortex (sgACC), which is an area of the brain that constitutes part of a circuit purported to be associated with emotional processing (Drevets *et al.*, 1997; Groenewold *et al.*, 2013). Moreover, these changes in sgACC activation were shown to normalise following treatment. These results suggest that altered activity of the brain circuitry involved in emotional processing is associated with some of the affective symptoms of major depressive disorder.

While this is a significant finding, there is still room for heterogeneity at the level of these neural mechanisms. A review by Drevets *et al.* (2008) notes that patients with major depressive disorder who had first-degree relatives with mania, alcoholism, or sociopathy did not differ from healthy controls with respect to sgACC glucose metabolism or volume. Similarly, a review by Roiser *et al.* (2012) suggests that some patients with major

depressive disorder have abnormal sgACC activation during emotional processing while other patients have normal baseline sgACC activation, with the former group showing most improvement with pharmacological treatment and the latter showing most improvement with psychological therapy. While this has potential prospects for treatment selection, it rests on the finding that there is variability of brain mechanisms among patients with major depressive disorder. A meta-analysis by Graham *et al.* (2013) implicates a variety of other brain areas associated with major depressive disorder, including the occipital cortex, insula, supplementary motor cortex, and the cerebellum, as well as finding contradictory data regarding the activity of the right amygdala, again highlighting the variability of brain mechanisms across cases. Another review suggests differences in the neurobiological correlates of different groups of people with major depressive disorder (Baumeister and Parker, 2012).

Another concern is that it is contested whether conceptualising major depressive disorder exclusively at the level of neural circuitry is sufficient for understanding some of the key features of its psychopathology. It has been claimed that neural mechanisms are of utmost interest because they are the proximal causes of behaviour (Roiser, 2015). However, there is no *a priori* reason to suppose that disorders must be defined by their most proximal causes. I argue that applying this neurocentrism universally can lead to trivial conclusions. For example, consider a patient presenting with a cough. Strictly speaking, the proximal cause of the cough *qua* behaviour is a neural mechanism, namely the stimulation of the medulla oblongata by afferent fibres in the vagus nerve leading to the subsequent firing of efferent fibres that innervate the respiratory muscles. While this may be true, it is too trivial to be of explanatory significance in the clinic. Rather, we want a diagnosis to capture the causal process, albeit a less proximal one, that is perpetuating this neural mechanism. That is to say, we want the diagnosis to tell us whether the cough reflex is being perpetuated by a tumour, an infection, an inflammatory condition, or

pulmonary oedema from heart failure, partly because this would tell us where we can therapeutically intervene. Similarly, in the case of major depressive disorder, the symptoms may indeed be mediated by proximal neural mechanisms, but conceptualising the disorder exclusively at the level of these neural mechanisms risks explanatory triviality, because it leaves out crucial information regarding the joint contribution of other causal processes on which the maintenance of these neurological mechanisms is contingent and, moreover, on which it may be possible to intervene. As with the cough example, it could be argued that an explanatory model of major depressive disorder would need to capture these processes to be of clinical value.

5.2.5 Psychology

In addition to attempts to characterise major depressive disorder at biological levels, numerous psychological theories of major depressive disorder have been proposed. Rather than focusing on biological phenomena such as genes, neurochemicals, and neural circuits, these theories aim to capture regularities at the level of intentional processes, although these processes may be realised by biological systems involving genes, neurochemicals, and neural circuits (Radden, 2003). Psychological accounts of major depressive disorder include psychodynamic, behavioural, and cognitive theories.

Psychodynamic theories explain personality and behaviour in terms of interactions of motivational drives, particularly unconscious drives, and how these are modified by childhood events. One of the earliest psychodynamic accounts of depression is Sigmund Freud's ([1917] 1946) theory that it is linked to early negative experiences, such as loss or rejection. He proposed that the repressed anger towards the person whom one has lost becomes inwardly directed, producing depression, or melancholia. Melanie Klein ([1957] 1984) further developed this idea and characterised depression as a state of intrapersonal confusion, whereby the conflicting unconscious drives within the person produce feelings

of a divided self. One of the purposes of psychodynamic therapy, then, is to increase insight into these unconscious operations underlying the symptoms, allowing the person to gain control over his or her behaviour.

In contrast to psychodynamic theories, behavioural theories focus on observable behaviour and the environmental conditions that influence the learning of this behaviour, rather than on intrapsychic drives. According to this approach, learned maladaptive behaviour has a crucial role in the onset and maintenance of depression. Charles Ferster (1973) and Peter Lewinsohn (1974) proposed that depressive symptoms are maintained through decreased positive reinforcement of healthy behaviour, increased reinforcement of passive behaviour, and punishment of healthy behaviour. Martin Seligman (1975) suggested that depression is a state of learned helplessness, whereby the person learns that he or she has a lack of control over the outcomes of situations after enduring inescapable aversive stimuli. More recent research has found empirical support for the role of avoidance behaviour in the development and persistence of major depressive disorder (Carvalho and Hopko, 2011).

Cognitive theories of depression go beyond overt behaviour, and focus on the content and processing of thoughts. Perhaps the most prominent cognitive theorist of depression is Aaron Beck (1967), who proposed that depression consists of maladaptive cognitive processes. While he acknowledges that adverse life events have important roles in the production and maintenance of these maladaptive cognitive processes, he considers the cognitive processes themselves to be the central features of depression. Beck's theory suggests that a depressed person has negative thoughts about the self ("I am worthless"), the world ("my life is terrible"), and the future ("things won't get better"), otherwise known as Beck's cognitive triad. The formation of these negative thoughts is influenced by deeply entrenched dysfunctional beliefs and cognitive biases, such as overgeneralisation, dichotomous thinking, and selective thinking. The negative

thoughts are purported to have causal connections with emotions, behaviours, and physiological processes, thus accounting for the other symptoms of major depressive disorder. In recent years, researchers have focused more specifically on the roles of negative cognitive biases in depression (Robinson and Sahakian, 2008; Gotlib and Joorman, 2010). Cognitive-behavioural therapy, which aims to challenge and modify the maladaptive cognitive and behavioural processes, has been shown to be effective as a treatment for major depressive disorder (Whitfield and Williams, 2003).

Attempts to conceptualise major depressive disorder at a psychological level, then, have yielded a diverse mix of theories that emphasise different features. As discussed above, psychodynamic theories emphasise motivational drives and negative early life experiences, behavioural theories emphasise learned maladaptive behaviours, and cognitive theories emphasise negative thoughts and cognitive biases. Whether or not these theories can ultimately be unified into a single theory is currently unclear, although there has been some attempt to integrate concepts from different theoretical perspectives (Street *et al.*, 1999).

As with the biological factors discussed earlier, there is significant variability with respect to the particular psychological factors associated with the development of major depressive disorder. Contrary to Beck's (1967) original hypothesis that there is a certain sort of depressive cognitive style that is characteristic of major depressive disorder, empirical studies suggest that episodes of major depressive disorder can be associated with a variety of different cognitive styles. For example, in a longitudinal study looking at inpatients with unipolar major depressive disorder without psychotic symptoms, Hamilton and Abramson (1983) report that the patients exhibit heterogeneity with respect to their cognitive patterns, which include their attributional styles, dysfunctional attitudes, and measures of hopelessness. Moreover, they report that around fifty percent of the cognitive styles exhibited by the depressed patients approximate the cognitive

styles of healthy individuals. In a later paper, Abramson *et al.* (1989) not only report that major depressive disorder can be associated with different attributional styles, but also that this is dependent on the differential responses of people with different attributional styles to different sorts of adverse life event. These studies indicate that major depressive disorder is not characterised by a singular kind of cognitive pattern, but can be associated with several different cognitive patterns that interact with various situational factors in different ways.

Another psychological measure that has been investigated in relation to major depressive disorder is personality. In particular, the personality style of neuroticism has been hypothesised, from both psychodynamic (Freud, [1917] 1946; Klein, [1957] 1984) and cognitive (Beck, 1967) perspectives, to be a central to the development of the depressive syndrome. Empirical research has yielded some support for an association between neuroticism and major depressive disorder, but there are also significant anomalies. For example, Kendler *et al.* (2006) report that high neuroticism is associated with the development of major depressive disorder in general, but also report that there is an inverse correlation between neuroticism and major depressive disorder in the subset of patients with melancholic features. These results support the idea of major depressive disorder as a psychologically heterogeneous category with different clinical variants having associations with different personality styles.

The psychological heterogeneity of major depressive disorder is also reflected by the differential responses of patients to different kinds of psychological intervention. In a recent randomised clinical trial, Driessen *et al.* (2016) examine the associations between the psychological profiles of patients with major depressive disorder and their responses to cognitive-behavioural therapy and psychodynamic therapy. These two kinds of therapy assume different theoretical frameworks and target different psychological processes, with cognitive-behavioural therapy targeting cognitive biases and avoidance behaviours,

and psychodynamic therapy targeting the interactions of partly unconscious motivational drives. The authors report different responses to the two kinds of therapy in different subsets of patients with major depressive disorder. Patients with high anxiety and shorter episode durations achieved better improvement of their depressive symptoms with cognitive-behavioural therapy, while patients with low anxiety and longer episode durations achieved better improvement with psychodynamic therapy. While it is acknowledged that there are many possible hypotheses that could account for these observations, the authors suggest that the results are likely to be attributable to the different subsets of patients having different cognitive styles and personality structures underlying their depressive episodes. For example, longer episode durations may indicate depressive episodes that are related to more pervasive personality structures that are more amenable to psychodynamic interventions, while high anxiety may indicate negative thought patterns that are more amenable to cognitive-behavioural interventions. There are also similar studies which report differential responses to cognitive-behavioural therapy and interpersonal therapy in different subsets of patients with major depressive disorder, again supporting the possibility that the psychological structures underlying depressive symptoms are heterogeneous (McBride *et al.*, 2006; Joyce *et al.*, 2007).

5.2.6 Social context

As noted in §5.2.2, it is understood that environmental factors have important roles in the development of major depressive disorder. In a classic study, Brown and Harris (1978) surveyed a sample of 458 women to investigate the connections between depressive episodes and social circumstances. Of the thirty-seven participants who had suffered from depression in the previous year, ninety percent had endured adverse life events or stressful social circumstances, compared to only thirty percent of the participants who had not suffered from depression in the previous year. Three kinds of

factor were identified, namely (i) provoking factors that trigger depression, such as bereavement or being in an abusive relationship, (ii) vulnerability factors that increase the risk of depression, such as maternal loss before the age of eleven and lack of a confiding relationship, and (iii) protective factors that decrease the risk of depression, such as employment and intimacy with one's spouse.

Of course, the claim that social factors are distal causes that contribute to the aetiology of major depressive disorder is not particularly contentious. After all, it is recognised that social factors are important contributors to the aetiologies of many medical disorders, including coronary heart disease, cerebrovascular disease, type II diabetes mellitus, and even several kinds of cancer (World Health Organisation, 2003). However, there are theorists who endorse the stronger claim that some cases of major depressive disorder are not merely caused by, but are partly constituted by social processes. In the philosophy of psychiatry, this is associated with externalism about psychiatric disorder, whereby the locus of a disorder is not confined within the body, but extends into the social sphere (Zachar and Kendler, 2007; Broome and Bortolotti, 2009; Fuchs, 2012; Davies, 2016).

Different arguments for externalism are on offer. Recall my argument in §5.2.4 that a clinically useful conceptualisation of a disorder cannot consist solely of its most proximal mechanism, but must capture the joint contribution of other causal processes on which the maintenance of this mechanism is contingent and on which it may be possible to intervene. Thomas Fuchs (2012: pp. 336–337) suggests that such causal processes need not be restricted to internal physiological processes, but could also include external social processes on the grounds that they may be processes that are actively perpetuating the patient's condition. He proposes that psychopathology cannot be understood as being detached from the interpersonal context, because the intrapersonal and interpersonal processes involved are continually intertwined in relations

of "horizontal circular causality". Other theorists suggest that the proximal mechanisms of psychiatric disorders themselves must be thought of as being partly constituted by interpersonal processes. For example, Broome and Bortolotti (2009) argue that unlike many symptoms of medical disorders, psychopathology often has intentional content whose meaning is supervenient on the social context in which the patient is situated. Hence, a description which is exclusively in terms of neural processes cannot account for the variance in intentional content across cases.

There is empirical evidence consistent with the idea that situational factors are partly responsible for the variance in depressive psychopathology. First, interpersonal therapy, which focuses on the interactions between symptoms and social stressors, has been shown to be an effective treatment for major depressive disorder (Klerman *et al.*, 1974; Weissman *et al.*, 1981). Second, in a study of 4,856 individuals with symptoms of major depressive disorder, Keller *et al.* (2007) reported that different sorts of social stressor were associated with different symptom profiles. For example, chronic stress was found to be associated with fatigue and hypersomnia, while bereavement and romantic longing were found to be associated with sadness, anhedonia, appetite loss, and guilt. These associations were shown not only to hold across different individuals with singular episodes, but also across different episodes within the same individual. And so, the above considerations suggest that there are good reasons to consider social contextual factors as partly determining the natures of depressive episodes. Moreover, they indicate that the ways in which these factors contribute to the various permutations of major depressive disorder are heterogeneous.

5.2.7 Summary

The evidence reviewed in the above paragraphs highlights two important features of major depressive disorder. The first feature is heterogeneity. The diagnostic category of

major depressive disorder does not correspond to a unitary kind of causal state, but subsumes a varied range of possible causal states. Moreover, this heterogeneity seems to be exhibited at every level of analysis, including genetics, neurobiology, psychology, and social context. The concern, then, is that major depressive disorder may lack unity as a diagnosis, as it is possible for different patients with major depressive disorder to instantiate very different causal structures (Poland *et al.*, 1994; Murphy, 2006; Hyman, 2010).

The second feature is complexity. Instances of depressive psychopathology are not generally attributable to singular causes acting individually, but to the complex interactions of several causal factors in varying combinations. Furthermore, as mentioned above, these causal factors belong to different levels of analysis, from the molecular to the interpersonal. The implication of this complexity is that there is no single privileged level at which major depressive disorder can be aetiologically defined (Kendler, 2012). Rather, a comprehensive understanding of the disorder requires the consideration of the various biological, psychological, and social processes that interact across levels.

5.3 Conceptualising psychiatric disorders

5.3.1 The limits of simple essentialism

A philosophical implication of the heterogeneity and complexity of major depressive disorder is that the diagnosis is not amenable to a simple essentialistic analysis. That is to say, there is no essential property that is instantiated by every case of major depressive disorder in the way that *T. pallidum* infection is instantiated by every case of syphilis or in the way that the abnormal CFTR ion transport system is instantiated by every case of cystic fibrosis. Couched in the two-dimensional semantic framework presented in Chapter 4, then, it turns out *a posteriori* that the secondary intension of "major depressive disorder" does not refer to an essentialistic kind that is determined by an invariant causal

structure, but to a heterogeneous kind that includes a variety of different causal structures. Moreover, in virtue of the heterogeneity of its symptoms and the multiple possible ways of satisfying the diagnostic criteria, it is also the case that the primary intension is not uniform, but disjunctive.

Nonetheless, the failure of simple essentialism regarding major depressive disorder does not entail that the category is arbitrary. Major depressive disorder may indeed be associated with a diverse range of factors, but these factors do seem to cluster together in statistically significant ways. With respect to symptom criteria, Zimmerman *et al.* (2015) note that of the 227 theoretically possible combinations that meet the diagnostic threshold, nine account for over forty percent of the actual observed cases. With respect to causes, there are statistical correlations between some of the factors discussed throughout §5.2. For example, HPA axis dysregulation has associations with chronic stress, altered serotonin receptor binding, and neuroplastic changes in areas of the brain purported to be associated with emotional processing (Drevets *et al.*, 2008; Palazidou, 2012). There are also plausible theoretical mechanisms for how some of these factors might be connected. Therefore, while the relations between the various factors are highly contingent, far from universal, and likely to vary across cases, there are good reasons to suppose that they are not merely accidental, but might reflect causal processes.

5.3.2 Homeostatic property clusters

In light of the above, there has recently been a move in the philosophy of psychiatry to conceptualise some psychiatric disorders as homeostatic property cluster kinds (Beebee and Sabbarton-Leary, 2010; Kendler *et al.*, 2011; Tsou, 2013; Kincaid, 2014). We previously encountered the concept of the homeostatic property cluster in Chapter 4. As noted in §4.3.3, it was developed by the philosopher of biology Richard Boyd (1999) to describe how the members of a biological species resemble each other. The members of

the species *Canis lupus*, for example, tend to share common properties, such as having four legs, two eyes, tails, prominent snouts, and sensitive olfaction (Ereshefsky, 2010: p. 259). However, there is significant variation within *C. lupus*, such that its members do not have to instantiate all of these properties and no single property is necessary or sufficient for membership of the species. For example, Pembroke Welsh corgis are sometimes born without tails and British bulldogs tend to have flat snouts.

Importantly, the associations between the properties proposed by Boyd are not accidental, but are due to homeostatic causal mechanisms. The properties tend to cluster together because they are contingently connected by causal processes. For example, the members of *C. lupus* breed with each other, have a common phylogenetic heritage, and are exposed to similar environmental influences. These processes sustain the stability of the cluster of similarities across the members of the species. However, as mentioned above, the connections between the properties in the cluster are contingent, and so an individual member may not fulfil all of the properties.

Applying this to major depressive disorder, a homeostatic property cluster conceptualisation accommodates the idea that the disorder is not determined by a single essential property, but involves multiple properties that tend to cluster together due to causal processes. Individual patients with major depressive disorder need not instantiate all of these properties and no single property is necessary for the development of the condition. Hence, different combinations of properties may be instantiated by different patients with the diagnosis.

Kendler *et al.* (2011: p. 1147) present two approaches to conceptualising psychiatric disorders as homeostatic property clusters, which I respectively call the aetiological property cluster approach and the symptom network approach. According to the aetiological property cluster approach, the various causal factors that produce the symptoms of a disorder interact and sustain each other in a stable cluster. For example,

Jonathan Tsou (2013) characterises major depressive disorder as an aetiological property cluster involving mostly neurobiological properties that tend to occur together due to causal relations between them. However, this may not be sufficient. As noted in §5.2.5 and §5.2.6, depressive psychopathology is highly contingent on psychological variables and social contextual factors in addition to neurobiological properties for its development and maintenance. Hence, if we are to conceptualise the causal profile of major depressive disorder as an aetiological property cluster, then we would need to include social and psychological, as well as biological, variables in the cluster. According to Kendler *et al.*, these would include "genes, cell receptors, neural systems, psychological states, environmental inputs and socio-cultural variables" (Kendler *et al.*, 2011: p. 1147). What this suggests is that the causal relations between the variables cross different levels of organisation, from the molecular to the social.

This approach presented by Kendler *et al.* has a number of strengths as a strategy for conceptualising major depressive disorder. First, it acknowledges the diverse range of causal factors that have been shown to contribute to the production of depressive symptoms, thus offering a more complete aetiological account of the disorder than a pure neurobiological account. Second, it accommodates the causal heterogeneity seen at many levels in major depressive disorder. Third, it offers an account of why, despite this heterogeneity, these diverse properties at various levels tend to cluster together in statistically significant ways. The associations between them are not accidental, but due to causal processes. Therefore, according to the aetiological property cluster approach, major depressive disorder is not an aetiologically neutral or arbitrary category, but one that is informed by causal considerations.

A different approach to conceptualising psychiatric disorders as homeostatic property clusters is the symptom network approach pioneered by the psychologist Denny Borsboom (2008). This focuses on the causal relations between the symptoms

themselves, rather between any underlying causes and the symptoms. It conceives a psychiatric disorder as consisting of a set of symptoms that causally reinforce each other in a stable cluster. Cramer *et al.* (2010) present such a symptom network model for major depressive disorder. In this model, "fatigue may lead to a lack of concentration, which may lead to thoughts of inferiority and worry, which may in turn lead to sleepless nights, thereby reinforcing fatigue" (Cramer *et al.*, 2010: 140–141). This accounts for why, in spite of there being numerous possible combinations of symptoms in major depressive disorder, certain symptoms tend to cluster together in statistically significant ways. Major depressive disorder is not just an arbitrary collection of symptoms, but a dynamic process in which causal mechanisms between various symptoms sustain the aggregation of these symptoms.

Furthermore, the symptom network model is presented by Cramer *et al.* (2010) as a way of accounting for the high degrees of comorbidity between different psychiatric disorders. Instead of postulating latent variables as common causes of the different disorders, they suggest that certain symptoms of one disorder may also have causal relations with certain symptoms of another disorder. For example, they propose that there are bridge symptoms shared by both major depressive disorder and generalised anxiety disorder, such as insomnia, fatigue, and diminished ability to concentrate. These bridge symptoms not only have causal connections with the other symptoms of major depressive disorder, but also with the other symptoms of generalised anxiety disorder, hence explaining why the two disorders tend to be associated.

Borsboom's (2008) idea that the tendencies of certain symptom clusters to recur may be due to causal relations between the symptoms is novel. Importantly, it challenges the traditional attitude that the clustering together of certain symptoms must imply commonalities with respect to their underlying causes. Nonetheless, as noted by Kendler *et al.* (2011), there is a way to interpret the symptom network approach as being a

homeostatic property cluster approach. As with the aetiological property cluster approach, the associations between the various properties in the network are not accidental, but causal. Moreover, these causal relations are statistical rather than deterministic, and so patients with the disorder do not have to instantiate all of the symptoms, but may have varying combinations of different symptoms.

The symptom network model of major depressive disorder is presented by Cramer *et al.* (2010) as being a stand-alone model. However, it could also be argued that the aetiological property cluster model and the symptom network model do not contradict, but complement, each other. As noted by Danks *et al.* (2010), the claim that the symptoms causally influence each other is entirely compatible with the claim that the factors that cause these symptoms sustain each other in clusters. Similarly, in the model suggested by Kendler *et al.* (2011: p. 1147), it is suggested that a series of psychological and biological causes interact with each other to produce the clinical features, and that in turn these clinical features causally interact with each other. Therefore, it may in principle be possible, indeed perhaps desirable, to integrate aetiological property cluster and symptom network approaches into a more comprehensive homeostatic property cluster model of the disorder.

It seems at least plausible that major depressive disorder could turn out to be characterisable as a homeostatic property cluster, as long as the properties include biological, psychological, and social variables, as well as the observable clinical features. However it should be acknowledged that such a conceptualisation currently remains promissory. While the empirical research reviewed throughout §5.2 has yielded knowledge of an array of causal variables associated with major depressive disorder, we are far from understanding the precise natures of the relations between many of these variables. Accordingly, attitudes among theorists in the philosophy of psychiatry towards the homeostatic property cluster model of major depressive disorder currently range

from optimistic (Tsou, 2013; Kincaid. 2014) to sceptical (Haslam, 2014). Nonetheless, as argued by Kendler *et al*., (2011), a good reason to take the theory seriously, aside from its plausibility, is that it directs research towards the practical goal of articulating the causal mechanisms that sustain psychopathology as a stable phenomenon without assuming simple essentialism.

5.3.3 Challenges

In spite of its plausibility, there still remain some challenges to a conceptualisation of major depressive disorder as a homeostatic property cluster that involves causal variables at different levels of organisation. The first challenge concerns how we make sense of the causal relations that cross different levels of organisation. The development of a unified theoretical model of a disorder can be relatively straightforward where the processes and mechanisms involved fall under a single explanatory perspective, but it is problematic in psychiatry where disorders are purported to involve interactions between different kinds of process that require different theoretical perspectives. For example, the processes involved in some bodily disorders such as myocardial infarction and acute appendicitis can be understood with a biological explanatory perspective, but understanding the different kinds of process involved in major depressive disorder requires a combination of biological, psychological, and social explanatory perspectives. The difficulty is how to integrate these different kinds of process.

One suggestion might be to try to reduce the higher-level processes to lower-level phenomena. In his work on the visual system, David Marr (1982) proposes that the same process can be viewed from three different levels of explanation, namely (i) computation, or what the system does in terms of problems and goals, (ii) algorithm, or how the system does what it does in terms of information inputs and outputs, and (iii) implementation, or how the system is realised by neurons in the brain. Given that these three levels are

supposed to represent the same process, we might expect that the higher-level cognitive generalisations could ultimately be replaced by lower-level facts about neurons, genes, and molecules.

However, Dominic Murphy (2008) argues that there are limits to this reductive approach in psychiatry. Whereas Marr's different levels are supposed to describe the same process, the processes described in higher-level terms in psychiatry are often different from the processes described in lower-level terms. For example, poor social relations and negative self-reinforcement are radically different kinds of process from underactive serotonin neurotransmission. Therefore, major depressive disorder presents a scenario where higher-level processes are not reduced to lower-level processes, but where different kinds of process from different levels causally influence each other.

Kendler and Campbell (2009) advocate a particular philosophical account of causation that accommodates the possibility of causal relations between radically different kinds of process, namely the interventionist theory of causation developed by James Woodward (2003). Broadly speaking, the interventionist theory of causation states that for *X* to count as a cause of *Y* is for there to be a regular response of *Y* on an intervention on *X* in at least some background circumstances. For example, if we want to know whether the drug fluoxetine causes remission of depressive symptoms, we can undertake a randomised controlled trial where the independent variable is the administration of fluoxetine, the dependent variable is the presence of depressive symptoms, and placebo control helps to ensure that the observed causal effect of the intervention flows through the pharmacological action of fluoxetine rather than through a different variable. This accomplishes the important task of distinguishing between merely correlatory and genuinely causal relations. Consider that variations in *X* are correlated with variations in *Y* and with variations in *Z*. Now consider that intervening on *Z* does not result in any changes in *X* or *Y*, intervening on *Y* does not result in any

changes in *X* or *Z*, but *X* results in regular responses in both *Y* and *Z*. It can be inferred from this that *Y* does not cause *Z* and vice versa, but *X* is the common cause of *Y* and *Z*.

The attraction for Kendler and Campbell of the interventionist theory is its permissiveness. It does not place restrictions on the kinds of variable that are allowed to feature in the characterisation of a disorder. There are no requirements for a variable to belong to a particular biological level of organisation, or even for variables to belong to the same level of explanation. In fact, in another publication, John Campbell (2008) argues that because no particular kind of variable is granted privileged status over other kinds, we can give up talk of hierarchical "levels of explanation" altogether and instead think of the characterisation of a disorder as being "many-sorted". And so, a homeostatic property cluster conceptualisation of major depressive disorder can include biological, psychological, and social variables, without considering any kind of variable to be more fundamental than the others. There is also no requirement for the specific mechanisms involved in the causal processes to be fully understood (Kendler and Campbell, 2009: p. 884). The inference that *X* causes *Y* is supported by the observation that intervening on *X* results in a regular response of *Y*, even if we do not know of a mechanism linking *X* and *Y*. For example, we can establish that social and psychological variables causally influence biological variables, and *vice versa*, even though we may not have clear ideas about precisely how they do so.

The interventionist theory of causation, then, accommodates the conceptualisation of major depressive disorder as a homeostatic property cluster involving different kinds of variable that are connected via causal relations. Because the theory avoids preconceptions about what kinds of thing can constitute causes and what sorts of mechanism are involved, it permits legitimately causal relations between biological, psychological, and social variables, even though we may currently be in the dark about what some of the mechanisms linking these variables might look like. A significant

implication of this is that the interventionist theory offers one way around the dichotomy, put forward by Karl Jaspers ([1913] 1997), between meaningful understanding and causal explanation in psychiatry. Under the interventionist theory, intentional states that contain particular meanings qualify as genuine causes of a variable if interventions on these states are shown to result in regular responses of the variable. Indeed, Woodward (2008: pp. 157–158) notes that this is precisely what happens in cognitive-behavioural therapy.

