What is Mental Disorder? Developing an Embodied, Embedded, and Enactive Psychopathology

Kristopher Nielsen

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Psychology)

Victoria University of Wellington

2020

Abstract

What we take mental disorder to be has implications for how researchers classify, explain, and treat mental disorders. It also shapes how the public treat those who are experiencing mental disorder. This is the often-underemphasized task of *conceptualization,* which sits at the foundation of psychopathology research. In this thesis I consider the nature of mental disorder through the lens of a growing perspective known as *embodied enactivism*. Embodied enactivism is a philosophical position on human functioning that holds the mind to be: *embodied* (non-cartesian, and constituted by both brain and body), *embedded* (richly influenced by the physical and social environment across development), and *enactive* (meaning and experience arise through the precarious organisms' interactions with the world). After overviewing a selection of current conceptual positions – present either as independent conceptual frameworks or within our classification systems – I move to presenting my own conceptual framework of mental disorder grounded in an embodied, embedded, and enactive view. Some implications of this framework for the task of classification are explored, and a meta-methodological framework for developing explanations of psychopathology is developed. It is shown that the concept of mental disorder developed: moves beyond the internalist bias of many current concepts, recognizes the normative nature of disorder, encourages consideration of cultural and individual variance, does not unduly prioritize brain-level explanations of human behaviour, and can sit comfortably within a wholly natural world view.

Acknowledgments

I want to give a huge thank-you to my supervisor, Dr. Tony Ward, for his knowledgeable guidance and feedback throughout the development of this thesis. Thanks also to the entirety of the EPC lab for your support; this thesis developed in the encouraging and analytical environment that your actions foster. I of course can't get away without offering my sincere thanks to Dr. Emma Ashcroft and Róisín Whelan (nor would I want to!). I could quite truthfully not have got through this process without you. I would also like to thank all my other friends at university who have provided help in countless ways. Among these friends I want to give particular thanks to Dr. Emma Tennent for battling through a much-appreciated proof reading. To my partner, Alice Leader, your love and support outside of university has seen me through and I am forever grateful for your patience while I have been working on this thesis. Finally, I must obviously thank my parents. Everything I do stems from what you taught me as I grew.

Publications Included in Text

Nielsen, K., & Ward, T. (in press). Phenomena Complexes as Targets of Explanation in Psychopathology: The Relational Analysis of Phenomena (RAP) Approach. *Theory & Psychology*. Copyright © 2019, Sage Publishing.

Nielsen, K., & Ward, T. (2018). Towards a New Conceptual Framework for Psychopathology: Embodiment, Enactivism and Embedment. *Theory & Psychology*, *8*(6), 800–822. Copyright © 2018, Sage Publishing. https://doi.org/10.1177/0959354318808394

Nielsen, K., & Ward, T. (2019). Mental Disorder as both Natural and Normative: Developing the Normative Dimension of the 3e Conceptual Framework for Psychopathology. *Journal of Theoretical and Philosophical Psychology*. Online first. Copyright © 2019, American Psychological Association. https://doi.org/10.1037/teo0000118

Theory and Psychology: Permission given by Administrative Assistant, Susan Blades, following consult with Editor, via email (2nd October 2019).

Journal of Theoretical and Philosophical Psychology: Permission granted via RightsLink (license number: 4699510308255, 31st October 2019).

Table of Contents

Abstract 3

Acknowledgments 4

Publications Included in Text 5

Table of Figures 11

Table of Tables 12

Chapter 1: Conceptualization as a Core Task of Psychopathology Research 13

Some General Questions to Get Started. 17

Structure and Argument of this Thesis 19

Chapter 2: Current Conceptual Models of Mental Disorder and an Observation 23

Structurally Oriented Concepts 23

Non-kinds/continua. 25

Natural/essentialist kinds. 25

Discrete kinds. 28

Fuzzy kinds. 28

Normatively Oriented Concepts 34

Anti-psychiatric/deflationary positions. 36

Statistical functionalism. 37

Evolutionary functionalism. 41

Evaluative concepts. 52

Practical kinds. 55

Some Preliminary Observations 57

Concerning conceptions of human functioning. 58

The normative gap may be artifactual. 59

Chapter 3: DSM, RDoC, and Frameworks of Human Functioning 61

Assumptions of the DSM and RDoC 62

Diagnostic and statistical manual of mental disorders (DSM). 62

Research domain criteria (RDoC). 65

A Possible Way Forward 71

Chapter 4: Questions of Structure 73

Embodied Enactivism 73

Previous Work in Embodied Enactive Psychopathology 79

Previous 3e explanatory models. 80

Embodied Enactivism and the Structure of Disordered Behaviour 83

Chapter 5: Questions Concerning Normativity 89

Recent Views on the Role of Normativity 89

Sample debate in this area. 90

What can be learnt here? 92

Groundwork for an Embodied Enactive Approach 94

'Functioning well' under embodied enactivism and the DCT. 95

What is a functional/natural norm? 96

Cultural embedment and normativity. 99

What Then Counts as Mental Disorder? 103

Evaluating this position. 105

A possible objection. 107

Conclusions and Summary 112

Chapter 6: A Concept of Mental Disorder and Two Challenges 115

Integrating into a Fuller Concept 116

Getting More Precise 118

Comparing Conceptual Models 125

Structural models 125

Normative models 127

Two Challenges 129

Operationalizing adaption. 130

Managing holism. 132

Chapter 7: The RAP Approach to Explanation 135

Explananda in Current Approaches 137

DSM based approaches 137

Research domain criteria (RDoC) based approaches. 138

Symptom network modeling [SNWM] based approaches 140

Summary 142

Groundwork for an Alternative Proposal 142

The Relational Analysis of Phenomena (RAP) Approach 144

Phase 1: List and map. 145

Phase 2: Focus and enrich. 147

Phase 3: Explain and evaluate. 153

Summary. 156

Limitations and counter-arguments. 157

Conclusions and Summary 159

Chapter 8: Summing Up and Moving Forward 161

Another Enactive Perspective 161

Biases in sense-making. 163

Demarcating pathology in sense-making. 165

An existential transformation vs. cultural embeddedness. 167

The value of different perspectives. 174

The Thesis in Brief 175

Disordered Eating as a Summary Example 177

Key Implications 180

Classification. 180

Explanation. 181

Further Limitations 183

Appropriate use. 183

Falsifiability/explanatory value. 184

Applicability. 184

Returning to our Starting Questions 185

Conclusions 186

Table of Figures

Figure 1. A Four Stage Model of Psychopathology Research. 16

Figure 2. The Constitutional View of Culture [CVC]. 101

Figure 3. Visualization of a Phenomena Complex [PC]..... 153

Figure 4. Schematic Representation of the RAP Process. 156

Table of Tables

Table 1. Hypothetical example of multi-scale description looking at the phenomenon of hyper-vigilance. 151

Table 2. How various conceptual positions may relate to conceptual and explanatory approuches in the study of anorexia nervosa. 178

Chapter 1: Conceptualization as a Core Task of Psychopathology Research

Mental disorders demand the development of effective treatments and management strategies as soon as we are able. In their various forms they negatively affect the lives of hundreds of millions of individuals, and represent a significant proportion of the global burden of disease (Kessler et al., 2009; E. R. Walker et al., 2015; Whiteford et al., 2015; World Health Organisation [WHO], 2016). Even if not affected ourselves, the vast majority of us will know someone who carries the weight of a mental disorder with them.

In order to develop effective *treatments* for mental disorders, we should ideally be working from a good understanding of the problems we are trying to address; we must have good *explanations* of mental disorders. Further, due to their complexity, explaining mental disorders necessitates coordinated action by researchers around the globe. Before it will be possible to explain mental disorders effectively then, there will likely need to be a common set of labels and concepts that ensure that researchers are seeking to explain the same things; we must have a way of *classifying* mental disorders[1]. These three tasks – classification, explanation, and treatment – are often seen as the three core tasks of psychopathology research.

The task of *classification* is concerned with finding some degree of order in the tangled and complex range of behaviors and experiences that appear to be disordered, so that we may diagnose and study them effectively (Berenbaum, 2013; L. A. Clark et al., 2017; Zachar & Kendler, 2017). The current dominant classification system is the Diagnostic and Statistical Manual of Mental Disorders (DSM), currently in its fifth edition[2] (American Psychiatric Association, 2013b). As will be reviewed in chapter four however, psychopathology classification is at a conceptual crossroads. It is increasingly becoming accepted that fundamental flaws in the DSM's underlying approach are

[1] Classification may not necessarily involve developing a typology of diagnoses ('diagnostic kinds'). There are current arguments for shifting away from diagnostic kinds all together and focusing on a wider set of 'psychiatric' kinds and their complex relations. The Research Domain Criteria [RDoC] represents one such shift that will be discussed in this thesis, but there are other flavours to this shift away from diagnostic kinds. See Tabb (2016) for a review.

[2] While a competitor with the DSM, the International Classification of Diseases [ICD] largely parallels the DSMs content, but using a prototype model of description rather than a list of criteria. Many of the critiques of the DSM presented also apply to the ICD. Throughout this thesis I therefore largely ignore the ICD in the interest of simplicity and brevity.

resulting in it struggling to pick out 'real' mental disorders as opposed to artificially selected clusters of symptoms (Lilienfeld & Treadway, 2016; Zachar & Kendler, 2017). Alternative classification systems are being developed, and the DSM's continued position as the bedrock document of psychopathology is in serious doubt (Casey et al., 2013; Cuthbert, 2014; Insel et al., 2010). Theoretical work within the field of classification is currently asking important questions such as: should our diagnostic systems simply give labels to patterns of signs and symptoms, or try to map onto the causal structures underlying disorder?; should our diagnostic systems attempt to be theoretically neutral, or be open and honest with their theoretical commitments?; and, how should our diagnostic systems be responsive to their political and social purposes outside of diagnosis and research (Zachar, 2018)?

The task of *explanation* meanwhile, concerns the postulation and validation of theories that make the behaviors and experiences observed in mental disorders less surprising and more comprehendible (Haig, 2014). Whether grounded in neuroscience, psychology, or some other discipline, good explanations of mental disorder point to opportunities for intervention by tracking factors that either *cause* or *maintain* mental disorder. Current and historic attempts at explanation have resulted in limited success. To illustrate this point very briefly, compare current understanding of the causal processes involved in bio-medical illnesses such as the common cold or cancer, to prototypical mental disorders such as depression and schizophrenia. We may not have 'cures' for any of these problems, but at least within the bio-medical examples we have some idea what is going on. Comparatively, almost all mental disorders lack agreed-upon underlying causal structures. Aside from the development of actual explanations, (meta-)theoretical work in the area of explanation and philosophy of science more broadly is currently asking questions such as: what are the role of 'mechanisms' in explanations of mental disorder (Glennan & Illari, 2017; Hartner & Theurer, 2018; Thomas & Sharp, 2019)?; what exactly should we be trying to explain – i.e. disorder syndromes, symptoms, brain malfunctions, phenomena, functional processes, or something else entirely (Elbau et al., 2019; Hawkins-Elder & Ward, in press; Insel et al., 2010; Nielsen & Ward, in press; T. Ward & Clack, 2019a)?; are detailed explanations always better than general ones (Craver & Kaplan, 2018; Potochnik, 2016, 2017)?; and

most generally, how might we go about explaining things as complex and unknown as mental disorders (Insel et al., 2010; Kendler, 2008, 2012a; Murphy, 2017)?

Finally, the task of *treatment* involves the development and validation of efficacious interventions for mental disorder; either pharmacological, psychotherapeutic, or through some other means. While obviously important, this task is less relevant to the current thesis. Suffice to say, developing targeted and efficacious treatments, as well as improving on current treatment approaches, will be much easier when the earlier tasks have been performed well. Well considered classification systems and valid explanations will provide a strong foundation for the task of treatment.

Notably, there is often overlap between these three tasks of classification, explanation, and treatment. As a science, and as individual researchers, we are often shifting backwards and forwards between them. The founding observation of this thesis however, is that this three-task model of psychopathology is incomplete. The elementary yet missing question seems to be: What exactly is mental disorder? Before we can classify mental disorders – or explain and treat them – we must have some concept of what counts as a mental disorder. What we take mental disorder to be, either explicitly or implicitly, directly informs how we go about the tasks of classification and explanation. Our understanding of the nature of mental disorder is a metaphysical commitment that will bias how we go about designing studies and reasoning about their findings (Hochstein, 2019). Conceptual links between the nature of mental disorder and the tasks of classification and explanation mean that elucidating the nature of mental disorder will likely help address some of the mentioned questions currently plaguing these areas. There is, therefore, a need to bring our understanding to the surface and study it directly, so that we may be aware of its biases (Hochstein, 2019). The task of *conceptualization* then, while sometimes taken as merely part of the task of classification, is better thought of as its own endeavor (see figure 1). It is primarily within this task of conceptualization that this thesis is situated.

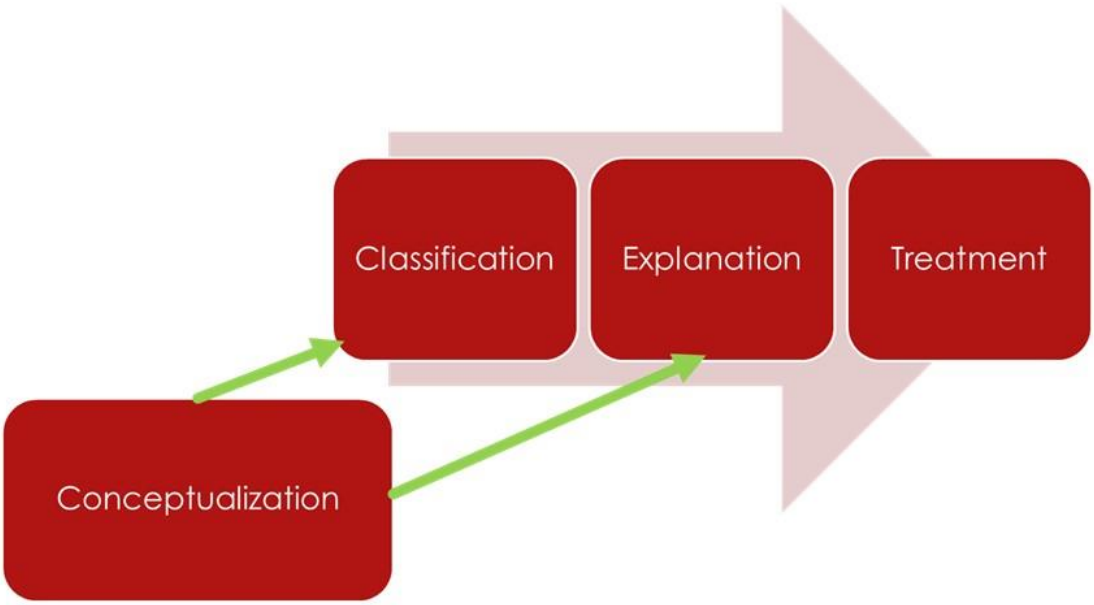

Figure 1. A Four Stage Model of Psychopathology Research. Each stage is represented by the red boxes. This thesis is concerned with the task of conceptualization (left), and its implications for classification and explanation. While receptive to work in later tasks, conceptualization is particularly important because it serves as the foundation for our efforts in classification, explanation, and treatment.

The sciences of psychopathology have a significant problem in the form of conceptual instability[3] (Sullivan, 2014). Simply put, current diagnostic labels capture large and highly variable populations (referred to as the problem of '*heterogeneity*'- this issue is discussed further in chapters three and seven). This means that researchers,

[3] Sullivan's broader work demonstrates that the current state of the mind-brain sciences *in general* is one of relative confusion and instability. Both our terms of reference and our methods often seem to pick out different phenomena across time, researchers, and disciplines (Sullivan, 2009, 2014, 2016b, 2016a, 2017). Arguably this shows the need for more rigorous conceptual work across the mind-brain sciences, not just in psychopathology. It is interesting to consider the conceptual difficulties within psychopathology as symptomatic of a wider lack of co-ordination in the mind-brain sciences.

even if they are all studying 'depression' for example, are actually often studying very different phenomena. This issue can be found across different areas of psychopathology (Contractor et al., 2017; Lilienfeld & Treadway, 2016; Monroe & Anderson, 2015; Olbert et al., 2014). There is therefore a need for greater co-ordination across the sciences of psychopathology, in terms of both the concepts and methods utilized, so as to stabilize the constructs under our study. This simply cannot be done without an understanding of what counts as mental disorder and what does not; classification is dependent upon conceptualization. The ideas in this thesis present one possible unified framework for understanding the nature of mental disorder. In chapter seven in particular, I develop a meta-methodological framework called the Relational Analysis of Phenomena ('the RAP'). The RAP formalizes some of the implications that the conceptual work of this thesis has for the task of explanation.

To be clear, the idea of focusing on the task of conceptualization is not itself novel. Much previous work has been done in this area, particularly with the development of conceptual models which I will review in chapter two. The claim I am making is that the value of conceptual work, as well as the pressing need for its continued development, is not sufficiently recognized in mainstream psychopathology, particularly by clinicians and empirical researchers. This will be apparent in chapter three when I review the conceptual paucity of the DSM, and some of the foundational issues plaguing the Research Domain Criteria [RDoC] – a funding system designed by the National Institute of Mental Health [NIMH] in America with the hope of developing an alternative or complimentary classification system to the DSM (Insel et al., 2010). Encouragingly, recognition of the need for good conceptual work does seem to be slowly on the rise, as seen in the emergence of 'philosophy of psychiatry' as an inter-disciplinary field over the last few decades (Fulford et al., 2013; Radden, 2006; Tekin & Bluhm, 2019).

Some General Questions to Get Started.

When people hear the term 'mental disorder', most seem to confidently assume that they know what this term means. A concern underlying this thesis is that this confidence may be somewhat misplaced. To briefly motivate recognition of the importance of this conceptual work then, it is useful to consider some fundamental

questions that those doing such work may try to answer. It is not my intention to provide answers here, although I do return to these questions in the closing chapter.

Firstly, are mental disorders something you *get* or something you *do*? In other words, does somebody 'have' depression or are they themselves depressed? This question is important because it has direct implications for how society and individuals, respond to someone experiencing/having/enacting a mental disorder. If a mental disorder is a disease or a lesion in someone's brain, the afflicted person is considered to have little control over it. It also then seems like the sort of thing that might be treated with *medication*. If, however, a mental disorder is something people do, the afflicted person is considered to have more control over their actions (Kvaale et al., 2013). Thus, they may be able to *learn to do things differently*, i.e. it is the sort of thing that might be treated with therapy.

Next, does a mental disorder exist inside someone's brain or is it dispersed across their brain, body and environment? For example, imagine someone is working in stressful conditions, and that this stress is maintaining their depression and anxiety. If they are taken out of this workplace, they may no longer be depressed and anxious. This raises the question, were they disordered or was their environment dysfunctional?

One final interesting question: are mental disorders defined by brute facts or by social norms and values? In the 1960s and 70s, psychiatrist and philosopher Thomas Szasz famously made the claim that mental disorder was a myth (Szasz, 1960, 1963, 1974). By this he meant that genuine 'disorder' is, by definition, medical/physical, thus leaving no space for disorders that are purely 'mental'. Rather, disorders with no physical basis are, according to Szasz, simply 'problems in living' and their medicalization a fantasy. For Szasz, this begged questions as to the function of this myth in society. Optimistically he considered whether this medicalization helped society believe in a naturally ordered state of life; one where significant problems-in-living are aberrations rather than the norm. Others in the anti-psychiatry movement however took a more pessimistic view, arguing that mental disorders are simply constructed labels for people that don't follow the unspoken rules of society, and viewing psychiatry as society's tool for dismissing those that refuse to conform (Foucault, 2003/1961). If, however, mental disorders are not based on social norms and values, instead picking out

real states or entities in the world, what exactly are they? Are genuine mental disorders required to be diseases or brain abnormalities, or may they be a different kind of thing entirely?