However, a problem with assuming the interventionist theory of causation is that it does not, on its own, seem to be adequate for satisfying the requirements for paradigmatic cases of causal explanation in medicine. As mentioned above, the interventionist theory does not require knowledge of mechanisms to establish the presence of a causal relation. In Chapter 3, though, I argued that a medical diagnosis explains a set of patient data *E* by specifying its cause *C*, but also that the intelligibility of the explanation also depends on theoretical knowledge of the mechanisms by which *C* produces *E*. Hence, if we are to assume the interventionist theory for the purposes of accounting for causation in psychiatry, then we would need to concede that causal explanation in psychiatry does not meet the standard of causal explanation in bodily medicine.

There are two ways of responding to this objection. One response is that the interventionist theory could be complemented by knowledge of mechanisms. This is the approach advocated by Kendler (2014), who suggests that a research programme can use the interventionist theory to establish the presence of a causal relation between variables, which in turn can be complemented by a further research programme aimed at specifying at least some of the mechanisms involved in this causal relation. Hence, according to this view, a mechanistic approach is not a challenge to the interventionist theory of causation, but a supplementation of it (Kendler and Campbell, 2008: p. 883). The other response,

which draws from Woodward (2002), is that that mechanisms can be reducible to causes and analysed counterfactually according to the interventionist theory. For example, we might establish whether the mechanism of angiotensin II production is causally relevant in the pathway by which heart failure causes leg oedema by intervening on this mechanism with an angiotensin-converting-enzyme inhibitor. According to this view, then, knowledge of the mechanisms linking *C* and *E* would amount to knowledge of a finer succession of causes between *C* and *E*. It may be that these sorts of response would not be particularly relevant to an interventionist like Campbell (2006), for whom part of the point of assuming an interventionist theory of causation is to circumvent altogether the desire for mechanistic knowledge. Furthermore, it may be that for some observed causal relations, especially those between variables at different levels of organisation, we may not be able to specify stable mechanisms. Nonetheless, the above responses do show that the interventionist theory of causation does not preclude a complementary research programme that is aimed at attaining knowledge of at least some of the mechanisms involved in some of the causal relations.

Before we move on to the second challenge to the homeostatic property cluster model of major depressive disorder, it is also worth acknowledging another rather different philosophical account of causation between biological and psychological states. The sophisticated account developed by Bolton and Hill (2004) develops the idea that nature does not just contain patterns exemplified by non-intentional causation, but also patterns exemplified by intentional causation. The former refers to causation according to the dynamics described by the laws of physics and chemistry, whereas the latter is characterised by its informational content. Bolton and Hill note that intentional causes are often invoked in biological explanations. For example, the regulation of blood pressure depends on the informational content encoded by the frequency of firing by arterial baroreceptors. They also argue that psychological explanations that involve

meanings are also intentional causal explanations, thus making them continuous with biological explanations. The authors broadly follow the view of Ludwig Wittgenstein (1953) that meaning is not reducible to a symbolic representation in the brain, but pertains to the guidance of activity that is embedded in social practices. Moreover, they suggest that a non-intentional causal process can interfere with an intentional causal process, such as when structural damage to the brain disrupts the comprehension of speech. The theory presented by Bolton and Hill, while more metaphysically ambitious than Woodward's interventionist theory of causation, does offer another possible framework for understanding how different sorts of causal factor can be integrated in the conceptualisation of a psychiatric disorder. The general upshot of the above discussion, then, is that there are appropriate philosophical theories of causation that can make sense of the idea of causal relations between biological and psychological factors.

The second challenge to the homeostatic property cluster model of major depressive disorder is conceptual and concerns where we draw the limits as to what can justifiably be called a homeostatic property cluster. As noted in §5.2.6 and §5.3.2, there is evidence indicating that social contextual factors have important roles in maintaining depressive psychopathology, and so a conceptualisation of major depressive disorder that meets the structural requirements for a homeostatic property cluster would need to include these social contextual factors as well as individual biological and psychological factors. This potentially expands the set of things that qualify as homeostatic property clusters, which is a worry for those who support the view that homeostatic property clusters are natural kinds. Such a point is made by Beebee and Sabbarton-Leary (2010: p. 22), who appeal to Paul Griffiths' (1999) observation that even a social convention like money could arguably be considered an homeostatic property cluster, because there are causal processes in society that sustain it as a stable phenomenon. However, in response, I argue that this is not so much a problem for the claim that major depressive disorder

can be conceptualised as a homeostatic property cluster, but rather a problem for the claim that it should be considered a natural kind. As I have stated in §5.1, this latter claim is not required for a clinically useful conceptualisation of major depressive disorder. It is possible to conceptualise major depressive disorder as involving a cluster of causal factors that sustain each other, while at the same time acknowledging that it is highly social.

The third challenge is epistemological. Even if we are able to conceptualise major depressive disorder as a homeostatic property cluster, such a conceptualisation may still be too broad for the diagnostic category to be explanatorily useful with respect to individual cases. Again, this is related to the contention that the development and maintenance of depressive psychopathology depend on highly contingent facts about individual constitution and environmental context, and so an aetiological conceptualisation of the disorder which meets the structural requirements for a stable homeostatic property cluster would need to include an extensive and diverse range of biological, psychological, and social factors. Rather than looking like a tight cluster of properties held together by robust processes, then, such a conceptualisation could look more like a loose network of diverse variables that are only probabilistically connected by highly contingent causal relations. Therefore, a homeostatic property cluster conceptualisation may not be enough to compensate for the causal heterogeneity of major depressive disorder in a manner that is clinically useful. Due to the looseness of the model, the diagnostic category could still subsume a vast array of different possible causal pathways that could result in the symptoms of major depressive disorder. I discuss further implications of this in Chapter 6.

In summary, a homeostatic property cluster conceptualisation of major depressive disorder seems plausible, although many of the details remain promissory. However, it is only plausible if social as well as biological and psychological factors are accommodated in the conceptualisation, and if we assume a philosophical attitude towards causation that

makes sense of causal relations between these different kinds of factor. A significant worry, though, is that even if the disorder does turn out to be characterisable as a homeostatic property cluster, there is no guarantee that this will significantly increase its epistemic utility, as the cluster may turn out to be too loose for the diagnostic category to be explanatorily valuable with respect to individual cases.

5.4 Other psychiatric disorders

5.4.1 Schizophrenia, bipolar disorder, and generalised anxiety disorder

So far, I have examined major depressive disorder as a paradigmatic example of a diagnostic category in psychiatry that is causally heterogeneous at multiple levels of analysis. Of course, major depressive disorder may not be representative of all psychiatric disorders, and so it is important to consider to what degree the above considerations are applicable to other diagnoses in psychiatry. This will be the focus of what is to follow. As we shall see, psychiatric disorders constitute a varied group of conditions that exhibit the attributes of causal heterogeneity and complexity to different degrees.

Current empirical evidence suggests that the above issues regarding major depressive disorder also apply to some of the more common major psychiatric disorders. Three of these are schizophrenia, bipolar disorder, and generalised anxiety disorder. Schizophrenia, which is typically associated with delusions, hallucinations, disorganised speech, catatonic behaviour, and diminished emotional expression, is heterogeneous with respect to both symptoms and causal factors. As with major depressive disorder, different combinations of symptoms can satisfy the diagnostic criteria for schizophrenia, with a minimum of two out of five being required (American Psychiatric Association, 2013: p. 99). Genetic factors increase the risk, but these likely involve a complex array of genes with small effect sizes (Kendler, 2006). Research suggests that dopamine dysregulation is associated with the disorder, but the evidence has been inconsistent and

only accounts for some aspects of the psychopathology (Guillin *et al.*, 2007). Correlates at the level of neural circuitry are also varied, with neuroimaging revealing a diverse range of structural connectivity alterations (Wheeler and Voineskos, 2014). Certain social circumstances are associated with increased incidence, including urbanicity and migration (van Os, 2004; Cantor-Graae and Selten, 2005). Moreover, a recent study suggests that social and cultural factors do not merely influence the contents of delusions and hallucinations, but the structural forms of the clinical presentations (McLean *et al.*, 2014).

Bipolar disorder, which is associated with alternating episodes of mania and depression, has a similar sort of profile. The causal factors associated with the disorder are summarised in a review by Maletic and Raison (2014). Although heritability is high with monozygotic twin concordance estimated between forty and seventy percent, this is again due to a vast and heterogeneous array of common genetic variants and epigenetic changes, each with small effect size. Research on the neurobiological correlates of bipolar disorder has yielded equivocal results with respect to neurochemistry and neural circuitry. Interestingly, fMRI results are among the least consistent. The authors make the bold conclusion that "from a neurobiological perspective, there is no such thing as bipolar disorder" (Maletic and Raison, 2014: p. 16). Rather, they suggest that the category of bipolar disorder subsumes many somewhat similar, but subtly different, causal structures.

Generalised anxiety disorder, which is associated with excessive anxiety in conjunction with various physiological and cognitive symptoms, also exhibits similar degrees of heterogeneity and complexity with respect to neurobiological, psychological, and social factors. However, heritability is estimated to be somewhat lower and stressful events are reported to be particularly significant. Moreover, the causal factors that have been implicated do not appear to be specific to generalised anxiety disorder, but have been shown to be associated with a range of anxiety and affective disorders (Cowen *et al.*, 2012: pp. 179–186).

Like major depressive disorder, then, the diagnostic categories of schizophrenia, bipolar disorder, and generalised anxiety disorder do not respectively reflect invariant causal structures, but each category subsumes a range of possible causal pathways involving combinations of biological, psychological, and social variables that tend to cluster in statistically significant ways. We can put the above in terms of the two-dimensional semantic framework presented in Chapter 4. It *a posteriori* turns out that the secondary intensions of "schizophrenia", "bipolar disorder", and "generalised anxiety disorder" do not refer to essentialistic kinds, but correspond respectively to heterogeneous kinds that include many different possible causal structures.

5.4.2 Dementias

It should be acknowledged that such causal heterogeneity may not be exhibited by all psychiatric disorders and that there are at least some diagnostic categories that reflect distinctive kinds of biological causal structure that are stable across cases. The dementias are neurodegenerative disorders that include Alzheimer's disease, vascular dementia, dementia with Lewy bodies, and frontotemporal dementia. They are associated with progressive and generally irreversible cognitive and neurological decline. The clinical presentations and risk factors can vary, but in all cases the symptoms are caused by distinctive kinds of neurodegenerative process. For example, in Alzheimer's disease they are caused by the accumulation of β-amyloid plaques and neurofibrillary tangles in the cerebral cortex, in vascular dementia they are caused by cumulative focal areas of ischaemic necrosis due to the occlusion of cerebral vasculature, in dementia with Lewy bodies they are caused by the abnormal aggregation of α-synuclein protein in cortical and subcortical areas, and in frontotemporal dementia they are caused by atrophy of the frontal and temporal lobes (Cowen *et al.*, 2012: pp. 326–333). In virtue of these determining properties, the different categories of dementia can be considered to

represent repeatable causal types, although it should be recognised that patients can exhibit features of more than one kind of dementia. Hence, the secondary intensions of "Alzheimer's disease", "vascular dementia", "dementia with Lewy bodies", and "frontotemporal dementia" refer respectively to the above mentioned neurodegenerative processes.

5.4.3 Panic disorder and obsessive-compulsive disorder

For some other disorders, there may be high degrees of causal heterogeneity at lower biological levels, but more stable causal regularities may be observed to emerge at higher psychological levels. Examples arguably include panic disorder and obsessive-compulsive disorder. Panic disorder, which is associated with recurrent and unexpected panic attacks, is highly complex and variable with respect to genetics, neurochemistry, and brain circuitry. Heritability is estimated at forty percent, although again this appears to be the result of combinations of genes whose effect sizes are small and contingent on other constitutional parameters. Several chemical abnormalities have been associated with the disorder, including serotonin dysregulation, noradrenaline hypersensitivity, γ-aminobutyric acid attenuation, and lactate hypersensitivity, but no single abnormality or particular combination of abnormalities has been shown to be present across all cases. Furthermore, neuroimaging has suggested changes in the amygdala and cingulate cortex, but the findings are not consistent (Cowen *et al.*, 2012: pp. 195–198).

However, a more stable pattern can be observed to emerge at a psychological level. David Clark (1986) proposes a cognitive model of panic disorder. The model posits causal connections between (i) a trigger stimulus (internal or external), (ii) perceived threat, (iii) apprehension, (iv) body sensations, and (v) interpretation of sensations as catastrophic (Clark, 1986: p. 463). The interpretation of sensations as catastrophic results in further perceived threat, and thus these variables reinforce each other in a cycle.

Clark's model describes a higher-level psychological process that is supposed to be instantiated by every case of panic disorder, while remaining neutral with respect to the lower-level biological processes in which these higher-level psychological processes could be grounded. Therefore, while panic disorder may lack unity with respect to its biology, it is characterised by a more stable causal structure at the level of its psychology.

A similar case could arguably be made regarding obsessive-compulsive disorder. Again, a number of neurobiological correlates have been discovered, including serotonin dyregulation, increased basal ganglia volume, and decreased orbitofrontal cortex volume, but the findings are not wholly consistent (Cowen *et al.*, 2012, pp. 199–204). At the psychological level, though, there appears to be a more stable causal pattern. An obsessional thought is purported to result when an intrusive thought is erroneously appraised as being salient or threatening due to the cognitive biases of overinflated responsibility (Salkovskis, 1985) and thought-action fusion (Rachman, 1993). Compulsive behaviour temporarily reduces the discomfort from the obsessional thought and is purported to be maintained via the process of negative reinforcement, as well as by strengthening the belief that the discomfort would have increased had the compulsion not been performed. This model is supported by evidence for the effectiveness of cognitive therapy in conjunction with exposure and response prevention, which are supposed to intervene on the above processes, for the treatment of obsessive-compulsive disorder (Veale, 2007).

And so, panic disorder and obsessive-compulsive disorder provide examples of psychiatric diagnoses that are biologically heterogeneous, but whose symptoms could nonetheless be explained by more stable causal regularities that emerge at psychological levels. In terms of two-dimensional semantics, it *a posteriori* turns out that the secondary intensions of "panic disorder" and "obsessive-compulsive disorder" correspond to stable underlying psychological processes, albeit psychological processes that can each be

realised by a range of different biological states. It is reasonable to suggest that this might also apply to some other monosymptomatic diagnoses in psychiatry, such as specific phobias and impulse control disorders.

5.4.4 Personality disorders

This brings us to a final group of disorders. While major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety are causally heterogeneous at multiple levels of organisation, we at least have evidence that the various causal factors involved in these disorders are likely to be loosely held together by contingent causal relations, and so we can be confident that the categories are not entirely causally arbitrary. However, for some other disorders, the symptoms may be the products of so many contingent circumstances that we may not be able to locate explanatorily relevant causal regularities that generalise even modestly across cases. This might apply to some of the personality disorders.

The personality disorders are typically characterised by persistent and pervasive patterns of behaviour that are problematic for the patients and for others around them. As with major depressive disorder, the diagnostic criteria for a personality disorder consists of a list of symptoms, of which a minimum number must be fulfilled. For example, a diagnosis of histrionic personality disorder requires at least five out of the following eight features:

> 1. Is uncomfortable in situations in which he or she is not the center of attention. 2. Interaction with others is often characterized by inappropriate sexually seductive or provocative behaviour. 3. Displays rapidly shifting and shallow expression of emotions. 4. Consistently uses physical appearance to draw attention to self. 5. Has a style of speech that is excessively expressionistic and lacking in detail. 6. Shows

> self-dramatization, theatricality, and exaggerated expression of emotion. 7. Is suggestible (i.e., easily influenced by others or circumstances). 8. Considers relationships to be more intimate than they actually are. (American Psychiatric Association, 2013: p. 667)

Because only five out of eight features are required for a diagnosis, different patients with histrionic disorder may exhibit different combinations of symptoms. Hence, as with major depressive disorder, the clinical presentations associated with histrionic personality disorder are heterogeneous. Interestingly, it has been suggested that this reflects the fact that personality traits are continuously distributed, which also accounts for why the boundaries between the different categories of personality disorder are so poorly defined (Cowen *et al*., 2012: p. 135). Accordingly, prior to the publication of *DSM*-5 (2013), the psychologist Thomas Widiger (2007) argued that a more appropriate way to classify personality disorders would be with a dimensional system, where the patient's personality would be assessed along several continuous dimensions rather than being placed into one of several distinct categories. However, the categorical system was ultimately retained in *DSM*-5.

As diagnostic categories, the personality disorders are notoriously controversial. First, as we shall see, there are doubts over whether they reflect distinctive kinds of causal process. Second, the categories are heavily shaped by moral and political values. Louis Charland (2004) argues that the personality disorders encompass behaviours considered in society to be morally bad that have been inappropriately medicalised as mental disorders. Peter Zachar notes that this marks a worry about a sort of psychiatric emotivism whereby "personality disorder is considered a name for unlikeable people who are highly neurotic" (Zachar, 2014: p. 197). Moreover, Nancy Potter (2004) proposes that some of the purported features of personality disorders, such as emotionality and

manipulativity, are influenced by misogynistic cultural assumptions regarding gender attributes. I do not spend more time on the second of these two issues in this chapter, instead focusing on the first.

It must be acknowledged that it is plausible that there are some personality disorders which are associated with certain causal factors in statistically significant ways. For example, antisocial personality disorder has been shown to be partly attributable to the effects of poor parental bonding and childhood physical abuse on social development, with the behavioural factor being associated with poor maternal care and the affective factor being associated with poor care from both parents (Gao *et al.*, 2010). Similarly, the development of borderline personality disorder, which is associated with affective lability, impulsivity, and interpersonal instability, has been partly attributed to the effect of childhood sexual abuse on the ability to modulate emotion (Winston, 2000). These two disorders also have associations with some genetic vulnerabilities and subtle neuroimaging changes in the areas of the brain purported to be involved in affective processing (Cowen *et al.*, 2012: pp. 144–145).

However, for other personality disorders, there may not be such statistically significant causal factors. A summary of evidence in the *Shorter Oxford Textbook of Psychiatry* suggests that this might be so for paranoid personality disorder, schizoid personality disorder, histrionic personality disorder, and avoidant personality disorder. The authors report that research into these disorders has generally failed to reveal causal factors that generalise even modestly across cases and that the results of studies have been inconsistent at best (Cowen *et al.*, 2012: pp. 143–146). Therefore, it is possible that at least some personality disorder diagnoses are not associated with any stable causal regularities.

Note that this is not to say that the symptoms of these disorders are uncaused, as it could arguably be contended that all behaviours have causes (Morse, 1999). Rather, it is

to say that while the symptoms of an individual patient with a given diagnosis may indeed be caused by a particular state of affairs, there may not be stable causal factors that are shared to even modest degrees by other patients with the same diagnosis. Hence, the scepticism is not about the presence of singular causation in the individual case, but about whether there are generalisable causal factors associated with the diagnostic category. There is a real possibility that the clinical features of the above mentioned personality disorders are not the products of repeatable causal processes, but of highly contingent combinations of circumstances that differ significantly across cases. In short, there may be many idiosyncratic reasons why people develop the kinds of personality they do. As I suggest later in Chapter 7, it may in principle be possible to discern some of these reasons via narrative exploration of the particular case, but the point is that these reasons may not be generalisable to other cases.

The above considerations, then, suggest that some diagnostic categories in psychiatry fail to correspond to even modestly repeatable causal types. Rather, it is possible that the clinical features associated with some of the personality disorders result from complex and highly contingent sets of circumstances that vary across cases. Due to the absence of such generalisable causal factors, it is unlikely that these disorders can be characterisable even as loosely construed homeostatic property cluster kinds.

Couched in the two-dimensional semantic framework presented in Chapter 4, the secondary intensions of "paranoid personality disorder", "schizoid personality disorder", "histrionic personality disorder", and "avoidant personality disorder" do not pick out characteristic causal structures. This suggests that these categorical diagnoses do not convey anything significantly informative regarding what might be causing the clinical features of patients. Because they are not particularly informative regarding causes, it appears that these diagnostic terms are little more than descriptive labels for sets of symptoms. That is to say, their useful semantic roles are in virtue of their primary

intensions. According to David Chalmers (1996: p. 62), it could be supposed that for such descriptive terms, as with other descriptive expressions like "doctor" and "square", the secondary intensions are simple copies of the primary intensions.

5.5 Conclusion

In this chapter, I have reviewed the empirical data from scientific research into the causal structures of some psychiatric disorders and explored some of the philosophical implications for theoretical conceptualisations of these disorders. Major depressive disorder presents a paradigmatic example of a psychiatric diagnosis that does not reflect an invariant causal type, but subsumes a heterogeneous range of possible causal structures that could produce the symptoms. Moreover, these causal structures involve combinations of diverse biological, psychological, and social factors that interact in complex ways.

An implication of such heterogeneity and complexity is that major depressive disorder cannot be captured with an essentialistic model. Instead, I considered the prospect of conceptualising major depressive disorder as a homeostatic property cluster. While it is plausible that the various factors associated with major depressive disorder tend to reinforce each other via probabilistic causal relations, I argued that the diversity of the factors involved and the highly contingent natures of the causal relations could make such a conceptualisation too loose to be of causal explanatory value as an undifferentiated diagnostic category.

I then examined how applicable these considerations are to other psychiatric diagnoses. I argued that some other major psychiatric diagnoses, such as schizophrenia, bipolar disorder, and generalised anxiety disorder, are likely to have causal profiles that exhibit degrees of heterogeneity and complexity comparable to that of major depressive disorder. However, a few diagnoses, such as the dementias, are characterised by

distinctive kinds of biological pathology. A few other diagnoses, such as panic disorder and obsessive-compulsive disorder, are heterogeneous and complex with respect to biological factors, but may be associated with more stable causal patterns that emerge at psychological levels. Finally, there are diagnoses, such as some of the personality disorders, whose symptoms are the products of so many contingent circumstances that we may not be able to locate stable causal factors or regularities that are generalisable across cases. For such conditions, it may be that the diagnostic terms are little more than descriptive labels for the sets of symptoms.

6. How Psychiatric Diagnoses Explain*

6.1 Introduction

The empirical data reviewed in Chapter 5 suggests that many of the current diagnostic categories in psychiatry are causally heterogeneous at every level of analysis. In this current chapter, I examine the implications of this for the explanatory statuses of psychiatric diagnoses. I shall argue that despite the problems of complexity and heterogeneity, some psychiatric diagnoses can still provide explanatory information that can be valuable for clinical purposes. Moreover, while these other sorts of explanation do not fit the standard model of causal explanation presented in Chapter 3, whereby a diagnosis specifies a distinctive cause *C* as being responsible for producing the patient data *E* via intelligible mechanisms, I shall show that they are nonetheless causal in satisfying ways.

The chapter proceeds as follows. I begin in §6.2 by distinguishing two kinds of explanatory question, which are the explanation of a syndrome in general and the explanation of the clinical presentation of a particular patient with appeal to a diagnosis. I explore the potential challenges that psychiatric disorders pose for these explanatory questions. While philosophers of psychiatry have offered promising approaches to the first kind of explanation that handle the challenges of heterogeneity and complexity, these problems continue to affect the second kind. Nonetheless, I argue in §6.3 that even though psychiatric diagnoses may not correspond to invariant causal types, there are other ways in which they can offer causal explanatory information. I suggest that some

---

* A version of this chapter has been published as: Maung, H. H. (2016a). "Diagnosis and Causal Explanation in Psychiatry". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 60: 15–24.

diagnoses can provide negative information that excludes certain causes, some diagnoses can provide disjunctive information about causal possibilities, and some diagnoses can provide information about the causal relations between the symptoms themselves.

6.2 Two kinds of explanatory question

6.2.1 Disease explanation

I had briefly noted earlier in Chapter 1, §1.2.2, that it is important to distinguish two kinds of explanatory question regarding diagnoses in medicine (Qiu, 1989: pp. 199–200; Thagard, 1999: p. 20). The first kind, which I call disease explanation, belongs to empirical research. This is related to the sort of scientific endeavour discussed in Chapter 5, where the *explanandum* is a given clinical syndrome in general, and the *explanans* involves developing a generalised model that brings together the relevant causal factors and mechanisms responsible for the syndrome in general. For example, the disorder characterised by swollen limbs and bleeding gums known as scurvy is explained by defective collagen synthesis due to ascorbic acid deficiency (Thagard, 1999: pp. 120–122). The second kind, which I call diagnostic explanation, occurs in clinical practice. This is the kind of explanation described in Chapter 3, where a patient presents to the clinical encounter with a set of symptoms and the clinician invokes a diagnosis to explain this set of symptoms. For example, the diagnosis of scurvy might be invoked as an explanation of an individual patient's symptoms of swollen limbs and bleeding gums. Here, the *explanandum* is not the clinical syndrome in general, but the clinical presentation of the particular patient, while the *explanans* is the diagnosis.

These two explanatory questions are connected. In diagnostic explanation, where a diagnosis is invoked to explain a patient's symptoms, the understanding of the condition denoted by the diagnosis comes from the generalised model that is constructed through disease explanation. For example, disease explanation informs us that myocardial

infarction in general involves rupture of an atherosclerotic plaque and thrombus formation leading to occlusion of a coronary artery and ischaemic necrosis of the myocardium, and it is in virtue of this knowledge that the diagnosis of myocardial infarction provides a causal explanation of the occurrence of chest pain in a particular patient. In other words, the diagnosis of myocardial infarction informs us that the particular patient instantiates the causal processes represented by the generalised model. This squares neatly with the proposal in Chapter 3 that the condition denoted by a medical diagnosis is often taken to be a repeatable type, of which individual cases are tokens. Hence, what the general model of a disorder looks like has implications for the explanatory function of the diagnosis in the particular case.

Much of the philosophical literature on explanation in psychiatry has focused on disease explanation, rather than diagnostic explanation. The high degrees of heterogeneity and complexity raised in Chapter 5 strongly suggest that most psychiatric disorders cannot be modelled essentialistically, and so may need to be conceptualised differently. For example, we considered in §5.3.2 the possibility of conceptualising some psychiatric disorders as homeostatic property clusters. However, there remains the methodological challenge of how to go about constructing such a generalised model of such a disorder given the problems posed by causal heterogeneity and multilevel complexity. Theorists in the philosophy of psychiatry have proposed idealisation (Murphy, 2006) and explanatory pluralism (Mitchell, 2008; Kendler, 2012) as solutions to these two problems, respectively.

In response to the problem of heterogeneity, Dominic Murphy (2006) suggests what when we try to explain a syndrome in general, what we are aiming to explain is an exemplar, which is an idealised theoretical representation of the syndrome. An exemplar *qua* idealisation is abstracted away from the idiosyncrasies of individual patients. Given the high degree of variability between cases of a given diagnosis, different patients may

resemble the exemplar in different respects and to varying degrees. According to Murphy, to explain a syndrome is to model the various causal relations and mechanisms that have been shown to contribute to the development of the idealised syndrome described in the exemplar. Again, such a model represents an idealised scenario, abstracted away from the actual happenings in particular cases. There may not be a single causal factor in the model that is instantiated by every case of the disorder and there may not be an actual case that instantiates all of the causal relations described in the model.

In response to the problem of multilevel complexity, Sandra Mitchell (2008) endorses a view called integrative pluralism, according to which a satisfactory explanation of a complex system like a psychiatric disorder requires the integration of causal components at multiple levels of organisation. As noted by Murphy (2008), these variables at different levels do not correspond to the same phenomenon described in different ways, but correspond respectively to different phenomena. Hence, it is insufficient to look for deterministic regularities exclusively at a single level, because whatever influence the variables at this level may have is heavily contingent on the joint contribution of variables at other levels. In the case of major depressive disorder, we might need to include information about genetic susceptibilities, neurochemical abnormalities, brain circuits, psychological vulnerabilities, and the ways in which these interact.

Similarly, Kenneth Kendler (2012) endorses an empirically-based pluralism. Given the diverse range of causal variables involved, he argues that there is no single privileged level at which a psychiatric disorder like major depressive disorder can be aetiologically defined. Rather, constructing a general model of the disorder requires the incorporation of research from different disciplines. Kendler (2014) suggests two philosophical approaches to causation that can guide this project. The first, visited earlier in Chapter 5, §5.3.3, is James Woodward's (2003) interventionist theory of causation, which

conceptualises causal factors as difference makers, without placing ontological restrictions on the kinds of variable that can be such difference makers. This allows the inclusion of different causal factors regardless of the explanatory levels to which they belong. The second, previously discussed in Chapter 3, §3.3.3, is the mechanistic approach to causation advocated by theorists such as Machamer *et al.* (2000), which focuses on specifying the mechanisms via which the identified difference makers interact to produce the clinical features of the disorder. According to Kendler (2014: pp. 934–935), it may even be possible to specify some, though perhaps not all, of the mechanisms that link causal factors from different levels, such as those between psychological stress and neurochemical changes in major depressive disorder. However, he acknowledges that this is likely to be a very challenging task, because the causal pathways that cross levels are often not linear, but bidirectional, recursive, and complex.

And so, with respect to disease explanation in psychiatry, there is recognition among contemporary theorists in the philosophy of psychiatry that general explanatory models of disorders are idealisations abstracted away from the heterogeneity of actual cases and that they involve the integration of diverse kinds of causal variable from different levels of organisation. However, significantly less has been written in the philosophical literature about diagnostic explanation in psychiatry. This is the focus of what is to follow.

6.2.2 Diagnostic explanation

Idealisation and explanatory pluralism are promising strategies for disease explanation in psychiatry. When we want to understand what causes depressive symptoms in general, we can conjure up an idealised general representation of the syndrome and model the causal factors that are known to contribute to the phenomenon described in the representation. The resulting model can be helpful for illuminating statistical generalisations and causal

regularities at a population level, in spite of the high degrees of heterogeneity seen in individual cases. However, I argue that heterogeneity remains problematic for diagnostic explanation, where the *explanandum* is not an idealised generalisation but a particular case. As we saw in Chapter 3 the paradigmatic case of diagnostic explanation in bodily medicine consists of a diagnosis indicating the cause of the patient's symptoms and this causal relation being made intelligible by background theoretical knowledge of the mechanisms involved. In such a case, the diagnosis *qua* category constitutes a successful causal explanation partly because it specifies a distinctive and stable kind of causal process which corresponds, with reasonable accuracy and precision, to the process occurring in the particular patient. The worry with a causally heterogeneous diagnostic category is its ambiguity or vagueness, that is, it does not provide such specification of a cause.

Of course, it must be conceded that there are some psychiatric diagnoses that are not beset by causal heterogeneity to such worrying degrees. In §5.4.2, I briefly mentioned the dementias, which are associated with distinctive kinds of neurodegenerative process. Hence, we can reasonably say that such diagnoses do serve as explanations of symptoms in the manner that a paradigmatic medical diagnosis explains a set of symptoms. For example, the diagnosis of Alzheimer's disease explains a patient's clinical presentation by indicating that it is caused by the accumulation of β-amyloid plaques and neurofibrillary tangles in the cerebral cortex, while the diagnosis of vascular dementia explains a patient's clinical presentation by indicating that it is caused by cumulative focal areas of ischaemic necrosis due to the occlusion of cerebral vasculature (Cowen *et al.*, 2012: pp. 326–333). In §5.4.3, I mentioned panic disorder and obsessive-compulsive disorder, which are heterogeneous at the lower biological levels of genetics, neurochemicals, and neural circuits, but are associated with more stable causal regularities at higher psychological levels. Therefore, such diagnoses arguably do explain patients' symptoms by specifying

the processes that are causing them, although the stable causal processes that are represented by the diagnostic categories in these cases are psychological, rather than biological. For example, the diagnosis of panic disorder explains a patient's clinical presentation by indicating that it is caused by a psychological process consisting of a trigger stimulus, perceived threat, apprehension, body sensations, and interpretation of sensations as catastrophic (Clark, 1986: p. 463). Similarly, the diagnosis of obsessive-compulsive disorder explains a patient's clinical presentation by indicating that it is caused by an intrusive thought being erroneously appraised as salient or threatening due to cognitive biases and maintenance of a behaviour via negative reinforcement (Salkovskis, 1985; Rachman, 1993; Veale, 2007).

However, such diagnoses are relatively rare in psychiatry. As noted in Chapter 5, most psychiatric diagnoses, including major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety disorder, are causally heterogeneous at every level of analysis. That is to say, they are variable with respect to the biological, psychological, and social processes involved. Hence, Murphy notes that the symptoms in different cases of major depressive disorder may be produced by different sets of causes:

> It seems unlikely that the same underlying causes explain an irritable adolescent who sleeps late, diets frantically, and lies around the house all day threatening to commit suicide on the one hand, and a sad middle-aged man who can not settle down to any of his normal hobbies, hardly sleeps, eats more and more, can not make love to his wife, and feels worthless. (Murphy, 2006: p. 329)

Similarly, Mitchell (2008: p. 30) suggests that there may be different routes leading to the same symptoms in different individuals. A general model of major depressive disorder,

then, would need to represent the multiple causal pathways that could be responsible for the development of depressive symptoms.

This has implications for what sort of causal information a psychiatric diagnosis like major depressive disorder conveys when a patient who presents to the clinic with mood symptoms is given the diagnosis. It would suggest that the diagnosis does not unequivocally specify a distinctive "disease entity" (Hucklenbroich, 2014) that is responsible for the patient's symptoms in the particular case. Rather, it subsumes a range of possible causal structures that could be instantiated by the patient. Another way to interpret this is to say that major depressive disorder is a disjunctive category. Take C_1, C_2 … C_n to be the diverse causal variables that have been implicated in its pathophysiology. These may interact in different combinations to produce different underlying pathological states, $S_1 = \{C_1 \dots C_x\}$, $S_2 = \{C_2 \dots C_y\}$ … $S_n = \{C_n \dots C_z\}$, each of which can produce the clinical syndrome that satisfies the diagnosis of major depressive disorder. Diagnosing a particular patient with major depressive disorder, then, indicates that the underlying state responsible for the patient's symptoms could be S_1 or S_2 … S_n, but does not provide further causal discrimination beyond this.

Furthermore, different cases of major depressive disorder may need to be understood with different theoretical frameworks. As noted in §6.2.1, the problem of multilevel complexity suggests that a general model of the disorder needs to integrate different kinds of causal variable (Mitchell, 2008; Kendler, 2012). With respect to individual cases, it is possible that the different combinations of variables instantiated by different patients with major depressive disorder may require different explanatory perspectives. For example, cognitive, psychodynamic, and social explanatory perspectives may be of more value for a patient with adverse social circumstances and a history of emotional trauma, while more emphasis may be placed on a neurobiological explanatory perspective for a patient with late-onset depression characterised by melancholic features.

This supports the contention that a psychiatric diagnosis like major depressive disorder lacks unity (Poland *et al.*, 1994). Not only can different patients diagnosed with major depressive disorder instantiate different underlying causal structures, but these different causal structures may need to be understood with appeal to different theoretical frameworks. Jeffrey Poland (2014) argues that this lack of unifying invariance makes the diagnostic categories in psychiatry poor tools for clinical practice. He suggests that a psychiatric diagnosis does not effectively contribute to serving important clinical functions because it "leaves most of the important clinical assessment work undone" (Poland, 2014: p.35). By subsuming different patients with diverse pathologies under the same category, a diagnosis masks information about individual variation that could be important for treatment selection and prognosis. For example, the undifferentiated diagnosis of major depressive disorder does not discriminate between the patient with a dramatic onset of melancholic symptoms, for whom a tricyclic antidepressant and electroconvulsive therapy may be warranted, and the patient with a history of emotional trauma, for whom psychotherapy may be more appropriate. Both patients would be subsumed under the same category of major depressive disorder.

The above criticism paints a rather pessimistic picture of psychiatric diagnoses. However, while I agree that the above mentioned problems significantly impact the clinical roles of diagnoses in psychiatry, I do not go as far as to say that the diagnoses contribute little or nothing of epistemic value to the clinical process. In §6.3, I argue that while most psychiatric diagnoses may not pick out specific causes, there are still ways in which they supply causal information that is of explanatory value.

6.3 Other sorts of diagnostic explanation

6.3.1 Negative causal information

One sort of causal information that can be provided by a psychiatric diagnosis is negative causal information. While a psychiatric diagnosis may not specify the precise causal process leading to the patient's symptom presentation, it nonetheless excludes certain causes. To better understand how this works in clinical practice, we need to look at the process of differential diagnosis, which is where the physician considers multiple possible diagnoses that could explain the patient's symptoms before selecting the diagnosis that best explains them. For example, after assessing a patient with chest pain, a physician may consider gastro-oesophageal reflux disease, pulmonary embolism, and myocardial infarction as possible causes, before inferring that myocardial infarction is the correct diagnosis.

For major depressive disorder, other conditions to be considered in the differential diagnosis include thyroid disorders, adrenal disorders, dementia, cerebral tumours, nutritional deficiencies, drug or alcohol intoxication, and other psychiatric disorders. When assessing a patient with depressive symptoms, it is recommended that he or she is appropriately investigated for these conditions. As stated in *Kaplan and Sadock's Concise Textbook of Clinical Psychiatry*:

> The workup should include tests for thyroid and adrenal functions because disorders of both of these endocrine systems can appear as depressive disorders. In substance-induced mood disorder, a reasonable rule of thumb is that any drug a depressed patient is taking should be considered a potential factor in the mood disorder. (Sadock and Sadock, 2008: p. 217)

Similarly, in *Common Medical Diagnoses: An Algorithmic Approach*, a guideline is presented for the assessment of the symptom of fatigue, according to which major depressive disorder would only be diagnosed once investigations have excluded anaemia, uraemia, diabetes mellitus, adrenal insufficiency, hypokalaemia, hyponatraemia, hepatitis, thyroid disorders, chronic infections, malignancies, and nutritional deficiencies as causes of the fatigue (Healey and Jacobson, 2006: p. 3). Accordingly, in the *Diagnostic and Statistical Manual of Mental Disorders*, it is recommended that major depressive disorder should only be diagnosed once these other medical diagnoses have been excluded:

> Such symptoms count towards a major depressive diagnosis except when they are clearly and fully attributable to a general medical condition. (American Psychiatric Association, 2013: p. 164)

This is not to say that extensive investigations are always performed whenever a patient presents with mood symptoms. Some conditions may be implicitly excluded due to their unlikelihood in the patient's demographic group, such as dementia or a cerebral tumour in a young and otherwise healthy patient with mild depressive symptoms. However, it is normally the case that a patient presenting to secondary care with new affective or psychotic symptoms would at least have blood and urine tests to exclude certain common conditions before a psychiatric diagnosis is established.

What the above highlights is that the diagnosis of major depressive disorder is not made solely on the basis of the relevant symptoms being present, but also requires certain medical causes for the symptoms to be ruled out. A patient who presents with depressive symptoms that turn out to be caused by a cerebral tumour, for example, would not be diagnosed with major depressive disorder, because the diagnosis is excluded by the fact that the symptoms are clearly and fully attributable to a general medical condition. In

virtue of this exclusion criterion, then, a psychiatric diagnosis provides information about what is not in the causal history of the patient's clinical presentation. A diagnosis of major depressive disorder may not pick out a specific cause of the patient's mood symptoms, but it does suggest that they are not being caused by hypothyroidism, drug intoxication, a tumour, and so on.

According to David Lewis (1986b), such exclusion of causes would still qualify as a legitimate sort of causal explanation. According to his account of causal explanation, "to explain an event is to provide some information about its causal history" (Lewis, 1986b: p. 217). This does not necessarily entail specifying a cause of the event, as there are other kinds of information one can give about an event's causal history, including information about what is not in its causal history. For Lewis (1986b: p.222), negative causal information can still be explanatorily relevant information, and so a psychiatric diagnosis can be explanatorily relevant by excluding certain causes, even if it does not itself cite a specific cause.

Helen Beebee (2004) offers a modal analysis of how negative causal information can be explanatorily relevant. She argues that information about the absence of an event provides information about the causal processes in counterfactual worlds where that event occurs. For example, consider that Flora normally waters the orchids regularly, but forgets on one occasion. According to Beebee, Flora's failure to water the orchids cannot be a cause, because it does not denote an event, but rather the absence of an event. Nonetheless, we still accept Flora's failure to water the orchids as an explanation of the orchids dying. This is because it provides information about the causal histories of the nearby possible worlds where Flora had not failed to water the orchids and how these causal histories differ from the causal history of the actual world. In these counterfactual worlds, the causal processes would have ensued in such a way that the orchids would have survived.

It is important to note that Beebee's analysis focuses specifically on the roles of absences in causal explanations, and so is not wholly analogous with my example of diagnostic explanation in psychiatry. Nonetheless, it highlights the general point that a causal explanation does not have to cite a specific cause, but can provide modal information about the possible causal histories of the *explanandum*. I suggest that a similar modal analysis can be applied to other cases of negative causal information, including diagnoses in psychiatry. By indicating that the patient's mood symptoms are not attributable a general medical disorder, the diagnosis of major depressive disorder is providing information about what would have been expected in the counterfactual worlds where the patient's mood symptoms are attributable to a general medical disorder. For example, in the actual world, the physician might only diagnose a patient with major depressive disorder after a thyroid function test yields a normal result, which suggests that the result from the thyroid function test would have been abnormal in the possible world where the patient is not diagnosed with major depressive disorder due to his or her mood symptoms being attributable to a thyroid disorder.

This negative causal explanation can be valuable in the clinical setting. First, it has utility in predicting prognosis and guiding therapeutic intervention. Indicating that a patient's mood symptoms are not due to hypothyroidism suggests that levothyroxine supplementation would not be a therapeutically effective intervention and indicating that they are not due to a cerebral tumour suggests that neurosurgical referral is not warranted. Hence, by excluding these causes, a diagnosis of major depressive disorder can inform clinical decisions. Second, even if it does not specify precisely what is causing the patient's symptoms, a psychiatric diagnosis can offer relief and reassurance by ruling out certain medical diagnoses. For example, when the family of a patient with a new onset of anhedonia, poor concentration, and psychomotor retardation want to know why the patient has developed such symptoms, they may find it extremely valuable to know that

they are not caused by dementia or by a cerebral tumour. In this sense, the diagnosis of major depressive disorder might be compared to the diagnosis of non-cardiac chest pain, which is a diagnosis that is made when a patient presents with central chest pain but investigations reveal no evidence of cardiac disease. The category does not pick out a specific disease kind, as it encompasses oesophageal, pleuritic, and musculoskeletal pathologies, but it is explanatorily valuable because it excludes cardiac causes of the chest pain.

This account of negative causal explanation, then, suggests that a psychiatric diagnosis does not need to identify a specific disease kind to be of causal explanatory value, but can be explanatorily valuable in virtue of the exclusion criterion that states that the symptoms must not be attributable to a general medical disorder. As noted by Beebee (2004), "*E* because *C*" is not equivalent to "*C* causes *E*". Hence, the causal claim "the patient's mood symptoms are caused by major depressive disorder" may indeed be misguided, we can still legitimately make the explanatory claim "the patient has mood symptoms because of major depressive disorder".

However, in spite of the usefulness of negative information in the clinical setting, the account of diagnostic explanation presented here has limits. One problem is that it sets the standard for an acceptable causal explanation too low. If all that is needed for a causal explanation is information about what is not the cause of the *explanandum*, then all sorts of claims that we would not normally consider to be explanations would qualify as causal explanations. For instance, "it's not asthma" would count as a causal explanation of a patient's chronic cough according to the negative causal explanation account. In response, one could propose that the strength of a negative causal explanation depends on how many causal possibilities are excluded by the *explanans*. Hence, major depressive disorder is a better explanation than "it's not asthma", because the former excludes several medical disorders while the latter only excludes asthma. Nonetheless, this would

still relegate psychiatric diagnoses to similar positions as the medically unexplained syndromes discussed in Chapter 2, §2.3.2, whose diagnostic criteria also exclude several medical disorders as causes of patients' symptoms yet are widely considered to be explanatorily unsatisfactory (Kirmayer *et al.*, 2004; Jutel, 2011; Cournoyea and Kennedy, 2014).

Another problem is that in practice there are many instances where psychiatric diagnoses are made without other medical disorders being excluded. In the above discussion, I have been considering an idealised case of differential diagnosis where a patient presents with a new onset of mood symptoms and different diagnoses are presented as possible explanations of the mood symptoms. Here, the diagnoses of major depressive disorder, hypothyroidism, and drug intoxication are presented as competing hypotheses, and the diagnosis of major depressive disorder is only established when the other diagnoses have been adequately excluded. However, there are also cases where a psychiatric disorder is not considered as a competing diagnosis, but as a comorbid diagnosis. For example, major depressive disorder is often treated as an additional comorbid diagnosis in patients with multiple sclerosis, even though it is recognised that in these cases the depressive symptoms may be caused by the pathology associated with the multiple sclerosis (Marrie *et al.*, 2009). Hence, in this sort of scenario, the diagnoses of major depressive disorder fails to exclude multiple sclerosis from the causal history of the patient's mood symptoms.

And so, while psychiatric diagnoses do sometimes provide valuable negative causal explanations, it is implausible that their entire explanatory worth lies only in their providing negative information. In the following subsections, I argue that they can also provide positive causal information. While these sorts of positive causal information fall short of picking out the specific causative pathologies in individual cases, they may nonetheless be explanatorily valuable in the clinical context.

6.3.2 Disjunctive causal information

The notion that psychiatric diagnoses do provide some positive causal information about patients' symptoms is corroborated by the fact that we have at least some scientific knowledge of the causal factors associated with certain disorders. As noted in §6.2.1, even though there is high heterogeneity among cases of major depressive disorder, we can seek to understand major depressive disorder in general by constructing an idealised model that is abstracted away from the idiosyncrasies of individual patients. Murphy (2014) argues that although patients may differ from the idealisation in different respects and to different degrees, the model can nonetheless provide at least an approximation of the causal processes in the individual case:

> The bet is that real patients will be similar to the exemplar in enough respects so that the explanation of the exemplar carries over to the patient. We assume that within the individual there are phenomena and causal relations that are relevantly similar to those worked out for the exemplar, but we cannot expect very precise predictions. (Murphy, 2014: p. 106)

The suggestion here is that while a psychiatric diagnosis *qua* idealised generalisation may not specify the precise causal structure underlying the patient's symptoms in a particular case, it does tell us about processes that are approximately similar to the actual causal processes in the patient's case. Hence, Murphy argues that a psychiatric diagnosis is explanatorily significant, because it gives us at least a vague idea of the sort of process that is producing the patient's symptoms.

However, I argue that things are more complicated than this. While the above picture acknowledges the high degree of variation between individuals, it rests on the assumption that cases of the disorder nonetheless share a similar sort of causal process

(Murphy, 2014: p. 106). As noted in §6.2.2, though, it is possible that there are different sets of causes leading to the same symptoms in different individuals, and so a general model of the disorder would need represent the different routes via which the syndrome can be produced. For example, it is possible that depressive symptoms are not caused by a single kind of causal structure, but may be associated with a disjunction of several underlying states, S_1 or S_2 … S_n, each produced by a different combination of interacting causal variables.

This might be viewed as problematic, because it is a matter of contention whether or not such disjunctive information can constitute an explanation. According to Jaegwon Kim (1998), it cannot. Kim argues that information about a disjunction of possible causes does not yield a single explanation with a disjunctive cause, but a disjunction of different possible explanations of which the correct explanation remains unknown. His example is the symptom of joint pain, which can be caused by a number of different disorders, including rheumatoid arthritis and systemic lupus erythematosus. Consider that a patient with joint pain undergoes a clinical test, the result of which suggests that he or she either has rheumatoid arthritis or systemic lupus erythematosus, but does not indicate which. Kim argues that we do not yet have an explanation of the patient's joint pain:

> I think there is a perfectly clear and intelligible sense in which we don't as yet have an explanation: what we have is a disjunction of two explanations, not a single disjunctive explanation. What I mean is this: we have two possible explanations, and we know that one or the other is the correct one but not which it is. What we have, I claim, is not an explanation with a "disjunctive cause", having rheumatoid arthritis or lupus. There are no such "disjunctive diseases". (Kim, 1998: p. 108)

Kim further qualifies this by arguing that "rheumatoid arthritis or systemic lupus erythematosus" *qua* disjunction does not specify a kind of event, and so is not eligible as a cause. Because it is not eligible as a cause, it cannot then be "citable as a cause in a causal explanation" (Kim, 1998: p. 109).

If Kim is right, then there is reason to suppose that major depressive disorder does not offer a positive causal explanation of a patient's mood symptoms, because it is associated with a range of many possible underlying causal structures but does not specify which one is actually the case in the patient. However, Kim's criteria for explanation are too restrictive. Even if a disjunctive category does not meet the explanatory ideal of picking out a specific cause, I argue that it can nonetheless provide some causal explanatory information. As noted in §6.3.1, it is not necessary to cite a specific cause of an *explanandum* in order to provide explanatorily relevant information about the *explanandum*'s causal history. For example, one could give information about the possible causal histories within which the *explanandum*'s actual causal history lies. I suggest that this is the sort of information a disjunctive category provides.

We can highlight the explanatory relevance of a disjunctive diagnosis by reframing the language in Kim's example. Suppose we say that the test result indicates that the patient has a multisystem autoimmune disease. This is a heterogeneous category that includes rheumatoid arthritis and systemic lupus erythematosus. Hence, stating that the patient has a multisystem autoimmune disease is equivalent to stating that that he or she has the disjunction "rheumatoid arthritis or systemic lupus erythematosus …" without specifying which of these disorders he or she actually has. Nonetheless, it is generally agreed that indicating that the patient has a multisystem autoimmune disease is still explanatorily relevant with respect to his or her joint pain (Rose and Mackay, 1985). Not only does it greatly narrow down the range of conditions in which the patient's actual condition could lie, but it also provides positive causal information about the conditions

that do fall within this range. The diagnosis of multisystem autoimmune disease tells us that the patient's joint pain could be caused by the erosion of the joint surfaces in the case of rheumatoid arthritis, or by systemic inflammation of the connective tissues in the case of systemic lupus erythematosus, and so on.

A disjunctive diagnosis, then, does not specify the actual cause of the patient's symptoms, but it nonetheless subsumes the actual cause within a tighter range of possible causal histories than otherwise would have been available and, moreover, provides some indication of the mechanisms involved in these possible causal histories. This information indicates differences between the causal histories of patients with the diagnosis and those of patients without the diagnosis that can inform further investigations and therapeutic interventions. For instance, stating that a patient has a multisystem autoimmune disease suggests that his or her condition is likely to respond to treatments that act on the immune system and provides a rational basis for further investigations, such as blood tests for specific autoantibodies, which can help specify whether he or she actually has rheumatoid arthritis or systemic lupus erythematosus. A similar example is the category of cancer. This is highly disjunctive, as it encompasses many different kinds of malignancy. Nonetheless, it is hard to deny that it is of causal explanatory value, as it narrows down possible causal histories, provides some indication of the mechanisms involved in these causal histories, and informs investigations to further specify the diagnosis.

The above analysis accommodates the notion that a psychiatric diagnosis *qua* disjunctive category could still provide explanatorily valuable information about a patient's symptoms, even if it does not specify the precise cause of these symptoms. The diagnosis of major depressive disorder, for instance, might be taken to suggest that the patient's symptoms could be due to a state involving underactive serotonin neurotransmission plus variables $C_1 \ldots C_n$, or by a state involving hypothalamus-pituitary-

adrenal axis dysregulation plus variables C_2 … C_{n+1}, and so on. The explanatory value of this disjunctive information is that it tells us some of the ways in which the possible causal structures that could be underlying the patient's symptoms might differ from the causal structure of the non-depressed state. In this sense, the explanatory role of a psychiatric diagnosis like major depressive disorder may be more akin to that of a superordinate category like multisystem autoimmune disorder than to that of a specific medical diagnosis like rheumatoid arthritis.

However, while this analysis shows that disjunctiveness does not necessarily preclude a diagnosis from being explanatory, this explanatory value is also contingent on other conditions. First, it is contingent on whether an exhaustive list of disjuncts can be specified. The superordinate category of cancer is explanatorily valuable, because we are able to specify the different kinds of malignancy that fall under the category. Moreover, we have impressive knowledge of the respective causal structures and mechanisms of these different kinds of malignancy. By contrast, we are far from being able to specify all the possible causal structures that fall under the category of major depressive disorder, or indeed say how many there are. As noted in Chapter 5, we may know a number of the causal variables that can be associated with major depressive disorder, but we still know little about how different combinations of these variables interact to produce symptoms in individual cases. Second, even if some of the disjuncts included in the category could be specified, one might argue that the explanatory value of the category is still contingent on whether we are capable of finding out precisely which disjunct is involved in any given case. This might be made possible with the discovery of biomarkers which indicate specific causal factors that may be potential targets for intervention. However, at the time of writing, such biomarker tests are conspicuously lacking in clinical psychiatry (Bolton, 2012: p. 10). Hence, we may currently be in a situation where research can discover

various causal factors associated with a psychiatric disorder, but we cannot match them to individual patients in the clinic. I shall return to this point in more detail in Chapter 7.

As a modest response to the above concerns, I suggest that a disjunctive category could still provide causal explanatory information of a statistical nature regarding the patient's condition. As noted by Harold Kincaid (2014), despite the diverse range of states that may be subsumed under the category of major depressive disorder, the diagnosis still indicates that the patient is a member of a class of individuals whose biopsychosocial makeups differ in a variety of possible ways from those of non-depressed individuals. Moreover, even if we cannot specify all of the possible disjuncts that fall under the category or find out which disjunct is involved in any given case, the diagnosis still indicates an increased probability of the patient having a given causal mechanism. This information could be clinically useful. For example, on the basis of the knowledge that a proportion of people with major depressive disorder have underactive serotonin neurotransmission, we can say that a given patient with a diagnosis of major depressive disorder has an increased probability of having underactive serotonin neurotransmission. This might provide some justification for a trial of antidepressant medication, which is presumed to exert its action by altering serotonin neurotransmission. Hence, even causal explanatory information that is of a statistical nature can provide some, albeit modest, justificatory support for the predictive and interventional functions of the diagnosis.

It must also be conceded, though, that such clinically useful probabilistic causal explanations are limited to a subset of diagnoses in psychiatry. Other diagnoses may turn out to be too causally heterogeneous to yield explanatorily significant information. I suggest that this may apply to the some of the personality disorders discussed in §5.4.4, whose symptoms likely result from highly contingent combinations of circumstances that differ across cases. With respect to the diagnosis of histrionic personality disorder, for

example, it may be that the lack of even modestly repeatable causal regularities leaves us unable to specify any stable disjuncts, or it may be that the vast number of contingent circumstances that could result in the syndrome translates to such an enormous number of disjuncts in the category that there is only a minute statistical association between each causal factor and the disorder. Therefore, it seems that there are some psychiatric diagnoses, including paranoid personality disorder, schizoid personality disorder, histrionic personality disorder, and avoidant personality disorder, which do not provide explanatorily significant causal information, even with a disjunctive analysis.