These are just some of the questions one might ask when considering the nature of mental disorder. In the following chapter I will review some proposed answers to such questions in the form of explicit conceptual models. This thesis is positioned in response to the need for greater focus on the task of conceptualization. I will also explore links to the later task of classification and explanation. In accordance with this purpose, the aim is to develop and argue for a novel concept of mental disorder and to explore ramifications for the later classificatory and explanatory tasks.

Structure and Argument of this Thesis

The underlying justification for this thesis can be broken down into three key points. Firstly, there are not yet good enough answers to questions such as 'what kind of things are mental disorders?' and 'why does a particular mental/behavioral phenomenon count as disordered?'. There is significant room for improvement in the conceptual understanding of mental disorders. Secondly, what we take mental disorder to be is conceptually related to our underlying assumptions about human functioning. In other words, if one understands humans to work in a particular way, then ones understanding of how humans can 'breakdown' is likely related to this. This raises the possibility that some understandings of human functioning might be more useful than others for generating understandings of mental disorder. Thirdly, the philosophical orientation known as *embodied enactivism* seems to be a good candidate for this role as a guiding framework of human functioning within psychopathology. This is because of its naturalistic orientation and its featuring of many useful conceptual tools.

In this thesis, I develop an understanding of what mental disorder is from the perspective of embodied enactivism. I argue that this perspective produces a rich and flexible understanding of mental disorder that can compete well against current popular approaches. I also point out two significant challenges that this approach will face if its full potential is to be met, and I provide one possible solution to one of these challenges (in chapter seven). Throughout this thesis, it is *not* my intention or purpose to argue for

embodied enactivism as a philosophy of mind, only its fruitfulness for considering the nature of mental disorder. Breaking this wider argument down by chapter, the thesis has the following structure:

In the current chapter I have introduced the general topic area and expressed the need for greater focus on conceptual work in the sciences of psychopathology. In chapter two I will explore current conceptual models of mental disorder and show that, while they all have their strengths, all have room for improvement. At the end of this chapter I argue that there is a need for a broader framework of human functioning in which to situate a concept of mental disorder. To put it simply, if we want to conceptualize *dysfunction* we must first formulate a concept of what it is to be *functional,* or otherwise not disordered.

In chapter three I review the underlying conceptual positions of the DSM and RDoC, exploring the implicit assumptions they make about human functioning. I overview some problems with the DSM conceptualization of mental disorder, and argue that, despite addressing some of these issues, RDoC has fundamental problems of its own. I end this chapter with a parallel claim to that of chapter two, that there is a need for a richer framework of human functioning in the sciences of psychopathology.

Chapter four begins the first novel and positive contribution of this thesis. First, I overview the position of embodied enactivism and argue that this position has potential to serve as the broader framework of human functioning needed. I then discuss past attempts to consider mental disorder from an embodied enactive view. I show that there are some problematic tendencies in this this area but highlight how we can learn from these attempts. Developing on from this I focus on what mental disorder is in a structural/ontic[4] sense when viewed through the lens of embodied enactivism. I argue that in terms of their structure, mental disorders can be seen as stable dynamic patterns of causal relations within the brain-body-environment system.

Chapter five focuses in on the normative domain rather than the structural, asking '*why* should something count as a mental disorder?'. Current normative

[4]'Ontic' is a term related to the term 'ontology' – it refers to the 'real' rather than the phenomenal or useful; to what exists.

perspectives on mental disorder are overviewed before I turn to how embodied enactivism can help answer this question. Here I argue that embodied enactivism, and in particular a component of it called the deep continuity thesis (DCT), contains an implicit commitment to natural normativity. I show how this allows for the development of an enriched systems functionalism which is able to successfully navigate many of the critiques that other normative models face. The main advantage of the perspective I develop here is that it allows the ascription of the disorder/dysfunction label to be made in the interest of the individual being diagnosed rather than on the basis of statistical, evolutionary, or societal norms.

In chapter six I combine the structural considerations of chapter four and the normative considerations of chapter five into a more complete model of mental disorder. By considering the structural and normative together, I argue that the embodied enactive view allows these complex patterns of causes we call mental disorders to be seen as fuzzy process structures within the agent-world system, working against the striving organisms attempts to adapt and self-maintain. I describe this concept in more detail using a conceptual taxonomy developed by Zachar and Kendler (2007). Using this taxonomy, I explore features of this concept such as how it simultaneously represents a realist and evaluativist position (i.e., through its system functionalism it holds mental disorder to be both a natural and normative phenomenon), and how, from this position, mental disorders may represent attractor basins in the human brain-body-environment system. I then make some comparisons to the most relevant of the conceptual models explored in chapter two, demonstrating how the embodied enactive concept holds its ground compared to extant models. At the end of this chapter I note two significant challenges that we face if we want to utilize an embodied enactive conception of mental disorder to its full potential. These challenges are: operationalizing the embodied enactive concept of 'adaption' (this challenge remains the most significant limitation of the thesis) and managing the holistic perspective that this concept demands.

Chapter seven then shifts away from considering the nature of mental disorder directly and attempts to respond to the challenge of managing holism within the task of explanation. I take the developed embodied enactive concept of mental disorder and ask

the question – 'given this view of what mental disorders are, what is it about mental disorders that we should seek to explain?'. I briefly overview some current approaches to identifying targets of explanation in psychopathology, before presenting the RAP. The RAP is a meta-methodological framework designed to support the development of explanations in accordance with the embodied enactive view of mental disorder (and views that share a similar structural perspective). The aim here is to explore the fruitfulness of the conceptual work in chapters four and five for thinking about the task of explanation; demonstrating that conceptual work can be useful for the development of more immediately practical ideas.

Finally, in chapter eight I summarize and draw conclusions. Firstly I overview another embodied enactive framework of mental disorder that I became aware of near the end of my time writing this thesis – de Haan (in press-b, in press-a, 2017) – and explore some of the differences between our frameworks. I argue that one relative strength of the framework developed in this thesis is its greater fertility for the task of explanation. I then bring the thesis full circle, summarizing the embodied enactive concept developed across this thesis and some of its implications for classification and explanation, as well as some of its limitations. I consider the benefits of my framework, but emphasize the need for continued conceptual refinement if the sciences of psychopathology are to progress.

Chapter 2: Current Conceptual Models of Mental Disorder and an Observation

In this chapter I review prominent conceptual models of mental disorder, commenting on their strengths and weaknesses. These are models that provide answers to the question 'what are mental disorders?'. Here I stick predominantly to the *formal* conceptual models – i.e. those presented as such. Models implicitly present in institutions and classification systems such as the DSM and the Research Domain Criteria [RDoC] are considered in chapter three. I have structured the presentation of these formal views in a way that highlights two different ways that we can understand the question 'what are mental disorders?'. I first present what I refer to as the *structurally oriented concepts*. These concepts focus on the nature of mental disorders in the ontic sense; on what mental disorders are in terms of their physical or causal structure. This is opposed to what I refer to as the *normatively oriented concepts,* which I present next. These normatively oriented concepts focus on why something should be (or should not be) considered a disorder. In closing this chapter, I make an observation as to a common need across most of the conceptual models discussed. A key role of this chapter is to demonstrate that while having a multitude of conceptual models at our disposal is useful (i.e. conceptual pluralism), this does not negate the need for conceptual refinement and the development of better models.

Structurally Oriented Concepts

Haslam (2002) presents a conceptual taxonomy that usefully organizes differing perspectives on the structural nature of psychopathology. Haslam ultimately argues for a conceptual pluralism, whereby different mental disorders are seen to likely have different structural natures; for example, that borderline personality disorder and bipolar disorder are not just different types of mental disorder*,* but different *kinds* of types, with the latter being much more homogenous and disease-like, and the former being much more heterogenous and socially weighted in its etiology. In accordance with this, Haslam sees pragmatic value in the plurality of structural views available, and his taxonomy is intended as a first pass attempt to collate the different kinds in a meta-structural way. He clusters the views under the labels: 'non-kinds/continua' (phenomena that don't form a kind but differ on a single spectrum, e.g.

colour/wavelength, neuroticism); 'practical kinds' (phenomena that can be clustered together because it is useful to do so, e.g., flying creatures, mood disorders); 'fuzzy kinds' (phenomena that can roughly be clustered together based on similarity even though all the instances aren't the same, e.g. board games, sandwiches); 'discrete kinds' (phenomena with no essences that can still be clearly identified as members or non-members most of the time, e.g., biological males[5]); and 'natural kinds' (phenomena with defined essences, e.g., atomic elements). I will unpack these labels further when discussing them below.

In this section I use an adaption of Haslam's (2002) taxonomy to organize my overview of the structurally oriented conceptual models. The key change I have made is that I have excluded 'practical kinds' from this section, instead discussing them in the following section on normatively oriented concepts. I give more room to the discussion of a fuzzy kind as this is a complicated concept which will be important in later chapters. I will further explain the differences between the kinds at the start of each sub-section. Note that all structural models discussed necessarily assume realism about mental disorders[6] (Kendler, 2016). Finally, I also note that the use of Haslam's taxonomy brings with it a focus on the degree of kindship/homogeneity of the underlying causal structures of mental disorder. This is as opposed to demarcating different conceptual positions by the etiological domains they emphasize (e.g. mental disorders are genetic diseases, neurological conditions, social problems)[7]. Where relevant I therefore point

[5] Biological sex is an arguable case of a discrete kind but is a good illustrative example in that it has no one essence, instead being composed of multiple related components (e.g., xx/xy chromosomes, hormone levels, internal and external physiology) that tend to bifurcate into male and female camps in *most* cases. This is not to deny the existence of intersex persons in anyway. One could also argue that biological males or females are examples of fuzzy kinds. I am less convinced that there is truly a clear demarcation between fuzzy and discrete kinds, but I include reference here to stay true to Haslam's taxonomy.

[6] 'Realism' refers to the view that there are ontic things in the world to which the label 'mental disorder' could refer, that these things, whatever form they take, are 'discovered' and exist independently of our attempts to classify them (i.e. they are not *entirely* socially constructed or pragmatic). I briefly discuss social constructionism and pragmatism in the following section on normatively oriented concepts. Socially constructed kinds could possibly be discussed in this section as, while they are constructed, they still have an ontic reality in the form of a pattern of behaviour (Mallon, 2016); for example see the controversial socio-cognitive model of dissociative identity disorder (Gleaves, 1996). I cover socio-constructionist models in the normative section due to their association with anti-psychiatry.

[7] By discussing two separate ideas/dimensions in proximity I risk conflating them here. The idea of a continuum of homogeneity (simple/essentialist – complex/emergent) and the idea of a 'continuum' of etiological domain (biological-social) are in fact separate ideas that are often conflated (although it is interesting to consider if there is actually a possible relationship between these dimensions). Also note

out recognized conceptual positions that are not only committed to a particular degree of homogeneity, but also to the primacy of particular etiological domains (e.g. biological essentialism, biopsychosocial holism).

Non-kinds/continua.

Haslam (2002) begins his taxonomy with a category that captures those concepts in psychiatry that *do not* count as kinds, i.e. things that are completely continuous and are therefore *non-kinds* or '*continua'*. A good example of a non-kind is neuroticism. There is no non-arbitrary level of neuroticism at which someone counts as 'neurotic' or not, rather people can be more or less neurotic, with no clear 'tipping point' at which one can be labeled. Neuroticism therefore is a case of a pure continuum rather than a kind.

Most concepts across psychology research are continuous in a certain sense. This also includes many diagnostic concepts, for example someone can be more or less depressed; depression comes in degrees. However, this level of continuity is subtly different to a non-kind where *no* meaningful point of demarcation or tipping point between members and non-members of the class is assumed to be present. There are few conceptual models of disorder that subscribe to this radical continuity, with most models assuming at least a fuzzy degree of categorical kindship across members of a class. The exceptions to this are some of the *practical kind* models which I will discuss in the section on normatively oriented concepts.

Natural/essentialist kinds.

Haslam (2002) draws a distinction between *natural kinds* proper and *discrete kinds* (which I will discuss next). Within his taxonomy, natural kinds have a clear common causal structure; a single 'latent variable', or 'essence' underlying them. From philosophy, the classic example of natural kinds in this strict sense are atomic elements which are clearly defined by the number of protons present, for example, gold always has seventy-nine protons while helium always has two. When referring to this kind

---

that the idea of particular mental disorders existing at *one place* on a organic-to-social continuum is a strongly criticized idea, mental disorders from schizophrenia to borderline personality are better seen as "dappled" across this spectrum, each with mechanisms at a variety of scales (Kendler, 2012b).

notion, I prefer to use the term *essentialist kinds*. The reasons for this choice of terminology are multiple. Firstly, my general use of the term 'natural kind' is a lot broader than Haslam's (2002) use. My use of 'natural kind' refers to a kind concept that picks out something real as opposed to conventional, selecting out a class of things which share properties to the degree that labeling them can be useful for our scientific purposes (i.e. correct application of the label to a thing allows for inductive inference as to other properties that the labeled thing may hold). This conception therefore encompasses both strictly natural and discrete kinds in Haslam's terms[8] (and even many 'fuzzy' kinds). Secondly, there is a lot of controversy over what authors actually mean when the term 'natural kind' is used, with some uses signaling a restrictive essentialist concept as in Haslam's taxonomy, and others a more open concept like my general use of the term (Bird & Tobin, 2018). Finally, sometimes there can be difficulty with the use of the term *natural* kind regarding whether such a concept can encompass social or mental phenomena. Rightly or wrongly, one criterion often discussed concerning natural kindship is that of 'mind independence'[9] (Khalidi, 2013). This is seemingly due to a false dichotomy intuitively drawn between what is 'natural' versus 'human' and can produce some difficulties when studying mental and social phenomena such as mental disorders.

Current conceptual models that propose mental disorders to be essentialist kinds tend to be those that model mental disorders on physical disorders, so called *biological essentialism*. These approaches assume that there are yet to be discovered biological disease processes or abnormalities underlying mental disorders. When uncovered, such biological lesions will reveal that mental disorders are essentially physical disorders (presumably of the brain) that manifest mental and behavioral symptoms. The idea is that revealing these latent biological variables will allow for clear and etiopathologically valid categorization. A structural conceptualization such as this can be implicitly seen in

[8] My orientation here is parallel to a natural kind position argued for by Boyd (1991) and by Magnus (Magnus, 2014b, 2014a), whereby some, but not all, natural kinds are MPCs (which will be discussed when covering fuzzy kinds).

[9] Khalidi (2013) offers a discussion of this issue, arguing for a shift away from mind independence as a criterion for natural kindship and toward consideration of whether a kind is categorized together based on causal relation/similarity versus categorized together as a matter of convention. Many social kinds (war, money, racism) can indeed be natural despite their mind dependence.

explanatory theories such as the – now highly contested – serotonin hypothesis concerning depression. This theory holds that depression is essentially a dysfunction in the serotonergic systems of the brain (Albert et al., 2012; Gardner & Boles, 2011). More explicitly, such essentialist conceptions can be seen in the work of authors like Insel and Cuthbert (2015), who – on the basis of the success of 'precision medicine' in areas such as oncology, where genotyping and targeting of specific cancer sub-types is becoming more common – argue for the need to make our diagnostic categories more precise. Up until this point Insel and Cuthbert's arguments represent a reasonably consensus view (as I will show in chapter three our current diagnostic categories are hopelessly heterogeneous). The essentialist (and theory-reductionist[10]) step these authors take is their next one, where they argue that the only way to achieve such precision is through adopting a biologically focused model of psychiatry; a model in which mental disorders are simply brain disorders with behavioral, cognitive, and emotional symptoms. Implicit in this step is the idea that, when it comes to mental disorders, the brain is where the money is; that there are undiscovered neurological essences to what we label (wrongfully in their mind) *mental* disorders[11]. These authors are part of the Research Domain Criteria [RDoC] project which I will return to in chapter three. Notably, biomedical notions of mental disorder seem to be gaining in popularity, both within psychopathology and with lay people (Lebowitz & Appelbaum, 2019).

Biological essentialism is not the only kind of essentialist position one could take in regard to mental disorder. For example, psychoanalytic approaches to the explanation of mental disorder represent an essentialist approach, but with the dominant latent variable being some underlying psychological factor (a 'neurosis'), rooted in past experience. The neurosis here, is in effect acting as a psychological essence and could therefore be termed a form of *psychological essentialism*. To use a more mainstream example, *cognitive models* of psychopathology – those that hold mental disorder to boil down to errors or biases in thinking – can also be understood as

[10] 'Theory-reductionism' is the view that the different domains of science can be reduced to the more 'fundamental' sciences, i.e. that psychology is applied biology, is applied chemistry, is applied physics, is applied maths.

[11] Another component of their argument is the need to unclip research efforts from current diagnostic categories. This is a point I agree with and will be covered more in chapters six and seven which are more focused on explanation.

examples of psychological essentialism. For example, think of therapists that utilize Cognitive Behavioral Therapy [CBT] with clear emphasis on the cognitive over the behavioral. Such therapists see behavioral interventions only as a tool to shift problematic patterns in cognition (to use a common turn of phrase, they do CBT with a capital 'C' and a small 'b'). Such therapists are implicitly taking a psychological essentialist position. Beck and Bredemeier's (2016) unified cognitive model of depression is a good example of a theory that also falls under this conceptual position. For the most part however, the idea that mental disorders are essentialist kinds tends co-occur with the idea that the essences in question lie within the brain.

Discrete kinds.

Haslam (2002) uses the term *discrete kinds* to distinguish things that feature clear membership conditions, but that – in contrast to essentialist kinds – are not defined by a single causal factor or essence. Instead, discrete kinds have complex underlying causal structures, but due to the dynamics of the causal structure in context they bifurcate into members and non-members of the kind. Thus, discrete kinds still produce a clear boundary with very few ambiguous cases. Haslam (2002) gives the example of melancholic depression. This is a diagnostic concept, present in the DSM-5 as a sub-type of depression, featuring dominant anhedonia and vegetative symptoms. Haslam cites taxometric evidence that melancholic depression is clearly categorical in nature but notes that this does not necessarily imply the existence of an underlying essence, instead arguing that this may be an example of a discrete kind. This is unfortunately the only diagnostic example Haslam mentions, and the concept of a discrete kind has not, to my knowledge, been picked up by other authors. It is also not clear what categorically separates a discrete kind from an essentialist kind with a particularly complex essence (or alternatively a reasonably homogenous MPC kind, discussed later). I mention it here as it remains an interesting idea, and to be true to Haslam's taxonomy.

Fuzzy kinds.

Fuzzy kinds are real and objective categories that exist in nature and are thereby very different to non-kinds/continua or practical kinds (discussed later). However, the

point of demarcation between what is and isn't counted as a token of the kind is blurry, or rather 'fuzzy'. Rather than a single tipping point, or 'joint' in nature, that separates members of a fuzzy kind and non-members, there is a *zone of ambiguity*; a gentle curve of demarcation rather than a defined point. Fuzzy kinds then, represent "real, discoverable discontinuities" in the world (Haslam, 2002, p. 208), and are therefore not non-kinds, but do admit to intermediate cases. As an example, the concept of a 'teddy-bear' is meaningful. There are clear cases of objects that are teddy-bears such as Mr. Bean's 'Teddy', and there are clear cases of objects that are not-teddy-bears such as my foot. However, there are also in-between cases such as a soft-toy Koala. Koalas are not proper bears yet are sometimes referred to as such. If I showed a soft-toy Koala to a selection of people, some would categorize it as a teddy-bear and some would not. But this does not mean that there is no meaningful difference between teddy-bears and other objects. Teddy-bears can therefore be said to be fuzzy, not just because of their texture, but because they admit ambiguous membership. It is important to note here that it is not the fact that people have difficulty identifying the members of a kind in itself that makes the kind fuzzy, but rather its *actual* in-between status. I am talking here about ontological fuzziness rather than epistemological fuzziness.

Interestingly, some concept being fuzzy suggests that the causal structures underlying the phenomena referenced by the concept are reasonably complex (Haslam, 2002). If some phenomenon is supported by a single causal factor or 'essence' then its identity tends to be clear cut (i.e., discrete or essential kinds). For example, a given atom either is an example of gold or it is not (depending on a single factor: the number of protons present). For fuzzy kinds, the existence of borderline cases suggests that more than one 'defining' factor is at play. For example, what counts as a teddy-bear is dependent on not just one factor but many: does it have a snout, is it cute, is it squishy, does it have round ears? While 'teddy-bear' is still a meaningful category, soft-toy Koalas also exist with enough of these properties to be meaningfully akin to teddy-bears, but to not quite be 'proper' teddy-bears. If a mental disorder (e.g. depression) differs meaningfully from both normality and other mental disorders (e.g. anxiety), yet there are messy in-between cases (e.g. anxious-depression, or people who are just a little bit

depressed) then the fuzzy kind label may be appropriate[12]. When considering mental disorders this idea seems appealing given that such a messy reality is exactly what we find; i.e. high rates of apparent artifactual co-morbidity and diagnostic ambiguity (Andrews et al., 2002; Lilienfeld & Treadway, 2016).

Given this association with complexity, a position intuitively associated with the idea of a fuzzy kind is the biopsychosocial movement (Borrell-Carrió et al., 2004; Engel, 1977). This movement is a broad approach to health and wellbeing, born in reaction to the growing biological reductionism of medicine in the middle of the twentieth century. Originally proposed by Engel (1977), the biopsychosocial movement emphasizes the need for holism, and the need to recognize that mental disorders (and physical disorders) generally arise from, or are influenced by, complex non-linear interactions between multiple factors, and that these factors range across different scales of analysis (from molecular to socio-cultural). The movement also emphasizes a congruent focus on the person above and beyond their disease during patient-professional interaction. Despite the value and importance of this approach, considering the biopsychosocial movement as a structural model of mental disorder is problematic. Doing so may seem like an attractive option. This is because the biopsychosocial movement is thoroughly anti-reductionistic and encourages broad and agentic considerations. Considering the structure of mental disorder through the biopsychosocial lens may therefore bring certain ethical advantages, perhaps producing a more compassionate psychiatry that is more mindful of the person-as-a-whole, rather than simply the mechanics of their disease processes. However, the only structural commitment this approach really makes is to the general facts that 1) factors across the different scales of analysis are likely relevant, and that 2) these factors may interact in complex ways. This is in no doubt true, certainly there is a need to recognize the complexity at hand. The problem here is that, in making no firm commitment to the nature of these interactions above and beyond their complexity, the biopsychosocial movement offers very little guidance for attempts at classification, explanation or treatment, other than to 'look at *all* the things' (Ghaemi, 2009). Considering this, it is not clear if there *is* such a thing as 'the

[12] The difficulty here is ruling out other possibilities such as anxious-depression being something different all together, or depression simply being radically continuous (i.e., a non-kind).

biopsychosocial model of mental disorder', or whether such a reference is better thought of simply as a call to widen our perspective and consider the complex reality of the phenomena we call mental disorders.

One structural model of mental disorder that puts the fuzzy and biopsychosocial ideas to work with greater specificity, is the view that mental disorders are *mechanistic property clusters* or 'MPC kinds'[13]. This model was applied to mental disorder by Kendler, Zachar and Craver (2011), building upon the philosophical work of Boyd (1991). MPC kinds are constituted by clusters of properties held together or caused by a mutually reinforcing *network of mechanisms*. For example, the kind 'sheep', in being a biological species, is often assumed to be a meaningful and categorical kind. But what makes a sheep a sheep? Well, for one, sheep are wooly, and have four legs. One problem with this answer is that if I have a three-legged sheep and shave it bare, it still seems like this poor creature, no matter its condition, is still a sheep in a meaningful sense. The properties of being wooly and having four legs then, don't seem to be the 'essence' of what it means to be a sheep. Boyd's answer to this problem was to change tack; not to look for the 'essence' of the sheep – the 'necessary and sufficient conditions' that define a sheep – but rather to propose that what makes a sheep a sheep is the fact that all sheep share an evolutionary lineage, representing overlap in the causal structures that led to any one sheep's existence. A slightly different example, given by Magnus (2012, 2014a, 2014b), would be pools of water. Pools of water do not necessarily share a causal lineage, e.g. a pool of water may form here on earth, as well as on a completely different planet. However, a very similar causal process underlies their formation (e.g. an affinity between H_20 molecules due to their dipole structure, processes of condensation, some process of containment). The mechanism (or set of mechanisms) that leads to the formation of such pools is the same or features significant similarity. Cases such as these are referred to as *type-causal* MPCs because the underlying causal pattern occurs multiple times; it is a 'type' of causal pattern that leads to members of the kind sharing properties. The previous example of a biological taxon (a sheep) is referred to as a *token-causal* MPC because there is a single causal cascade (in this case an evolutionary

[13] Following Boyd (1991), the philosophical terminology is homeostatic property cluster (HPC), but here I use Kendler et al.'s label (MPCs) as this is conventional in the psychopathology literature.

history) shared by all members and leading to their overlapping properties (Magnus, 2012, 2014b, 2014a).

On this MPC view then, mental disorders are fuzzy sets of properties (i.e. properties of people, presumably signs and symptoms) and a network of causal mechanisms that holds these properties together in a wider possibility space (Kendler et al., 2011). This causal network may consist of the symptoms themselves, as well as underlying states and processes. Importantly, the factors playing a role in this causal network may cross boundaries of scale – evolutionary, physiological, psychological, social, etc. – with no *a priori* privilege given (Kendler et al., 2011). Kendler et al. also highlight the flexibility of this position, leaving room for more or less homogenous MPC kinds:

> "In the limit of simplicity and determinacy, MPCs tend toward essences, with properties and mechanisms common to all and only members of the kind. At the other extreme, cluster kinds tend toward constructed or practical kinds, where the boundaries of categories are often defined with respect to the classificatory practices of some interested party." (Kendler et al., 2011, p. 1146)

Note that more homogenous MPC kinds would likely be captured by Haslam's concept of a discrete kind (Haslam, 2002). The MPC concept is therefore very flexible in its reference.

The MPC view is currently popular when considering the structural nature of mental disorder. It offers a possible reason why no dominating causal factors or clearly defined causal networks underlying any modern mental disorders have been found. Mirroring the study of physical disorder and disease, it has been historically assumed that the discovery of a such 'essences' is the ultimate goal of psychopathology research. The MPC view, and other such 'fuzzy' models, suggest that maybe the reason we are failing to find such essences is that they simply may not exist. Fuzzy models allow us to consider this without giving up on kindship altogether, instead suggesting that mental disorders may be different to many physical disorders, not just because they concern behaviour and 'the mind', but because of their complexity. In other words, that they may be heterogeneous categories with no definable essence but that meaningful and useful

patterns can still be found. The major issue facing the MPC and other fuzzy views is parallel to that faced by the biopsychosocial approach. If we recognize this degree of complexity, where do we start? Will some scales of analysis be more useful than others? Which mechanisms should be focused on? Despite being more specified than the biopsychosocial approach, the MPC view still does not offer much *guidance* in this respect. As will be seen in later chapters, the concept of mental disorder developed in this thesis is structurally very similar to an MPC view. The perspective developed attempts to address this issue with guidance, not by prioritizing any scale of analysis *a priori*, but through consideration of the normative dimension of mental disorder and its intersection with the structural.

Before moving on, one currently popular idea that attempts to put the notion of an MPC to work is that of the Symptom Network Model of mental disorders [SNWM]. The SNWM approach assumes that many mental disorders are best understood as *networks of symptoms,* which can be statistically modeled. Symptoms within these networks are hypothesized to cause each other, with recursive feedback resulting in the relative stability of the network over time (Borsboom et al., 2018; Cramer et al., 2010; McNally, 2016). Recent years have seen a significant increase in SNWM research, with many examples being used successfully in empirical studies (Fried et al., 2017). This approach is presented by its proponents as a radically new way of conceptualizing psychopathology; as a model of mental disorder that rejects the search for underlying cause/s of psychopathology, i.e., the essentialist or latent variable model (Borsboom et al., 2018). However, there is considerable debate over whether this is the case, or whether SNWM is simply a new and promising measurement tool (Bringmann & Eronen, 2018; Epskamp et al., 2017; Fried & Cramer, 2017; Haig & Vertue, 2010; Humphry & McGrane, 2010; Molenaar, 2010; T. Ward & Fischer, 2019). These concerns seem warranted, especially given that, conceptually, the SNWM seems very much like an MPC model that restricts itself to the level of signs and symptoms. I will return to this idea in more detail in chapter six, where I consider SNWM as a valid and interesting approach to modeling and attempting to explain mental disorders, rather than as a novel conceptual model of what mental disorders are. I will now shift to overviewing a selection of normative conceptual models.

Normatively Oriented Concepts

The conceptual models covered in this section focus on *why* something should be considered a mental disorder and are mostly not covered by Haslam's (2002) taxonomy as this was oriented predominantly towards structural concepts. Another way to think of these normatively oriented models is that they try to provide understandings of mental disorder with 'conceptual validity' (Wakefield, 2014a). Conceptual validity refers to the ability of a concept or framework to correctly distinguish between 'normal' functioning on one side and *dis*order, *dys*function, or pathology, on the other[14]. The use of 'correctly' here comes from Wakefield's definition and I take it to be synonymous with 'well-reasoned/justified'. To label someone's thoughts and behavior's as 'broken' or 'bad' in anyway invites stigma and has a huge impact on people's lives and self-understandings. As the arbiters of such labels, psychiatry and clinical psychology need explicit ethical guidance, a necessary part of which is a clear understanding of what counts as mental disorder and what doesn't. For this and many other reasons[15], the conceptual pluralism prescribed when discussing the structural nature of mental disorder can seem less applicable when discussing the normative nature of mental disorder. By this I mean that if we are going to label someone as 'dis'-anything, we ought to be able to provide good reasons for doing so, and we ought to seek to be correct in making this distinction (whatever that may turn out to mean).

Even if there is 'one correct' way to understand the normative nature of mental disorders, conceptual pluralism may still be the best way forward given the complexity at hand. Fulford and Colombo (2004) give the analogy of a complex mural on the wall in a dark room, with the mural representing the 'correct' concept of mental disorder. There are six people in the room and each one is given a flash-light. The beam of each flash-light, through taking a different perspective, reveals a different facet of the mural. With enough flash-lights we may hope to perceive the entire mural, but each individual flash-light likely has value in this task. I would add to this however, that given the ethical

[14] This is not to pre-suppose a categorical difference. In fact the divide seems likely to be continuous.

[15] See Telles-Correia, Saraiva, and Gonçalves (2018) and Wakefield (1992a, 2007)for discussions surrounding the need for a precise definition. Contrarywise see Bingham and Banner (2014)

weight of our task alluded to above, critical care is required; we need to make sure that someone isn't pointing their flash light at the wrong wall.

In what follows I overview some of the conceptual models offered as justification for use of a mental disorder label, or those that attempt to offer guidance as to what should count as mental disorder. It is not my intention to cover all normatively oriented models available as this is not a comprehensive review. For example, I do not cover models that see mental disorder as an entirely moral or religious concept, not do I cover those reason-based models that see mental disorder as defined in some way by irrationality[16] (Graham, 2013; Megone, 1998). I also do not cover Roschian models that hold mental disorder to be a multi-dimensional cluster concept, centered around a prototype rather than necessary and sufficient conditions[17] (Lilienfeld & Marino, 1995; M. J. Walker & Rogers, 2018). I focus instead on families of conceptual models that are currently or recently popular, and that together offer the reader a general overview of the conceptual landscape. I first briefly cover anti-psychiatric or *deflationary* positions as these historically provided the impetus for the development of the other models in this section. I then cover *statistical functionalism*, followed by *evolutionary functionalism*. I give extra room to discussing evolutionary functionalism as it is quite a popular position and the critiques of this position are reasonably complex. I then discuss *evaluative* concepts, and finally *practical kinds*. Note that some of these normatively oriented models draw from the philosophy of medicine, and are often concerned with disorder, dysfunction, or disease in general rather than just mental disorder. Because of this I occasionally draw on examples across both physical and psychiatric medicine.

---

[16] Briefly, my key issue with these reason-based-models is that they commit to an understanding of the 'rational man' as an ideal from which to contrast disorder. This seems very culturally specific, and it seems there is a risk that this may illegitimately pathologize cultural variance. Megone's (1998)model in particular is also reliant on unfavourable ideas such as Aristotelian teleology (final causes as a function of essence), and human exceptionalism (the idea of a unique and vital difference between humans and animals).

[17] Briefly the issue with these Roschian/Wittgensteinian models is that they are overly flexible, thereby providing very little specificity or guidance. This is a similar weakness to the pragmatic concepts that I will discuss. I will briefly return to Roschian models in the final chapter.

Anti-psychiatric/deflationary positions.

In overviewing understandings of what makes mental disorder 'disordered', it would be remiss to not highlight those views that hold the label of disorder to be unjustified and/or unethical. Because of their use by persons and groups opposed to the institution of psychiatry through the latter half of the 20th century, these positions are often referred to as *anti-psychiatric*. However, 'anti-psychiatry' is quite a loaded term, and it is important to distinguish between opposition to psychiatry as a whole, and principled disagreement with the concept of mental disorder. For these reasons it may be better to refer to these positions as *deflationary*. These deflationary positions are responsible for much of the debate concerning the normative justification for the mental disorder label as they represent the null hypothesis: that in important ways the label 'mental disorder' fails to refer to anything in nature.

The psychiatrist and philosopher Thomas Szasz is responsible for the most famous of these deflationary positions (Szasz, 1960). The core of Szasz's position is that real illness or disorder is necessarily a bodily phenomenon. If this is assumed, then the category 'mental disorder' seems problematic. What we refer to as mental disorders will either turn out to have a physiological cause – and thus be disorders of the brain or body – or they will turn out to have no basis in the body, and therefore not qualify as genuine instances of illness/disorder. For Szasz then, 'mental' disorder is an impossibility and our use of the term must be a 'myth'. While, in public discussion, Szasz is often implied to be some sort of radical social constructionist, his issue with the concept of mental disorder actually stems from a position of biological disease realism. Szasz's use of the word 'myth' is very intentional and has a double meaning. On one side he is referring to the apparent impossibility of *mental* disorder (as explained), and on the other he is speculating that we use the notion of mental illness/disorder to distance ourselves from the harsh realities of our society. The idea here is that the labeling of genuine but normal 'problems in living' as medical issues, and thereby as uncontrollable deviances from the norm, allows us to believe that the society we have constructed is kinder than it really is.

Another famous deflationary position is that of the philosopher Michel Foucault (2003/1961). Foucault's study of the development of the concept of madness in Europe

lead him to the conclusion that the modern label of mental disorder is primarily a label for social deviance, and a tool for controlling those whom society disvalues. While we have come to see a categorical difference between those that suffer mental disorder and those that do not, Foucault's analysis suggests that such objectification of these differences has in part arisen because of the way we have historically separated those viewed as 'mad' – alongside political dissidents and criminals – from the rest of society through the practice of institutionalization.

While neither of these views are currently popular in the mainstream psychopathology literature[18], it is somewhat unfair to say they have failed simply because the institution of psychiatry still stands. Many of the normatively oriented concepts I will explore in this section were conceived of as responses to the concerns of these deflationary positions. These deflationary views helped to highlight why the sciences of psychopathology need a strong conceptual base, including a principled reason to demarcate the disordered from the benign. Without such a reason, those of us currently working with mental health diagnoses are practicing on the basis of a non-natural and/or unjustified conceptual framework. In other words, these deflationary positions demonstrate that without a convincing positive understanding of what mental disorders are, psychologists and psychiatrists lack sufficient ethical justification for their practice.

Statistical functionalism.

One common understanding of what counts as mental disorder is that it has something to do with deviation from the statistical norm. This view is apparent when we use the term 'abnormal psychology' as synonymous with 'dysfunctional' or 'disordered' psychology. Unfortunately, by itself such a view does not get us very far. This is because it cannot distinguish between 'good' and 'bad' forms of abnormality, e.g. being abnormally good at maths or abnormally good at giving speeches does not seem to count as a mental disorder. For this reason, conceptual models of what counts as mental disorder based around typicality have to further specify what kind of abnormalities or

[18] Such views are expressed elsewhere in academia. One notable example from within psychopathology is the Power Threat Meaning Framework (Johnstone et al., 2018) which takes a similar deflationary perspective on mental disorder.

typicalities are relevant to disorder and why. *Functionalism* of some stripe or another often fills this position and will be discussed in the current section. In the following sections I will also discuss models that use *values* or *pragmatics* to fill this space.

The most well-known position of the *statistical functionalist* variety is the Bio-Statistical Theory of Health (BST) developed by Christopher Boorse (1975, 1977, 2014). This is a conceptual model of health and 'disease' in general but can be used to inform a view of mental disorder. Under the BST, a disease is an *internal* state that impairs health by bringing about reduced efficiency of so-called *normal functions* relative to a *reference class*. Reference classes are members of the same species, sex, and age group[19], thus making normal functions effectively things that others like you can do that contribute to survival or reproduction (Boorse, 1977; Nordenfelt, 2007). If you go bald at the age of 13 while other teenaged humans of your sex do not, then this would count as disease under the BST (so long as hair can be assumed to serve a biological function such as keeping the sun off your head and/or helping to attract mates). The general gist of the BST is that "diseases are internal states that interfere with functions in the species design" (Boorse, 1977, p. 558). Boorse developed this concept to be explicitly value-free; as a concept that sees diseases as empirical facts rather than value-based distinctions[20]. For Boorse then, 'disease' is a theoretical/technical concept and should be distinguished from a more general sense of 'illness' which he does see as value-laden[21]. In other writings he has used the alternative term 'pathology' to refer to disease/disorder (Wakefield, 2014b).

---

[19] Boorse indicates that ethnicity should sometime be considered in-so-far as the differences in functional design across ethnic groups are relevant (Boorse, 1977).

[20] Both Kingma (2007) and Varga (2011) counter Boorse's claim that the BST is in fact value-free by pointing that the use of sex, age, and ethnicity to define the reference class is not itself based on empirical fact but on intuition, and thereby is likely importing value into the process. For example, one common criticism of the BST is that is seems to define homosexuality as a disease on the basis of its statistical deviance and the resulting lower rates of reproduction. Kingma points out that the addition of sexual orientation to the defining attribute of the reference classes would change this entirely. Those that include sexual orientation in the reference class selection would view homosexuality as entirely normal, and those that do not would view it as a disease. Really the BST is only potentially value-free post the selection of a reference class.

[21] Fulford (2001) criticizes the BST, for one arguing that, even if it does produce an internally consistent value-free concept of disease it fails to recognize that the term 'disease' is *used* evaluatively, even by Boorse himself.

While he does not make it a focus of the theory it is important to note that Boorse (1977) limits the kinds of things that can count as diseases under the BST to inefficiencies/difficulties with *physiological* functions. Thus I refer to the BST as an example of *physiological statistical functionalism*. For example, someone with abnormally high blood pressure relative to a standard developed by measuring the blood pressure of others of the same sex and age could be said to have a disease (hypertension) under the BST, whereas someone with abnormally low empathy would not *necessarily* be seen to have a disease under the BST. In order to be seen as diseases under the BST an assumption has to be made that abnormal mental conditions are causally supported by an abnormal physiological structure (usually in the brain). On this view then, mental disorders are not 'mental diseases' but rather physiological diseases, not yet understood, that happen to feature mental and behavioral outcomes (hence why they are sometimes referred to as 'disease models'). The BST, and other (*physiological*) *statistical functionalist* views (such as: Reitschel (2014), the RDoC movement – see chapter three), are typically associated with a clearly categorical or even essentialist structural view, whereby mental disorders are assumed to have a yet to be discovered dominant causal factor or essence. It is this exclusion of the possibility of independent mental dysfunction/disorder (mental difficulties without a physiological abnormality as a basis) that opens such views to charges of reductionism.