In summary, a disjunctive analysis accommodates the possibility of a heterogeneous diagnostic category being of causal explanatory value. However, this explanatory value is also dependent on other considerations, including the degree of causal heterogeneity, whether the disjuncts can be exhaustively specified, and whether we are able to find out which causal variables are involved in any given case. Given the ongoing challenges for research into causal pathways and biomarkers in psychiatry, it must be conceded that at present the positive causal explanatory value of a psychiatric diagnosis *qua* disjunctive category is modest at best and, moreover, that there may be some diagnoses which are too heterogeneous to provide any causal information that is of explanatory value.

6.3.3 Symptom networks

The third sort of causal information a psychiatric diagnosis can provide is information about the causal relations that occur between the symptoms and sustain them as a stable cluster. This draws on the symptom network approach to psychiatric disorders advocated by Denny Borsboom (2008) and Cramer *et al.* (2010), which we previously encountered in Chapter 5, §5.3.2, in my discussion of homeostatic property cluster conceptualisations of psychiatric disorders. Recall that according to this approach, a psychiatric disorder is

conceptualised as a network of symptoms that reinforce each other via causal relations. For example, in the case of major depressive disorder, fatigue may result in poor concentration, which may trigger thoughts of inferiority and worry, which in turn may impair sleep, thus reinforcing the fatigue (Cramer *et al.*, 2010: pp. 140–141).

By emphasising the causal relations between the symptoms themselves, the symptom network approach accounts for why the symptoms associated with a given psychiatric diagnosis tend to cluster together in a statistically significant way, without the need to invoke an underlying latent pathology as the cause of these symptoms. Fatigue, poor concentration, worry, and insomnia cluster together because they causally reinforce each other, not because they are caused by a common underlying pathology. Hence, by defining a psychiatric disorder at the level of its symptoms rather than at the level of underlying biological causal factors, advocates of the symptom network approach can sidestep the problems of heterogeneity and complexity that affect these underlying causal factors.

Conceptualising the disorder at the level of its symptoms has implications for diagnostic explanation. In their commentary on Cramer *et al.*'s (2010) paper, Hood and Lovett (2010) present an argument, reminiscent of Thomas Szasz (1960: p. 15), suggesting that a logical consequence of excluding underlying causes from the conceptualisation of a psychiatric disorder is that the disorder cannot then function as a causal explanation of a patient's symptoms. If it turns out that the diagnosis of major depressive disorder, for example, does not refer to anything over and above the symptoms of low mood, anhedonia, fatigue, and so forth, then to invoke the diagnosis of major depressive disorder as an explanation of why these symptoms occur in a particular patient would be circular. Hence, we would be faced again with the conceptual problem regarding the explanatory status of a psychiatric diagnosis, which I introduced in Chapter 1, §1.1.2, and addressed in Chapter 4. However, even if Hood and Lovett may be right in

claiming that something cannot be the cause of a set of symptoms if it is nothing over and above these symptoms, I argue that the symptom network approach enables a psychiatric diagnosis to provide causal information of a different sort. In particular, it provides information about the above mentioned causal relations between the symptoms themselves. It is in virtue of this causal information that the symptom network approach distinguishes between an arbitrary grouping of symptoms and a grouping of symptoms that reflect the causal structure of the world. As argued by Borsboom and Cramer:

> In addition, network modeling has the philosophical advantage of dropping the unrealistic idea that symptoms of a single disorder share a single causal background, while it simultaneously avoids the relativistic consequence that disorders are merely labels for an arbitrary set of symptoms … (Borsboom and Cramer, 2013: p. 93)

This suggests that although the symptom network model defines a psychiatric diagnosis at the level of its symptoms, the diagnosis does not merely serve as a descriptive label for these symptoms, but also provides additional information about the causal relations that sustain these symptoms as a stable cluster.

Consider the patient who presents to the clinic with low mood, poor concentration, fatigue, and insomnia. According to the symptom network approach, the diagnosis of major depressive disorder indicates that these symptoms constitute a dynamically stable system held together by causal relations. Again, this does not meet the standard model of explanation where a diagnosis picks out an underlying pathology that is causing the patient's symptoms, but there is nonetheless good reason to think of it as being a sort of causal explanation. In particular, it explains why the patient's symptoms of occur concomitantly. By positing causal relations between the symptoms, the diagnosis of

major depressive disorder explains why they have aggregated and persisted as they have, regardless of what pathological processes may be underlying them in the particular case.

Hence, if the symptom network approach is assumed, a psychiatric diagnosis can provide some causal explanatory information about a patient's symptoms, even if the underlying causes of the symptoms vary across cases. However, it is causal explanatory information of a different sort from that provided by a medical diagnosis like myocardial infarction, which picks out an underlying cause of the patient's chest pain. Again, a claim such as "the patient's mood symptoms are caused by major depressive disorder" is misguided, this time because the symptom network model suggests that major depressive disorder does not refer to a latent underlying pathology responsible for the symptoms, but we can still claim that the diagnosis of major depressive disorder causally explains the patient's symptoms on the grounds that it refers to the causal structure by which the symptoms induce and reinforce each other.

The claim that a psychiatric diagnoses provide information about the causal structures by which sets of symptoms are maintained sits well with the fact that specific therapies for some psychiatric disorders often achieve reductions in some symptoms by optimally intervening on others (Borsboom and Cramer, 2013: p. 98). For example, cognitive-behavioural therapy for major depressive disorder employs the notion that thoughts, actions, emotions, and bodily symptoms can all influence one another. The idea is that intervening on the patient's negative thoughts and level of activity through cognitive restructuring and behavioural activation might then lead to improvements in his or her mood and interest level. Therefore, under the symptom network approach, the causal information conveyed by a psychiatric diagnosis can support therapeutic intervention.

The symptom network approach, then, makes it possible for a psychiatric diagnosis to convey causal explanatory information about a patient's symptoms without specifying

an underlying causative pathology. However, a limitation of the approach is that it may turn out not to be applicable to all major psychiatric diagnoses. For instance, it is not obvious why, in the case of schizophrenia, hallucinations and delusions should be causally connected to blunted affect and catatonic behaviour. Similarly, in the case of bipolar disorder, it is not obvious how mania and depression are supposed to causally induce each other. It appears that in these cases we need to appeal to additional causal variables, such as underlying neurobiological processes, in order to make the link between hallucinations and affective blunting, and the link between mania and depression intelligible. Therefore, while there are plausibly some psychiatric diagnoses that provide causal explanatory information about symptoms without needing to invoke information about the underlying processes, it is unlikely that this is the case for all psychiatric diagnoses.

6.4 Conclusion

We should take seriously the possibility that many major psychiatric disorders may turn out to exhibit high degrees of causal heterogeneity and complexity. This chapter has examined some of the implications of this for the diagnostic explanation in psychiatry. If it turns out that a given diagnostic category subsumes a variety of different underlying causal structures, then this would suggests that diagnostic explanation in psychiatry falls short of the standard model of causal explanation where a diagnosis specifies the causative pathology responsible for the patient's symptoms.

Nonetheless, I have argued that some psychiatric diagnoses can still provide other sorts of causal information that can be explanatorily relevant. First, in virtue of the exclusion criteria, a psychiatric diagnosis can sometimes provide negative causal information by ruling out other medical causes. Second, in virtue of our scientific knowledge of some of the various causal factors implicated in psychiatric disorders, a

diagnosis can provide some probabilistic or disjunctive information about the possible causal processes that might be relevant to the patient, although this information is likely to be vague and partial given our limited scientific understanding of how these various factors come together. Third, in virtue of the causal relations between the symptoms themselves, a psychiatric diagnosis can provide information about why the patient's symptoms occur together and persist as they do. I have also shown how these causal explanatory functions of psychiatric diagnoses might still be useful in supporting some of the other functions of the diagnoses, even if they do not indicate specific causal processes. The negative causal information can exclude certain avenues for intervention and offer reassurance to patients. The probabilistic information about possible causal processes can occasionally support therapeutic decisions, although it must be conceded that we are far from being able to specify pathways and biomarkers that could allow for powerful interventions. The information about the causal relations between symptoms can support therapeutic interventions that target particular symptoms to optimally reduce others. However, it must also be conceded that not all psychiatric diagnoses provide all three sorts of explanatory information and that there are likely to be some diagnoses that do not provide any causal information of explanatory significance.

The above considerations have normative implications for clinical practice. These include implications for how psychiatric diagnoses are communicated in clinical discourse, the validity of current psychiatric classification, and the respective roles of categorical diagnoses and individualised formulations in psychiatry. I shall lay out some of these implications in Chapter 7.

7. Normative and Practical Implications

7.1 Introduction

As we saw in Chapter 6, many psychiatric diagnoses do not meet the explanatory ideal in medicine of the diagnosis that indicates a specific cause of the patient's symptoms. Nonetheless, I have argued that they can still convey other sorts of causal information that are explanatorily relevant, including negative causal information, disjunctive causal information, and information about causal relations in symptom networks. In this chapter, I explore some of the normative implications of the above for clinical psychiatric practice.

I look at three areas of practice that are affected by these epistemological issues. These are the ways in which psychiatric diagnoses are communicated, the validity of current psychiatric classification, and the complementary roles of categorical diagnoses and individualised formulations. In §7.2, I examine the implications of the above for how diagnoses ought to be communicated in psychiatric discourse. I suggest that the ways in which they are currently portrayed in some clinical texts amounts to problematic essentialisation and that more nuanced language is required. In §7.3, I address whether the above concerns warrant revision of our current psychiatric classification system so that the diagnostic categories reflect distinctive causal structures. While I agree that there are good reasons to aim for a causal classification system, I argue that there remain significant challenges to devising and implementing such a classification system that make the prospects of one coming into use in the immediate future unlikely. In §7.4, I explore whether, given the concern about causal heterogeneity, the diagnostic process in clinical psychiatry ought to involve a more individualised approach than merely invoking a categorical diagnosis. This is a strategy recommended by the World Psychiatric Association's (2003) International Guidelines for Diagnostic Assessment (IGDA)

workgroup. The idea is that an individualised formulation can complement a categorical diagnosis to attain more precise knowledge of the causal factors that pertain to the particular patient, thus arriving at a more clinically useful explanation of his or her symptoms than would have otherwise been provided by the categorical diagnosis on its own.

7.2 Communicating psychiatric diagnoses

7.2.1 The problem of essentialisation

As noted in Chapter 1, psychiatric diagnoses are often communicated in clinical discourse as if they refer to distinctive kinds of condition that are causes of sets of symptoms. This is apparent in the passages from clinical textbooks and health information resources quoted in §1.1.1. The sort of language used in such passages reflects the influence of the "disease entity" model in medicine, whereby diagnoses are taken to correspond to distinctive and repeatable causal types (Hucklenbroich, 2014). This model has had reasonable success with many, though by no means all, diagnoses in bodily medicine. Given the historical and cultural underpinnings of psychiatry as a medical discipline, it is unsurprising that such a model remains influential in psychiatric discourse (Poland, 2014: pp. 31–33).

However, the findings of Chapter 5 and Chapter 6 suggest that this sort of language regarding psychiatric diagnoses is misguided. With the exceptions of the dementias and a few disorders with stable psychological causal structures, most psychiatric diagnoses do not correspond to distinctive and repeatable causal types, but to heterogeneous categories involving variable combinations of diverse causal factors. Hence, a diagnostic category in psychiatry does not refer to a distinctive kind of condition that causes a set of symptoms as per the above mentioned clinical textbooks and health information resources, but at best corresponds to a disjunction of possible

causal structures. This suggests that some of the portrayals of psychiatric diagnoses in clinical discourse amount to problematic essentialisation.

Such essentialisation is not only misleading, but it has been argued that it is potentially harmful. Nick Haslam (2014) presents evidence suggesting that the essentialisation of psychiatric disorders can encourage damaging stigma. It might appear intuitive that attributing disordered behaviour to a distinctive biological cause such as a brain abnormality or a neurotransmitter imbalance would increase sympathy and reduce blame, but studies have shown that it is actually associated with negative attitudes towards the patient, including the desire for greater social distance from him or her, greater perceived dangerousness, and lower expectations that he or she will recover (Mehta and Farina, 1997; Lam *et al.*, 2005; Phelan, 2005). According to Haslam, essentialisation encourages these attitudes because it "represents sufferers as categorically abnormal, immutably afflicted, and essentially different" (Haslam, 2014: p.25).

7.2.2 Modifying clinical discourse

The above considerations suggest the need for more caution regarding the ways psychiatric diagnoses are communicated in clinical discourse. Instead of portraying a psychiatric diagnosis as specifying a distinctive kind of pathology that causes a set of symptoms, it ought to be made more explicit in clinical textbooks, health information resources, and communicative exchanges that a psychiatric diagnosis refers to a heterogeneous category associated with a cluster of symptoms that could be caused by a variety of possible causal pathways. This would encourage a more nuanced and empirically accurate conception of the disorder which acknowledges its causal basis while avoiding the problem of essentialisation. Hence, in virtue of the causal knowledge attained from empirical research, the diagnosis is not a mere label for an arbitrary collection of symptoms, yet the complexity and heterogeneity of its causal profile

indicates that the category does not pick out an invariant type, but subsumes a variety of causal processes.

I also recommend that the multifactorial natures of the causal pathways ought to be acknowledged more explicitly. As noted by France *et al.* (2007), the dominant cultural narrative regarding major depressive disorder is that it is a chemical imbalance, a narrative that is partly attributable to the influence of the pharmaceutical industry's marketing antidepressant medication as restoring chemical balance. Another popular narrative is the claim that psychiatric disorders are brain disorders, a narrative that is motivated by recent research in neuroscience (Insel *et al.*, 2010). However, as shown in Chapter 5, there is evidence that the causal profiles of many psychiatric disorders involve complex interactions between diverse biological, psychological, and social factors. Hence, neurocentrism leaves out many important aspects of psychopathology and avenues for intervention. Instead, I suggest that we ought to endorse an explanatory pluralism that does not privilege any single level of analysis and explicitly acknowledge that psychiatric disorders are constituted by combinations of biological, psychological, and social processes. This can encourage more a more nuanced understanding of psychiatric disorder that reflects its multifactorial nature. Such a notion is not new to psychiatry and can be traced at least as far back as George Engel's (1977) biopsychosocial model of health. The original version of the model has recently been criticised by Nassir Ghaemi (2009) for its empirically questionable claim that biological, psychological, and social factors are always equally involved. Nonetheless, as noted by Derek Bolton (2013), we can still maintain a version of the model that is consistent with the recent developments in psychiatric science which states that the causal pathways and interventions for psychiatric disorders may involve biological, psychological, and social levels, without claiming that all of them are always involved to equal degrees. Kenneth Kendler (2012)

also argues this sort of pluralism can also enable more unbiased and empirically rigorous research into psychiatric aetiology and treatment.

While I have argued that it is mistaken to speak of many psychiatric diagnoses as if they correspond to distinctive kinds of pathology that cause sets of symptoms, I suggest that it can nonetheless be defensible, except perhaps for those diagnoses that are not associated with even modestly repeatable causal factors, to speak of symptoms being explained by psychiatric diagnoses. As noted in Chapter 6, Helen Beebee (2004) argues that a causal explanation involves providing information about the *explanandum*'s causal history but does not necessarily require the precise cause of the *explanandum* to be cited, and so "*E* because *C*" is not equivalent to "*C* causes *E*". A psychiatric diagnosis like major depressive disorder may not specify the precise cause of the patient's symptoms, but it does convey other sorts of causal information as argued previously. Hence, we would be mistaken to say "the patient's anhedonia and fatigue are caused by major depressive disorder", but we could still say "the patient has anhedonia and fatigue because of major depressive disorder". However, I suggest that it should also be clarified what sort of explanatory information is being provided when such a diagnosis is being communicated.

I concede that there are likely to be difficulties with implementing the above recommended changes in the communicative practices regarding psychiatric diagnoses. As argued by Peter Zachar (2014), the ways we tend to think about disorders are influenced by pervasive essentialist biases that are difficult to overcome. Furthermore, as previously noted, the "disease entity" model, which is a dominant paradigm in modern medicine, is shaped by such essentialistic thinking (Hucklenbroich, 2014). Hence, conceptualisations of psychiatric diagnoses that acknowledge their causal complexity and heterogeneity not only run counter to our habits and inclinations regarding disorders, but challenge deeply entrenched medical tropes and traditions. Nonetheless, I remain

optimistic that causal complexity and heterogeneity can be successfully communicated in terms that are comprehensible and acceptable to the public. A good example can be found in the recently published health information leaflet on major depressive disorder by the Royal College of Psychiatrists:

> Why does it happen? … There is often more than one reason, and these will be different for different people. They include: Things that happen in our lives … Circumstances … Physical illness … Personality … Alcohol … Gender … Genes … (Royal College of Psychiatrists, 2015)

While it does not go into a lot of detail, this leaflet is commendable. First, it acknowledges that major depressive disorder can have different causal structures in different patients. Second, it acknowledges that these causal structures can involve combinations of factors at biological, psychological, and social levels, without privileging any particular level over the others. Third, it achieves the above without compromising the idea that major depressive disorder is a valid clinical condition. I suggest that other portrayals of psychiatric diagnoses should follow this authoritative example.

7.3 Classificatory revision

7.3.1 Current classification in context

In addition to the need to modify the ways in which psychiatric diagnoses are communicated in clinical discourse, there is the more radical question of whether the issues raised above warrant the revision of the diagnostic classification system in psychiatry so that its categories reflect invariant causal types. As noted in Chapter 1, §1.1.3, the dominant classification system in use today is the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), which is currently in its fifth edition (American

Psychiatric Association, 2013). Since *DSM*-III (1980), the definitions of and operational criteria for the diagnoses have largely been based on observable symptoms, rather than underlying causes. A key motivation for this atheoretical approach to classification is to enable use by and increase diagnostic reliability among practitioners from different theoretical backgrounds. Of course, it is worth noting that whether or not *DSM*-III actually succeeds at being genuinely atheoretical is contested. For example, Fulford *et al.* (2006: pp. 289–313) suggest that *DSM*-III cannot be atheoretical but must be at least implicitly theory-laden, on the basis that observations in general are theory-laden. Nonetheless, what is important is that the diagnoses in *DSM*-III are explicitly classified and defined on the basis of symptoms, while remaining neutral about causes. This also applies to *DSM*-IV (1994) and *DSM*-5 (2013), despite the latter having a revised chapter organisation whereby disorders that are believed to have similar aetiologies are placed adjacent to each other.

A consequence of such an aetiologically neutral nosology is that it permits the possibility of diagnoses that are causally heterogeneous. For example, as mentioned in Chapter 5, there appear to be several possible causal pathways that could produce the symptoms of major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety disorder. For other diagnoses, such as some of the personality disorders, there may even be so many contingent reasons why the syndromes arise that it may not be possible to discern repeatable causal structures. The likelihood of *DSM*-5 including causally heterogeneous diagnostic categories is further compounded by the fact that the classification has also been shaped by influences other than scientific research. Rachel Cooper (2005) notes that a variety of social and political forces, including insurance companies, the pharmaceutical industry, and lobbyist groups, have influenced the diagnostic categories in the recent editions of the *DSM*. Moreover, as noted in Chapter 5, §5.4.4, it has been argued that some diagnostic categories, especially the personality

disorders, have been heavily shaped by moral judgements and cultural attitudes regarding gender (Charland, 2004; Potter, 2004). Of course, it is by no means necessarily the case that all diagnostic categories that are shaped by such social and political forces are causally heterogeneous. However, the suggestion is that the inclusion of a given diagnostic category in the *DSM* may not be entirely based on a scientifically informed expectation that instances of the diagnosis share the same kind of causal structure.

7.3.2 Towards a causal classification

Several theorists in the philosophy of psychiatry find the *DSM*'s symptom-based approach to classification unsatisfactory, because it permits diagnostic categories that are causally heterogeneous (Poland *et al.*, 1994; Bentall, 2003; Murphy, 2006; Haslam, 2014). According to Dominic Murphy (2006: pp. 323–324), this places psychiatry at odds with the rest of medicine, where diagnoses correspond to the causal antecedents of symptoms. Accordingly, there have been recent calls to revise diagnostic classification in psychiatry so that its diagnostic categories correspond to distinctive and repeatable causes.

Different authors suggest different approaches to classificatory revision. Poland *et al.* (1994) argue that the focus of investigation should be the individual problems of patients, such as elemental cognitive deficits, behavioural skills deficits, and problematic social interactions, which are not adequately captured by the *DSM* categories. Similarly, Richard Bentall (2003) proposes that we abandon the current diagnostic categories and instead try to locate regularities at a finer level. For example, he suggests that the category of schizophrenia should be discarded, and that instead we should separate out and investigate the individual problems, such as auditory hallucination and thought disorder, in isolation. According to Bentall, these individual problems are what are likely to yield stable causal structures, whereas the current category of schizophrenia merely represents a contingent and variable conjunction of these problems. For instance, he hypothesises

that thought disorder has a stable causal structure consisting of disturbances in working memory due to emotional arousal interacting with other deficits in semantic memory and introspective monitoring.

A current research programme that follows a similar approach is the National Institute of Mental Health's Research Domain Criteria (RDoC) programme (Insel *et al.*, 2010). The RDoC programme "asks investigators to step back from diagnoses based on heterogeneous clusters of symptoms and, instead, to focus on basic dimensions of functioning across the wellness spectrum that might relate to various aspects of symptoms" (Ford *et al.*, 2014: p. S295). Rather than beginning with the current *DSM*-5 categories, it suggests beginning with "what is known about healthy, adaptive behavioral and neural circuit functioning, and then to understand how alterations in these systems could eventuate in various types of symptoms and impairments" (Ford *et al.*, 2014: p. S296). For example, instead of beginning with a syndromic category like schizophrenia, the RDoC might begin with a symptom like auditory hallucination, and proceed to study the associated neurobiology in clinical and non-clinical populations. The aim is to discover the causal mechanisms that can produce auditory hallucination in clinical and non-clinical populations, which in turn can inform new ways of classifying psychopathology for research purposes.

The above described approach of searching for causal regularities at the level of individual symptoms is dubbed by Dominic Murphy as the "zooming-in" approach (Murphy, 2010: p. 607). In contrast to this approach, Murphy endorses a "zooming-out" approach. Rather than isolating individual symptoms as the objects of investigation, he suggests constructing exemplars, which are idealised theoretical representations of syndromes abstracted away from the idiosyncrasies of individual cases. Empirical research can then enable us to build models of the various causal factors and mechanisms that can contribute to the phenomena described by the exemplars.

The idea, according to Murphy, is that the knowledge of causes provided by these models can inform a new aetiologically-based classification system. If multiple syndromes turn out to result from the same causal pathway, then this might support the lumping together of these syndromes into a single category. For example, in general medicine, the syndromes that were known as phthisis, consumption, and scrofula were eventually lumped together and subsumed under the diagnosis of tuberculosis when it was discovered that their causal structures all involve *Mycobacterium tuberculosis* infection. With respect to psychiatric disorders, Murphy (2006: p. 355) makes the tentative suggestion of lumping specific phobias and panic disorder together as threat response system disorders. Conversely, if there turn out to be a number of causal pathways that can produce a given syndrome, then this might support the splitting of the old diagnostic category into new diagnostic categories, each representing one of the causal pathways. Murphy (2006: pp. 352–354) discusses the case of paediatric autoimmune neuropsychiatric disorder associated with *Streptococcus* (PANDAS) as an example of a category that arose from such splitting. Among the population of children diagnosed with obsessive-compulsive disorder, some cases were observed to have developed following infections by group A β-haemolytic *Streptococcus pyogenes* (Swedo *et al.*, 1998). These cases were subsequently discovered to involve a distinctive pathological process, antibody-mediated inflammation of the basal ganglia triggered by *S. pyogenes* antigens, which is markedly different from the cognitive basis of classical obsessive-compulsive disorder. The discovery of this distinctive causal pathway led to PANDAS being recognised as a separate diagnostic category from standard obsessive-compulsive disorder.

Importantly, Murphy (2006: pp. 357–359) contends that a distinctive causal pathway does not entail a specific aetiological factor such as a single gene, but can consist of a proximal generalisation involving the complex interaction of multiple variables, provided this process is sufficiently stable across cases. Furthermore, these variables may

occur at different levels of organisation and belong to different theoretical frameworks. For example, a causal process might involve interactions of genes, cognitive disruptions, psychodynamic factors, and social factors like marital difficulties (Murphy, 2006: p. 351).

In view of these divergent theoretical frameworks in psychiatry, Jonathan Tsou (2015) proposes that the move to an aetiologically-based diagnostic classification would benefit from theoretical pluralism, both in the revision process and at the taxonomic level. He suggests that the *DSM* revision process ought to involve dialogues between and accommodate the research of investigators from diverse theoretical backgrounds, including literature reviews from investigators working outside the *DSM* revision process. Furthermore, given that there are multiple purposes for classifying disorders, he suggests that researchers and clinicians should develop alternative classifications, which can then inform the future *DSM* categories. The hope is that the theoretical knowledge about the causes of psychiatric syndromes provided by these alternative classifications can be incorporated into the new *DSM* classification system.

Although their suggested methods differ, the theorists mentioned in this subsection describe the shared goal of a classification system whose diagnostic categories correspond to distinctive causal processes that are sufficiently stable across cases. Ultimately, such an aetiologically-based classification system would require us to discard or further subtype diagnostic categories that are causally heterogeneous. This would certainly not be without precedent in the history of medicine. For example, when diabetes mellitus was discovered to be associated with two different pathologies, namely autoimmune destruction of pancreatic β-cells and insulin resistance, it was subtyped into type I diabetes mellitus and type II diabetes mellitus, respectively (Zajac *et al.*, 2010). When bronchial carcinoma was discovered to include neoplasms with varying histological origins, it was subtyped into small cell carcinoma, large cell carcinoma, squamous cell carcinoma, and adenocarcinoma (Travis, 2012). When dropsy was discovered to subsume a number of different

pathological processes, including heart failure and renal failure, the category of dropsy was discarded in favour of more specific diagnoses that referred to these pathologies (Kendell, 1989; Peitzman, 2007). Also, such an aetiologically-based classification system may require us to relocate a diagnostic category if new information arises regarding the kind of causative pathology with which it is associated. For example, Creutzfeldt-Jakob disease was once thought to be a viral disease, but it was subsequently reclassified as a prional disease after it was discovered that the responsible agents lacked nucleic acids (Thagard, 1999).

A revisionary approach to psychiatric classification, then, suggests that some of the current *DSM*-5 diagnoses, perhaps including major depressive disorder and schizophrenia, can be likened to the archaic diagnosis of dropsy, because each subsumes a heterogeneous range of causal pathways. Under an aetiologically-based classification system, their replacement with or division into newer categories that reflect more precise causal pathways might be warranted. Other diagnostic categories may need to be reshuffled if an aetiologically-based classification is assumed, such that those with associated with similar kinds of causal process are placed closer together. Finally, other diagnoses, such as some of the personality disorders, may not be associated with even modestly repeatable causal pathways, and so an aetiologically-based classification system may require us to discard them altogether.

7.3.3 Critical discussion

In theory, there are good epistemic reasons to prefer an aetiologically-based diagnostic classification to the current symptom-based classification. If diagnostic categories are made to correspond to distinctive causal structures, then this would make them more precise causal explanations of patients' symptoms than the current *DSM*-5 diagnoses. While current *DSM*-5 diagnoses like major depressive disorder and schizophrenia are

associated with heterogeneous ranges of possible causal processes, the aim is that a diagnosis under a successful aetiologically-based classification would specify the causative pathology responsible for the patient's symptom presentation, and so meet what is often considered the diagnostic ideal in medicine (Schwartz and Elstein, 2008; Cournoyea and Kennedy, 2014).