Not all views that could be labeled as varieties of statistical functionalism are restricted to physiological deviations. For example, Bergner (1997, 2004) – continuing the original work of Ossorio (1985) – proposes a *disability concept* of mental disorder[22]. A key part of their definition is that mental disorder involves significant restriction in a person's ability to engage in deliberate behaviors that that they *ought* to be able to engage in. Regarding this use of 'ought', Bergner (1997) explains that 1) this is purposefully ambiguous in order to accommodate clinical judgement, but also that 2) the idea is that the behaviors one 'ought' to be able to engage in are specified in a sense that is "highly developmental and highly contextual" (p. 240). The essence of what Bergner is claiming seems to be that mental disorder concerns *deliberate behaviors that*

[22] For further (empirical) support of this disability view see Bergner & Bunford (2017), for a critique see Wakefield (1997c).

others can typically perform but that the sufferer cannot, while excluding any such restrictions on behaviour that can be explained in reference to contextual factors (e.g. age, culture, immigrant status, physical environment). Direct parallels are clear here to the BST and the idea of relativizing disease to a reference class (although the 'reference class' in this model is much more specific). It is for this reason that I consider Bergner to be proposing a form of *behavioral statistical functionalism*[23].

The key difficulty with statistical functionalism applied to mental disorder can be summed up by the question 'why should being normal matter?' In both varieties of statistical functionalism espoused here, the typicality of some state or action is used to infer that this state is the way that our bodies *ought* to be, or that this action is the way we *ought* to act. Problematically, the link from the 'is' of the statistical norm, to the 'ought' of claiming that a biological state of affairs is *better or worse* than another – what I will refer to as *the normative gap*[24] – seems reasonably thin and unclear. For Bergner, this normative gap goes virtually unrecognized, while for Boorse, the (tentative) link has to do with the normal state representing species design/baseline health: "…the normal is the natural" (Boorse, 1977, p. 555). This does not seem like a big problem when considering physical disorders because at this level what is 'good' versus 'not good' is generally quite clear. As a simple example, most people agree that a heart attack is just plain bad. When speaking of behaviour, thought, and emotion however, there is not always one right way to function. In explicitly evaluative words unavailable to these authors, there is a diversity of legitimate values in the psychological realm that is not present in the physiological (Fulford, 2001). For example, statistical functionalism is often argued to erroneously capture homosexuality under the banner of mental disorder given it is statistically deviant and results in less offspring. This all suggests

[23] This label is by no means a perfect fit, for example, I am not sure whether Bergner and Ossorio would agree with the use of 'functionalism' here. I could label it *contextualised behavioural statisticalism* or something similar. However, in so far as behaviours one 'ought' to be able to do can be referred to as functions the label used seems acceptable. The current label also highlights important similarities across divergent views; just as the BST contrasts the individual's physiology against a reference class, this view contrasts the individual's capacities against similar others in similar contexts. Further, my sense is that Bergner would disagree that context can ever really be sufficiently captured by use of a reference class nor any statistical means, and that therefore clinical judgement will always be required in diagnosis. He is probably right, but how do we go beyond the statistics while maintaining clarity, rigour, and a common language? This is another reason why a richer conceptual model/framework is required.

[24] This normative gap is of course nothing new – it is simply the domain-local version of Hume's 'ought-from-an-is' problem (Hume, 1978/1738)

very strongly that the use of statistical normality, even if applicable to the definition of dysfunctional physiology, is not applicable in the definition of dysfunctional psychology. I will discuss this diversity of functionality of behaviour in greater detail in chapter five.

At this impasse there are two options standardly recognized: 1) move away from statistical normality and attempt to plug the normative gap with a better story of how functions can naturally arise. I will explore this option in the next section on evolutionary functionalism. Alternatively, 2) recognize that values do play a role in defining mental disorder, as explored in the following section of value-laden concepts. At the end of this chapter I will suggest that there is another, less recognized, option available to us.

Evolutionary functionalism.

Under evolutionary functionalism, what is disordered is that which fails to perform its evolved function. Rather than deriving ideas of function from that which is statistically normal as above however, this position holds that functions are capacities that parts of the body or mind have, *due to their being selected for across the evolution of the organism*. Evolutionary functionalism then, attempts to plug the normative gap using evolutionary theory. The most well-known conceptual model of this type is Jerome Wakefield's harmful dysfunction (HD) analysis, or more specifically the 'dysfunction' component of this model (1992b, 2007, 2014b). The HD analysis is a two-part model. It holds that mental disorder is 'dysfunction' plus 'harm'. In this section I will primarily discuss the dysfunction component of Wakefield's HD analysis as it is a good example of the pitfalls that arise for the evolutionary functionalist, despite the positions intuitive appeal (I will explore the harm component in the value-laden concepts section).

On the HD view then, dysfunction is a necessary but not sufficient component of disorder (Wakefield, 1992b, 2007, 2014b). This is contrary to the BST in which dysfunction by itself is sufficient for attributing disorder (or rather disease/pathology in BST terminology). The dysfunction component of the HD analysis is defined evolutionarily, requiring that mental disorders include a part or behaviour of the organism that doesn't do what it has been selected to do by the evolutionary process: "A

"dysfunction" exists when an internal mechanism is unable to perform one of its natural functions" (Wakefield, 2007, p. 152). Comparing to the BST once again, the key difference here is the use of the term '*natural* function' as opposed to '*normal* function'. The former are products of random mutation and natural selection across time, and the latter are statistically derived (Boorse, 1977; Wakefield, 1992b). Specific to mental disorder, Wakefield describes the internal mechanisms concerned as 'mental mechanisms'; as evolved tendencies and capacities in behaviour, motivation, cognition, perception, or emotion, that have been selected for due to their serving the survival and reproduction of the species and their ancestors[25]. Mental dysfunction within the HD analysis then, is when evolved mental mechanisms don't function as designed by natural selection (with *disorder* being ascribed when the dysfunction results in socio-culturally defined harm). For example, genuine cases of depression, for Wakefield, represent a malfunction in the psychological mechanisms evolved to regulate emotion, leading to a set of behaviors and experiences society deems harmful (Wakefield, 1997a). Hence, Wakefield's well-known criticism of the removal of the bereavement exclusion in the DSM-5 depression criteria: grief following bereavement is not a dysfunction, but rather an evolved mechanism acting as it should (Wakefield, 2013).

Despite the popularity of the HD analysis, many critiques have been made of this approach to understanding dysfunction. Here I summarize the most successful points from these critiques (as well as Wakefield's responses), structuring them by their tendency to target the HD notion of dysfunction at three different theoretic levels. The first approach to criticism simply attempts to generate *counter-examples* to the HD notion of dysfunction. The second approach is *epistemological and theoretical*, targeting Wakefield's use of evolutionary theory and the idea that we can really come to know an attributes evolutionary 'purpose'. Finally, the third approach is *methodological and practical*. On the basis of the epistemological issues highlighted in the second criticism, this third approach attempts to undercut the claim that HD-style dysfunction is value-free, arguing that, because of the inherent difficulties with figuring out something's evolutionary function, values will always permeate in the actual application

[25] This use of 'mechanism' is again bio-functional, a common intent. Broader definitions of mechanism are in use so it is important to specify (Andersen, 2014a, 2014b; Garson, 2017; Illari & Glennan, 2017)

of the HD analysis. The critiques by counter-example represent by far the most popular approach to criticizing the HD-analysis, and these criticisms have also been the source of significant debate. As I will also show, not one of the suggested counter-examples is a clear-cut case, rather the impact of these counter-examples against the HD notion of dysfunction is cumulative. For these reasons I have given the counter-example approach significantly more room in the discussion.

Critiques by Counter-example. Some authors make the claim that the HD construal of dysfunction excludes cases of genuine disorder and is thereby overly exclusive. The most common variant of this approach refers to cases when disorder arises due to a mismatch between the current and ancestral environments, and cases when disorder arises due to 'pathogenic input' into a normally acting mechanism[26] (Lilienfeld & Marino, 1995; Murphy & Woolfolk, 2000; Nesse, 2001; Varga, 2011).

The mismatch cases can be exampled well by depression. Some current explanations of depression posit that, rather than a being a dysfunction of an evolutionary mechanism, the suite of behaviors and cognitive tendencies labeled depression may represent an historically adaptive response to a loss of an important resource. The idea is that, in our evolutionary past, following such things as the loss of social status or reproductive partners, retreating away from others and generally reducing levels of activity may have been a good strategy to recuperate and garner sympathy (Beck & Bredemeier, 2016; Price et al., 1994). Depression then, especially in mild to moderate cases, may not be a dysfunction in the HD sense. Instead these behaviors may represent evolved mechanisms acting as 'designed', but not working for us in a modern context. Another example would be blood/injury phobias. Contrary to other phobic responses, a phobic response to blood or injury typically involves acute reduction in heart rate and blood pressure; presumably an evolutionary adaption to help reduce blood loss. Yet, this response can be genuinely maladaptive in a modern context where we have the medical capability to manage injury (Lilienfeld & Marino, 1995).

[26] Another case of potential exclusion is where the function is culturally shaped rather than evolutionarily selected, such as in the ability to read (Lilienfeld & Marino, 1995). Wakefield (1999a) circumvents such cases by assuming that cases of genuine disorder feature a dysfunction of a sub-mechanism, required for the function, that is itself evolutionarily selected. i.e. genuine dyslexia is taken to involve some brain dysfunction rather than simply a lack of practice reading.

The 'pathogenic input' case can be exampled by Conduct Disorder (CD) and/or Oppositional Defiant Disorder (ODD); DSM diagnoses that describe serious misbehavior and social norm breaking in children. Many theories hold that these patterns of behaviour are a result of *normal learning processes* occurring within a family environment that is inadvertently training the child to misbehave – i.e. 'pathogenic input' (Chang & Shaw, 2016; Labella & Masten, 2018; Smith et al., 2014). In such cases there does not seem to necessarily be a *failure* in an evolved mental mechanism. Both mismatch and pathogenic input cases then, demonstrate that disorders seem to sometimes exist in the absence of evolutionary malfunction[27], bringing into question whether dysfunction in Wakefield's evolutionary sense is in fact necessary for ascribing mental disorder (Murphy & Woolfolk, 2000; Varga, 2011).

Wakefield (1999b, 2000b) has responded to these cases of environmental mismatch and pathogenic input. In the cases of mismatch, he points out that the relevancy of these cases is conditional upon the explanations by ancestral-context-mismatch turning out to be correct. In the event that such explanations turn out to be well-founded – i.e. that some conditions currently called disorder turns out to be evolved mechanisms misaligned with current context – Wakefield stands by the HD analysis and suggests we will stop referring to such conditions as disordered. Wakefield uses the example of a fever here, which used to be understood as pathological but is now understood as a functional immune response to a pathogen.

To the cases of pathogenic input into a normally functioning mechanism, Wakefield's (2000) response has two prongs. First, he suggests that, much like the mismatch cases, apparent disorders arising from pathogenic input should often not be considered disordered:

[27] PTSD/Trauma is another good example of disorder where dysfunction (in Wakefield's sense) arguably does not seem to be present. The common 'symptoms' of hyper-vigilance and aggression seen following trauma can be understood as attempts to adapt to dangerous/hostile environments (therefore representing normal and adaptive learning). This is also representative of a mismatch case, as this kind of response was probably evolutionarily adaptive when our environments were less predictable than the modern context. Alternatively, we might consider the trauma as 'pathogenic input' which exceeds our design limitations. Analogous to drowning when placed underwater, perhaps extreme trauma is just not something our evolutionary design is capable of coping with.

> "If a condition is a direct learned response to ongoing environmental reinforcers and involves no consequent dysfunctions, we do not consider the condition a disorder" (Wakefield, 2000b, p. 261).

Here Wakefield cites evidence that when children meet criteria for DSM disorders of conduct, clinicians tend not to consider them 'disordered' if the oppositional behaviour seems to be a learned response to their environmental context[28] (Kirk et al., 1999). The second prong of Wakefield's response is to claim that there are cases where HD-style dysfunction can occur despite a mechanism *acting* 'as designed'. In cases of pathogenic input for example, the mental mechanism is *acting* 'as designed' by evolution, and therefore is not 'broken', but the *function* the mechanism is meant to serve is still compromised due to inappropriate input (e.g., cars don't drive if you put water in the tank instead of petrol, and while this doesn't constitute a malfunction of the engine it does describe the engine being unable to serve its function). This seems like a successful rebuttal given Wakefield's definition of dysfunction as "…when an internal mechanism is unable to *perform* one of its natural functions" (Wakefield, 2007, p. 152, emphasis added). It is not actually the breaking of the particular mechanism that Wakefield's sense of dysfunction is tried to, but rather the ability of the mechanisms to serve its evolutionarily 'designed' function[29].

Another variant of the counter-example approach is through reference to particular evolutionary phenomena that somehow cast doubt on the notion of function emerging through natural selection. For example, Lilienfeld and Marino (1995) raise the case of exaptations. Exaptations are features of an organism that originated to serve a particular purpose, but have since been co-opted by the organism's design to serve a different one (Gould, 1991). Feathers are a classic example of an exaptation as they

---

[28] I find this use of Kirk, et al. (1999) a little convenient here, and potentially misleading. Firstly, participants in the cited study were trainee social workers, potentially biased by the ecological focus of their chosen profession. Secondly, situations involving children often discourage the use of a disordered label anyway, as children are (rightfully) seen as 'unfinished' and more environmentally dependant. It would be interesting to see if the same tendency to not ascribe an internal dysfunction is observed with adult cases of anti-social personality disorder, or PTSD, which are also disorders that demand an ecological focus.

[29] There are questions to be asked here about whether an evolved/selected mechanism can really have an 'designed/intended' function over and above what is does and the conditions it does this in; this seems to grant a questionable amount of agency to the evolutionary process (Lilienfeld & Marino, 1995).

originally evolved to provide warmth rather than to support flight (Murphy & Woolfolk, 2000). The supposed issue for the HD analysis here is that that many features of human mental functioning seem likely to be exaptations (i.e. not used for their original evolutionary purpose). Wakefield circumnavigates this issue quite easily however, essentially by saying that we should refer to the most recent evolutionary purpose that the mechanism has served. For example, while the original purpose of feathers may have been for warmth, they certainly now serve the purpose of assisting with flight, and an inability to serve this function would therefore be a dysfunction. One variant of this approach, particularly relevant for mental disorder, is the case of cultural exaptations (sometimes called cultural spandrels; see footnote 26). Cultural exaptations are where the reutilization of a feature to serve an alternative function is not achieved via differential reproduction, but by a cultural process (Lilienfeld & Marino, 1999). The paradigm case of this is reading. The ability to read has appeared much too recently in human history to be an evolved function/specific brain mechanism, rather it is a culturally transmitted skill, built from basic functions such as visual attention, which were selected for different purposes (Heyes, 2018). Lilienfeld and Marino (1995, 1999) question how dyslexia (a selective difficulty with reading) can be seen as a dysfunction in the HD sense if the ability to read is not an evolved function. These authors also example other disabilities concerning culturally transmitted skills such as amusia and acalculia (tone deafness, and difficulties with maths). Wakefield (1999a) however, circumvents such cases, by assuming that cases of genuine disorder feature a dysfunction of a sub-mechanism; one of the basic functions from which cultural exaptations are built[30].

In all these cases of potentially unwarranted exclusion – environmental mismatch, pathogenic input, exaptations (cultural or otherwise) – it is ultimately not clear whether Wakefield's rebuttals are successful or not. Overall, Wakefield seems satisfied with his rebuttals, but his critics remain unconvinced. This may point to

[30] Another form this counter argument takes is through talk of spandrels (evolutionary by-products that serve no functional purpose) and vestigial features (features that used to serve a function but are now just left-over parts – such as the appendix). Having no function it is not clear how these features can be dysfunctional (Murphy & Woolfolk, 2000). Wakefield's responses to these cases are very similar – he shifts down an organizational layer and says that a component function/part is not acting as it is evolutionarily intended, and is therefore dysfunctional (Wakefield, 2000b).

difficulties in Wakefield's central method of *conceptual analysis,* where 'conventional' cases of disorder or non-disorder are used as litmus tests for the conceptual model of disorder under study. This method seems potentially circular, and also assumes the conceptual structure of mental disorder to take a classical 'necessary and sufficient' form[31] (M. J. Walker & Rogers, 2018). By attempting criticism through raising potential counter-examples, these authors are essentially playing Wakefield at his own game. What can be concluded at this point however, is that the complexity and number of the arguments surrounding these possible counter-examples is itself a concern. The HD analysis is intended to provide clear guidance as to what counts as disorder and what doesn't, but its complex notion of dysfunction seems to be getting in the way of achieving this.

Epistemological/Theoretical Critique. Rather than attempting to find cases of genuine disorder that the HD analysis notion of dysfunction excludes, this line of criticism is targeted at the core theoretical workings of the HD analysis and whether it can really do the work that Wakefield claims it can. For this reason, it seems much more convincing. Critiques at this level tend to feature the same overarching gist. This is that: 1) evolutionary processes are extremely complicated, 2) the relevant processes occurred in the past and over a very long time, making them hard to gain knowledge of, and 3) Wakefield's use of evolutionary theory to attribute functions seems simplistic and convenient in light of these points.

Sadler and Agich (1995) for example, directly accuse Wakefield of misrepresenting evolutionary theory. In particular, they are concerned that Wakefield anthropomorphizes the evolutionary process and gives natural selection a sense of purposiveness through use of teleological terms like 'design'. Such terms they argue, erroneously inject intentionality into the processes of evolution. This concern is also strengthened through Wakefield's choice of analogy when explicating how natural

[31] Some authors suggest that disorder may be better captured as a Roschian/Wittgensteinian cluster/family/prototype concept that does not feature necessary and sufficient conditions, but where tokens of the concept are clustered together by similarities across multiple dimensions (Lilienfeld & Marino, 1995, 1999; M. J. Walker & Rogers, 2018). I have no doubt that the concept of mental disorder present in the public mind, or even in medical professionals, would be best represented in such a way (a descriptive notion), the more interesting question – and the interest of this thesis – is what the concept of mental disorder *should* represent in a prescriptive sense (Muders, 2014).

functions arise via natural selection; he draws a parallel to human artifacts which really are designed – i.e., watches really do have an intended function to measure time. In pointing this out, Sadler and Agich are criticizing the underlying idea present in Wakefield's position, that a part or behaviour of an organism can clearly be a 'natural' or 'proper' function in a literal sense, simply because it arises due to natural selection. In fact, Sadler and Agich (and others) disagree with the term 'natural selection' entirely, instead preferring the term 'differential reproduction'. 'Selection', they claim, is an agential term, instilling a sense of purpose into what is ultimately a large-scale natural/causal process. To clarify here, Sadler and Agich do not seem to be making the ontological claim that natural functions *cannot* arise through the evolutionary process. Rather they claim that by anthropomorphizing the evolutionary process, Wakefield makes it seem as if our ability to come to know the function of a part/behaviour of an organism is easier than it really is; as if it is parallel to recognizing the function of a watch.

Similar concerns are raised by Murphy and Woolfolk (2000), who highlight the speculative nature of evolutionary psychology in its present state, and by Lilienfeld and Marino (1995), who point out that in many cases natural selection results in a within-species-diversity of adaptive strategies rather than a single mode of functioning against which dysfunction can be contrasted. All of these points are attempts to demonstrate the serious limitations on our ability to come to know the evolutionary function of an organism's behaviour (in order to infer deviation from this as dysfunctional). Speculatively, further criticisms of this style could be developed. While I am not aware of any current criticisms that attempt to do so, any selective pressure that highlights the contingency of evolutionary success (e.g. genetic drift), or selects for traits that may actually hamper survival (e.g. the sexual selection of 'handicaps' such as peacock's tails or male risk-taking behaviour) could potentially be used to demonstrate that Wakefield often paints evolution with an idealized brush[32].

[32] Relatedly, one could also challenge Wakefield on his assertion that natural selection is the only known source of natural functions: "…natural selection is the only such process." (Wakefield, 1999b, p. 472). There are arguably other possible sources, such as the emergence of functional norms through the precariousness of autonomous and adaptive systems (Christensen & Bickhard, 2002). This idea will play a key role in chapter five. I hope to show that using this idea as a basis will constitute a significantly different form of 'functionalism'; one that is richer and explicitly evaluative.