Potential practical benefits of this enhanced causal explanatory function include stronger predictions and targeted therapeutic interventions. As noted by Robert Kendell, it was "only after physicians had learned to distinguish between the renal and cardiac forms of dropsy that it was possible to predict which patients were likely to benefit from digitalis" (Kendell, 1989: p. 47). While a diagnosis that subsumes a range of possible causal pathways can still support probabilistic estimates, its heterogeneity makes it unlikely to yield very precise predictions. However, a diagnosis that corresponds to a specific causal structure is capable of supporting more precise predictions about clinical outcomes and guiding more effective treatment decisions. This is for two reasons. First, as argued by Richard Boyd (1999), a causally homogeneous category is more projectable than a causally heterogeneous one, because its members share causal similarities that can ground inductive inferences. Second, a diagnostic category that reflects a distinctive causative pathology rather than a range of possible causal pathways provides more precise knowledge of the mechanisms producing the patient's symptoms, which identifies targets for therapeutic interventions. In other words, by indicating the cause of the patient's symptoms, such a diagnosis tells the clinician where to intervene.

Of course, basing diagnostic categories on causes is not the only reasonable way of revising psychiatric classification. Another alternative might be to classify diagnoses based on treatment effects, an approach Jennifer Radden (2003) calls "drug cartography". The idea is to group together conditions that respond to the same kind of treatment, such as placing conditions that are alleviated by antidepressant drugs in the same category. In

their recent paper, "Carving Bipolarity Using a Lithium Sword" (2014), Malhi and Geddes suggest classifying affective disorders based on responsiveness to lithium. An obvious strength of this approach is its therapeutic utility. A diagnosis based on treatment response tells the clinician what therapeutic intervention is likely to alleviate the patient's condition. Moreover, diagnostic categories whose respective members are unified by their treatment effects can potentially support inductive inferences that can inform evidence-based treatment guidelines. However, a drawback of such a treatment-based classification is that like the current symptom-based classification, its diagnostic categories have limited explanatory value. Although they may permit inductive inferences about therapeutic outcomes, they do not explain why these therapeutic outcomes occur, nor do they indicate the specific causes of the patients' symptoms. By contrast, an advantage of an aetiologically-based diagnostic category is that it provides a causal story about why the patient has a certain set of symptoms and why they are likely to respond to a certain treatment.

In spite of the above mentioned benefits of classificatory revision, I argue that there are significant empirical, conceptual, and bureaucratic challenges to devising and implementing a successful aetiologically-based classification system that make one unlikely to in the immediate future. A problem with the "zooming-in" approach of relocating the categories at the level of individual problems is that there is no guarantee that these individual problems will be significantly less causally heterogeneous than the current syndromic categories. For example, discarding the syndromic category of schizophrenia and examining the individual problem of delusion in isolation may seem like a promising way to focus in on a more stable unit, but this rests on the assumption that the category of delusion has a significantly more stable causal structure than schizophrenia. However, as argued by Murphy (2014), the concept of delusion is not defined by a causal signature but by folk psychological assumptions about how the mind

works, and so it would not be surprising if there turns out to be numerous possible causal pathways that could lead to what we call a delusion.

Moreover, I argue that it is possible that an individual problem studied in isolation could actually turn out to be more causally heterogeneous than it would have been if it had been studied as part of a broader syndrome. Consider the example of headache as a symptom of migraine. This is currently understood to be caused by the disturbance of the subcortical aminergic sensory modulatory systems (Goadsby, 2012). However, headache can also occur in many other disorders and has different causes depending on the diagnosis. For example, headaches in meningitis and subarachnoid haemorrhage are caused by irritation of the nociceptors in the meninges, headache in a cerebral tumour is caused by raised intracranial pressure, and tension headache is caused by activation of the peripheral nerves in the head and neck (Bogduk, 1995). Ignoring the diagnostic categories and studying headache as an independent category, then, would yield much greater causal heterogeneity than studying headache as part of a given syndrome. The diagnostic category of migraine in this case narrowed down the type of headache being investigated, and so allowed the discovery of a distinctive causal structure. It is therefore important to consider the possibility that some syndromic categories in psychiatry might also narrow down the scope of investigation, such that studying a certain symptom as it occurs in a given syndrome could yield less causal heterogeneity than studying the symptom regardless of the diagnosis.

Another problem with the "zooming-in" approach, noted by Murphy (2010: p. 608), is that even if stable causal processes are found at the level of individual problems, it does not account for how these processes interact to produce the various clinical presentations of actual patients. He notes that individual psychiatric symptoms are not independent of each other, but tend to occur together in statistically significant clusters. This suggests that psychiatric syndromes are not just static conjunctions of individual

problems with each its own distinctive causal structure, but depend on complex and dynamic interactions between diverse causal variables.

The "zooming-out" also faces challenges. As noted in §7.3.1, Murphy (2006) suggests revising diagnostic categories based on the causal processes that account for the phenomena described by idealised representations of psychiatric syndromes. However, given that psychiatric disorders involve heterogeneous combinations of different sorts of factor interacting across multiple levels of organisation, it is not clear how such a cornucopia of diverse and interacting causal variables can be reconciled with a simple and stable classificatory system to be used by clinicians and researchers (Bolton, 2007). In view of this causal complexity, it is plausible that there may be multiple ways to categorise disorders based on their causes. This is recognised by Kendler *et al.*, who argue that knowledge of causes on its own "does not tell us how or whether to privilege one set of mechanisms over another" (Kendler *et al.*, 2011: p. 1149), and so is insufficient for determining the lumping or splitting of categories.

A plausible idea is that in addition to knowledge of causes, explanatory interests help determine the lumping and splitting of categories. This is hinted at by Beebee and Sabbarton-Leary (2010: p. 24), who argue that if schizophrenia were to turn out to have two distinct neurological bases N_1 and N_2, there would only be grounds to split the category of schizophrenia into N_1 and N_2 if these separate categories yield better predictions than the old category of schizophrenia, but not if it turns out that whether one has N_1 or N_2 makes no significant difference to prognosis or treatment. A problem, however, is that different people may have different explanatory interests. As we saw in In §7.3.1, Tsou (2015) recommends that in light of their different explanatory interests and theoretical backgrounds, clinicians and researchers develop alternative classifications that can then inform future revisions of the *DSM*. The challenge, then, would be how to successfully reconcile these different classifications in view of these divergent interests

and theoretical perspectives. It may turn out to be the case that a universal and aetiologically-based classification system that satisfies both researchers and clinicians of different theoretical orientations is not feasible.

A final challenge to classificatory revision in psychiatry is that there are external processes that make the *DSM* very resistant to change. Rachel Cooper (2015) provides an account of these processes with appeal to the concepts of path-dependence and lock-in (Bowker and Star, 2000). These refer to when a classification that at an initial time is accepted as part of the information infrastructure of science facilitates processes that further reinforce the use of that classification, until the classification becomes extremely hard to dislodge. With respect to the *DSM*, Cooper suggests that path-dependence and lock-in have occurred at two levels. First, the American Psychiatric Association's past success in publishing the *DSM* facilitates its future success. The sales of *DSM*-III and *DSM*-IV brought in substantial profits for the American Psychiatric Association, allowing them to invest more in future editions and build up the bureaucratic structures that enable their production. This makes it very difficult for any other professional body to produce an alternative classification to rival the *DSM*. Second, Cooper argues that it has become very difficult for the American Psychiatric Association to radically change the *DSM* categories. Since the *DSM* categories became accepted, they have been extensively used in research, as well as tied to other bureaucratic structures such as the World Health Organisation's *International Classification of Diseases* and health insurance companies. The need to maintain acceptability with these external organisations sets complex constraints on the ways in which the *DSM* categories can be revised. Hence, despite Kupfer *et al.*'s (2002) vision of *DSM*-5 marking a paradigm shift in diagnostic classification, its actual publication revealed its categories to have undergone very few changes from those of *DSM*-IV.

In this subsection, I have presented some of the challenges to classificatory revision in psychiatry. While in principle an aetiologically-based classification system whose diagnostic categories correspond to distinctive causal processes would have significant epistemic benefits, there are substantial empirical, conceptual, and bureaucratic issues that make the development and implementation of such a classification system very difficult. This is by no means saying that such a classification system can never be achieved or that the efforts towards developing one are futile. To the contrary, I argue that the knowledge gained from such research efforts is of tremendous benefit for clinical practice and progress in psychiatric science. Nonetheless, as it stands, these challenges do suggest a successful aetiologically-based classification system for diagnoses in psychiatry is unlikely to come into use in the immediate future. In the following section, I look at how the causal explanatory practices of psychiatrists can be improved in clinical practice while working with the current *DSM*-5 diagnoses. Moreover, I show how the knowledge gained from the above mentioned research efforts towards an aetiologically-based classification can still be put into use without having to abandon the current diagnostic categories.

7.4 Psychiatric formulations

7.4.1 Working with the current diagnostic categories

In the present absence of an aetiologically-based alternative to *DSM*-5, there would be costs associated with giving up the current diagnostic categories in clinical psychiatry. As argued in Chapter 6, psychiatric diagnoses may not meet the explanatory ideal in medicine of the diagnosis that indicates a specific cause of the patient's symptoms, but some can still provide other sorts of causal explanatory information that enable them to serve clinically useful functions. These include conveying that the patient's symptoms are not caused by a general medical condition, signalling the possible causal factors that could be relevant to the patient's case, and providing some indication of why the symptoms

cluster together as they do. Such explanatory information can in some cases provide support for predictions and interventions. Hence, I argue that such diagnoses are not, as Jeffrey Poland (2014: p. 34) suggests, "free riders" that add little or nothing further to the knowledge of the individual symptoms.

Another useful function is denotation. Even in cases where the diagnoses do not provide useful causal explanatory information, such as with some of the personality disorders, they can still function as shorthand descriptive labels for collections of symptoms, which can be useful for facilitating communicative exchanges between clinicians. The diagnoses in *DSM*-5 offer standardised definitions of psychiatric disorders, thus providing a common language with which clinicians can communicate information about patients' problems (Tsou, 2015). Moreover, the fact that certain symptoms cluster with others in statistically significant ways means that the diagnostic labels do not just correspond to arbitrary sets of symptoms. Some combinations of symptoms occur with greater frequencies than others, and so it is useful to have shorthand labels for those more frequent combinations. Such labels can be predictively valuable. For instance, a diagnosis of paranoid personality disorder quickly and concisely tells the clinician that the patient exhibits suspiciousness, sensitivity to criticism, and a tendency to bear grudges, knowledge which is useful to the clinician for planning his or her approach towards the patient.

It is also worth acknowledging the useful social functions served by psychiatric diagnoses, including the mobilisation of therapeutic resources and authorisation of financial support. As noted in §7.3.3, the *DSM* has become tied to bureaucratic structures, such as health insurance companies and other funding bodies. For instance, in the United States of America, a diagnosis is required for a patient to be allocated state-funded care. Of course, we saw in Chapter 2 that whether or not these social functions are justified in light of the epistemic concerns about psychiatric diagnoses is a matter of

contention (Ingleby, 1982; Moncrieff, 2010), although I suggest that the other sorts of explanatory information supplied by psychiatric diagnoses could provide at least some justificatory support for some of these social responses. For example, negative causal information that excludes general medical causes supports the allocation of the patient to a mental health service team rather than to a general medical team, while probabilistic information about the possible psychological and social processes that could be involved in the patient's condition provides some support for the mobilisation of certain psychotherapeutic and supportive resources. Nonetheless, regardless of the concerns about whether these responses are epistemically justified, the very fact that certain institutional arrangements are dependent on and tied up with the *DSM* categories highlights further costs to giving up our current diagnostic practices in psychiatry.

And so, while most of the current psychiatric diagnoses do not meet the explanatory ideal in medicine, they nonetheless still serve some useful epistemic and instrumental functions. Given that we currently do not have a successful alternative classification system, it is worthwhile exploring how we might improve the causal explanatory practices in clinical psychiatry in a way that continues to utilise the current diagnostic categories. In what is to follow, I look at how this can be achieved by complementing categorical diagnoses with individualised formulations.

7.4.2 Individualised formulation

While a diagnosis assigns a patient's condition to a generalised category, a formulation is supposed to be an individualised account that pertains to the idiosyncratic circumstances and problems of the particular patient. Over the past two decades, practitioners have emphasised the importance of the individualised formulation in psychiatry (Weerasekera, 1996; Aveline, 1999; Mace and Binyon, 2005; MacNeil *et al.*, 2012). This is motivated by the recognition that the *DSM* categorical diagnoses on their own miss out much

information that is important for explaining patients' symptoms, planning therapy, and understanding the individual circumstances of patients. For example, Mark Aveline (1999: p. 200) argues that formulations can provide clearer guides to aetiology, prognosis, and treatment than diagnoses on their own, particularly for patients with personality disorders. Similarly, MacNeil *et al.* (2012) argue that a categorical diagnosis tells us little about the actual causation of the symptoms in a particular case, while an individualised formulation can yield more precise understanding of which causal factors pertain to the patient in question. In their discussion of the individualised formulation in psychodynamic psychotherapy, Mace and Binyon (2005: p. 418) argue that personalised information about the patient's defence style can sometimes predict prognosis better than a categorical diagnosis.

Although the above authors argue that individualised formulations can supply more clinical information about patients than categorical diagnoses, they do not suggest abandoning categorical diagnoses in favour of individualised formulations. Rather, they propose that categorical diagnoses and individualised formulations are complementary, and that both have useful clinical roles. This is consistent with the recommendations in *DSM*-5, which states that the "ultimate goal of a clinical case formulation is to use the available contextual and diagnostic information in developing a comprehensive treatment plan that is informed by the individual's cultural and social context" (American Psychiatric Association, 2013: p. 19).

The idea, then, is that a categorical diagnosis and an individualised formulation both feature in the assessment of a patient. This is made explicit by the World Psychiatric Association's (2003) IGDA workgroup. As part of the Institutional Program on Psychiatry for the Person, the IGDA workgroup advocate a comprehensive model of assessment that involves "the articulation of two diagnostic levels", the first being "a standardised multi-axial diagnostic formulation, which describes the patient's illness and

clinical condition through standardised typologies and scales", and the second being "an idiographic diagnostic formulation with a personalised and flexible statement" (World Psychiatric Association, 2003: p. S55).

The first level, the "standardised multi-axial diagnostic formulation", is supposed to include a categorical psychiatric diagnosis, numerical ratings for disabilities in four domains, codes for contextual factors, and a standardised score for quality of life. The IGDA workgroup recommend the use of *ICD*-10 for the categorical diagnosis and the contextual factors (World Psychiatric Association, 2003: pp. S52–S53). Interestingly, *DSM*-IV (1994) also recommended diagnostic assessment in multiple domains or axes, but this was removed in *DSM*-5 (2013) in favour of categorical diagnosis in a single axis. The second level, the "idiographic diagnostic formulation", is supposed to reflect the integrated perspectives of the clinician, the patient, and the patient's family. It is supposed to include information about the causal factors relevant to the patient's condition, the positive factors of the patient, and the expectations on the restoral of health. The information about the causal factors includes "the biological (e.g. genetic, molecular, toxic), psychological (e.g. psychodynamic, behavioural, cognitive) and social (e.g. support, cultural) factors that are relevant to that condition" (World Psychiatric Association, 2003: p. S55). Positive factors might include personality traits, skills, social resources, personal aspirations, and spiritual beliefs. Expectations on the restoral of health include specific expectations about the types of treatment and aspirations about health status.

Other authors vary somewhat in their recommendations for what should be included in the individualised formulation. These differences partly reflect their different theoretical orientations. Mark Aveline (1999) emphasises the role of the formulation in explorative psychotherapy, and so suggests that it includes information about intrapsychic conflict, the effects of the problem on others, coping strategies, biological vulnerabilities,

social stressors, motivation to change, and so on. By contrast, Priyanthy Weerasekera (1996) and MacNeil *et al.* (2010) present the formulation as a more general tool that serves a broad range of functions in the assessment and management of a patient, including identifying aetiological factors, understanding key difficulties, guiding therapeutic interventions, and anticipating challenges. They suggest the inclusion of the "five Ps", which are the presenting problem, predisposing factors, precipitating factors, perpetuating factors, and protective factors. Despite their differences, authors share the general idea that the formulation is supposed to identify the factors that contribute to the development and maintenance of the symptoms in the particular patient, as well as the internal and external factors that are likely to influence the outcome.

7.4.3 Idiographic understanding

Advocates of the individualised formulation in psychiatry often emphasise one of its roles as providing understanding of the patient's condition in the context of his or her individual perspective, values, and experiences. The IGDA states that one function of the formulation is to offer a "thorough, contextualised and interactive understanding of a clinical condition and of the wholeness of the person who presents for evaluation and care" (World Psychiatric Association, 2003: p. S55). Similarly, Mace and Binyon suggest that the formulation "is concerned with why events have followed one another and the meaning of these for the patient" (Mace and Binyon, 2005: p. 417). Such emphasis on the contextualised understanding of the patient's condition recalls Karl Jaspers ([1913] 1997), who taught that explanation in psychiatry should be complemented by understanding. The former accounts for a psychic state in terms of causes, while the latter is the empathic insight into the meaningful connections between thoughts.

Tim Thornton (2010) is a contemporary philosopher who interprets the aim of the individualised formulation as the provision of meaningful understanding over and above

the sort of explanation provided by the criteriological diagnosis. In his analysis of the World Psychiatric Association's comprehensive model, he compares the IGDA's "standardised multi-axial diagnostic formulation" and "idiographic diagnostic formulation" respectively to Wilhelm Windelband's ([1894] 1980) nomothetic and idiographic approaches to understanding. A nomothetic approach applies laws that are generalisable to several instances, thus emphasising the similarities between instances of a particular kind. By contrast, an idiographic approach investigates the contingent and unique characteristics of an individual case, thus emphasising the properties that set this case apart from others. Thornton suggests that a successful idiographic approach in psychiatry is one that provides a judgement about the patient that is "epistemically independent of all other judgements" (Thornton, 2010: p. 255). In other words, it is to provide a judgment that does not involve comparison with other cases.

However, Thornton argues that the sort of idiographic judgement outlined in the IGDA's publication ultimately falls short of being epistemically individualised in this way. He appeals to what Wilfrid Sellars ([1956] 1997) calls the "myth of the given". This characterises a form of foundationalism, according to which perception can supply non-inferential knowledge of a fact and this non-inferential knowledge presupposes no other knowledge of further facts. Sellars accepts the former claim, but rejects the latter on the grounds that the knowledge expressed in a perceptual report depends on one's overall worldview, which includes knowledge that the perceptual report is reliable and the knowledge that a specific type of perceptual report corresponds to a specific type of state of affairs (Thornton, 2010: p. 256). And so, Thornton argues that the information that is supposed to be included in the IGDA's idiographic formulation, such as information about biographical contingencies, does not constitute epistemically individualised understanding because any judgement concerning these factors is dependent on

knowledge of how these factors normally relate to patients more generally. Rather, the information just provides a further, albeit more detailed, generalisation.

Instead, Thornton proposes that epistemically individualised understanding can be provided by a narrative account of the patient's condition. This is also suggested by the psychiatrists Juan Mezzich (2005) and James Phillips (2005) in their discussions of the IGDA. Such a narrative account, according to Thornton, answers "to a different kind of internal logic to non-normative nomological accounts" (Thornton, 2010: p. 259). Drawing on the vocabulary of Wilfrid Sellars ([1956] 1997) and John McDowell (1994), he states that the internal logic of a narrative account belongs to the "space of reasons", whereas that of a nomothetic account belongs to the "realm of law". In other words, whereas a nomothetic account is couched in terms of law-like generalisations and causal connections, a narrative account is couched in terms of rational connections between propositional attitudes. According to Thornton, such a narrative account allows the clinician to make a normative yet individualised account. It is normative because it relies on our norms of rationality, but it is individualised because it does not involve "subsuming symptoms under kinds which fit into law-like patterns of disease aetiology and prognosis" (Thornton, 2010: p. 259). Therefore, Thornton proposes that a narrative account can add something epistemically novel to the assessment of a patient over and above the explanation provided by a nomothetic account, insofar as the former provides information about reasons and the latter provides information about causes.

I agree that complementing a categorical diagnosis with an individualised formulation can provide meaningful understanding over and above the sort of explanation provided by the categorical diagnosis. However, I argue that Thornton is too quick to dismiss the sort of non-narrative idiographic judgement that is suggested in the IGDA's model. While it may not constitute epistemically individualised understanding, I propose that it nonetheless provides important information that is not captured by the

categorical diagnosis on its own. More specifically, it complements the diagnosis by enabling a more precise causal explanation of the patient's symptoms that is informed by the findings from empirical research into psychiatric aetiology.

7.4.4 Individualised causal explanation

The idea that an individualised formulation can provide relevant causal explanatory information about the patient's clinical presentation is often mentioned in the literature on psychiatric formulation. As noted in §7.4.2, the IGDA proposes that the formulation should include "the biological (e.g. genetic, molecular, toxic), psychological (e.g. psychodynamic, behavioural, cognitive) and social (e.g. support, cultural) factors that are relevant to that condition" (World Psychiatric Association, 2003: p. S55). MacNeil *et al.* propose that one function of the formulation is "understanding significant etiological factors that have influenced the person's presentation", including "possible biological contributors (for example, organic brain injury and birth difficulties), genetic vulnerabilities (including family history of mental health difficulties), environmental factors (such as socio-economic status, trauma, or attachment history) and psychological or personality factors (including core beliefs or personality factors)" (MacNeil *et al.*, 2012: p. 2). The Royal College of Psychiatrists, in their curriculum for specialist training in psychiatry, state that psychiatrists constructing a formulation should describe "the various biological, psychological and social factors involved in the predisposition to, the onset of and the maintenance of common psychiatric disorders" (Royal College of Psychiatrists, 2013: p. 44).

In addition to providing meaningful understanding of the patient's condition, then, the information in an individualised formulation can complement the categorical diagnosis by more precisely specifying the relevant causal factors that contribute to the particular patient's presentation. I suggest that this process can utilise the findings of the

research efforts into the causes of psychiatric syndromes mentioned in §7.3.2, while still involving the *DSM*-5 or *ICD*-10 categorical diagnoses. The idea is as follows. A categorical diagnosis in psychiatry is highly causally heterogeneous, and so only provides imprecise causal explanatory information regarding an individual patient's symptoms. Nonetheless, empirical research has revealed a range of biological, psychological, and social factors that can causally contribute to the symptoms of the disorder in general, at the population level. When assessing a patient with such symptoms, the clinician can use this theoretical knowledge of the causal factors associated with the diagnosis in general to guide and focus further enquiry, in order to specify the causal factors that are actually instantiated by the particular case. The result is a formulation that provides an individualised causal explanation of the patient's symptoms.

Although he does not link it to psychiatric formulation, Murphy (2006) sketches the outline of a similar approach. He writes:

> A clinician can treat the causal pathways of interest as schema to be filled in with the specific details of interest in the particular case. Each path through the model can be realized by numerous different causal histories, so the clinician can use the pathway to a given symptom as a way to look for and organize the relevant details of the patient's life. (Murphy, 2006: p. 369)

It must be conceded that with the current state of our scientific knowledge in psychiatry, we do not yet have knowledge of specific causal pathways for many disorders, and so the suggestion as put by Murphy may seem overly optimistic in the present day. However, we do, as mentioned above, have knowledge of several of the biological, psychological, and social factors that contribute to these disorders, even if we do not know how these factors interact within specific pathways. In view of this, I suggest that it is more

reasonable to expect a psychiatric formulation to provide a looser sort of causal explanation of a patient's symptoms, whereby the relevant causal factors and known mechanisms are specified, but not necessarily unified into a coherent pathway. Nonetheless, such explanatory information would still be an improvement over what is provided by the categorical diagnosis on its own, because it specifies the causal factors that pertain to the patient rather than suggesting causal possibilities.

It might be supposed that if a formulation does indeed provide a better explanation than a diagnosis, then this would make the diagnosis redundant (Johnstone, 2006: p. 275). However, I argue that this is not the case. First, the diagnosis can still serve the communicative and administrative functions described in §7.4.1. Second, whatever limited explanatory information is supplied by a diagnosis can be instrumental in the process of generating the formulation. In virtue of our theoretical knowledge of the various possible causal factors that can contribute to a syndrome, a categorical diagnosis functions as a heuristic tool that informs the selection of what information is relevant to include in the formulation. That is to say, the causal search does not stop at the diagnosis, but is guided by it. Therefore, while the individualised formulation potentially has more explanatory value than the categorical diagnosis, this does not make the latter superfluous. By signalling the causal possibilities that could be associated with the patient's condition, the categorical diagnosis frames the formulation process so that each patient does not need to be treated as a "first instance" (Johnstone, 2006: p.277).

It is also worth noting the information gathered in the formulation can also influence the diagnosis. Consider a patient presenting with symptoms that initially suggest a provisional diagnosis of major depressive disorder. If a more detailed formulation reveals causal factors and biographical details that are normally more closely associated with a different syndrome, then this might provide grounds to rethink the diagnosis. For example, if the formulation reveals that the patient has a strong family history of bipolar

disorder, a history of poor responses to antidepressants, and previous episodes marked by prominent psychomotor changes, then this might warrant the consideration of bipolar disorder, even in the absence of symptoms that meet the *DSM* threshold for hypomania. Therefore, it is helpful to think of the relation between the diagnosis and the formulation being bidirectional.

7.4.5 An example formulation

To illustrate how a diagnosis and a formulation can work together in practice to causally explain a patient's symptoms, let us consider an example formulation by Priyanthy Weerasekera (2009). The case concerns Antoinette, a young woman diagnosed with anorexia nervosa. The history includes a history of an eating disorder in her mother, a history of weight problems in her father, the divorce of her parents when she was a child, perfectionistic personality traits, and high achievement at school. The formulation includes the following:

> From a biological perspective, she is vulnerable to anorexia given that her mother has suffered from this condition. In addition, her father struggles with weight issues indicating further biological vulnerabilities towards weight instability. … The client's early developmental years indicate a mother who wished her daughter to follow in her footsteps: to be thin and to be a ballerina. Antoinette, however, was unable to pursue her mother's dream, for she was seen has having a muscular physique … It is possible that the rejection of her body type set the stage for her obsessional preoccupations with thinness, perfectionism, ritualistic eating and exercising. … Her father and brother leaving shattered her family as she knew it, leading her to feel even more helpless about her ability to control the world around her. Her ritualistic eating behaviour and struggle to be perfect may be seen as her

> attempt to exert some control over a life which is perceived as unpredictable and chaotic. (Weeraseksera, 2009: pp. 149–150)

The formulation then goes on to discuss the factors maintaining Antoinette's condition, positive and negative prognostic factors, and a treatment plan that is informed by consideration of these factors.

I bring attention to three notable features of Weerasekera's formulation. First, in keeping with the suggestions made in §7.4.4, the causal information that is selected for inclusion in the formulation is partly guided by the categorical diagnosis of anorexia nervosa and the theoretical knowledge of the possible causal factors that can be associated with the condition. Weerasekera explicitly highlights the role of this theoretical knowledge in the generation of the formulation:

> The variables chosen for inclusion were those that related to the predisposing, precipitating, perpetuating and protective factors, and the coping-response style since this this framework is used in multiperspective case formulation. … Research tells us what has been empirically investigated and what is the most plausible hypothesis or explanation of the patient's current difficulties. … Prospective research particularly informs us about variables that may be important in the development of a condition. (Weerasekera, 2009: pp. 146–147)

As with most psychiatric diagnoses, anorexia nervosa is a causally heterogeneous category. Nonetheless, empirical research has implicated a number of causal contributory factors associated with the disorder at the population level, including genetic factors, perfectionistic personality traits, disturbed family relationships with enmeshment and rigid parenting, and cultural expectations regarding body image (Cowen *et al.*, 2012: pp.