Before moving on, it is useful to consider an evolutionary functionalist position different to that of HD-style dysfunction; that of Troisi and Macguire (2002). I mention this here because, in generating their own position of 'Darwinian Psychiatry', these authors demonstrate awareness of some of the mentioned epistemological issues with evolutionary functionalism that hamper Wakefield's analysis of dysfunction. In particular, Troisi and Macguire point out the vital role of phenotypic variability in the evolutionary process, as well as that the evolutionary fitness of a behaviour is highly contingent and nigh on impossible to measure directly. In doing so they acknowledge our epistemological limits concerning the evolved functionality of a behaviour. As such they suggest a need to measure *functional consequences in the individual* rather than inferring whether they were adaptive for the species in the ancestral context. The problem with this of course is that 'functional consequences' in a Darwinian frame boil down to the number and quality of the offspring produced. Due to obvious time constraints we can't sit around and wait while counting the number of off-spring someone has and/or how long they live. Troisi and Macguire' solution is to suggest the use of 'the achievement of short-term biological goals' as a proxy measurement for evolutionary success. 'Darwinian Psychiatry' then is a much more successful but much less ambitious variation of evolutionary functionalism in comparison to the HD analysis. More importantly for the current discussion however, the limitations these authors place on themselves stem directly from their understanding of the messy realities of evolution. These limitations highlight nicely where Wakefield's concept of dysfunction arguably oversteps what evolutionary theory can truly provide.

The Methodological/Value-Creep Critique. What is interesting about the line of criticism above is that, at times, Wakefield himself seems to agree that there is a significant epistemological challenge that can be made against his framework: "discovering what in fact is natural or dysfunctional may be extraordinarily difficult" (Wakefield, 1992a, p. 236). How then does Wakefield claim that we can overcome this challenge; to know the function of a part or behaviour of an organism, or that a proposed mental disorder involves a dysfunction? Wakefield's approach to this typically involves inferring the presence of an underlying dysfunction based on indirect evidence, in particular, whether the behaviors displayed are 'normal' or 'proportionate' given the

persons context (Murphy & Woolfolk, 2000). The problem here is that 'normal', and 'proportionate' are not evolutionary concepts, suggesting some alternative source of normativity is at play. This gets us to this to the current line of criticism; the concern that, due to the epistemological limits described above, Wakefield's notion of dysfunction opens up room for the covert importation of values into our demarcation of the disordered from the benign (Murphy & Woolfolk, 2000; Sadler & Agich, 1995).

Wakefield generally attempts to navigate claims concerning the value-laden-ness of his notion of dysfunction through a concept he has labeled *black-box essentialism* (Wakefield, 1997b, 2000a). This is the idea that dysfunction itself has an essence – that it exists as a purely factual thing in nature – but that we are at a stage of discovery where we cannot yet reliably detect this essence. In explaining this Wakefield draws an analogy to water, now known to have the 'essence' of being constituted by the compound H_2O. Before we knew this, claims Wakefield, we recognized water by the properties of a 'base set' of things that were most prototypically 'water'. Properties like being a clear liquid, being thirst-quenching, and being present in rivers and lakes, formed a metaphorical black box within which we assumed there was some essence shared by all instances of water (which we now know to be H_2O). Wakefield therefore sees himself as attempting to distil the essence of what it means to be dysfunctional, and holds his evolutionary concept of dysfunction as the best attempt currently available. As we come to understand the evolutionary processes underlying the development of human behaviour in more detail, we will be more and more able to separate out genuine dysfunction. This offers Wakefield a solid conceptual defense against claims that his concept of dysfunction is value-laden; it is simply one of his core assumptions that it is not, and cases of value-laden-ness are simply a methodological rather than conceptual error.

This however misses the point of criticisms raised by Murphy and Woolfolk (2000), and by Sadler and Agich (1995), whose criticisms are not (only) conceptual, but rather *are* methodological in nature. Given our current (and likely future) inability to confidently know the evolutionary functions of a behaviour, the HD notion of dysfunction can offer very little guidance in practice, or worse, encourage us to generate evolutionary stories that implicitly align with our values and biases. We need guidance as to what to label dysfunctional now, not at some unknown time in the future when we

understand the evolutionary processes underlying behaviour. For example, homosexuality could conceivably be considered a dysfunction in Wakefield's sense, given it presumably leads to lower reproductive success. While there are evolutionary theories as to the possible adaptive function of homosexuality, these are (and likely will continue to be) speculative and contested. The HD notion of dysfunction therefore offers little guidance as to whether homosexuality should be considered a dysfunction, because it relies on information that we do not have access too.

Wakefield's response to such criticisms has been to claim that this is a non-issue, holding that neither the epistemological issues, the resulting lack of guidance, nor the susceptibility to value-creep are relevant to the validity of the dysfunction concept itself, merely to its implementation (Wakefield, 1999b). The reality is however, that a fundamental value of conceptual work is its ability to provide guidance in later tasks. A concept of mental disorder that provides principle yet guides no praxis renders its own principle impotent. Similar points have been raised by Sadler (1999) who claims that Wakefield has boxed himself into irrelevance for the entire field of psychopathology. If the HD notion of dysfunction reduces in practice to a covert statistical functionalism (where we derive function from what deviates from the norm) then it is not clear how Wakefield's position differs from the BST, nor what justifies the extra evolutionary baggage of his formulation. If, however, evolutionary dysfunction reduces in practice to a covertly value-laden concept, then it would seem a much better idea to make these values explicit so that we may consider them honestly; removing the evolutionary or naturalized trappings entirely. This is the approach that will be explored in the next section[33].

[33] There is also a wider criticism that can be put to evolutionary functionalism that is worth briefly considering. This is simply that life seems to be about more than reproduction and survival. For example, why should two cases of symptomatically identical depression be treated differently, simply because one is considered an evolutionarily adaptive response to some trigger, and the other is considered a misfiring of that response? Both instances represent similar degrees of harm and functional impairment (in a wider non-evolutionary sense), and both have the same capacity to restrict the sufferer's access to 'the good life' (whatever that is for them). As I will aim to show in chapter five, claim's such as Wakefield's, that natural selection is the only known source of 'natural function' are highly debatable.

Evaluative concepts.

The normative conceptual models explored so far have all been attempts at *naturalizing* mental disorder; of limiting the normative scope of the concept to exclude values, especially individual and culturally specific values[34]. Many authors argue however, that attempting to do so is futile and we should instead be open and honest about the role of values in psychiatric diagnosis (Doust et al., 2017; Fulford, 2002; Sadler & Agich, 1995; Stier, 2013). Metaphorically, these positions are bridging the normative gap with values; sourcing their claims about the 'goodness' or 'badness' of human thought and behaviour from socio-cultural value structures. Moreover, those who hold this position tend to claim that everyone else is doing this too, only without realizing it. Positions that recognize the role of values in this way are broadly known as *evaluative* in nature. In contrast, the collective term for those who attempt to naturalize mental disorder – to see it as purely factual – are most typically known as *descriptivists* (Fulford, 2002). In line with Zachar and Kendler (2007) however, I will refer to this position as *objectivism* in order to avoid using multiple senses of 'descriptivist' across this thesis.

Generally speaking, evaluativists are motivated by two observations. The first of these observations is that values are almost certainly playing a role in the conception and application of current diagnostic concepts (Foucault, 2003; Sadler, 2005; Stier, 2013; Szasz, 1960). If this is true, this means that when a clinician or psychiatrist makes a diagnosis, there seems to be a very real sense in which they are evaluating the client rather than simply describing their state. Objectivists find this conclusion unsettling, preferring that diagnosis be a purely factual matter (for example see; Hucklenbroich, 2014). A workable objectivist rebuttal here is that evidencing the value-laden nature of current concepts and diagnostic practice speaks only to an understanding of concepts and practice *as they are*, not necessarily *as they should be* (Muders, 2014). This thereby leaves room for the possibility that, despite the role of values in current diagnostic

[34] The popularity of such naturalized value-free models may well be a reaction to the arguments of the anti-psychiatry movement who questioned the concept of mental disorder predominantly on the basis of its evaluative (and therefore on their view non-scientific) conceptual nature (Varga, 2011).

concepts, there is a way to consider them as wholly objective and that perhaps such a way is preferable.

The second observation that often motivates evaluativism is simply that popular objectivist approaches, such as the two brands of functionalism explored above, seem to fail to distinguish between disorder and non-disorder effectively. For example, Doust, et al. (2017) explore three examples of conventionally accepted medical disorders and demonstrate that functionalism offers very little guidance as to where the boundaries of disorder should be placed. Instead, they propose, the answer to this question seems to revolve around the values at play. Therefore, they argue that our conceptual models should openly recognize the role of values in demarcating disorder. If they do not do so, we meet the same problem we saw with the HD notion of dysfunction where values may creep in unannounced and therefore unconsidered. Problematically however, Doust, et al. offer no framework for how this recognition of the role of values could be achieved.

There are generally three different evaluative stances, taken in response to the acceptance of these observations, as to what a concept of mental disorder should be. I refer to these stances as: *weak-evaluativism*, *strong-evaluativism*, and *anti-psychiatric evaluativism*.

Weak-evaluativism simply recognizes that terms like dysfunction and disorder are evaluative *in a limited sense*. Specifically, weak-evaluativism does not prescribe the inclusion of socio-culturally and individually specific values in consideration of what counts as disorder. According to the weak-evaluativist then, cases where socio-cultural values are playing a role in diagnosis – e.g., see Stier (2013) – are in error. Under weak evaluativism, the values at play are assumed to be universal and therefore not particularly contentious. As I will explore further in chapter five, this brand of evaluativism seems potentially workable for bio-medical disorders where values are relatively agreed upon – e.g., it doesn't seem contentious to say that brain tumors are bad – but seems much less workable in the domain of mental disorder where values are exponentially more diverse (Fulford, 2002).

Strong-evaluativism, in contrast to the weak form, accepts that socio-cultural and individual values should and do play a role in demarcating disorder. The immediate

problem with this position however, is that it introduces a high degree of relativism (Jefferson, 2014). This is where what counts as disorder changes across cultures and time periods, dependent on the local value set. For example, under a strong-evaluativism, the labeling of homosexuality as disordered within the bounds of a conservative culture seems concerningly uncontestable. This relativism also opens-up boundary issues, i.e., how do we know whose values to use, and where does one culture stop and another start? It is potentially due to these issues of relativism that very few strongly-evaluativist concepts have been proposed as formal conceptual models of mental disorder.

Finally, the third evaluativist position that can be taken is anti-psychiatric evaluativism. This position holds that concepts of mental disorder are so value-laden that they do not refer to anything 'real', that they are ethically unacceptable, and that we should therefore discontinue their use. Foucault's (2003/1961) position mentioned in the deflationary section would be an example of this kind of evaluativism.

One unique approach to strong-evaluativism that seems to successfully contain the threat of relativism is the HD analysis (Wakefield, 1992b, 2007, 2014b). By specifying that both harm and dysfunction are necessary for an attribution of disorder, but that neither are individually sufficient, Wakefield incorporates socio-cultural values into his conceptual model while staving off unconstrained relativism. Under the HD analysis, harm is considered in explicitly culturally relative terms:

> "…disorder lies on the boundary between the given natural world and the socially constructed world; a disorder exists when the failure of a person's internal mechanisms to perform their functions as designed by nature impinges harmfully on the person's wellbeing as defined by social values and meanings." (Wakefield, 1992b, p. 373).

The general gist of this idea – how it utilizes both components to constrain the other – is regarded highly. For example, renowned author in this area, Peter Zachar, refers to the HD idea as "parsimonious, elegant, and useful" (2014, p. 121); three descriptive terms from which I would certainly agree with the first two. The issue, as we saw in the previous section, is primarily with the workability of the dysfunction

component. It is not clear whether this notion of dysfunction represents an acceptable use of evolutionary theory, nor whether we can ever obtain the deep knowledge of evolutionary processes required to utilize it. Hence, with the dysfunction component virtually defunct, the parsimony of the HD idea, and how it attempts to put strong-evaluativism to work in a suitably constrained manner, ultimately falls flat.

Before moving on I should note that a core assumption of this thesis is that, in the demarcation of disorder, the question of whether norms and values have a role to play *at all* is somewhat trivial. At its simplest, a diagnosis is a claim that something is *wrong* with a person. On my view it is therefore *necessarily* normative/evaluative, and I therefore reject total objectivism (although not, as I will show, the allure of naturalization). In chapter five, I will attempt to carve new ground between the weakly and strongly evaluative positions. The resulting view will include certain socio-cultural values as relevant to mental disorder on a principled basis, while maintaining a thorough going naturalism. This will be achieved through the use of a framework that subscribes to value-inclusive naturalism, allowing us to move beyond the dichotomy of objectivist versus evaluativist positions (Thornton, 2000).

Practical kinds.

Faced with the many competing normatively oriented concepts explored above, some authors have suggested turning to pragmatism for solutions. A *pure* or *radical* pragmatic view holds that the underlying structure of mental disorders is either that of 1) non-kinds and therefore continuous with normal human behaviour, or 2) totally socially constructed. Nonetheless the pragmatist holds that it is *useful* for our purposes as explainers and clinicians (who work within socio-legal environments that often demand categorical identifiers) to treat them as more 'real' and categorical than they may be. On this view then, it is the *usefulness* of mental disorder concepts that justifies their use, despite the fact that they may not refer to any real kind in nature (Haslam, 2002; Kendler, 2016). To return to our metaphor, the pragmatists are skipping over the normative gap and saying 'let's just do what seems useful'.

In this radical form, pragmatism risks total nominalism (nominalism in the sense that they have no referent in the natural world and are thereby empty labels[35]). This where what counts as mental disorder are simply those things that we, or a particular group, *label* as mental disorders. For example, O'Connor (2017) presents the idea that mental disorders are practical psychiatric kinds. By this he means that mental disorders are those categories that psychiatry invents because they are useful for psychiatry's purpose of helping people. This position is not intended to be a deflationary one; rather than define psychiatry as the profession that treats mental disorder, O'Connor flips this around and defines mental disorder as that which psychiatry treats. Psychiatry in turn is defined in a broader sense as the profession that aims to "…help those with emotional or psychological impairments who seem to be unable to help themselves." (O'Connor, 2017, p. E-8)[36]. This position rejects naturalism about mental disorder, both in the sense that mental disorder may represent natural dysfunction/s, and in the sense that mental disorders may be understood structurally as natural kinds. Rather for O'Connor, mental disorder concepts are the products – and tools – of psychiatric practice which, in turn, he seems to see as a broadly moral enterprise. While this may represent a valid – if slightly disparaging – perspective on the nature of current diagnostic concepts in mental health, it still leaves mental disorders as totally nominal entities and thus provides next to no guidance as to what kinds of things we should or shouldn't count as mental disorder.

In response to this issue of nominalism, some pragmatist positions take only a *partially* pragmatic approach by incorporating other normative or structural elements. One such model would be Zachar's Practical/MPC hybrid model (2015). This model combines the concept of a fuzzy MPC kind with pragmatism:

---

[35] Note that this use of the term 'nominalism' differs from its use in philosophy/metaphysics. My use of the term in this way is preceded by Zachar and Kendler (2007).

[36] There is a charge of circularity that can be made against this position. For example, what exactly defines an 'emotional or psychological *impairment*'? This seems to be another term for a mental disorder. I take this to be representative of O'Connor's point – on his view mental disorder is a conceptually thin notion, constructed through the practice of a morally defined institution.

> "Concepts for psychiatric disorders are constituted by discoveries *and* decisions. There is an interaction between what the world produces and what we find useful to notice." (Zachar, 2015, p. 289).

Under this model, paradigm mental disorders are seen to be likely tracking MPC like structures in human behaviour. The fuzzy nature of MPCs provides instances of ontological indeterminacy, in response to which classificatory decisions are made in accordance with our pragmatic purposes. For example, if, for the moment, we assume that depression and its melancholic subtype are MPC kinds whose properties overlap, there is a genuine sense in which the decision to treat these entities as having a type-subtype relation is somewhat arbitrary. We could alternatively treat them as different entities with similar symptom profiles. This is not a totally nominalistic position as there are structures in nature to which mental disorder labels are thought to refer, but Zachar's model highlights that many such arbitrary or pragmatic decisions have, over time, shaped our diagnostic systems[37].

Again however, a pressing issue with Zachar's (2015) model concerns the lack of guidance it provides. It is undeniable that our current diagnostic concepts are partially 'historical' in nature; that their current form is contingent upon past human affairs and decisions rather than representing naturally separable phenomena. Pragmatism helps us recognize this but doesn't necessarily treat it as a problem, let alone provide a solution. This is because, other than their usefulness, pragmatism doesn't commit to any particular notion of what a diagnosis of mental disorder *should* represent. Pragmatic notions of mental disorder seem too thin in that they fail to provide an ideal; they are 'unambitious' in this way (Kendler, 2016). If tomorrow, we discover a new putative mental disorder, pragmatism offers us very little help in deciding whether to include it in our diagnostic systems or not.

Some Preliminary Observations

This concludes the review of the dominant positions available when considering the conceptual nature of mental disorder. All of the models presented can tell us

[37] Zachar explicitly recognises this partial nominalism/historicism in his Imperfect Community Model, where mental disorders are seen to be clustered under a single banner partially due to genuine family resemblance, but partially due to pragmatic and historical factors (Zachar, 2014).

something interesting about the nature of mental disorder, but all face significant problems. Before moving on to an overview of the DSM and RDoC, I will make two observations that help motivate this thesis.

Concerning conceptions of human functioning.

The first observation is that the concept of mental disorder that an individual subscribes to tends to track the individual's *conception of human functioning* in general. Put more simply, someone's understanding of how the human mind works seems to inform their understanding of the human mind as not working properly. This points to an important conceptual co-determinacy between frameworks of human functioning and frameworks of mental disorder. As a slightly contrived example, if I got in a time-machine, visited Rene Descartes, and asked him what mental disorders are, I assume that his answer would be grounded in his dualistic understanding of the mind-body. Perhaps he would suggest that mental disorders represent corruptions of the soul, or perhaps he would suggest they represent some sort of mechanistic breakdown in the soul communicating through the body.

We can see this conceptual co-determinacy between what someone understands mental disorder to be and how they understand human functioning in the positions explored in this chapter. Foucault, for example, was interested in the relation between individuals and society, believing that behaviour is strongly regulated by socially generated norms and concepts (and therefore that the production of these norms and concepts is where true power lies in society). His understanding of mental disorder as a socially constructed label for certain kinds of deviance makes sense in light of this. As a further example, consider Insel and Cuthbert (2015) who argue for a biologically focused model of mental disorder as a route to precision medicine in psychiatry. Note how their essentialist assumptions make perfect sense given the medically minded and brain-focused approach to human functioning that they ground themselves in.

This same conceptual co-determinacy is most clear when considering the functionalist positions. The very idea of these positions is to contrast disorder against an understanding of the things humans should be able to do if they are functioning normally. For the statistical functionalist these things are derived from an

understanding of what most others can do, for the evolutionary functionalist these things are derived from an understanding of what is evolutionarily successful. The connection is also clear in the evaluativist position. The evaluativist's central claim is that all objectivist positions fail because they miss the irreducible role of values in our lives. In essence they are saying something like 'we are more than our statistical normality, more that our ability to pass on our genes; we have values'. The claim then is that the objectivist does not hold a rich enough (i.e. value-inclusive) understanding of human functioning by which to contrast mental disorder.

This observation opens up the question of what happens if we consciously position ourselves within an understanding of human functioning that seems fit for purpose. Rather than considering humans as simply units in an evolutionary process, as brains driving our bodies around like cars, or as leaves on the wind of social processes, perhaps we should seek to consider human functioning in a richer and more integrative way? Perhaps if we do so we may come to a more comprehensive understanding of what mental disorder is. This is the underlying idea that inspired this thesis.

The normative gap may be artifactual.