353–355). Weerasekera's above formulation applies this theoretical knowledge of anorexia nervosa to identify, organise, and elaborate on the causal factors that are actually relevant to Antoinette's particular case. The result is an individualised causal explanation of Antoinette's symptoms.

Second, Weerasekera's formulation integrates multiple theoretical perspectives. As noted in §7.3.2, most psychiatric disorders involve the complex interactions of multiple causal variables that belong to different levels of organisation. Accordingly, there has been a call for psychiatry to engage theoretical pluralism in the conceptualisation of such disorders (Mitchell, 2008; Kendler, 2012; Tsou, 2015). In her formulation, Weerasekera uses such a pluralistic approach to explain Antoinette's symptoms. By bringing together information about Antoinette's genetic vulnerabilities, cognitive schema, coping style, and family environment, the formulation integrates biological, cognitive, psychodynamic, and social theoretical perspectives.

This is not to say that all formulations are theoretically pluralistic. While Weerasekera explicitly endorses a multiperspective approach to formulation, other formulations can lean towards particular theoretical orientations. For example, Aveline (1999), and Mace and Binyon (2005) emphasise the role of the formulation in psychodynamic therapy, and so focus more on the psychodynamic factors that are relevant to patients' conditions. Hence, one's choice of theoretical orientation and selection of the causal factors to include in the formulation depends on one's therapeutic interests. Nonetheless, Weerasekera's formulation of Antoinette's case demonstrates that individualised psychiatric formulations can accommodate the theoretical pluralism that has been encouraged in the research into psychiatric aetiology.

Third, the formulation contains narrative strands. The passages about Antoinette's struggle to be perfect and attempts to exert control over a life perceived as chaotic are very much couched in terms of meaningful connections between propositional attitudes.

Her ritualistic eating behaviour is interpreted in light of her experiences, values, and desires. Therefore, in addition to providing causal explanatory information, the formulation can also provide the sort of individualised understanding of reasons that Thornton (2010) endorses.

7.4.6 Critical discussion

To briefly recapitulate, I have proposed that a categorical diagnosis and an individualised formulation can have complementary roles in the causal explanation of a patient's symptoms in psychiatry. In virtue of the theoretical knowledge of the causal possibilities that are associated with a given syndrome, a categorical diagnosis narrows down the sorts of information to be considered in the formulation. The formulation particularises this information to the individual case, thus providing an account of the causal factors that are actually instantiated by the patient. Hence, the complementing a diagnosis with a formulation can overcome the problem of causal heterogeneity that affects the former.

I should stress that I am in no way suggesting that causal explanation is the sole purpose of the psychiatric formulation. As noted in §7.4.2, MacNeil *et al.* (2012) emphasise that the formulation is a versatile tool that can serve a broad range of functions in the clinic. One such function, discussed in §7.4.3, is to provide reason-based understanding of the patient's condition in light of his or her experiences, values, and desires. Other important functions include identifying the key difficulties the patient is facing, guiding therapeutic interventions, anticipating challenges, and predicting prognosis. My claims are rather that causal explanation of the patient's presentation is one of the useful functions served by an individualised formulation and that a categorical diagnosis can assist in this process.

It should also be noted that the above described process whereby the categorical diagnosis narrows down the causal information to be considered in the individualised

formulation may not be applicable to those psychiatric diagnoses that do not give any useful causal information, such as paranoid personality disorder, schizoid personality disorder, histrionic personality disorder, and avoidant personality disorder. Because these diagnostic categories are not associated with stable causal factors, they do not contribute to clinical assessments in the ways described above. However, I argue that the lack of useful causal information supplied by the categorical diagnoses makes individualised formulations even more crucial for these conditions. As I noted in Chapter 5, §5.4.4, while there may not be causal factors that generalise across cases, it is still possible, via narrative exploration, to uncover factors and processes contributing to the development and maintenance of the patient's problems in a particular case. This is a line of thought endorsed by Aveline (1999: p. 200), who argues that the formulation is a better guide to aetiology, prognosis, and treatment than the diagnosis in the case of personality disorder, because it explores various factors and the links between them in the context of the patient's particular developmental history. Moreover, although the categorical diagnosis may not contribute causal information in such a case, I argue that it can still contribute descriptive information that is useful for the process of constructing the formulation. For example, as mentioned in §7.4.1, a personality disorder diagnosis informs the clinician that the patient is likely to display certain sorts of interactive behaviours, which can be useful to know for the purposes of planning the style of consultation.

Practising psychiatrists may find the above unsurprising on the basis that it describes what they have already been doing all along. Nonetheless, I argue that my discussion makes some important philosophical contributions. First, it offers support to a deflationary approach to the classification problem in psychiatry. As I discussed in §7.3, there have been calls to revise diagnostic classification in psychiatry, with particular emphasis on moving towards an aetiologically-based classification system. While this may be an epistemically respectable endeavour, there are significant challenges that make a

successful aetiologically-based classification unlikely in the near future. However, in his discussion of the classification problem, Derek Bolton (2012) argues that this should not be too much of a worry, because classification is not the main point of psychiatric science. Rather, he suggests that the main aims of science are "*prediction*, refined by *causal explanatory models*, and, on that basis, if cause-effect relationships are sufficiently strong, making *technological applications* including interventions" (Bolton, 2012: pp. 6–7). Psychiatric science, according to Bolton, has come to over-value classification as a goal in itself, when it should be seen as an instrument whose purpose is to help us achieve the more important goals of prediction, causal explanation, and intervention.

My analysis of the psychiatric formulation supports the sort of deflationary approach offered by Bolton, because it shows that psychiatry has other resources, aside from diagnostic categories, that contribute to predictions, causal explanations, and interventions in clinical practice. Due to its causal heterogeneity, a psychiatric diagnosis on its own may not meet the explanatory ideal in medicine of the diagnosis that specifies the causative pathology and mechanisms responsible for the patient's symptoms, but the psychiatrist has another resource, namely the formulation, which can complement the diagnosis to support stronger predictions, inform targeted interventions, and supply a more precise causal explanation that is particularised to the patient's individual case. Hence, in virtue of its complementary relation with the formulation, a psychiatric diagnosis can still have an important role in the production of such a causal explanation, even if the diagnostic categories in the current classification system do not reflect distinct and homogeneous kinds.

Second, my analysis provides a clarification of the epistemic relations between scientific research, categorical diagnosis, individualised formulation, theoretical pluralism, and causal explanation in psychiatry. Research into psychiatric aetiology yields empirical knowledge of the various possible causal factors that can be associated with a given

diagnosis. In virtue of this empirical knowledge, the diagnosis informs what aspects of the patient's history are relevant causal factors to include in the formulation. Given that these causal factors may belong to different explanatory levels, the formulation may integrate multiple theoretical perspectives. The result is a formulation that can provide, among other things, an individualised yet evidence-based causal explanation of the patient's symptoms.

This individualised causal explanation can, in theory, enable more targeted treatment, as it identifies the relevant factors contributing to the syndrome that could potentially be modified by therapeutic interventions. In the formulation of Antoinette's case presented in §7.4.5, perfectionistic attitudes, perceived loss of control, parental issues, biological vulnerability to weight instability, and rejection of her body type are identified as relevant causal factors in the development of her symptoms. Accordingly, the treatment plan proposes cognitive-behavioural therapy, experiential techniques to facilitate the expression of affect, family therapy, a selective serotonin reputable inhibitor to stabilise eating behaviour, and ongoing social activities to reinforce confidence as therapeutic interventions to target these respective causal factors (Weerasekera 2009: pp. 153–154).

There is also the suggestion that the individualised causal explanation provided by a formulation may be more epistemically satisfactory to the patient than a categorical diagnosis on its own, although it must be conceded that there is very little qualitative data to determine whether this is the case in practice. One study, by Chadwick *et al.* (2003), investigated the perspectives of patients diagnosed with psychotic disorders. The majority of participants reported increased hope and understanding after their formulations, which could be interpreted as supporting the claim that individualised explanations can be of some intrinsic value to patients. However, some participants also described finding the

information upsetting and worrying, which suggests that patients' affective and cognitive responses to receiving explanations of their conditions are complex.

While a psychiatric formulation can in principle provide a resource for individualised and evidence-based causal explanation, I do not mean to present it as a panacea. Given the current status of scientific knowledge in psychiatry, the full epistemic value of the psychiatric formulation remains tentative. In other words, given that a formulation can only draw on the best available scientific evidence, our incomplete scientific understanding of a certain disorder poses a limit on the causal explanatory strength of the formulation. As noted in §7.4.4, while we have substantial empirical knowledge of the various causal factors and mechanisms that occur in many disorders, we do not yet have precise understanding of the complex interactions between these variables. Therefore, for such disorders, formulations might only be able to provide loose causal explanations which specify the relevant causal factors and mechanisms without unifying them into complete causal pathways. The hope is that as the causal explanatory strengths of psychiatric formulations will improve as the empirical research into psychiatric aetiology and mechanisms progresses.

A related challenge, which I briefly raised in Chapter 6, §6.3.2, is that even if empirical research has revealed various causal factors that can be associated with a diagnostic category, we currently lack biomarker tests that enable us to match these specific causal factors to individual patients (Bolton, 2012: p. 10). Of course, it is possible to identify some causal factors through the clinical interview, mental state examination, and collateral history, particularly the psychological and social factors. For instance, it is possible to establish the presence of a causally relevant adverse social context by simply asking the patient. It may even be possible to infer a heritable component to the disorder by asking about family history. However, for many biological factors, such as specific genetic variants, neurochemistry, and neural circuitry, tests are not readily available for

use in the clinic. For example, a number of peripheral biomarkers for inflammation and oxidative stress in major depressive disorder have been explored, but these are neither sufficiently sensitive nor specific to be used in isolation (Lopresti *et al.*, 2014). Recent functional neuroimaging data has suggested some potential avenues for biomarker research, such as the review by Roiser *et al.* (2012), mentioned earlier in Chapter 5, §5.2.4, which found differential responses to pharmacological treatment and psychological therapy for depressed patients with and without abnormal anterior cingulate cortex activity. However, such tests are usually reserved for research, rather than clinical, purposes. With respect to biomarker tests that could be readily used in clinical practice and not just in the research laboratory, psychiatry has fallen short of other medical specialties. Therefore, an individualised formulation may attain more precise causal information than the categorical diagnosis on its own with respect to certain psychological, social, and heritable factors that pertain to the patient's case, but the absence of readily available biomarker tests makes us unable to attain greater precision with respect to many of the neurobiological factors.

A final challenge, raised by Lucy Johnstone (2006), is that while for a given disorder there may be a considerable evidence base regarding causal factors from which the psychiatric formulation can draw, there is currently little empirical evidence for the therapeutic effectiveness of the psychiatric formulation as a specific intervention. As mentioned earlier, in theory, a psychiatric formulation could draw on scientific knowledge about causal factors and mechanisms to inform targeted interventions. However, whether this actually makes a difference to the therapeutic outcome in practice is not clearly established. Studies investigating therapeutic effectiveness have yielded equivocal results. For example, Schulte *et al.* (1992) compared two groups of people diagnosed with phobias, with the first group receiving standardised behavioural therapy and the second group receiving tailored therapy based on individualised formulations. In this study, it

was the standardised treatment group that showed more improvement. Emmelkamp *et al.* (1994) performed a similar study comparing two groups of people diagnosed with obsessive-compulsive disorder, with the first group receiving standardised exposure therapy and the second group receiving tailored cognitive-behavioural therapy based on individualised formulations. Both groups showed improvements, but no significant differences were found between their respective outcomes.

In their review of these studies, Tarrier and Calam (2002) argue that the results have poor generalisability because the sample sizes were too small to demonstrate statistically significant differences in the effect sizes. The study by Emmelkamp *et al.* used a sample size of 22, but Tarrier and Calam estimate that the sample sizes that would be needed to show significant differences with 80% power and 0.05 significance level would be 25 for the Rational Behaviour Inventory, 560 for the Symptom Check List-90-Revised, 800 for the Self-Rating Depression Scale and Inventory of Interpersonal Symptoms, 4,000 for the Maudsley Obsessional-Compulsive Inventory, and over 15,000 for measures of anxiety and discomfort (Tarrier and Calam, 2002: p. 316). Therefore, studies on the therapeutic effectiveness of the psychiatric formulation would require much larger sample sizes than the extant studies to yield statistically significant results.

There is also a gap in the literature on therapeutic effectiveness with respect to the range of conditions and purposes for which formulations are constructed. The studies mentioned above look at formulations constructed for specific psychotherapeutic purposes in patients with phobias and obsessive-compulsive disorder. However, studies on the therapeutic effectiveness of the psychiatric formulation as a more general assessment tool for a broader range of psychiatric disorders are lacking at the time of writing.

And so, as well as the multidisciplinary research into the causes and mechanisms associated with psychiatric disorders, the scientific respectability of the psychiatric

formulation would also depend on support from additional clinical studies investigating its practical utility. Outcomes to investigate might include prognostic power, patient acceptability, and therapeutic effectiveness compared to the categorical diagnosis on its own. However, given the very large sample sizes that would be required to show significant results, such research is likely to involve serious methodological and logistical challenges. A possible solution might be to utilise practice research networks consisting of clinicians who collaborate to gather data from actual practice rather than from orchestrated clinical trials (Zarin *et al.*, 1996; Margison *et al.*, 2000; Audin *et al.*, 2001). This inclusion of practice-based evidence could potentially provide datasets large enough to achieve statistically significant results, which in turn could reciprocally inform evidence-based practice.

7.5 Conclusion

To sum up, the heterogeneous causal profiles of psychiatric diagnoses suggest the need for more caution regarding the ways in which psychiatric diagnoses are communicated in clinical discourse. More specifically, portrayals of the diagnoses should explicitly acknowledge the variable and multifactorial natures of their causal pathways, and care must be taken not to misleadingly essentialise the disorders. The issue of heterogeneity has also led to claims that the current diagnostic categories are unsatisfactory and has inspired calls to replace them with an aetiologically-based classification system. While this may be an epistemically respectable endeavour, there are serious conceptual, empirical, and bureaucratic challenges that make the development and implementation of a successful aetiologically-based classification system unlikely in the near future.

Nonetheless, this should not cause too much worry, because there are other resources in clinical psychiatry apart from categorical diagnoses that serve useful epistemic functions. One such resource is the individualised psychiatric formulation. In

addition to providing meaningful understanding of the patient's condition, the formulation can draw on empirical knowledge concerning psychiatric aetiology to provide an individualised causal explanation of the patient's symptoms. Furthermore, in virtue of the theoretical knowledge of the causal possibilities that can be associated with a given syndrome, the categorical diagnosis complements this process by narrowing down the sorts of information to consider in the formulation. Therefore, despite their high degrees of causal heterogeneity, categorical diagnoses can continue to have important roles in the generation of causal explanations of patients' symptoms. However, while the individualised causal explanations provided by formulations could in theory inform more targeted therapeutic interventions, there remain significant empirical challenges. First, more research is required to better understand the various causal pathways of disorders and to identify biomarkers that could help us match these pathways to individual patients. Second, more evidence is required to assess the therapeutic effectiveness of the individualised formulation in practice.

8. Conclusion

We have reached a good point to sum up the main points of this thesis. I began my investigation with some observations on the roles that diagnoses have in clinical practice and how they are used in medical discourse. In bodily medicine, diagnoses often, though by no means always, serve explanatory functions. That is to say, when a patient presents to the clinic with a set of symptoms, the physician infers a diagnosis to explain why he or she has these symptoms. I have argued that this normally constitutes a causal explanation, whereby the diagnosis indicates the causative process responsible for the patient's symptoms and this relation is understood within a broader theoretical framework of mechanisms. I have also argued that this causal explanatory function is important, because it informs therapeutic interventions, supports predictions about prognosis, and conveys understanding to the patient. In psychiatry, diagnoses are sometimes presented in clinical texts and discourse as if they also serve as such causal explanations of patients' symptoms. This is unsurprising, given that psychiatric practice occurs in a context shaped by medical roles and traditions. However, there are serious conceptual and ontological problems that cast doubt on whether psychiatric diagnoses actually do serve these causal explanatory functions. Over the course of the thesis, I have addressed these problems in order to arrive at a clearer understanding of the causal explanatory roles of diagnoses in clinical psychiatry.

The conceptual problem is that according to formal diagnostic manuals, psychiatric diagnoses are defined by their symptoms. This suggests that invoking a psychiatric diagnosis as an explanation of a patient's symptoms, when it is merely a descriptive label for them, amounts to problematic circularity. To address this problem, I explored how diagnostic terms secure their meanings according to theories of reference in the philosophy of language, and showed how both kinds of talk regarding psychiatric

diagnoses, namely diagnoses *qua* descriptions of symptoms versus diagnoses *qua* causes of symptoms, are accommodated by a conceptual framework based on two-dimensional semantics, which integrates descriptive and causal considerations. According to this two-dimensional semantic framework, a diagnostic term has a complex semantic value involving both a primary intension and a secondary intension, which respectively correspond to the descriptive and causal conceptions of the diagnosis. These are not to be taken as exhausting the meaning of the term, but as capturing different aspects of the term's complex semantic value that have useful epistemic roles. This allows us to take seriously the descriptive definitions in diagnostic manuals as necessary criteria for the diagnoses, yet still talk about the diagnoses as referring to the causes of the symptoms that make up these definitions. Therefore, the fact that psychiatric diagnoses are defined by their symptoms does not necessarily preclude the possibility of appealing to them in causal explanations of these symptoms.

This brings us to the ontological problem. Even though my two-dimensional semantic framework accommodates the possibility of a diagnostic term being taken to denote the causal process responsible for a cluster of symptoms, the epistemic value of a diagnosis in the individual case also depends on whether the category *a posteriori* corresponds to a causal structure that is stable enough to be citable in a causal explanation. I have argued that one reason why a diagnosis in bodily medicine can function as a good causal explanation of a patient's symptoms in the individual case is because the category captures a distinctive kind of causal structure that is sufficiently invariant across cases in the appropriate respects. In virtue of this invariance, the disease type that is denoted by a medical diagnosis can be considered to represent with accuracy the actual causal structures instantiated by the particular patients with the diagnosis. However, this sort of invariance is not to be found in many of the current diagnostic categories in psychiatry. The findings from empirical research into psychiatric aetiology

indicate that, with the possible exceptions of the dementias and some disorders with stable psychological profiles, most psychiatric diagnoses are highly causally heterogeneous. That is to say, in different patients with the same diagnosis, the symptoms may be produced by different causal processes. Furthermore, these causal processes typically involve complex interactions between diverse variables at different levels of organisation, which suggests that there is no single privileged level of organisation at which a psychiatric diagnosis can be aetiologically defined and that theoretical pluralism is desirable in the causal conception of a psychiatric disorder. Hence, a typical diagnostic category in psychiatry is not associated with a distinctive causal structure that is invariant across cases, but a heterogeneous range of possible causal structures, each involving the complex interactions of biological, psychological, and social variables.

Due to their high degrees of causal heterogeneity, many psychiatric diagnoses do not meet the explanatory ideal in bodily medicine of the diagnosis that picks out a specific causative process responsible for the patient's symptoms. The dementias are exceptions, as the categories respectively correspond to reasonably distinctive and stable kinds of neuropathological mechanism that are causally responsible for the clinical presentations. I have also conceded that some diagnoses, such as panic disorder and obsessive-compulsive disorder, may be heterogeneous at biological levels, but are characterised by more stable causal regularities at psychological levels. Therefore, while these diagnoses may not specify what biological processes are involved, they could serve as psychological causal explanations of patients' symptoms. However, it must be acknowledged that these sorts of diagnosis are relatively rare in psychiatry and that many other psychiatric diagnoses, including major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety disorder, appear to be heterogeneous all the way down.

Still, I have argued that some of the complex and heterogeneous diagnoses in psychiatry still convey other sorts of causal information about patients' symptoms that

can be explanatorily relevant. First, a psychiatric diagnosis might convey negative causal information. It is recommended in clinical texts that symptoms count towards a psychiatric diagnosis only if other medical causes have been satisfactorily excluded, and so a psychiatric diagnosis contains an implicit supposition that the patient's symptoms are not caused by any of these other medical conditions. Second, a psychiatric diagnosis might convey some disjunctive information about the causal possibilities that may be relevant to the patient's symptoms. Although our current diagnostic categories in psychiatry have turned out not to respectively correspond to homogeneous causal structures, empirical research into psychiatric aetiology has yielded knowledge of various causal factors that can contribute to the development of the clinical syndromes. In virtue of this empirical knowledge, a psychiatric diagnosis can provide some disjunctive information about the possible causal processes that might be contributing to the patient's symptoms. However, we can only expect this information to be partial and imprecise, given the limited extent of our current empirical knowledge. Third, a psychiatric diagnosis might convey information about the causal relations that occur between the symptoms themselves. Recent theorists suggest that the symptoms of a psychiatric disorder constitute a network and reinforce each other via reciprocal causal relations. Hence, while it may not indicate a specific pathology underlying the symptoms, a psychiatric diagnosis can still, in virtue of the causal relations between the symptoms, provide some sort of explanation of why the patient's symptoms aggregate and persist as they do. I have suggested, then, that there are ways in which a psychiatric diagnosis can convey some causal explanatory information about a patient's symptoms, albeit information that falls short of the explanatory ideal of specifying a particular causal process that is responsible for the symptoms.

While the above considerations apply to many major psychiatric diagnoses, such as major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety

disorder, I have argued that there remain some psychiatric diagnoses that may not even serve these limited causal explanatory functions. I have suggested some of the personality disorders as examples. The problem is that the behavioural features associated with these diagnoses could possibly result from various combinations of highly contingent circumstances that we may not be able to locate causal regularities that generalise even modestly across cases. Hence, the diagnostic categories fail to supply explanatorily significant information about what factors are likely to have caused the patients' symptoms. Nonetheless, such diagnoses may still be clinically useful for descriptive purposes.

Finally, I looked at some strategies for achieving better causal explanations of patients' symptoms in psychiatric practice. In view of the fact that many of the current diagnostic categories in psychiatry do not correspond to distinctive and homogeneous kinds of pathological process, some theorists in the philosophy of psychiatry advocate revising the diagnostic classification system so that the categories do respectively reflect more stable causal structures. While a move to a classification system based on causes is an epistemically respectable endeavour, I have argued that there are significant conceptual, empirical, and bureaucratic challenges that make the development and implementation of such a classification system unlikely in the near future. Nonetheless, I have tried to deflate this concern by emphasising that clinical psychiatry has another resource, the individualised formulation, which can serve a complementary epistemic role to the categorical diagnosis. In addition to conveying meaningful understanding of the patient's predicament in terms of reasons, a formulation can utilise the knowledge gained from empirical research and supplement the categorical diagnosis to provide a more satisfactory causal explanation of the patient's symptoms that can inform therapeutic interventions in the individual case. In virtue of the knowledge acquired from empirical research of the causal possibilities that can be associated with a given diagnostic category,

the diagnosis indicates what sorts of causal information would be relevant to consider in the formulation. The formulation then specifies the causal factors that are actually instantiated in the particular case, the result being an individualised causal explanation of the patient's symptoms. This can also accommodate the theoretical pluralism that is considered desirable in the understanding of a psychiatric disorder. Therefore, despite its high causal heterogeneity, a categorical diagnosis in psychiatry can still make an important contribution to the development of a causal explanation of a patient's symptoms in a way that is clinically useful. However, I made some concessions. First, given that the formulation draws on our empirical knowledge of causes and mechanisms in psychiatry, and that this empirical knowledge is currently limited, we can only expect a formulation to provide a loose and incomplete sort of causal explanation that might specify only some of the relevant causal factors and mechanisms without unifying them into a complete causal pathway. Second, more clinical studies are also required to assess the therapeutic effectiveness of the formulation as a specific intervention in practice.

Through my philosophical investigation, I hope to have contributed some novel insights to our understanding of the epistemic roles and uses of diagnoses in medicine and psychiatry. More specifically, I would like to believe that this thesis has enhanced our understanding of the explanatory relations between diagnoses and symptoms, how the uses of diagnostic terms reflect their complex semantic values, and the implications of causal heterogeneity and complexity for the explanatory roles of psychiatric diagnoses. This thesis also has normative implications for clinical practice and research, such as how psychiatric diagnoses ought to be communicated in ways that do not amount to problematic essentialisation, how categorical diagnoses need to be complemented by individualised formulations in causal explanations of patients' symptoms, and how the problem of diagnostic classification must be approached with the recognition that classification is not the primary purpose of psychiatry. Parts of this thesis suggest some

degree of scepticism regarding certain aspects of psychiatric diagnoses. However, this is at most a modest scepticism limited to the idea of psychiatric diagnoses referring to stable disease types and the sorts of causal explanation they are sometimes portrayed as providing in clinical discourse. Nowhere do I dispute the distress experienced by patients with the diagnoses or the appropriateness of managing such distress in a health care setting.

As we draw to a finish, I would like to briefly reflect on the implications of my discussion for Thomas Szasz's critique of psychiatry in "The Myth of Mental Illness" (1960), which I cited at the beginning of this thesis as one of the motivations for my investigation, and on some potential areas for future philosophical research. First, Szasz argues that mental illness cannot be invoked as an explanation of certain behaviour, because it is just a shorthand label for this behaviour. Second, he argues that unlike bodily illness, mental illness is not defined by a pathophysiological lesion, but by the deviation from moral and social norms. Regarding the first argument, this can be dispelled by the two-dimensional semantic framework I put forward. A diagnostic term can have a complex semantic value, and so a descriptive definition based on a cluster of symptoms does not necessarily preclude it from referring to the causal profile underlying these symptoms. Regarding the second argument, the implications of my investigation are more nuanced. Indeed, if we take a pathophysiological lesion to mean, as Szasz does, a distinctive morphological abnormality, then it would seem that, with the exceptions of the dementias, he is correct that psychiatric diagnoses are not associated with pathophysiological lesions. However, I argue that this does not warrant his claim that they are therefore not genuine illnesses, but rather incentivises us to acknowledge that there may be other kinds of causal profile with which illnesses can be associated. While most psychiatric disorders are not constituted by distinctive morphological abnormalities, empirical evidence suggests that they are characterised by interactions between biological,

psychological, and social factors that are heterogeneous but still tend to aggregate in statistically significant ways.

Therefore, for many common psychiatric disorders, including major depressive disorder, schizophrenia, bipolar disorder, and generalised anxiety disorder, Szasz's arguments can be effectively countered. While major depressive disorder, for example, is defined descriptively through its symptoms, the diagnosis can still, under the two-dimensional semantic framework I have presented, refer to the causal profile responsible for its symptoms. Of course, we need to concede that the causal profile associated with the diagnostic category is comprised by a complex and heterogeneous array of diverse variables rather than a distinctive pathophysiological lesion, but I have shown that there are ways in which this can be explanatorily valuable. However, I accept that Szasz's arguments could still apply to those diagnoses that do not appear to be associated with even modestly generalisable causal factors, such as some of the personality disorders. Given these diagnoses do not convey anything significantly informative regarding what might be causing the behaviours of patients, it appears that they are little more than shorthand descriptive labels for these behaviours.

With respect to Szasz's proposal that the attribution of mental illness depends on consideration of moral and social norms, I have not written anything in this thesis which disputes this. As I conceded at the beginning of this thesis, my investigation focuses on just one aspect of diagnosis, that is, the diagnosis *qua* explanatory hypothesis about the patient's clinical presentation as examined through the lens of analytic philosophy of science. As such, I have not had the opportunity to explore in detail the other interesting and important philosophical aspects of diagnosis in psychiatry that are raised by Szasz's above proposal. These include the distinction between disorder and non-disorder, the roles of values in diagnosis, the historical and cultural dynamics that have shaped our diagnostic categories, and the performative roles of diagnoses in the social context. While

these issues run orthogonal to my analysis, there will nonetheless be areas where they meet. Such areas would provide rich grounds for future research. For example, it would be worthwhile to investigate whether the nature of the explanation provided by a diagnostic category has any bearing on whether the condition denoted by the category is considered a medical disorder or another kind of problem, such as a moral or a social problem, as this would be relevant to discussions about what conditions to include and exclude in future diagnostic classification systems. It would also be worthwhile to examine the roles that values have in the development of a diagnostic category, the assessment of symptoms in the diagnostic process, and the judgement about whether the patient warrants a diagnostic label, as these considerations have the potential to influence clinical practice. Finally, it would be worthwhile to further explore the relation between the explanatory status of a diagnosis and its use as a social device to legitimise certain activities, as this could have political and legal implications. These are tasks for another day, but such philosophical research would be welcome to integrate the epistemological contributions of this thesis with these other important issues, and so attain a more comprehensive understanding of the roles of diagnoses in medicine and psychiatry.

Bibliography

Abramson, L. Y., Alloy, L. B., and Metalsky, G. I. (1989). "Hopelessness Depression: A Theory-Based Subtype of Depression". *Psychological Review*, 96: 358–372.

Achinstein, P. (1983). *The Nature of Explanation*. New York: Oxford University Press.

Aliseda, A. and Leonides, L. (2013). "Hypothesis Testing in Adaptive Logics: An Application to Medical Diagnosis". *Logic Journal of the IGPL*, 21: 915–930.

American Psychiatric Association (1952). *Diagnostic and Statistical Manual: Mental Disorders*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders*, 2nd edition. Washington, DC: American Psychiatric Association.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edition. Washington, DC: American Psychiatric Association.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition, Washington, DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Washington, DC: American Psychiatric Association.

Andersen, H. (2012). "Mechanisms: What are they Evidence for in Evidence-Based Medicine?" *Journal of Evaluation in Clinical Practice*, 18: 992–999.

Aquilina, C. and Warner, J. (2004). *A Guide to Psychiatric Examination*. Cheshire: PasTest.

Arborelius, L., Owens, M. J., Plotsky, P. M., and Nemeroff, C. B. (1999). "The Role of Corticotropin-Releasing Factor in Depression and Anxiety Disorders". *Journal of Endocrinology*, 160: 1–12.

Audin, K., Mellor-Clark, J., Barkham, M., Margison, F., McGrath, G., Lewis, S., Cann, L., Duffy, J., and Parry, G. (2001). "Practice Research Networks for Effective Psychological Therapies". *Journal of Mental Health*, 10: 241–251.

Aveline, M. (1999). "The Advantages of Formulation over Categorical Diagnosis in Explorative Psychotherapy and Psychodynamic Management". *European Journal of Psychotherapy and Counselling*, 2: 199–216.

Azam, M., Qureshi, M., and Kinnair, D. (2016). *Psychiatry: A Clinical Handbook*. Banbury: Scion Publishing.

Balint, M. (1964). *The Doctor, His Patient, and the Illness*, 2nd edition. London: Churchill Livingstone.

Barnett, D. (2000). "Is Water Necessarily Identical to H_2O?" *Philosophical Studies*, 98: 99–112.

Baumeister, H. and Parker, G. (2012). "Meta-Review of Depressive Subtyping Models". *Journal of Affective Disorders*, 139: 126–140.

Beck, A. T. (1967). *Depression: Clinical, Experimental, and Theoretical Aspects*. London: Staples Press.

Beebee, H. (2004). "Causing and Nothingness". In L. A. Paul, E. J. Hall, and J. Collins (eds.), *Causation and Counterfactuals*, 291–308. Cambridge, MA: MIT Press.

Beebee, H. (2013). "How to Carve Nature across the Joints without Abandoning Kripke-Putnam Semantics". In S. Mumford and M. Tugby (eds.), *Metaphysics and Science*, 141–163. Oxford: Oxford University Press.

Beebee, H. and Sabbarton-Leary, N. (2010). "Are Psychiatric Kinds 'Real'?" *European Journal of Analytic Philosophy*, 6: 11–27.

Belmaker, R. H. and Agam, G. (2008). "Major Depressive Disorder". *New England Journal of Medicine*, 358: 55–68.

Bentall, R. P. (2003). *Madness Explained: Psychosis and Human Nature*. London: Penguin.

Benzi, M. (2011). "Medical Diagnosis and Actual Causation". *L&PS – Logic and Philosophy of Science*, 9: 365–372.

Bird, A. (2000). *Thomas Kuhn*. Chesham: Acumen.

Bird, A. (2002). "Kuhn's Wrong Turning". *Studies in History and Philosophy of Science*, 33: 443–463.

Bird, A. (2004). "Kuhn on Reference and Essence". *Philosophia Scientiae*, 8: 39–71.

Bird, A. (2011). "What Can Philosophy Tell us about Evidence-Based Medicine? An Assessment of Jeremy Howick's *The Philosophy of Evidence-Based Medicine*". *International Journal of Person Centred Medicine*, 1: 642–648.

Blaxter, M. (1978). "Diagnosis as Category and Process: The Case of Alcoholism". *Social Science and Medicine*, 12: 9–17.

Bluhm, R. (2005). "From Hierarchy to Network: A Richer View of Evidence for Evidence-Based Medicine". *Perspectives in Biology and Medicine*, 48: 535–547.

Bogduk, N. (1995). "Anatomy and Physiology of Headache". *Biomedicine and Pharmacotherapy*, 49: 435–445.

Boissier, M.C., Semerano, L., Challal, S., Saidenberg-Kermanac'h, N., and Falgarone, G. (2012). "Rheumatoid Arthritis: From Autoimmunity to Synovitis and Joint Destruction". *Journal of Autoimmunity*, 39: 222–228.

Bolton, D. (2007). Review of *Psychiatry in the Scientific Image* by Dominic Murphy. *British Journal of Psychiatry*, 191: 273.

Bolton, D. (2008). *What is Mental Disorder? An Essay in Philosophy, Science, and Values*. Oxford: Oxford University Press.

Bolton, D. (2012). "Classification and Causal Mechanisms: A Deflationary Approach to the Classification Problem". In K S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry II: Nosology*, 6–11. Oxford: Oxford University Press.

Bolton, D. (2013). "Should Mental Disorders be Regarded as Brain Disorders? 21st Century Mental Health Sciences and Implications for Research and Training". *World Psychiatry*, 12: 24–25.

Bolton, D. and Hill, J. (2004). *Mind, Meaning and Mental Disorder: The Nature of Causal Explanation in Psychology and Psychiatry*, 2nd edition. Oxford: Oxford University Press.

BonJour, L. (2010). "Against Materialism". In R. C. Koons and G. Bealer (eds.), *The Waning of Materialism*, 3–23. Oxford: Oxford University Press.

Boorse, C. (1977). "Health as a Theoretical Concept". *Philosophy of Science*, 44: 542–573.

Borsboom, D. (2008). "Psychometric Perspectives on Diagnostic Systems". *Journal of Clinical Psychology*, 64: 1089–1108.

Borsboom, D. and Cramer, A. O. J. (2013). "Network Analysis: An Integrative Approach to the Structure of Psychopathology". *Annual Review of Clinical Psychology*, 9: 91–121.

Bowker, G. and Star, S. (2000). *Sorting Things Out*. Cambridge, MA: MIT Press.

Boyd, R. (1999). "Homeostasis, Species, and Higher Taxa". In R. A. Wilson (ed.), *Species: New Interdisciplinary Essays*, 141–185. Cambridge, MA: MIT Press.

Braman, S. S. (2006). "Chronic Cough Due to Chronic Bronchitis: ACCP Evidence-Based Clinical Practice Guidelines". *Chest*, 129: S104–S115.

Brinkmann, S. (2014). "Psychiatric Diagnoses as Semiotic Mediators: The Case of ADHD". *Nordic Psychology*, 66: 121–134.

Broome, M. R. and Bortolotti, L. (2009). "Mental Illness as Mental: In Defence of Psychological Realism". *Humana Mente*, 11: 25–43.

Brown, G. and Harris, T. (1978). *Social Origins of Depression: A Study of Psychiatric Disorder in Women*. Cambridge: Cambridge University Press.

Campbell, J. (2006). "An Interventionist Approach to Causation in Psychology". In A. Gopnik and L. J. Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation*, 58–66. Oxford: Oxford University Press.

Campbell, J. (2008). "Causation in Psychiatry". In K. S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*, 199–216. Baltimore: Johns Hopkins University Press.

Cantor-Graae, E. and Selten, J. P. (2005). "Schizophrenia and Migration: A Meta-Analysis and Review". *American Journal of Psychiatry*, 162: 12–24.

Cartwright, N. (2005). "How can we know what Made the Ratman Sick? Singular Causes and Population Probabilities". In A. Jokić (ed.), *Philosophy of Religion, Physics, and*

Psychology: Essays in Honor of Adolf Grünbaum, 377–390. Amherst, NY: Prometheus Books.

Carvalho, F. P. and Hopko, D. R. (2011). "Behavioral Theory of Depression: Reinforcement as a Mediating Variable between Avoidance and Depression". *Journal of Behavior Therapy and Experimental Psychiatry*, 42: 154–162.

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., and Poulton, R. (2003). "Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene". *Science*, 301: 386–389.

Chadwick, P., Williams, C., and Mackenzie, J. (2003). "Impact of Case Formulation in Cognitive Behaviour Therapy for Psychosis". *Behaviour Research and Therapy*, 41: 671–680.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. J. (2002). "On Sense and Intension". *Noûs*, 36: 135–182.

Chalmers, D. J. (2010). *The Character of Consciousness*. New York: Oxford University Press.

Chalmers, D. J. (2012). *Constructing the World*. New York: Oxford University Press.

Charland, L. (2004). "Character: Moral Treatment and the Personality Disorders". In J. Radden (ed.), *The Philosophy of Psychiatry: A Companion*, 64–77. Oxford: Oxford University Press.

Chiong, W. (2004). "Diagnosing and Defining Disease". In A. L. Caplan, J. J. McCartney, and D. A. Sisti (eds.), *Health, Disease, and Illness: Concepts in Medicine*, 128–135. Washington, DC: Georgetown University Press.

Clark, D. M. (1986). "A Cognitive Approach to Panic". *Behaviour Research and Therapy*, 24: 461–470.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2014). "Mechanisms and the Evidence Hierarchy". *Topoi*, 33: 339–360.

Collier, J., Longmore, M., and Amarakone, K. (2013). *Oxford Handbook of Clinical Specialties*, 9th edition. Oxford: Oxford University Press.

Cooper, R. (2005). *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*. Dordrecht: Springer.

Cooper, R. (2012). "Is Psychiatric Classification a Good Thing?" In K. S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry II: Nosology*, 61–70. Oxford: Oxford University Press.

Cooper, R. (2014). *Diagnosing the Diagnostic and Statistical Manual of Mental Disorders*. London: Karnac Books.

Cooper, R. (2015). "Why is the *Diagnostic and Statistical Manual of Mental Disorders* So Hard to Revise? Path-Dependence and 'Lock-in' in Classification". *Studies in History and Philosophy of Biology and Biomedical Sciences*, 51: 1–10.

Coppen, A. (1967). "The Biochemistry of Affective Disorders". *British Journal of Psychiatry*, 113: 1237–1264.

Coppen, A., Rao, V. A. R., Ruthven, C. R., Goodwin, B. L., and Sandler, M. (1979). "Urinary 4-Hydroxy-3-Methoxyphenylglycol is not a Predictor for Clinical Response to Amitriptyline in Depressive Illness". *Psychopharmacology* (Berlin), 64: 95–97.

Cournoyea, M. and Kennedy, A. G. (2014). "Causal Explanatory Pluralism and Medically Unexplained Physical Symptoms". *Journal of Evaluation in Clinical Practice*, 20: 928–933.

Cowen, P., Harrison, P., and Burns, T. (2012). *Shorter Oxford Textbook of Psychiatry*, 6th edition. Oxford: Oxford University Press.

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., and Borsboom, D. (2010). "Comorbidity: A Network Perspective". *Behavioral and Brain Sciences*, 33: 137–150.

Danks, D., Fancsali, S., Glymour, C., and Scheines, R. (2010). "Comorbid Science?" *Behavioral and Brain Sciences*, 33: 153–155.

Darden, L. (2013). "Mechanisms versus Causes in Biology and Medicine". In H. Chao, S. Chen, and R. L. Millstein (eds.), *Mechanism and Causality in Biology and Economics*, 19–34. Dordrecht: Springer.

David, A. and Wessely, S. (1993). "Chronic Fatigue, ME, and ICD-10". *The Lancet*, 342: 1247–1248.

Davidson, D. ([1967] 2001). "The Logical Form of Action Sentences". In *Essays on Actions and Events*, 2nd edition, 105–148. Oxford: Oxford University Press.

Davies, W. (2016). "Externalist Psychiatry". *Analysis*, 76: 290–296.

Deale, A. and Wessely, S. (2001). "Patients' Perceptions of Medical Care in Chronic Fatigue Syndrome". *Social Science in Medicine*, 52: 1859–1864.

Delgado, P. L. (2011). "The Tryptophan Depletion Challenge Test in Medical Research: Unresolved Issues and Broader Implications for the Use of Physiological Challenges to Investigate and Categorize Disease". *Biological Psychiatry*, 69: 718–719.

Denolin, H., Kuhn, H., Krayenbuehl, H. P., Loogen, F., and Reale, A. (1983). "The Definition of Heart Failure". *European Heart Journal*, 4: 445–448.

Descartes, R. ([1641] 1999). *Meditations on First Philosophy*. Cambridge: Cambridge University Press.

Dragulinescu, S. (2011). "Kuhnian Paradigms: On Meaning and Communication Breakdown in Medicine". *Medicine Studies*, 2: 245–263.

Dragulinescu, S. (2012). "On Anti Humeanism and Medical Singular Causation". *Acta Analytica*, 27: 265–292.

Drevets, W. C., Price, J. L., Simpson, J. R., Todd, R. D., Reich, T., Vannier, M., and Raichle, M. E. (1997). "Subgenual Prefrontal Cortex Abnormalities in Mood Disorders". *Nature*, 386: 824–827.

Drevets, W. C., Savitz, J., and Trimble, M. (2008). "The Subgenual Anterior Cingulate Cortex in Mood Disorders. *CNS Spectrums*, 13: 663–681.

Driessen, E., Smits. N., Dekker, J. J., Peen, J., Don, F. J., Kool, S., Westra, D., Hendriksen, M., Cuijpers, P., and Van, H. L. (2016). "Differential Efficacy of Cognitive Behavioral Therapy and Psychodynamic Therapy for Major Depression: A Study of Prescriptive Factors". *Psychological Medicine*, 46: 731–744.

Dumit, J. (2006). "Illnesses you have to Fight to Get: Facts as Forces in Uncertain, Emergent Illnesses". *Social Science and Medicine*, 62: 577–590.

Elliott, C. (1999). *A Philosophical Disease: Bioethics, Culture and Identity*. New York: Routledge.

Ellis, B. (2001). *Scientific Essentialism*. Cambridge: Cambridge University Press.

Emmelkamp, P. M. G., Bouman, T. K., and Blaauw, E. (1994). "Individualised versus Standardised Therapy: A Comparative Evaluation with Obsessive-Compulsive Patients". *Clinical Psychology and Psychotherapy*, 1: 95–100.

Engel, G. L. (1977). "The Need for a New Medical Model: A Challenge for Biomedicine". *Science*, 196: 129–136.

Ereshefsky, M. (2010). "Species, Taxonomy, and Systematics". In A. Rosenberg and A. Arp (eds.), *Philosophy of Biology: An Anthology*, 255–271. Chichester: Wiley-Blackwell.

Evans, G. (1973). "The Causal Theory of Names". *Proceedings of the Aristotelian Society*, 47: S187–S225.

Evidence-Based Medicine Working Group (1992). "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine". *Journal of the American Medical Association*, 268: 2420–2425.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., and Munoz, R. (1972). "Diagnostic Criteria for Use in Psychiatric Research". *Archives of General Psychiatry*, 26: 57–63.

Ferster, C. B. (1973). "A Functional Analysis of Depression". *American Psychologist*, 28: 857–870.

Feyerabend, P. (1962). "Explanation, Reduction and Empiricism". In H. Feigl and G. Maxwell (eds.), *Scientific Explanation, Space, and Time*, 28–97. Minneapolis, MN: University of Minneapolis Press.

First, M. B. (2009). "Harmonization of ICD-11 and DSM-5: Opportunities and Challenges". *British Journal of Psychiatry*, 195: 382–390.

Fitzgerald, P. J. (2013). "Black Bile: Are Elevated Monoamines an Etiological Factor in Some Cases of Major Depression?" *Medical Hypotheses*, 80: 823–826.

Fleck, L. ([1935] 1981). *Genesis and Development of a Scientific Fact*. F. Bradley and T. J. Trenn (trans.). Chicago: Chicago University Press.

Ford, J. M., Morris, S. E., Hoffman, R. E., Sommer, I., Waters, F., McCarthy-Jones, S., Thoma, R. J., Turner, J. A., Keedy, S. K., Badcock, J. C., and Cuthbert, B. N. (2014). "Studying Hallucinations within the NIMH RDoC Framework". *Schizophrenia Bulletin*, 40: S295–S304.

France, C. M., Lysaker, P. H., and Robinson, R. P. (2007). "The 'Chemical Imbalance' Explanation for Depression: Origins, Lay Endorsement, and Clinical Implications". *Professional Psychology: Research and Practice*, 38: 411–420.

Frege, F. L. G. ([1892] 1952). "On Sense and Reference". In P. Geach and M. Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*, 56–78. Oxford: Blackwell.

Freis, E. D. (1954). "Mental Depression in Hypertensive Patients Treated for Long Periods with Large Doses of Reserpine". *New England Journal of Medicine*, 251: 1006–1008.

Freud, S. ([1917] 1946). "Mourning and Melancholia". In *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, volume XIV: 237–258. London: The Hogarth Press.

Fried, E. I. and Nesse, R. M. (2015). "Depression Sum-Scores don't Add Up: Why Analyzing Specific Depression Symptoms is Essential". *BMC Medicine*, 13: 72.

Foucault, M. ([1961] 1964). *Madness and Civilization: A History of Insanity in the Age of Reason*. R. Howard (trans.). London: Tavistock.

Fuchs, T. (2005). "Delusional Mood and Delusional Perception – A Phenomenological Analysis". *Psychopathology*, 38: 133–139.

Fuchs, T. (2012). "Are Mental Illnesses Diseases of the Brain?" In S. Choudhury and J. Slaby (eds.), *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience*, 331–344. Chichester: Wiley-Blackwell.

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., and Komaroff, A, (1994). "The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study". *Annals of Internal Medicine*, 121: 953–959.

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

Fulford, K. W. M., Thornton, T., and Graham, G. (2006). *Oxford Textbook of Philosophy and Psychiatry*. Oxford: Oxford University Press.

Fürst, M. (2011). "What Mary's Aboutness is about". *Acta Analytica*, 26: 63–74.

Gao, Y., Raine, A., Chan, F., Venables, P. H., and Mednick, S. A. (2010). "Early Maternal and Paternal Bonding, Childhood Physical Abuse and Adult Psychopathic Personality". *Psychological Medicine*, 40: 1007–1016.

Ghaemi, S. N. (2009). "The Rise and Fall of the Biopsychosocial Model". *British Journal of Psychiatry*, 195: 3–4.

Glennan, S. S. (1996). "Mechanisms and the Nature of Causation". *Erkenntnis*, 44: 49–71.

Goadsby, P. J. (2012). "Pathophysiology of Migraine". *Annals of Indian Academy of Neurology*, 15: S15–S22.

Goodyer, I. M., Herbert, J., Altham, P. M. E., Pearson, J., Secher, J. S. M., and Shiers, H. M. (1996). "Adrenal Secretion during Major Depression in 8- to 16-Year-Olds, I. Altered Diurnal Rhythms in Salivary Cortisol and Dehydroepiandrosterone (DHEA) at Presentation". *Psychological Medicine*, 26: 245–256.

Gorovitz, S. and MacIntyre, A. (1975). "Toward a Theory of Medical Fallibility". *The Hastings Center Report*, 5: 13–23.

Gotlib, I. H. and Joorman, J. (2010). "Cognition and Depression: Current Status and Future Directions". *Annual Review of Clinical Psychology*, 6: 285–312.

Graham, J., Salimi-Khorshidi, Hagan, C., Walsh, N., Goodyer, I., Lennox, B., and Suckling, J. (2013). "Meta-Analytic Evidence for Neuroimaging Models of Depression: State or Trait?" *Journal of Affective Disorders*, 151: 423–431.

Great Britain (1957). *Homicide Act 1957*. London: The Stationery Office.

Great Britain (1983). *Mental Health Act 1983*. London: The Stationery Office.

Great Britain (2009). *Coroners and Justice Act 2009*. London: The Stationery Office.

Great Britain (2014). *Mesothelioma Act 2014*. London: The Stationary Office.

Green, B. (2009). *Problem-Based Psychiatry*, 2nd edition. Oxford: Radcliffe Publishing.

Griffiths, P. (1999). "Squaring the Circle: Natural Kinds with Historical Essences". In R. Wilson (ed.), *Species: New Interdisciplinary Essays*, 209–228). Cambridge, MA: MIT Press.

Groenewold, N. A., Opmeer, E. M., de Jonge, P., Aleman, A., and Costafreda, S. G. (2013). "Emotional Valence Modulates Brain Functional Abnormalities in Depression: Evidence from a Meta-Analysis of fMRI Studies". *Neuroscience and Behavioural Reviews*, 37: 152–163.

Guillin, O., Abi-Dargham, A., and Laurelle, M. (2007). "Neurobiology of Dopamine in Schizophrenia". *International Review of Neurobiology*, 78: 1–39.

Gulati, G., Lynall, M. E., and Saunders, K. E. A. (2014). *Lecture Notes: Psychiatry*, 11th edition. Chichester: Wiley-Blackwell.

Hacking, I. (1999). *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hamilton, E. W. and Abramson, L. Y. (1983). "Cognitive Patterns and Major Depressive Disorder: A Longitudinal Study in a Hospital Setting". *Journal of Abnormal Psychology*, 92: 173–184.

Haslam, N. (2014). "Natural Kinds in Psychiatry: Conceptually Implausible, Empirically Questionable, and Stigmatizing". In H. Kincaid and J. Sullivan (eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, 11–28. Cambridge, MA: MIT Press.

Haslanger, S. (2006). "Philosophical Analysis and Social Kinds: What Good are Our Intuitions?" *Proceedings of the Aristotelian Society*, 80: S89–S118.

Haukioja, J. (2015). "On Deriving Essentialism from the Theory of Reference". *Philosophical Studies*, 172: 2141–2151.

Healey, P. M. and Jacobson, E. J. (2006). *Common Medical Diagnoses: An Algorithmic Approach*, 4th edition. Philadelphia: Saunders.

Hempel, C. G. (1965a). "Aspects of Scientific Explanation". In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 331–496. New York: The Free Press.

Hempel, C. G. (1965b). "Fundamentals of Taxonomy". In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 137–154. New York: The Free Press.

Hoffmann, B. (2016). "Obesity as a Socially Defined Disease: Philosophical Considerations and Implications for Policy and Care". *Health Care Analysis*, 24: 86–100.

Hood, S. B. and Lovett, B. J. (2010). "Network Models of Psychopathology and Comorbidity: Philosophical and Pragmatic Considerations". *Behavioral and Brain Sciences*, 33: 159–160.

Howick, J. (2011). *The Philosophy of Evidence-Based Medicine*. Chichester: Wiley-Blackwell.

Hucklenbroich, P. (2014). "'Disease Entity' as the Key Theoretical Concept of Medicine". *Journal of Medicine and Philosophy*, 39: 609–633.

Hume, D. ([1748] 2000). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.

Hyman, S. (2010). "The Diagnosis of Mental Disorders: The Problem of Reification". *Annual Review of Clinical Psychology*, 6: 155–179.

Ingleby, D. (1982). "The Social Construction of Mental Illness". In P. Wright and A. Treacher (eds.), *The Problem of Medical Knowledge: Examining the Social Construction of Medicine*, 123–143. Edinburgh: Edinburgh University Press.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., and Wang, P. (2010). "Research Domain Criteria (RDoC): Toward a New Classification

Framework for Research on Mental Disorders". *American Journal of Psychiatry*, 167: 748–751.

Jackson, F. (1982). "Epiphenomenal Qualia". *Philosophical Quarterly*, 32: 127–136.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jason, L. A., Katz, B. Z., Shiraishi, Y., Mears, C. J., Im, Y., and Taylor, R. R. (2014). "Predictors of Post-Infectious Chronic Fatigue Syndrome in Adolescents". *Health Psychology and Behavioral Medicine*, 2: 41–51.

Jaspers, K. ([1913] 1997). *General Psychopathology*. J. Hoenig and M. W. Hamilton (trans), 7th edition. Baltimore: Johns Hopkins University Press.

Johnstone, L. (2006). "Controversies and Debates about Formulation". In L. Johnstone and R. Dallos (eds.), *Formulation in Psychology and Psychotherapy: Making Sense of People's Problems*, 208–235. London: Routledge.

Joyce, P., McKenzie, J., Carter, J., Rae, A., Luty, S., Frampton, C., and Mulder, R. (2007). "Temperament, Character, and Personality Disorders as Predictors of Response to Interpersonal Psychotherapy and Cognitive-Behavioural Therapy for Depression". *British Journal of Psychiatry*, 190: 503–508.

Jutel, A. G. (2011). *Putting a Name to It: Diagnosis in Contemporary Society*. Baltimore: Johns Hopkins University Press.

Kant, I. ([1781] 1998). *A Critique of Pure Reason*. Cambridge: Cambridge University Press.

Keller, M. C., Neale, M. C., and Kendler, K. S. (2007). "Association of Different Adverse Life Events with Distinct Patterns of Depressive Symptoms". *American Journal of Psychiatry*, 164: 1521–1529.

Kendell, R. E. (1975). "The Concept of Disease and Its Implications for Psychiatry". *British Journal of Psychiatry*, 127: 305–315.

Kendell, R. E. (1989). "Clinical Validity". *Psychological Medicine*, 19: 45–55.

Kendell, R. E. and Jablensky, A. (2003). "Distinguishing Between the Validity and Utility of Psychiatric Diagnoses". *American Journal of Psychiatry*, 160: 4–12.

Kendler, K. S. (2006). "Reflections on the Relationship between Psychiatric Genetics and Psychiatric Nosology". *American Journal of Psychiatry*, 163: 1138–1146.

Kendler, K. S. (2008). "Explanatory Models for Psychiatric Illness". *American Journal of Psychiatry*, 165: 695–702.

Kendler, K. S. (2012). "The Dappled Nature of Causes of Psychiatric Illness: Replacing the Organic-Functional/Hardware-Software Dichotomy with Empirically Based Pluralism". *Molecular Psychiatry*, 17: 377–388.

Kendler, K. S. (2014). "The Structure of Psychiatric Science". *American Journal of Psychiatry*, 171: 931–938.

Kendler, K. S. and Campbell, J. (2009). "Interventionist Causal Models in Psychiatry: Repositioning the Mind-Body Problem". *Psychological Medicine*, 39: 881–887.

Kendler, K. S., Gatz, M., Gardner, C. O., and Pedersen, N. L. (2006). "A Swedish National Twin Study of Lifetime Major Depression". *American Journal of Psychiatry*, 163: 109–114.

Kendler, K. S., Zachar, P., and Craver, C. (2011). "What Kinds of Things are Psychiatric Disorders?" *Psychological Medicine*, 41: 1143–1150.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kincaid, H. (2014). "Defensible Natural Kinds in the Study of Psychopathology". In H. Kincaid and J. Sullivan (eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, 145–173. Cambridge, MA: MIT Press.