The second observation is that the 'normative gap' observed between simply describing human behaviour and being able to say that some behaviors are disordered or bad in some way, may in-part be an artifact of how we talk about values. Typically, we talk about values as if they are entities that somehow transcend maters-of-fact, but assuming naturalism this simply cannot be the case. This observation has been made before, and put in much clearer terms by Thornton (2000). Thornton considers the debate between those who see mental disorder as necessarily evaluative (e.g. Fulford, Sadler, Stier) and those that are attempting to 'naturalize' mental disorder through the concept of a natural function (e.g. Boorse, and [regarding his concept of dysfunction] Wakefield). The functionalists think, very roughly, that incorporating values into the concept of disorder/dysfunction is to admit that it is not a natural/scientific phenomenon. Hence, they are trying to show they can *reduce* this notion of mental disorder to a more basic, purely factual language. The evaluativists meanwhile disagree, believing that there is an irreducibly evaluative element to mental disorder. Thornton however, points out that in doing so, both sides tend to agree that *values are not*

natural. Thornton's proposal is that a non-reductionistic understanding of naturalism does not rule out an understanding of values as part of the natural world: "…although mental illness cannot be reduced to the realm of law, it is no less real for that." (2000, p. 75). While he does not go into detail, what Thornton is implying here is that 'values' may be real things in the world, emergent at levels of organization higher than physics or chemistry. Further, he seems to be suggesting that the adoption of a naturalized but non-reductionistic world-view may help to resolve, or in other ways navigate, the apparent evaluative-objective divide.

What this is calling for is a naturalized but non-reductionistic conception of human functioning; one that can incorporate the obvious fact that humans have values. Such a framework could conceivably plug the normative gap in a naturalistic way without leaving us making do with an impoverished notion of what it means to be human. This second observation then, is pointing in a similar direction to the first. If we want a richer understanding of mental disorder, we need to situate ourselves within a richer understanding of human functioning. One framework that may be able to serve this role is *embodied enactivism*, which I will introduce in chapter four. First, however, we should consider the understandings of mental disorder implicit in institutions such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the Research Domain Criteria (RDoC).

Chapter 3: DSM, RDoC, and Frameworks of Human Functioning

As mentioned in chapter one, the classification of mental disorders is a central task of psychopathology science. The classification of an individuals' clinical presentation into dimensions or categories, such as depression, facilitates communication and access to relevant scientific literature, thereby guiding both data exploration and theory generation (Haig, 2014; T. Ward et al., 2016). The impact of a diagnostic system only increases as we move up from the study of psychopathology in individuals, to populations, where in many ways diagnostic systems shape the landscape in which psychopathology research is done. Diagnostic systems constrain both which putative disorders get studied and – most importantly for my purpose in this thesis – they shape how we conceive of disorders in the first place. For many researchers and clinicians, the answer to 'what is mental disorder?' is in fact simple – 'mental disorder' refers to the diagnostic labels listed in the DSM. Contrasting this simplistic nominalism[38] with some of the complex issues explored in the previous chapter, it should be apparent that this is not really a good enough answer. It is vital then to understand the conceptual models of mental disorder implicitly or explicitly represented within our current diagnostic systems. While underspecified compared to the models discussed in chapter two, in many ways the models present in our classification systems represent the status quo ways of thinking about mental disorder.

In this chapter I first briefly overview some of the conceptual assumptions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the problems that arise from these. I then explore in more detail how the Research Domain Criteria (RDoC) attempts to address these challenges. I argue that in doing so, RDoC makes some problematic assumptions concerning the nature of psychopathology and human functioning in general. On the basis of the work of Murphy (2017) I propose that these underlying assumptions reflect a commitment to a view of the mind known as *eliminative materialism,* or at least *material-reductionism*. While theoretical commitments about the way the mind is organized and its relationship to the body and

[38] Again, note that I am using 'nominalism' in a particular way here, contrary to its use in philosophy. Here it refers to diagnostic labels failing to refer to kinds demarcated in nature, and instead being defined by their pragmatic or historic use (Zachar, 2014; Zachar & Kendler, 2007).

environment are needed to support any psychopathology classification system, I propose that *eliminative materialism* has led RDoC researchers astray. This chapter closes with a call parallel to that made at the end of chapter two – that the development of a richer understanding of the nature of mental disorders itself requires a foundation within a rich understanding of human functioning.

Assumptions of the DSM and RDoC

Diagnostic and statistical manual of mental disorders (DSM).

The DSM is often referred to as a 'signs and symptoms' approach to classification. It works from the reasonable assumption that the causal processes supporting psychopathology are complex and hard to obtain knowledge of. The DSM sidesteps this difficulty by being 'atheoretical' with regards to etiology. This means that rather than basing diagnostic constructs on a set of causes or underlying processes as is the case in other areas of medicine (causalism), diagnostic constructs are inferred from observed patterns of clinical features across the relevant population (descriptivism). Under the DSM's descriptivist model, signs and symptoms observed to co-occur and be associated with harm and/or functional impairment are given a label of *mental disorder*. The central issue of relevance is that this model is solely focused on the *reliability* of diagnosis, and not whether these diagnostic constructs pick out common causal processes; whether they are etiopathologically *valid* (Lilienfeld & Treadway, 2016; Zachar & Kendler, 2017).

Through its organizational structure the DSM also represents a *categorical* approach. 'Categorical' refers to the constructs having clear boundaries between both normal functioning and each other (e.g. you either do or do not have Major Depressive Disorder (MDD) rather than being more, or less, depressed; MDD is seen as totally separate from anxiety). DSM diagnoses are represented as lists of criteria, usually given a letter label, all of which have to be met to justify a diagnosis. Diagnosis under the DSM is 'algorithmic' in this way, where for example you need to meet criteria A through E to be diagnosed with Major Depressive Disorder (American Psychiatric Association, 2013b). However, many of these criteria themselves refer to lists of signs and symptoms. Not all signs and symptoms need to be present to meet a particular criterion, rather a

certain threshold number is required. For example, you need five out of the ten listed symptoms to meet criterion A for Major Depressive Disorder. This element of the diagnostic structure, where a criterion is met by having a certain number out of a wider list of signs and symptoms, is referred to as 'polythetic'. This overarching algorithmic and polythetic diagnostic structure is what makes the DSM categorical; you either meet the criteria or you do not[39].

One strength of the DSM model is that it makes an attempt at what Wakefield (2014a) calls *conceptual validity*. Since the publication of the DSM III a mental disorder diagnosis has required the presence of some degree of harm or functional impairment for the individual concerned (American Psychiatric Association, 1980). This is vital as it helps to justify why a particular set of signs and symptoms should be labelled as a disorder rather than just an atypical variant of what is essentially normal functioning. Some authors have noted, however, that this requirement has been watered down in DSM-5, with a change of wording in the preamble concerning the definition of disorder. This now states "Mental disorders are *usually* associated with significant distress or disability in social, occupational, or other important activities" (American Psychiatric Association, 2013b p. 20, emphasis added), alongside a removal of a harm criterion from many diagnoses (Cooper, 2013a). For more on the DSMs definition of mental disorder see Lee (2012) and Stein et al. (2010).

Beyond the DSM's categorical nature, its descriptivism, and its general statement that mental disorders should (usually) be associated with harm, the DSM fails to paint a rich and coherent picture of what it takes mental disorder to be. It is interesting to consider whether this conceptual paucity relates to the DSMs intention to be atheoretical. Conceivably, and as alluded to at the end of the previous chapter, it is difficult to go beyond a surface level conception of mental disorder without making some firm commitment as to what it means for humans to be functioning well, or to otherwise *not* be experiencing mental disorder.

As a core institutional pillar of psychiatry and clinical psychology around the world, it is also important to consider, not just what the DSM is formally committed to,

[39] This categorical system does not fit with best evidence which calls for recognition of the fuzzy or even continuous boundaries (Haslam et al., 2012; Markon et al., 2011)

but how the DSM is actually used. As stated, the DSM's model of disorder is descriptivist rather than causal. In practice however, often encouraged by the DSM's algorithmic, polythetic, and categorical diagnostic structure, constructs are often treated as essential and objective entities. Drawing back to chapter two, *essential* refers to the constructs being 'real', discoverable, and very similar across different instances, and *objective* refers to the constructs being factually-based and not concerning values. To put things simply, DSM constructs are treated by many practitioners, researchers, and by the public, as more 'real' than is warranted (i.e. the problem of reification[40]). This can be seen in the common assumption that DSM diagnostic constructs such as depression must have some yet to be discovered cause (see 'essentialist kinds' in chapter 1).

There are in fact many recognized issues with the DSM model that give us reason to doubt the etiopathological validity and objectivity of its diagnostic constructs (For review of these issues see; Karter & Kamens, 2019; Lilienfeld & Treadway, 2016; Zachar & Kendler, 2017). Those most relevant to the question of validity include the issues of artefactual co-morbidity[41] (Andrews et al., 2002), symptomatic and etiological heterogeneity[42] (Lilienfeld, 2014), false positives[43] (Cooper, 2013a; Wakefield, 2015), concept creep[44] (Haslam, 2016), and the above mentioned problem of reification (Hyman, 2010). Taking these issues together, there is good reason to believe that the DSM does *not* adequately pick out valid clinical entities with stable causal structures for us to go about discovering. This firstly has important ramifications for both research

[40] 'Reification' is a Marxist term meaning 'a process of making the ideal real'. Hyman (2010) applies this term to mental disorder referring to the general tendency for DSM constructs (known to have issues with validity and to be reasonably artificial/constructed) to come to be seen as real entities through their use.

[41] Co-morbidity refers to when an individual has more than one mental health diagnosis at one time. Under the DSM, this occurs at much higher rates than would be expected if mental disorders were independent phenomena, suggesting that this may be an artefact of how we conceive of and measure our diagnostic concepts. Note that there is continuing debate on this issue.

[42] Heterogeneity refers to diagnostic constructs being too 'large', capturing meaningfully different individuals under the same label. This can include individuals with very different symptom profiles (symptomatic), and/or disorders with very different causes/constitutions (etiological). Under the DSM this occurs frequently (Contractor et al., 2017; Dickinson et al., 2017; Galatzer-Levy & Bryant, 2013; Hawkins-Elder & Ward, in press; Monroe & Anderson, 2015; Olbert et al., 2014).

[43] False positive refers to when people are diagnosed as having a disorder but probably do not have the disorder/a genuine problem.

[44] Concept creep refers to the observed tendency for our concepts of harm to grow over the last hundred years or so. I include this here as the cited paper by Haslam includes many examples from the DSM. If DSM concepts can expand (or contract) with social mores, this brings into question their objective nature.

and treatment, where we want to be able to assume that a disorder has similar causes and solutions respectively. But secondly, note the tension between the DSMs stated descriptive and atheoretical stance on one side, and how its constructs are used in practice and by the public on the other (as essential and objective). It is interesting to consider whether the lack of conceptual and theoretical commitment, explicitly fostered during the development of the DSM (to increase its applicability across professionals with diverse theoretical orientations), alongside the natural human bias to default to essentialist style thinking (Gelman, 2003) has actually played a role in encouraging this contrary use and interpretation of DSM constructs.

Research domain criteria (RDoC).

RDoC is a research funding framework proposed by the US National Institute of Mental Health (NIMH) in direct response to the acknowledged problems with the DSM. One of RDoC's keys goals is to shift attention from surface features to the underlying casual processes that generate signs and symptoms; it is a causalist model (Insel et al., 2010). RDoC adopts a central organizing structure in the form of a two-dimensional grid, with the horizontal axis containing seven 'units of analysis' which are largely structural in nature, and the vertical axis containing five domains/constructs, also referred to as systems, which are functional (Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Lilienfeld & Treadway, 2016; Morris & Cuthbert, 2012).

While the RDoC is not a diagnostic system, it is intended to lay the groundwork for one (Insel et al., 2010), although it is explicitly uncommitted to the form that this diagnostic system might take (Cuthbert & Kozak, 2013). However, it is reasonable to presume that diagnostic entities in the system will represent dysfunctions of the identified functional systems, observed at or through the lens of the various units of analysis.

The three foundational postulates of the RDoC are stated clearly by Morris and Cuthbert (2012):

1. Psychiatric disorders are dysfunctions of brain circuits.
2. The tools of neuroscience can identify these dysfunctions.

3. Clinical neuroscience alongside genetics research will yield bio-signatures of dysfunction that will augment classical clinical signs and symptoms of disorder.

Within the RDoC the explanatory focus has been shifted from DSM diagnostic entities (clusters of signs and symptoms) to transdiagnostic mechanisms that are thought to underlie them (Cuthbert & Insel, 2013; Hoffman & Zachar, 2017). The hope is that the identification of such mechanisms will allow for faster scientific progress, translation across levels of analysis, more precise medication and treatment, and perhaps even lead to the development of reliable bio-markers of psychopathology (Cuthbert, 2014; Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Insel et al., 2010; Morris & Cuthbert, 2012). Hoffman and Zachar (2017) also point out that the narrowing of scientific attention from disorders to transdiagnostic mechanisms (while being mindful of the original purpose for seeking an explanation) may hopefully provide a better understanding of the relationships between levels of analysis; the logic being that diverse/more separated phenomena at the macro level will be constituted by simpler and more homogenous sets of mechanisms at the micro level.

When applying for a grant through the RDoC system, researchers cross-reference the two dimensions (structural and functional) to specify the target/s of their study. The explanatory aim within RDoC funded research, therefore, is to study how some observation at a particular unit/level (e.g. higher levels of striatal dopamine, lower dendritic spine density in brain area X) affect the degree to which the functional construct is achieved (e.g. response to acute threat, approach motivation). *A priori,* an observation resulting in functional impairment is assumed to reflect a 'dysfunction' at the level of brain circuitry (from foundational postulate number one). Under RDoC, therefore, 'transdiagnostic mechanisms' refer to neural circuit abnormalities that negatively affect the specified functional domains. To put it simply, RDoC conceptualizes mental disorder as, or at least constituted by, *dysfunctional mechanisms* (mechanism here being used in an evolutionary/functional sense[45]).

[45] This being opposed to a more permissive/minimal sense of 'mechanism' that I will use in chapter 6. For a discussion of the different kinds and meanings of 'mechanism' refer to Glennan and Illari (2017), and for a discussion focused on this function vs. minimal distinction see Garson (2017).

In summation then, and drawing on the terminology introduced in chapter two, RDoC seems to reflect a form of *statistical functionalism* in the normative dimension, and something akin to *biological essentialism* in the structural (note that RDoc is also committed to the idea that the dysfunctions it reveals will come in degrees of severity; that they will be largely continuous with normal human behaviour). However, its central unit of interest is no longer the syndromes of mental disorder with which we are familiar, rather its focus is on hypothesized neural-level 'dysfunctions' which are assumed to be components of current DSM-style syndromes. To further clarify the RDoC's conceptual commitments, I will evaluate it against a conceptual taxonomy presented by Zachar and Kendler (2007). This taxonomy features six important factors upon which conceptions of psychopathology often differ and offers a concise way of sketching out conceptual positions in this area. Note that I interpret these factors as continuums rather than as dichotomies. Many of these terms have been introduced in chapter two – the exceptions being the last two listed – but I include Zachar and Kendler's definitions here for reference and to highlight where their use of the relevant terms may subtly differ from my own. Along the way I will state some key criticisms of the RDoC approach.

Causalism/descriptivism. This factor relates to the question "Should psychiatric disorders be categorized as a function of their causes (causalism) or their clinical characteristics (descriptivism)?" (Zachar & Kendler, 2007, p. 557). The primary motivation for RDoC is to shift to a causal model, one that picks out etiologically valid constructs in a way that the descriptivist DSM does not.

Categories/continua. This factor relates to the question "Are psychiatric disorders best understood as illnesses with discrete boundaries (categorical) or the pathological ends of functional dimensions (continuous)?" (Zachar & Kendler, 2007, p. 559). RDoC is committed to viewing the symptoms of psychopathology in dimensional terms, in opposition to the categorical DSM approach (Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Lilienfeld, 2014; Lilienfeld & Treadway, 2016). This means that features of psychopathology are viewed as quantitative extensions of normal behaviors or biological states and therefore exist in degrees. This is a significant strength of RDoC

and aligns with current evidence for the vast majority of constructs in the field of psychopathology (Haslam et al., 2012; Markon et al., 2011).

Essentialism/nominalism. This factor relates to the question "Are categories of psychiatric disorder defined by their underlying nature (essentialism), or are they practical categories identified by humans for particular uses (nominalism)?" (Zachar & Kendler, 2007, p. 558). In order to first locate the hypothesized transdiagnostic mechanisms it is a primary intention of RDoC to reverse the DSM psychopathology research model of noting clusters of signs and systems within the population and then investigating them. The resulting research model is one whereby abnormalities across the units of analysis are discerned during the study of both typical and atypical populations, and it is later observed how these atypicalities may be serving as causal mechanisms in dysfunction (Cuthbert & Insel, 2013). Given the stated commitment to a dimensional conception of disorder, if mental disorders exist in degrees, then research methods that cover the whole population and capture such continuous variation are needed. For this model to make sense, the assumption has to be made that the atypicalities exist in nature, waiting to be discovered. In this way there is an essentialist element to RDoC. However, it should be noted that in RDoCs current nascent state of development it is not clear whether this is the full story. If these observed atypicalities themselves constitute a disorder in the classification system that evolves from RDoC, then this would indeed seem quite essentialist. However, if these mechanisms – or perhaps regularly observed clusters of mechanisms – are labelled as 'disorder', and this is done for value-based or practical reasons, then this would situate the resulting diagnostic system as *moderately nominalist* in Zachar and Kendler's terms (aligning with evaluativism or pragmatism respectively).

Objectivism/evaluativism. This factor relates to the question "Is deciding whether or not something is a psychiatric disorder a simple factual matter ("something is broken and needs to be fixed") (objectivism), or does it inevitably involve a value laden judgement (evaluativism)?" (Zachar & Kendler, 2007, p. 558). Given RDoC's empirical intentions to work from the bottom up noting atypicalities across the population and from this inferring disorder, it seems very likely that the RDoC belongs in the objectivism camp.

This, and the discussion around essentialism versus nominalism above, is relevant to a criticism of RDoC raised by Wakefield (2014a). In his paper Wakefield criticizes the RDoC for its lack of conceptual validity, arguing that it has great difficulty explaining why a set of phenomena should be labelled a disorder. In its current form the RDoC relies merely on the abnormality of a phenomenon and its probabilistic association with some poorly defined harmful outcome (hence my labeling RDoC as statistical functionalist). Grounded in his *Harmful Dysfunction* model Wakefield demonstrates that this is not sufficient, and is actually a step backwards from the DSM, which at least attempts to delineate the disordered from the simply atypical. While, as seen in chapter two, Wakefield's model faces significant issues, this critique of RDoC's statistical functionalism rings true. Abnormality is not disorder. Diagnoses should be given with the interest of the client in mind, and we should not pathologize behavior unless there is evidence of dysfunction or harm.

Internalism/externalism. This factor relates to the question "Should psychiatric disorders be defined solely by processes that occur inside the body (internalism) or can external events also play an important (or exclusive) defining role (externalism)?" (Zachar & Kendler, 2007, p. 558). The RDoC would certainly come under the internalism banner. Of the seven units of analysis, five are "beneath the skin", with the other two being behavior and self-report, which are still focused on the individual rather than on interpersonal or situational factors. Further, explicit conceptual focus is given to the 'brain-circuit' level, at which disorders are primarily located. The privileged status of neural circuits in the explanation of mental disorders makes it vulnerable to the criticism of being overly "neurocentric", and reductionistic (Berenbaum, 2013; Hershenberg & Goldfried, 2015; Hoffman & Zachar, 2017; Kirmayer & Crafa, 2014; Lilienfeld, 2014; Lilienfeld & Treadway, 2016).

Entities/agents. This factor relates to the question "Should psychiatric disorders be considered to be things people get, or are they inseparable from an individual's personal subjective make up?" (Zachar & Kendler, 2007, p. 559). The RDoC would seem to fall somewhere in the middle of these two possibilities, perhaps leaning slightly towards the entity view. Given its stated focus on lower units of analysis it seems to lack the holism required to encapsulate an agential and purposive perspective. At the

same time RDoC does not see sufferers entirely as mere vehicles of mental pathogens. Rather, through its functional axis, RDoC alludes to the biological norms of an organism and includes behavior as a unit of analysis, which suggests an awareness of an interplay between disease processes and individual agents.

Conclusions regarding the RDoC. In this discussion I have focused on two key criticisms which I see to be fundamental, and therefore intractable without radically changing the RDoC's central assumptions. Firstly, the claim that the RDoC is too neurocentric (Berenbaum, 2013; Hershenberg & Goldfried, 2015; Hoffman & Zachar, 2017; Kirmayer & Crafa, 2014; Lilienfeld, 2014; Lilienfeld & Treadway, 2016). RDoC authors have attempted to rebut the claim of neurocentricism, but it remains a popular criticism (Cuthbert & Kozak, 2013). Secondly, the argument that it lacks conceptual validity; it is not clear why an atypicality noted in the RDoC framework should be seen as a dysfunction/disorder (Wakefield, 2014a).

I propose that these two problems stem from the same root; RDoC's underlying assumptions concerning the nature of the mind. Murphy (2017) has recently argued that RDoC is grounded in a wider framework of human functioning known as *eliminative materialism,* albeit in a moderate form. This is the view that phenomena at higher levels – such as human cognition and behavior – are ultimately reducible to lower levels such as the biological or molecular, and that 'folk psychology' explanations rooted in higher levels will be eliminated or heavily revised as science progresses. Whether one agrees with Murphy's particular labeling or not, it is at least fair to say that he demonstrates that RDoC has reductive and materialist aspirations. This observation is all that is required for my argument to stand. Such reductive aspirations make sense of RDoC's neurocentricism, and the conceptual validity issue. This is because the labelling of a phenomenon as a disorder seems unjustified without reference to norms and values from which to demarcate harm or dysfunction. As I will show in following chapters, the normativity of concern in mental disorders is best conceptualized as an emergent property of the entire organism and is therefore difficult to account for under a reductionist framework. This of course begs the question of whether there is an alternative to eliminative materialism. The following chapter argues that *embodied*

enactivism is such an alternative, and offers a sketch of what psychiatric disorder might look like from this perspective.

Before moving to this however, I should briefly mention that I am aware there will be disagreement over how RDoC is represented here. Regarding the view that RDoC sees mental disorders as brain disorders, Cuthbert and Kozak (2013) state that: "…this controversial assumption is neither essential nor inherent to the RDoC initiative…" (2013, p. 931). They make the alternative claim that: "… statements [which appear neurocentric] are interpretable as an expression of the need to move beyond symptom-based nosologies for mental disorders…" (2013, p. 931).

Such a position aligns with the more integrative aspirations alluded to on the RDoC website regarding the need for developmental and environmental considerations. However, it is directly contradictory to the core assumptions of the RDoC stated by Insel et al. (2010), and Morris and Cuthbert (2012), which are clearly brain focused (listed earlier in this chapter). This inconsistency suggests variation in the conceptual positions of the RDoC authors, with some positions being less neurocentric than others. For clarity this thesis assumes RDoC to take the more neurocentric position as per its stated core assumptions. However, the general argument is still applicable if subscribing to RDoC a more moderate position, despite the fact that such views may not constitute *eliminative materialism* proper. While the claim of neurocentricism is partially weakened against these more moderate views, the criticism of lacking conceptual validity still holds. The RDoC of Cuthbert and Kozak (2013) fares much better than that of the more neurocentric parties, but still underrepresents both socio-cultural factors and the normative nature of diagnosis.

A Possible Way Forward

In this chapter I have attempted to briefly demonstrate the growing consensus that the DSM, despite its good intentions and institutional status, is not fit for purpose. I have tried to show that its weaknesses run to its conceptual roots in that it does not paint a clear picture of human functioning, and thereby the DSMs notion of 'dysfunction' remains conceptually thin. The RDoC, composed in reaction to the recognized issues with the DSM, does provide us with an understanding of human

functioning, but it is a neurocentric view that struggles to explain dysfunction/disorder in a wider sense. In chapter six I will show that I think RDoC will still be a useful tool in the efforts to explain mental disorders. Now however, we can begin our task of assuming a rich understanding of human functioning and observing what mental disorder looks like from such a perspective. As noted, one such view – and the focus of this thesis – is the philosophical orientation of embodied enactivism.

Chapter 4: Questions of Structure

In this chapter I propose that a philosophical orientation referred to as *embodied enactivism* – seeing the mind as embodied, embedded, and enactive (alternatively labeled *3e Cognition*) – has the potential to provide a more integrative and richer framework of human functioning within which to study mental disorders. I first introduce embodied enactivism, before looking at some previous attempts to take an enactive/embodied perspective on particular mental disorders. I comment on the strengths and failings of such approaches before moving on to highlight some of the theoretical tools embodied enactivism provides. I then present the beginnings of a conceptual sketch as to what disorders look like from the perspective of embodied enactivism. Following the format I used in chapter two, I am here focusing on the *structural nature* of mental disorder, leaving normative considerations aside till chapter five.

Embodied Enactivism

By embodied enactivism I firstly refer to the view that the mind is fully material, and that it is constituted by not just the brain, but the brain-body system; we are *embodied* beings. The mind then is not a thing above and beyond the organism, neither in the Cartesian mind-substance sense (i.e., we are not made of 'mind stuff'), nor in an information-theory sense (i.e., our mind cannot be uploaded to a computer[46]). More than this, interactions with the physical and social environments within which the organism is situated provide necessary conditions for the development of the mind over time; we are *embedded*. To explain human behaviour then, we are not just interested in the brain-body system, but the brain-body-environment system. Finally, we are *enactive* creatures. According to enactivism[47], organisms are intrinsically purposive; more specifically they strive to *self-maintain* and *adapt* to changing circumstance (Di Paolo,

[46] To do so would only ever produce a copy, and one of a fundamentally different kind given its disembodied nature. In a similar vein see Thompson and Cosmelli (2011) regarding the brain-in-a-vat hypothesis as an argument for embodiment.

[47] The specific version of enactivism being described here is sometimes referred to as autopoietic enactivism, due to the central inspiration/metaphor for this position being the autopoietic process observed in cells. This is contrasted against more strictly anti-representational brands of enactivism – i.e. Radically Enactive Cognition or REC (Hutto & Myin, 2012) – as well as theories focused on perception – i.e. Sensori-Motor Enactivism (O'Regan & Noë, 2001). For discussion of these labels see Ward, Silverman, and Villalobos (2017).

2005; Thompson, 2007). This striving, inherent to all life, sets up a natural normativity and grounds the development of *meaning for the organism* through its needful relationship with its environment (Colombetti, 2014; Thompson, 2007).

In this section I will attempt to briefly sketch out the embodied enactive position. My intention is simply to offer an outline of embodied enactivism, not to argue for it. For a presentation and defense of the embodied enactive viewpoint see: Colombetti, (2014); Durt, Fuchs, and Tewes, (2017); Fuchs, (2017); Gallagher, (2006, 2017); Gibbs, (2005); Maiese, (2016); Thompson, (2007); Hutto and Myin (2012, 2017) and Varela, Thompson, and Rosch (2017/1991).

To start with a classic example, consider a simple life form such as a bacterium. Bacteria have an interesting tendency to move towards concentrations of sugar (their food source) and away from certain substances that are toxic to them. This is achieved through a simple mechanism by which the motions of their flagella are responsive to the concentrations of sugar and some toxic substances. Embodied enactivism highlights that this is an *evaluative* process; one by which the bacterium is acting *in the interest* of its own survival. The claim here is not that simple bacteria are conscious, but that sugar has an embodied *meaning* for the bacterium as 'good/food', and that there is therefore a simple *mindedness* at play here, whereby the organism is responsive to the conditions required for its own survival. What the embodied enactivist notices in this situation is that this dynamic is present in, and perhaps definitional of, all life – that *mind is in life* (Thompson, 2007).

This connection between the structures of life and mind is referred to as the deep continuity thesis (DCT), and I will return to it in more detail in the next chapter. For now, suffice to recognize that there is a sense in which things can be good or bad for bacteria, trees, tigers, and people, in a way that things can't be good or bad for a pile of rocks. This is because, essentially, it is easy for these life forms to die, and hard for them to keep living; they are *precarious* in that they are 'far-from-equilibrium' systems. The process of self-maintenance requires a metabolism, and therefore energy, which is sourced from the environment (at the cellular level this is referred to as auto-poesies; literally 'self-creation'). For the enactivist this precariousness, and the needful relation it

establishes between the organism and its environment, is the root from which meaning develops.

Within embodied enactivism literature 'cognition' and 'perception' are seen as reasonably continuous and are often referred to as *sense-making*. This highlights their *relational* nature. We can see this in the above example of the bacteria; the sugar and toxins have meaning *in relation to* the bacteria's precarious situation as a life form. By responding differentially to these two substances in its environment, and in a way that accords with its self-maintenance, there is a very basic sense in which the bacteria as a system is *making sense* of the world.

As another example of sense-making and relational qualities, take the colour red. Redness is what is referred to in philosophy as a 'secondary quality' because redness is not *in* an object the same way something's mass is. Redness rather, is subjective and experiential, phenomenal rather than noumenal (although as we will see, enactivism helps us move beyond this dichotomy). If there were no experiencing agents in the universe then there would still be objects, these objects would have mass, and light-waves would bounce off those objects in certain ways. But there would be no 'red'. So where does red come from? According to traditional cognitivist thought, redness is 'in the mind'; redness is an experience/neural-code in our mind/brain, hallucinated in response to a particular pattern of activation in the optic nerve. Something about this seems rather absurd. Embodied enactivism provides a different answer- redness is *relational* (Fuchs, 2017). It exists *between* the agent and the world, generated[48] – or 'enacted' – by the organism, to help it *make sense* of the world in accordance with its needs. Those from diverse theoretical orientations can likely all agree that certain organisms have evolved to experience red (and other colours) as they do because it is useful and helps them survive to pass on their genes. Under embodied enactivism though, redness and other colour exist *for* the organism, not as part of a model or hallucination of the world in the mind/brain, but as a learnt and evolved mode of experiencing the world directly. Under embodied enactivism there is a veridical world, but there is also an *Umwelt*; that same world as experienced from a concerned point of

[48] I use the word 'generated' here very tentatively. I do not want to imply that enaction/sense-making is always an active/conscious process.

view, or the world *for* the organism. The umwelt then is a world of immanent meaning and valence (Thompson, 2007).

Embodied enactivists also have a very different understanding of emotionality. From this perspective the affective nature of our experience – the meaning that is immediately apparent in the world around us or what Maiese (2016) calls *affective framing* – can be seen as real and thorough-going (Colombetti, 2014; Colombetti & Thompson, 2008). If I feel angry, this emotion is not simply a brain-state that 'signals' something about the world, but a dynamic pattern of states and processes cascading throughout my body, developed across evolutionary and life-span development, which primes me for certain actions such as punching and yelling. Through socio-cultural experiences across my development I have learnt to recognize the experience of this embodied cascade as "anger". Anger and other emotions then, are not separate to cognition, or deviations from some idealized rationality, but are part and parcel of living in a world that has meaning and valence for us. Emotions are temporary fluctuations of intensity, not against a back drop of neutral rationality, but in an organism-umwelt system that is *primordially affective* given our precarious situation as biological organisms (Colombetti, 2014).

Thus through embodied enactivism the mind can be seen, not as a linear symbol processing machine with a defined inputs and outputs, but as *mindedness*; an emergent property of the whole organism arising from interactions in the brain-body-environment system to better serve the organism's *self-maintenance* and *adaption* (Di Paolo, 2005; Maiese, 2016; Thompson, 2007). The mind then is not well modelled by computer metaphors and reference to such things as 'representation' and 'processing', but rather is better modeled by life itself[49]:

> "a natural cognitive agent – an organism, animal or person – does not process information in a context-independent sense. Rather it brings forth or enacts meaning in structural coupling with its environment." (Thompson, 2007, p. 58).

[49] This recognition that the core idea of enactivism is to shift from a computer analogy of the mind to a 'life-form analogy' is based on a comment made by Dan Hutto (personal communication).

Representing a sub-type of enactivism, some authors – so-called 'RECers' (REC: radically enactive cognition) – attempt to forgo ideas of 'representational content' all together[50] (Hutto & Myin, 2012).

When seeking to understand someone's behaviour, the embodied enactivist takes as their central unit of analysis the sense-making organism within its context; i.e. they consider the whole *brain-body-environment* as a dynamic system. Embodied enactivism is thereby strongly anti-reductionistic. By this I mean that embodied enactivism is incompatible with ideas of theory-reduction, where explanations of 'higher-level' phenomena (such as human behaviour) are seen as in-principle reducible to the language of 'lower-level' theory (such as genetics)[51](Andersen, 2016; Brigandt, 2013). This is because, implicit in the embodied enactive view is a commitment to 'down-ward causation'; the view that phenomena at higher levels can influence the

[50] While being slightly tangential, I think it is important to explain where I stand on the issue of representation. The core motivation of this RECer position is referred to the 'hard problem of content'. This is the observation that information (understood as a difference that makes a difference) can't really carry meaning itself – i.e. it doesn't have "content". Rather, the information has meaning *for* the receiver. Harvey (2015) explains this complex idea very well when demonstrating that the issue also applies to language. If I say to you "that is a very cool coat you are wearing", then this sentence only has meaning within the local linguistic-cultural community; those that speak English. An enactivist position would hold that the sound waves I generate in this example stimulate an enaction of meaning in the listener because of our shared developmental histories. The words have meaning *for* someone. Turning this idea 'downward' into the brain, an important question arises: who is the agent for which a 'representation' in the brain has meaning? My position on this, very tentatively, is that if the person is consciously aware of these representations as 'thoughts' then perhaps they have meaning for the person in question, but if they are not aware of them (i.e., if they are sub-personal), then these 'representations' only seem to be information in a deflated sense – where they are just differences in a subsystem of the brain that make a difference for other subsystem of the brain. To return to the language analogy, perhaps patterns of neural impulses are like the phonemes of language, themselves devoid of meaning until combined in certain ways that have meaning for the organism. For many enactivists then, when we try to interpret what is going on in someone's brain, the apparent observation of 'representations' is an artifact of our third-person point of view. Ultimately though, I am not taking a side in this debate concerning representation. Firstly, there is not room in this thesis, and I am not a philosopher of mind. Secondly, while on the enactive view the idea of representation is problematic in an ontological sense, I do wonder if it is perhaps still a useful way of trying to understand or model the complex dynamics of the brain in a way that is interpretable by the meaning-loving creatures such as we are (i.e. maybe representation works in a pragmatic epistemological sense). My sense is that there seems to be some confusion as to what different people mean by 'representation', and that a contentless understanding of sub-personal 'representation' may be coherent with an (autopoietic) enactive worldview.

[51] While theory-reduction is certainly at odds with the 3e world view, I do not see a good reason why the embodied enactive idea is incompatible with ideas of explanatory reduction/causal-mechanistic explanation – where wholes are broken down into parts to try and understand some property of the whole (so long as the holistic perspective is not sacrificed). In fact I think this style of explanation may be complimentary to the traditionally dynamical approach (see; Bechtel, 2009a; Brigandt, 2013; D. M. Kaplan, 2015). I will return to this in chapter six and seven.

behaviour of entities at lower levels[52]. I will return to this concept later in the chapter when I pull out some core theoretical tools that are present in the embodied enactive view. The reason I mention this here is to explain a terminological shift that I would like to make at this point in the thesis. Rather than the traditional ontological 'level', I will from this point refer to 'scale'. This is to highlight that, in a world featuring down-ward causation, simply because a phenomenon exists at a smaller scale, this does not make it somehow more fundamental or important. This shift is more in-line with the embodied enactive view (for further reasons for this shift in terminology see; Potochnik, 2010; Potochnik & McGill, 2012).

So, to summarize, under embodied enactivism mental processes are necessarily *embodied* in the brain, nervous system, and all other biological systems of the body – they are things that we do (i.e. embodiment). These processes necessarily occur within an environment with which we are richly and bi-directionally causally connected (i.e. embedment). For social-cultural creatures such as ourselves this includes a social environment, which we as a group constitute. Phenomenological experience and meaning emerges (i.e. is enacted) by virtue of the organism making sense of and adapting to the world (Di Paolo, 2005); it *is* the body making sense of itself and the world (Fuchs, 2017). Ultimately the enactive/embodied conception of human functioning is based on a relatively simple idea: psychological functioning and sense of meaning is shaped in fundamental ways by our nature as biological organisms. As striving organisms we have needs, but further, we have a way of achieving these needs, within our contexts, based on our personal, cultural, and species-level histories (Gallagher, 2006; Thompson, 2007; Varela et al., 2017).

Before moving on, note that I am using the term 3e as an alternative for the embodied enactivism label, when the term 4e is often used. I do so because I do not subscribe to the fourth 'e' - *extension* (where the mind is seen as partially constituted by the external environment; A. Clark & Chalmers, 1998). My reasons for this are multiple but I will briefly allude to them. Firstly, it is doubtable that full extension is compatible

[52] Although it should be said that some enactivists would even reject this label – they fear that 'down-ward' still implies a hierarchy of importance rather than merely scale. This is a reasonable concern but I still find the imagery useful, hence I have compromised by using the term 'scale'.

with enactivism and embodiment given that the latter two emphasize the process of continual separation between organism and environment (self-maintenance), while extension de-emphasizes this (Maiese, 2017). Secondly, enactivism holds that meaning is always relational – it is generated by an organism through its needful relation with the world (Fuchs, 2017; Thompson, 2007). The constitutional boundaries of the organism become blurry and ever-changing under extension (Maiese, 2017), and this seems to make the nature of the enactive relation very unclear. Thirdly, for our purposes at least, subscription to embeddedness (rich and necessary causal relations between organism and environment), as opposed to extension (constitutional expansion), can achieve much of the same conceptual ends while allowing for clearer explanations, e.g. it would be very hard to explain the depression of some client 'John' if we spend our time trying to decide where 'John-the-system' ended and his environment began. Fourthly, many brands of extension seem to rely on an information-processing account that I disagree with due to their running clearly afoul of the hard problem of content – see footnote 50 (for more on this see: Harvey, 2015; Hutto & Myin, 2012). Finally, Thompson and Stapleton (2009) show that once the concept of extension is cut to size in-order to fit with embodied enactivism then genuine extension of the mind becomes a much less remarkable and quite rare phenomenon.

Previous Work in Embodied Enactive Psychopathology

Limited work has been done to bring conceptual analysis of the nature of mental disorders together with the embodied enactive perspective. Drayson (2009) lays down the challenge:

> "…for embodied cognitive science…to…come up with an explanatory model of the origin and development of psychiatric disorders that can adequately compete with the current orthodox model" (Drayson, 2009, p. 339)[53].

Drayson (2009) argues that such a model would potentially show great promise for those disorders with large bodily components such as impairments of mood or eating, but questions how it could be applied to psychiatric problems in which representational

[53] Note that in a sense Drayson is jumping ahead here, before we can explain mental disorders from this viewpoint surely we need a clear concept of what it means for something to be a disorder under this worldview.

content such as delusional disorders have a prominent role. I view these as open questions.

Fuchs (2009), offers some insight into what such a conceptual framework could look like. His main conclusion is that it will necessitate *multi-scale analysis* and a focus on *circular causality* operating across all scales. As a consequence, in contrast to the cognitively dominated and reasonably brain-bound orthodox models of psychopathology, Fuchs prescribes greater focus on perception and action as these are the primary modes by which we are coupled with our environment. Taking such a multi-scale approach allows for a more comprehensive view across brain, body, and environment, as required by the 3e framework (also see; Fuchs & Schlimme, 2009).

Previous 3e explanatory models.

Some authors have attempted to generate embodied and/or enactive explanatory models of particular disorders. These models are often grounded in a phenomenological approach which features a rich but often confusing terminology. I have attempted to translate these ideas into standard psychological concepts but acknowledge that in the process of doing so some coherency and richness may be lost.

Zautra (2015) presents an enactive account of addiction. He describes how current models fail to offer an account of the first-person experience of addiction, how they do not give sufficient weight to interactions with the social and physical environment, and how they tend to be based on an entity conception of addiction and fail to recognize the agency of the individual suffering. Zautra addresses these problems by emphasizing the "lived experience" of addiction. His model describes how exposure to the drug has developed a need for it within individuals, and how they meet this and other needs in accordance with the affordances and constraints of their environment. Further, the model emphasizes how having this dominant need changes the agent's embodied experience in a multitude of ways; from attentional processes, through impulsivity, to their relationships with emotion.