Kirmayer, L. J., Groleau, D., Looper, K. J., and Dao, M. D. (2004). "Explaining Medically Unexplained Symptoms". *Canadian Journal of Psychiatry*, 49: 663–671.

Klein, M. ([1957] 1984). "Envy and Gratitude". In *Envy and Gratitude and Other Works 1946–1963*, 176–235. London: The Hogarth Press.

Klerman, G. L., DiMascio, A., Weissman, M., Prusoff, B., and Paykel, E. S. (1974) "Treatment of Depression by Drugs and Psychotherapy". *American Journal of Psychiatry*, 131: 186–191.

Kripke, S. ([1972] 1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

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

Kuhn, T. S. (2000). *The Road since Structure*. Chicago: Chicago University Press.

Kupfer, D. J., First, M. B., and Regier, D. A. (2002). "Introduction". In D. J. Kupfer, M. B. First, and D. A. Regier (eds.), *A Research Agenda for DSM-V*, xv–xxiii. Washington, DC: American Psychiatric Association.

Laing, R. D. (1967). *The Politics of Experience and the Bird of Paradise*. Harmondsworth: Penguin.

Lam, D. C., Salkovskis, P. M., and Warwick, H. (2005). "An Experimental Investigation of the Impact of Biological versus Psychological Explanations of the Cause of 'Mental Illness'". Journal of Mental Health, 14: 453–464.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Ledley, R. S. and Lusted, L. B. (1959). "Reasoning Foundations of Medical Diagnosis". *Science*, 130: 9–21.

Levine, J. (1983). "Materialism and Qualia: The Explanatory Gap". *Pacific Philosophical Quarterly*, 64: 354–361.

Lewinsohn, P. M. (1974). "A Behavioral Approach to Depression". In R. J. Friedman and M. M. Katz (eds.), *The Psychology of Depression: Contemporary Theory and Research*, 157–178. New York: John Wiley & Sons.

Lewis, D. K. (1986a). "Causation". In *Philosophical Papers*, volume II, 159–213.

Lewis, D. K. (1986b). "Causal Explanation". In *Philosophical Papers*, volume II, 214–240. Oxford: Oxford University Press.

Lipton, P. (2004). *Inference to the Best Explanation*, 2nd edition. London: Routledge.

Loas, G., Noisette, C., Legrand, A., and Boyer, P. (2000). "Is Anhedonia a Specific Dimension in Chronic Schizophrenia?" *Schizophrenia Bulletin*, 26: 495–506.

Longmore, M., Wilkinson, I. B., Baldwin, A., and Wallin, E. (2014). *Oxford Handbook of Clinical Medicine*, 9th edition. Oxford: Oxford University Press.

Lopresti, A. L., Maker, G. L., Hood, S. D., and Drummond, P. D. (2014). "A Review of Peripheral Biomarkers in Major Depression: The Potential of Inflammatory and Oxidative Stress Biomarkers". *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 48: 102–111.

Luyten, P., van Houdenhove, B., Pae, C., Kempke, S., and van Wambeke, P. (2008). "Treatment of Chronic Fatigue Syndrome: Findings, Principles and Strategies". *Psychiatry Investigation*, 5: 209–212.

Mace, C. and Binyon, S. (2005). "Teaching Psychodynamic Formulation to Psychiatric Trainees, Part 1: Basics of Formulation". *Advances in Psychiatric Treatment*, 11: 416–423.

Machamer, P., Darden, L., and Craver, C. F. (2000). "Thinking about Mechanisms". *Philosophy of Science*, 67: 1–25.

MacNeil, C. A., Hasty, M. K., Conus, P., and Berk, M. (2012). "Is Diagnosis Enough to Guide Interventions in Mental Health? Using Case Formulation in Clinical Practice". *BMC Medicine*, 10. DOI: 10.1186/1741-7015-10-111.

MacQueen, G. M., Campbell, S., McEwen, B. S., Macdonald, K., Amano, S., Joffe, R. T., Nahmias, C., and Young, L. T. (2003). "Course of Illness, Hippocampal Function, and Hippocampal Volume in Major Depression". *Proceedings of the National Academy of Sciences of the United States of America*, 100: 1387–1392.

Maletic, V. and Raison, C. (2014). "Integrated Neurobiology of Bipolar Disorder". *Frontiers in Psychiatry*, 5. DOI: 10.3389/fpsyt.2014.00098.

Malhi, G. S. and Geddes, J. R. (2014). "Carving Bipolarity Using a Lithium Sword". *British Journal of Psychiatry*, 205: 337–339.

Marconi, D. (2005). "Two-Dimensional Semantics and the Articulation Problem". *Synthese*, 143: 321–349.

Margison, F.R., McGrath, G., Barkham, M., Mellor Clark, J., Audin, K., Connell, J., and Evans, C. (2000). "Measurement and Psychotherapy: Evidence-Based Practice and Practice-Based Evidence". *British Journal of Psychiatry*, 177: 123–130.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Marrie, R. A., Horwitz, R., Cutter, G., Tyry, T., Campagnolo, D., and Voillmer, T. (2009). "The Burden of Mental Comorbidity in Multiple Sclerosis: Frequent, Underdiagnosed, and Undertreated". *Multiple Sclerosis*, 15: 385–392.

Massart, R., Mongeau, R., and Lanfumey, L. (2012). "Beyond the Monoaminergic Hypothesis: Neuroplasticity and Epigenetic Changes in a Transgenic Mouse Model of Depression". *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367: 2485–2494.

Maung, H. H. (2016a). "Diagnosis and Causal Explanation in Psychiatry". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 60: 15–24.

Maung, H. H. (2016b). "The Causal Explanatory Functions of Medical Diagnoses". *Theoretical Medicine and Bioethics*. Published online first 16th September 2016. DOI: 10.1007/s11017-016-9377-5.

Maung, H. H. (2016c). "To What Do Psychiatric Diagnoses Refer? A Two-Dimensional Semantic Analysis of Diagnostic Terms". *Studies in History and Philosophy of Biological and Biomedical Sciences*, 55: 1–10.

McBride, C., Atkinson, L., Quilty, L., and Bagby, R. (2006). "Attachment as Moderator of Treatment Outcome in Major Depression: A Randomized Control Trial of Interpersonal Psychotherapy versus Cognitive Behavior Therapy". *Journal of Consulting and Clinical Psychology*, 74: 1041–1054.

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

McGorm, K., Burton, C., Weller, D., Murray, G., and Sharpe, M. (2010). "Patients Repeatedly Referred to Secondary Care with Symptoms Unexplained by Organic Disease: Prevalence, Characteristics and Referral Pattern". *Family Practice*, 27: 479–486.

McLean, D., Thara, R., John, S., Barrett, R., Loa, P., McGrath, J., and Mowry, B. (2014). "DSM-IV 'Criterion A' Schizophrenia Symptoms across Ethnically Different Populations: Evidence for Differing Psychotic Symptom Content of Structural Organization?" *Culture, Medicine, and Psychiatry*, 38: 408–426.

Megone, C. (1998). "Aristotle's Function Argument and the Concept of Mental Illness". *Philosophy, Psychiatry, and Psychology*, 7: 45–65.

Mehta, S. and Farina, A. (1997). "Is being Sick Really Better? Effect of the Disease View of Mental Disorder on Stigma". *Journal of Social and Clinical Psychology*, 16: 405–419.

Mezzick, J. E. (2005). "Values and Comprehensive Diagnosis". *World Psychiatry*, 4: 91–92.

Millikan, R. G. (1999). "Historical Kinds and the 'Special Sciences'". *Philosophical Studies*, 95: 45–65.

Mitchell, S. (2008). "Explaining Complex Behavior". In K. S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*, 19–38. Baltimore: Johns Hopkins University Press.

Moncrieff, J. (2010). "Psychiatric Diagnosis as a Political Device". *Social Theory and Health*, 8: 370–382.

Morse, S. J. (1999). "Craziness and Criminal Responsibility". *Behavioral Sciences and the Law*, 17: 147–164.

Murphy, D. (2006). *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press.

Murphy, D. (2008). "Levels of Explanation in Psychiatry". In K. S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*, 102–125. Baltimore: Johns Hopkins University Press.

Murphy, D. (2011). "Explanation in Psychiatry". *Philosophy Compass*, 5: 602–610.

Murphy, D. (2014). "Natural Kinds in Folk Psychology and in Psychiatry". In H. Kincaid and J. Sullivan (eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, 105–122. Cambridge, MA: MIT Press.

Nagel, T. (1974). "What is it Like to Be a Bat?" *The Philosophical Review*, 83: 435–450.

Nemeroff, C. B., Owens, M. J., Bissette, G., Andorn, A. C., and Stanley, M. (1988). "Reduced Corticotropin-Releasing Factor Binding Sites in the Frontal Cortex of Suicide Victims". *Archives of General Psychiatry*, 45: 577–579.

Nervi, M. (2010). "Mechanisms, Malfunctions and Explanation in Medicine". *Biology and Philosophy*, 25: 215–228.

Nettleton, S., O'Malley, L., Watt, I., and Duffey, P. (2004). "Enigmatic Illness: Narratives of Patients who Live with Medically Unexplained Symptoms". *Social Theory and Health*, 2: 47–66.

Nettleton, S., Watt, I., O'Malley, L., and Duffey, P. (2005). "Understanding the Narratives of People who Live with Medically Unexplained Illness". *Patient Education and Counseling*, 56: 205–210.

Newby, D. E., Grubb, N. R., and Bradbury, A. (2010). "Cardiovascular Disease". In N. R. Colledge, B. R. Walker, and S. H. Ralston (eds.), *Davidson's Principles and Practice of Medicine*, 21st edition, 521–640. Edinburgh: Churchill Livingstone.

NHS Choices (2014a). *Clinical Depression* [online]. Available at: http://www.nhs.uk/Conditions/Depression/Pages/Introduction.aspx (accessed 4th April 2016).

NHS Choices (2014b). *Generalised Anxiety Disorder in Adults* [online]. Available at: http://www.nhs.uk/conditions/anxiety/pages/introduction.aspx (accessed 4th April 2016).

Palazidou, E. (2012). "The Neurobiology of Depression". *British Medical Bulletin*, 101: 127–145.

Parsons, T. (1951). *The Social System*. Glencoe, IL: The Free Press.

Patient.info (2013). *Schizophrenia* [online]. Available at: http://patient.info/health/schizophrenia-leaflet (accessed 4th April 2016).

Peitzman, S. J. (2007). *Dropsy, Dialysis, Transplant: A Short History of Failing Kidneys*. Baltimore: Johns Hopkins University Press.

Phelan, J. (2005). "Geneticization of Deviant Behavior and Consequences for Stigma: The Case of Mental Illness". *Journal of Health and Social Behavior*, 46: 307–322.

Phillips, J. (2005). "Idiographic Formulations, Symbols, Narratives, Context and Meaning". *Psychopathology*, 38: 180–184.

Poland, J. (2014). "Deeply Rooted Sources of Error and Bias in Psychiatric Classification". In H. Kincaid and J. Sullivan (eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*, 29–64. Cambridge, MA: MIT Press.

Poland, J., von Eckhardt, B., and Spaulding, W. (1994). "Problems with the DSM Approach to Classifying Psychopathology". In G. Graham and G. L. Stephens (eds.), *Philosophical Psychopathology*, 235–260. Cambridge, MA: MIT Press.

Potter, N. (2004). "Gender". In J. Radden (ed.), *The Philosophy of Psychiatry: A Companion*, 237–244. Oxford: Oxford University Press.

Putnam, H. (1975a). "The Meaning of 'Meaning'". In *Mind, Language and Reality: Philosophical Papers*, volume II, 215–271. New York: Cambridge University Press.

Putnam, H. (1975b). "Brains and Behavior". In *Mind, Language and Reality: Philosophical Papers*, volume II, 325–341. New York: Cambridge University Press.

Qiu, R. Z. (1989). "Models of Explanation and Explanation in Medicine". *International Studies in the Philosophy of Science*, 3: 199–212.

Rachman, S. J. (1993). "Obsessions, Responsibility and Guilt". *Behaviour Research and Therapy*, 31: 149–154.

Radden, J. (1982). "Diseases as Excuses: Durham and the Insanity Plea". *Philosophical Studies*, 42: 349–362.

Radden, J. (2003). "Is This Dame Melancholy? Equating today's Depression and Past Melancholia". *Philosophy, Psychiatry, and Psychology*, 10: 37–52.

Rendle-Short, J. and Gray, O. P. (1967). *A Synopsis of Children's Diseases*, 4th edition. Bristol: John Wright and Sons.

Reznek, L. (1991). *The Philosophical Defence of Psychiatry*. London: Routledge.

Risch, N., Herrell, R., Lehner, T., Liang, K. Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., and Merikangas, K. R. (2009). "Interaction between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-Analysis." *Journal of the American Medical Association*, 301: 2462–2471.

Rizzi, D. A. (1994). "Causal Reasoning and the Diagnostic Process". *Theoretical Medicine*, 15: 315–333.

Robinson, O. J. and Sahakian, B. J. (2008). "Recurrence in Major Depressive Disorder: A Neurocognitive Perspective". *Psychological Medicine*, 38: 315–318.

Roiser, J. (2015). "What has Neuroscience Ever Done for us?" *The Psychologist*, 28: 284–287.

Roiser, J., Elliott, R., and Sahakian, B. J. (2012). "Cognitive Mechanisms of Treatment in Depression". *Neuropsychopharmacology*, 37: 117–136.

Rose, N. R. and Mackay, I. R. (1985). "Genetic Predisposition to Autoimmune Diseases". In N. R. Rose and I. R. Mackay (eds.), *The Autoimmune Diseases*, 1–29. Orlando: Academic Press.

Roth, N. L. and Nelson, M. S. (1997). "HIV Diagnosis Rituals and Identity Narratives". *AIDS Care*, 9: 161–179.

Royal College of Psychiatrists (2013). *A Competency-Based Curriculum for Specialist Core Training in Psychiatry*. Available at: http://www.rcpsych.ac.uk/pdf/Core%20Curriculum_FINAL%20Version_July2013_updatedJun15.pdf (accessed 3rd February 2016).

Royal College of Psychiatrists (2015). *Depression*. Available at: http://www.rcpsych.ac.uk/healthadvice/problemsdisorders/depression.aspx (accessed 3rd October 2016).

Russell, B. (1905). "On Denoting". *Mind*, 16: 479–493.

Russo, F. and Williamson, J. (2007). "Interpreting Causality in the Health Sciences". *International Studies in the Philosophy of Science*, 21: 157–170.

Rzepiński, T. M. (2007). "The Structure of Diagnosis in Medicine: Introduction to Interrogative Characteristics". *Theoretical Medicine and Bioethics*, 28: 63–81.

Sadegh-Zadeh, K. (2012). *Handbook of Analytic Philosophy of Medicine*. Dordrecht: Springer.

Sadler, J. Z. (2005). *Values and Psychiatric Diagnosis*. Oxford: Oxford University Press.

Sadock, B. J. and Sadock, V. A. (2008). *Kaplan and Sadock's Concise Textbook of Clinical Psychiatry*, 3rd edition. Philadelphia, PA: Lippincott Williams and Wilkins.

Salkovskis, P. M. (1985). "Obsessive–Compulsive Problems: A Cognitive–Behavioural Analysis". *Behaviour Research and Therapy*, 23: 571–583.

Salmon, W. C. ([1975] 1998). "Causal and Theoretical Explanation". In *Causality and Explanation*, 108–124. Oxford: Oxford University Press.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sankey, H. (2009). "Scientific Realism and the Semantic Incommensurability Thesis". *Studies in History and Philosophy of Science*, 40: 196–202.

Saul, R. (2014). *ADHD Does Not Exist*. New York: HarperCollins.

Schaffner, K. F. (1986). "Exemplar Reasoning about Biological Models and Diseases: A Relation between the Philosophy of Medicine and Philosophy of Science". *Journal of Medicine and Philosophy*, 11: 63–80.

Schaffner, K. F. (2008). "Etiological Models in Psychiatry: Reductive and Nonreductive Approaches". In K. S. Kendler and J. Parnas (eds.), *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*, 48–90. Baltimore: Johns Hopkins University Press.

Schildkraut, J. J. (1965). "The Catecholamine Hypothesis of Affective Disorders: A Review of Supporting Evidence". *American Journal of Psychiatry*, 122: 509–522.

Schneider, R. K. and Levenson, J. L. (2008). *Psychiatry Essentials for Primary Care*. Philadelphia, PA: American College of Physicians.

Schramme, T. (2013). "'I Hope I get Old before I Die': Ageing and the Concept of Disease". *Theoretical Medicine and Bioethics*, 34: 171–187.

Schulte, D., Kunzel, R., Pepping, G., and Schulte-Bahrenberg, T. (1992). "Tailor-Made versus Standardised Therapy of Phobic Patients". *Advances in Behaviour Research and Therapy*, 14: 67–92.

Schwartz, A. and Elstein, A. S. (2008). "Clinical Reasoning in Medicine". In J. Higgs, M. A. Jones, S. Loftus, and N. Christensen (eds.), *Clinical Reasoning in the Health Professions*, 3rd edition, 223–234. Amsterdam: Elsevier.

Seligman, M. E. P. (1975). *Helplessness: On Depression, Development, and Death*. San Francisco: W. H. Freeman.

Sellars, W. ([1956] 1997). *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

Sethi, S. (2008). *Textbook of Psychiatry*. New Delhi: Elsevier.

Shaw, D. M., O'Keeffe, R., MacSweeney, D. A., Brooksbank, B. W., Noguera, R., and Coppen, A. (1973). "3-Methoxy-4-Hydroxyphenylglycol in Depression". *Psychological Medicine*, 3: 333–336.

Shorter, E. (2011). "Still Tilting at Windmills. Commentary on… the Myth of Mental Illness". *The Psychiatrist*, 35: 183–184.

Shyn, S. I. and Hamilton, S. P. (2010). "The Genetics of Major Depression: Moving beyond the Monoamine Hypothesis". *Psychiatric Clinics of North America*, 33: 125–140.

Simon, J. (2008). "Constructive Realism and Medicine: An Approach to Medical Ontology". *Perspectives in Biology and Medicine*, 51: 353–366.

Skern, T. (2010). "100 Years Poliovirus: From Discovery to Eradication". *Archives of Virology*, 155: 1371–1381.

Soames, S. (2005). *Reference and Description: The Case against Two-Dimensionalism*. Princeton, NJ: Princeton University Press.

Sparr, L. F. (2009). "Personality Disorders and Criminal Law: An International Perspective". *Journal of the Academy of Psychiatry and the Law*, 37: 168–181.

Stalnaker, R. (1978). "Assertion". *Syntax and Semantics*, 9: 315–332.

Stanley, D. E. and Campos, D. G. (2013). "The Logic of Medical Diagnosis". *Perspectives in Biology and Medicine*, 56: 300–315.

Stevens, L. and Rodin, I. (2010). *Psychiatry: An Illustrated Colour Text*, 2nd edition. London: Churchill Livingstone.

Street, H., Sheeran, P., and Orbell, S. (1999). "Conceptualizing Depression: An Integration of 27 Theories". *Clinical Psychology and Psychotherapy*, 6: 175–193.

Strickland, P. L., Deakin, J. F. W., Percival, C., Dixon, J., Gater, R. A., and Goldberg, D. P. (2002). "Bio-Social Origins of Depression in the Community: Interactions between Social Adversity, Cortisol and Serotonin Neurotransmission". *British Journal of Psychiatry*, 180: 168–173.

Sullivan, P. F., Neale, M. C., and Kendler, K. S. (2000). "Genetic Epidemiology of Major Depression: Review and Meta-Analysis". *American Journal of Psychiatry*, 157: 1552–1562.

Summerfield, D. (2001). "The Invention of Post-Traumatic Stress Disorder and the Social Usefulness of a Psychiatric Category". *British Medical Journal*, 322: 95–98.

Swain, M. (1980). "Causation and Distinct Events". In P. van Inwagen (ed.), *Time and Cause: Essays Presented to Richard Taylor*, 155–169. Dordrecht: Reidel.

Swedo, S. E., Leonard, H. L., Garvey, M., Mittleman, B., Allen, A. J., Perlmutter, S., Lougee, L., Dow, S., Zamkoff, J., and Dubbert, B. K. (1998). "Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections: Clinical Description of the First 50 Cases". *American Journal of Psychiatry*, 155: 264–271.

Szasz, T. (1960). "The Myth of Mental Illness". *American Psychologist*, 15: 113–118.

Tarrier, N. and Calam, R. (2002). "New Developments in Cognitive-Behavioural Case Formulation. Epidemiological, Systemic and Social Context: An Integrative Approach". *Behavioural and Cognitive Psychotherapy*, 30: 311–328.

Thagard, P. (1999). *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.

Thornton, T. (2007). *Essential Philosophy of Psychiatry*. Oxford: Oxford University Press.

Thornton, T. (2010). "Narrative rather than Idiographic Approaches as Counterpart to the Nomothetic Approach to Assessment". *Psychopathology*, 43: 252–261.

Travis, W. D. (2012). "Update on Small Cell Carcinoma and Its Differentiation from Squamous Cell Carcinoma and Other Non-Small Cell Carcinomas". *Modern Pathology*, 25: S18–S30.

Tsou, J. Y. (2013). "Depression and Suicide are Natural Kinds: Implications for Physician-Assisted Suicide". *International Journal of Law and Psychiatry*, 36: 461–470.

Tsou, J. Y. (2015). "DSM-5 and Psychiatry's Second Revolution: Descriptive vs. Theoretical Approaches to Psychiatric Classification". In S. Demazeux and P. Singy (eds.), *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*, 43–62. Dordrecht: Springer.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

van Os, J. (2004). "Does the Urban Environment Cause Psychosis?" *British Journal of Psychiatry*, 184: 287–288.

Veale, D. (2007). "Cognitive-Behavioural Therapy for Obsessive-Compulsive Disorder". *Advances in Psychiatric Treatment*, 13: 438–446.

Wakefield, J. C. (1992). "On the Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values". *American Psychologist*, 47: 373–388.

Ware, N. C. (1992). "Suffering and the Social Construction of Illness: The Delegitimation of Illness Experience in Chronic Fatigue Syndrome". *Medical Anthropology Quarterly*, 6: 347–361.

Weerasekera, P. (1996). *Multiperspective Case Formulation: A Step towards Treatment Integration*. Malabar, FL: Krieger.

Weerasekera, P. (2009). "A Formulation of the Case of Antoinette: A Multiperspective Approach". In P. Sturmey (ed.), *Clinical Case Formulation: Varieties of Approaches*, 145–156. Chichester: Wiley-Blackwell.

Weissman, M. M., Klerman, G. L., Prusoff, B. A., Sholomskas, D., and Padian, N. (1981). "Depressed Outpatients: Results One Year after Treatment with Drugs and/or Interpersonal Psychotherapy". *Archives of General Psychiatry*, 38: 52–55.

Westmeyer, H. (1975). "The Diagnostic Process as a Statistical-Causal Analysis". *Theory and Decision*, 6: 57–86.

Wheeler, A. L., and Voineskos, A. N. (2014). "A Review of Structural Neuroimaging in Schizophrenia: from Connectivity to Connectomics". *Frontiers in Human Neuroscience*, 8. DOI: 10.3389/fnhum.2014.00653.

Whitbeck, C. (1981). "What is Diagnosis? Some Critical Reflections". *Metamedicine*, 2: 319–329.

Whitfield, G. and Williams, C. (2003). "The Evidence Base for Cognitive-Behavioural Therapy in Depression: Delivery in Busy Clinical Settings". *Advances in Psychiatric Treatment*, 9: 21–30.

Widiger, T. A. (2007). "Dimensional Models of Personality Disorder". *World Psychiatry*, 6: 79–83.

Wilde, M. I. and Benfield, P. (1995). "Tianeptine: A Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Depression and Coexisting Anxiety and Depression". *Drugs*, 49: 411–439.

Wilkinson, S. (2000). "Is 'Normal Grief' and Mental Disorder?" *The Philosophical Quarterly*, 50: 289–304.

Williams, N. E. (2011a). "Putnam's Traditional Neo-Essentialism". *The Philosophical Quarterly*, 61: 151–170.

Williams, N. E. (2011b). "Arthritis and Nature's Joints". In J. K. Campbell, M. O'Rourke, and M. H. Slater (eds.), *Carving Nature at its Joints*, 199–230. Cambridge, MA: MIT Press.

Willis, B. H., Beebee, H., and Lasserson, D. S. (2013). "Philosophy of Science and the Diagnostic Process". *Family Practice*, 30: 501–505.

Wilson, S. and Adshead, G. (2004). Criminal Responsibility. In J. Radden (ed.), *The Philosophy of Psychiatry: A Companion*, 296–311. New York: Oxford University Press.

Windelband, W. ([1894] 1980). "Rectorial Address, Strasbourg, 1894". *History and Theory*, 19: 169–185.

Winston, A. P. (2000). "Recent Developments in Borderline Personality Disorder". *Advances in Psychiatric Treatment*, 6: 211–218.

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

Woodward, J. (2002). "What is a Mechanism? A Counterfactual Account". *Philosophy of Science*, 69: S366–S377.

Woodward, J. (2003). *Making Things Happen*. New York: Oxford University Press.

Woodward, J. (2008). "Cause and Explanation in Psychiatry: An Interventionist Perspective". In K. S. Kendler and J. Parnas (eds.), Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology, 132–184. Baltimore: Johns Hopkins University Press.

World Health Organisation (1992). *International Classification of Diseases: Classification of Mental and Behavioural Disorders*, 10th revision. Geneva: World Health Organisation.

World Health Organisation (2003). *Social Determinants of Health: The Solid Facts*, 2nd edition. Copenhagen: World Health Organisation.

World Psychiatric Association (2003). "Essentials of the World Psychiatric Association's International Guidelines for Diagnostic Assessment (IGDA)". *British Journal of Psychiatry*, 182: S37–S66.

Worrall, J. (2007). "Evidence in Medicine and Evidence-Based Medicine". *Philosophy Compass*, 2: 981–1022.

Wright, B., Dave, S., and Dogra, N. (2010). *100 Cases in Psychiatry*. Boca Raton, FL: CRC Press.

Young, S., Bramham, J., Gray, K., and Rose, E. (2008). "The Experience of Receiving a Diagnosis and Treatment of ADHD in Adulthood". *Journal of Attention Disorders*, 11: 493–503.

Zachar, P. (2000). "Psychiatric Disorders are Not Natural Kinds". *Philosophy, Psychiatry, and Psychology*, 7: 167–182.

Zachar, P. (2014). *A Metaphysics of Psychopathology*. Cambridge, MA: MIT Press.

Zachar, P. and Kendler, K. S. (2007). "Psychiatric Disorders: A Conceptual Taxonomy". *American Journal of Psychiatry*, 164: 557–565.

Zajac, J., Shrestha, A., Patel, P., and Poretsky, L. (2010). "The Main Events in the History of Diabetes Mellitus". In L. Poretsky (ed.), *Principles of Diabetes Mellitus*, 2nd edition, 3–16. New York: Springer.

Zarin, D. A., West, J. C., Pincus, H. A., and McIntyre, J. S. (1996). "The American Psychiatric Association Practice Research Network". In L. I. Sederer and B. Dickey (eds.), *Outcomes Assessment in Clinical Practice*, 146–155. Baltimore, MD: Williams & Wilkins.

Zimmerman, M., Ellison, W., Young, D., Chelminski, I., and Dalrymple, K. (2015). "How Many Different Ways do Patients Meet the Diagnostic Criteria for Major Depressive Disorder?" *Comprehensive Psychiatry*, 56: 29–34.