While Zautra's work provides a valuable *description* of addiction, one that is compassionate and that emphasizes agency, there are many issues with this model. Greater detail is needed concerning how the relevant need is constituted within the

individual, how this is triggered by the drug, and why this need becomes dominant for some and not others. Essentially, whilst offering an insightful perspective on the *experience* of addiction, the model does not offer the multi-scale view that Fuchs (2009) suggests a 3e perspective should generate. This analysis of Zautra's model suggests two conclusions. Firstly, an optimal model of disorder from the 3e perspective should align with the subjective experience of the sufferer, and thereby both generate compassion and define the explanandum. Secondly, as an optimal model of mental disorder in general, it should offer mechanistic insights at multiple scales above and below[54] the scale of the first-person perspective in order to best guide treatment.

Fuchs and Röhricht (2017) and Maiese (2016) present embodied and enactive accounts of schizophrenia. Within these models the primary dysfunction is seen as a breakdown in the experience of the basic or bodily self, also referred to as the experience of *ipseity*; the first-person 'givenness' of all experience. This breakdown results in a lack of unity of perception, action, and thought, whereby the relation to objects in the world, thoughts in the mind, as well as body parts and their actions, lack qualities of wholeness and 'for-me-ness'. Response to these experiences produces hyper-reflexive self-observation and feelings of isolation and detachment. Trust in others thereby becomes very fragile, and the shared understanding of the world is damaged by this, fostering the development of delusions. Maiese's model differs slightly, with greater focus on the role of affective relations. Both offer rich subjective accounts of schizophrenia, with some explanatory value in that they account for many signs and symptoms of schizophrenia as understandable psychological responses to a central feature. However, as with Zautra's (2015) account of addiction, these are not fully explanatory models. It is unclear what the origins of the disruption to the experience of the bodily self are, and how this disruption to experience is constituted at lower scales of analysis. Finally, more specific detail is also required concerning the role of interpersonal, social, and cultural factors.

All three models offer first person accounts of their respective disorders which have epistemic and pragmatic value. However, they miss factors situated at scales of analysis above or below the first-person experience. This seems to be due to their

[54] I use the terms 'above' and 'below' here for convenience, not to imply a hierarchy of anything other than scale.

grounding in phenomenology, which is largely descriptive and concentrates on subjective experience. Hence these accounts focus on the *experience* of embodiment but de-emphasize the lower scales of analysis that constitute individuals, and factors from higher scales that constrain human functioning (e.g., social institutions); seemingly because both elements require a third person perspective. In order to offer a sufficiently comprehensive causal account, cross-scale analysis is required (Kendler, 2012b, 2012a; Kinderman, 2005).

Kyselo (2016) makes a similar observation concerning enactive accounts of schizophrenia. She outlines a model developed by Parnas and Sass (2010) which is centered around dysfunction in the experience of ipseity, and notes that it is descriptive rather than causal. Alongside this analysis she explores a proposal by Ebisch and Gallese (2015), based on an empirical review of recent neuroscientific evidence, that disturbances in the experience of ipseity and the distinction between self and others may be partly caused by a disruption in multi-sensory integration within the ventral pre-motor cortex. Kyselo argues that these two perspectives can be seen as complementary, one as a description, the other as a potential account of underlying causes. Kyselo goes on to offer two criticisms of importance. Firstly, she states that both models' stress on the first-person perspective results in a failure to emphasize social elements. She argues that such an individualistic focus does not make room for incidence and prognostic factors like social support and socio-economic status (Agerbo et al., 2015; Bhavsar et al., 2014; Buchanan, 1995; Lim et al., 2017; Tsai et al., 2014), nor for consideration of cultural factors that may shape both contexts and the individuals' understanding of their experiences. Secondly, she criticizes both papers for defining disorder simply in opposition to the normal, rather than in terms of the norms and values of the individual. This is related to the idea of conceptual validity mentioned earlier. A conceptual model of psychiatric disorder ought to make sense of why a cluster of clinical phenomena[55] should be labelled a mental disorder, and not do so simply on the basis of deviation from the normal. Kyselo goes further, presenting a view of psychopathology as "an

[55] I have used the terminology 'cluster of clinical phenomena' here. As I hope to show in chapters six and seven, 'phenomenon' (singular) may not be appropriate as it implies a stable and singular explanatory target. 'Clinical' is used to highlight that the phenomena investigated should be of clinical relevance.

altered form of striving for quasi-equilibrium in the organization of a person's self." (Kyselo, 2016, p. 607). This refers to seeking a balance between connecting to and differentiating ourselves from others, and is based on Kyselo's socially defined conception of the self. Problematically, it is not clear how this model accounts for the phenomena which currently define the schizophrenia construct such as hallucinations, delusions, and negative symptoms. I do not therefore view Kyselo's model as explanatory either. However, a particular strength of her model is that it attempts to define disorder on the basis of the norms of the individual rather than by the abnormality of the observed behaviour.[56]

Reflecting on these previous attempts, three conclusions can be drawn. Firstly, previous 3e models of mental disorders have typically focused on a *subjective mode* of explanation, likely due to their roots in phenomenology. While this approach is of value, comprehensive explanatory models need to employ a broader multi-scale analysis (Kendler, 2012b, 2012a; Kinderman, 2005). Such a perspective is required given our embodiment and embedment as biological and social organisms (Fuchs, 2009; Fuchs & Schlimme, 2009). Secondly, as alluded to by Kyselo (2016), and in alignment with the criticism of RDoC raised by Wakefield (2014a) concerning the need for conceptual validity, disorders need to be defined by more than just abnormality. Thirdly, previous conceptual and specific explanatory attempts which draw on embodied and enactive perspectives have not attempted to integrate current literature surrounding conceptions of psychopathology[57].

Embodied Enactivism and the Structure of Disordered Behaviour

As a field *embodied enactivism* is consistent with naturalist and non-dualist assumptions. One key strength of the 3e view however, is *how* it meets these assumptions. It does this in a way that places equal value on physiological processes and on first personal and interpersonal scales of explanation, because they are all different elements of the dynamic whole – an agent standing in relation to their environment. In accordance with the comprehensive view prescribed by Fuchs (2009), genes and

[56] Other explanatory models have been composed from a 3e or related perspective, one notable example is that of Autism Spectrum Disorder (De Jaegher, 2013).

[57] De Haan (in press-b) would be one exception here, which I will discuss in the final chapter.

neuronal networks are vitally important for understanding the etiology and symptoms of psychiatric disorder, but so are emotional regulation skills, interpersonal relationships, and culture. Further than this widening of the lens across scales, 3e cognition also broadens our view laterally, to biological factors that have been historically overlooked because they lie outside the central nervous system. A good example of the relevance of extended biological processes are recent findings concerning the importance of the gut biome and nutrition for mental health (B. J. Kaplan et al., 2015; Rucklidge & Kaplan, 2013).

There are also important theoretical ideas contained within embodied enactivism, such as those of emergence and constitution. *Emergence* is the view that a whole may gain properties from the interaction of its parts rather than being simply the sum of them. A classic example of this is the phenomenon of starling murmurations, where birds have been shown to respond to the seven or so birds in their local environment, resulting in what appears to be coordinated behavior of the whole flock, confusing predators (King & Sumpter, 2012). A simpler example is water. The property of water being a liquid is not held by a single H^2O molecule, rather it is emergent from the interaction of multiple H^2O molecules repelling each other due to their dipole structure.

Constitution is the idea that wholes can be made up of parts without the whole being eliminated or becoming meaningless as an explanatory entity. For example, if you build a tower of lego, both the form of the tower and the lego blocks exist and can be useful in an explanation of why the tower fell over under certain conditions. During the time that the lego blocks constitute the tower the blocks *are* the tower. Similarly, organisms are made up of many parts, and derive properties, such as mindedness, from the interactions between these parts. Both the parts and the organism are no less real because of the knowledge we gain about their parts and how they manage to constitute a minded creature. These conceptual tools are not available to more reductionistic perspectives such as *eliminative materialism*, yet they are arguably necessary for a comprehensive conception of psychopathology where an understanding of both wholes and parts, as well as the interactions between them, is required.

Related to its roots in dynamic systems theory and evolutionary theory, a final strength of a 3e cognition framework is that it provides a way to account for something akin to what Aristotle referred to as *final causes* (Falcon, 2015), and Kant called *purposiveness* (Ginsborg, 2014). Intuitively, the idea that a desired end state can cause an action seems to be a teleological error where the future acts backwards on the past, and this seems at odds with a mechanistic view of the universe. 3e cognition offers an elegant solution: the organism system is shaped by its ontogenetic and phylogenetic past to act in accordance with its needs, and to do so in accordance with the constraints of the environments that shaped it. This allows an organism to live in a valenced world; to have purposes, goals, and even values inherent as tendencies within the organism system (Maiese, 2016; Thompson, 2007). This is achieved without the future acting backward on the past, thus granting a sort of naturalized teleology. To clarify, I do not wish to endorse final causes as a function of essence as per Aristotle (i.e. birds fly because it is part of what it means to be a bird to fly), rather I wish to suggest that life forms can be seen as having purposes and goals in a non-trivial sense, in so far as they have been naturally selected across time to self-maintain and adapt (Thompson, 2007).

To briefly sum up what human functioning looks like from the embodied enactive perspective, multi-scale explanations are required to thoroughly account for behavioral phenomena, with no preference given to any particular scale simply because it is higher or lower. *Particular scales* of explanation may be *of specific importance in any instance*, while the behavior itself – from both a first- and third-person perspective – is obviously of import as the explanandum (i.e., explanatory target). Relations between scales can explicitly be constitutional, a point left unclear in current formulations of the RDoC (Hoffman & Zachar, 2017); and phenomena can emerge at higher scales that would be impossible to predict from an understanding of lower scale structures and processes. Furthermore, higher scale processes can act *downward* to constrain/enslave processes at lower scales. Finally, the explanatory tools outlined above allow us to see how, shaped by evolutionary and developmental histories, behaviors can have purpose in supporting the continuation and flourishing of the organism – they can have *meaning* for the organism and understanding this is vital if we really want to understand a

behaviour. This is a potential source of conceptual validity that I will return to in the next chapter.

For now, however, consider what the *structure* of mental disorders looks like from this perspective. If you can, try to ignore that they are 'bad' in anyway – ignore the normative concerns. From this purely structural view, mental disorders are simply repetitive patterns and tendencies in 'behaviour' that seem to occur in a similar manner across individuals. Note that by 'behaviour' I am referring to actions, thoughts, emotions, sense-making, basically everything that an organism does (from an embodied enactive view, even the act of perceiving is often seen as a behaviour; O'Regan & Noë, 2001). Considering this, and the view of human function I have just summarized, an image of a complex causal/process structure starts to appear. For an embodied enactivist, complex behaviors cannot be accounted for by simply looking at neural processes, nor is it likely a simple function of some social-level factor. We are bodily organisms richly embedded in a physical and social world. The body, the physical and socio-cultural environment, as well as considerations of evolution, development, and the meaning we find in the world, are vital for understanding both why behavior is performed, and why it takes the form that it does.

In the absence of some dominant causal factor, such as those we have failed to find for mental disorders, we can make a further inference. Given the inflexible nature of disorder – the fact that mental disorders often represent *repetitive* patterns of behaviour– we can infer that the causal structures supporting these patterns of behaviour are likely *locked-in* in some way; that they themselves are *stable dynamic patterns of causal relations within the brain-body-environment system*. For example, consider the differences between Parkinson's Disease and Depression. In Parkinson's we can observe a behaviour – shaking and loss of motor control, among others– but these are tied to a relatively homogenous set of known causal factors within motor areas of the brain, such atrophy of dopaminergic neurons in the substantia nigra. The reason for the maintenance of this behaviour then is reasonably 'in the brain'[58]. Compare this to

[58] This is not to say that the wider pattern of difficulties that people with organic diseases experience cannot be fruitfully analysed though a system-wide lens. Such an analysis would also likely highlight a complex network of causal relations impinging on a sufferer's wellbeing, but we can visualise the network in this instance as being much more centralized around the core pathogenic process in the

depression, where we are not aware of any dominant causal factors within the brain. Assuming that we continue to 'fail' in this search for the 'essence' of depression, we do not need to relegate depression (nor other mental disorders) as merely problems in living. Instead, embodied enactivism, with its view of the entire brain-body-environment as a dynamic system, highlights the possibility that the maintenance of dysfunctional behaviour may be emergent from a network of factors and feedback-loops across the system. Rather than representing an undiscovered disease process in the brain, mental disorder syndromes may represent circular multi-scale networks of causal relations. In short, from an embodied enactive view, the causal structure supporting repetitive patterns of behaviour (remembering that I mean this in a wider sense, inclusive of tendencies in emotions and thoughts etc.) starts to look a lot like the fuzzy MPC kinds we explored in chapter two (Kendler et al., 2011). This is what I take Fuchs (2009, 2017) to be highlighting when he discusses the concept of circular causality. One key distinction to make between the MPC view and the view expressed here, however, is that on the embodied enactive view this MPC causal structure is necessarily existing within the adaptive processes of an agent-in-relation-to-the-world. The epistemological distinction/isolation/abstraction of the 'disorder' from this agent-environment system is performed on a normative basis, and this is what I will explore in the following chapter.

In this chapter I have introduced embodied enactivism, highlighted some of the theoretical tools in brings to the table, and attempted to begin a sketch of what mental disorder may look like from the embodied enactive perspective, focusing on the structural dimension. So far, it may seem that this complex exercise of trying on a new framework of human functioning has not got us very far. What we have discovered seems reasonably supportive of an extant concept; that of mental disorders as MPC kinds. In the next chapter, I will shift to exploring the normative dimension of mental disorder through the lens of embodied enactivism. Here I hope to show that embodied enactivism brings a novel perspective as to *why* some behaviors should be considered disordered. Further, in Chapter Six I aim to discuss the structural and normative

---

brain – as being denser in the middle. The claim I am making is that the network supporting depression and other mental illnesses is likely more diffuse (although there may well be hubs of causal connections, within the brain or elsewhere). The structures of mental disorder and physical disorder seem continuous in this way.

considerations together and demonstrate that the resulting concept of mental disorder does in fact provide us with novel insights.

Chapter 5: Questions Concerning Normativity

In this chapter I return to the question foreshadowed by Thornton (2000) and explored at the end of chapter two: Can assuming a richer and non-reductionistic worldview allow us to see beyond the evaluativist/objectivist dichotomy and understand the role of values/normativity within the concept of mental disorder in naturalistic terms? In other words, what I am exploring here is whether embodied enactivism affords us a way to see values and normativity as a natural part of the world and thereby collapse the normative gap[59]. I will be arguing that it can perform this function.

Throughout this chapter I am primarily concerned with what is normatively required for something to be considered a mental disorder, and in developing an answer to this question from an embodied enactive perspective. In line with this purpose I largely set aside structural and epistemological issues. Broadly, I first explore a recent debate concerning the role of normativity within the concept of mental disorder and use this debate to sketch out some requirements of a successful normative formulation. Following the listing of these requirements I lay the groundwork for an embodied enactive perspective by demonstrating that a 'natural normativity' is present within the deep continuity thesis of life and meaning [DCT] – which as we saw in the last chapter is an inherent part of enactivism. In doing so I explore ideas of natural normativity that are external to embodied enactivism but are coherent with this world view due to their common roots in dynamic systems theory. From all this groundwork I then pull together an understanding of what counts as mental disorder that fits within the embodied enactive view. Finally, the strengths and weaknesses of the position developed are discussed, and an addendum proposed in response to a foreseeable counter-argument.

Recent Views on the Role of Normativity

As explored in chapter two, there is much debate about the role that values should play in diagnosis. Most generally this has been a two-sided argument in the form of 'values in' (evaluativist) versus 'values out' (objectivist) positions. The former position argues that the evaluative nature of a diagnosis is inescapable, while the latter proposes

[59] Much of this chapter is directly parallel to Nielsen and Ward (2019) and is reproduced here with permission; Copyright © 2019, American Psychological Association.

that diagnostic claims are purely factual in nature (Fulford, 2002). Ultimately, the question of whether norms and values have a role to play at all seems somewhat trivial; at its simplest, a diagnosis is a claim that something is wrong with a person. On my view it is therefore *necessarily* normative, and I therefore assume an evaluativist position of some degree (others do disagree, see; Hucklenbroich, 2014). The more interesting question seems to be around what kinds of norms and values demarcate disorders from benign conditions, and how should they be employed to do so. Particularly contentious is the question of whether *social* and *cultural* norms should play a role or not (denoting strong and weak evaluativism respectively).

In this section, I sample some current and representative work in this area. In order to streamline discussion I concentrate on an article by Stier (2013) and a selection of responses. I have chosen this formulation because it manages to capture the core issues at play in a succinct manner. I have also drawn this debate from mainstream psychological literature in order to give the reader a sense of how engagement with ideas from philosophy of medicine can be a little lacking[60] (I have applied some descriptive labels *post hoc* in order to connect back to the positions explored in chapter two). My local aim is to draw out what is required of a framework attempting to conceptualize the role of normativity in demarcating mental disorder.

Sample debate in this area.

Stier (2013) makes the claim that with the progression of neuroscience the medical model is gaining increasing traction within psychiatry. With the rise of a biologically based psychiatry, Stier argues that we are disregarding the obviously normative nature of assessing human behavior and making diagnostic claims. On his view, our growing knowledge of the brain is leading us to mistakenly conclude that disorder itself is always reducible to a brain abnormality. Even if we assume that all behavior and experience stems from the brain (a counter-embodiment position he assumes within the context of his argument), the label of 'disorder' relies on assessment of the experience and behavior of the individual as pathological. Mental disorder therefore, cannot be identified at a purely physiological scale. According to Stier, there

[60] Muders (2014) and Jefferson (2014) would be noted exceptions here, however these authors also happen to be philosophers.

are many normative frames of reference against which psychiatry makes a diagnostic judgement: the personal values of the diagnostician, cultural expectations, generalizations about human nature, and the concepts of harm and disturbance. The examples he uses suggest a strong form of evaluativism (i.e. one inclusive of social and cultural norms). Stier goes on to explore some further normative concepts that play vital roles in psychiatry, but for the purposes of this discussion what we have covered here will suffice. Stier concludes that the prevalence of such normative factors within psychiatry as a practice supports his earlier argument that whether or not something is a mental disorder can only be determined on the psychological (including behavioral) scale.

Responding to Stier's (2013) claims, Muders (2014) raises two key criticisms. First, that Stier seems to be talking about the practice of psychiatry as it is done, rather than arguing for how it should be done. In doing so, he misses the possibility that while we currently rely on these normative frames of reference, this may actually be an error and thereby not suggestive of what the concept of mental disorder *should* be. Second, Muders suggests that Stier fails to unpack what it means for something to be normative. While the position I will eventually argue for is in line with Stier's claim regarding the irreducibility of mental disorder to a brain state, Muders' criticisms are valid. A framework circumscribing the role of normativity in the concept of mental disorder needs to be clear about *what kind of norms are at issue* and *where they come from*. It should also make a distinction between the concept as evident in the current process of diagnosis and the 'ideal' concept – how mental disorder should be thought of[61].

Jefferson (2014) also responds to Stier (2013). In the second half of her paper, she turns to the role that Stier describes normativity as having in the act of diagnosis, highlighting that his position is more than mere weak evaluativism. Rather, Stier's assertion is that diagnostic claims in psychiatry are directly and pervasively influenced by moral, social, and cultural norms – a strong form of evaluativism – thereby introducing a large degree of relativity. Jefferson argues that this is problematic because

[61] Regarding this last point, I wish to make it clear that throughout this thesis I am attempting to aim for the later; to develop a concept of what diagnosis should ideally represent within the normative domain.

it does not seem acceptable that what counts as disorder in one culture changes if somebody was uprooted to another culture with differing standards. While she accepts that some degree of vagueness is inescapable, Jefferson argues that we should strive for objectivity in diagnosis. She calls for "…a standard according to which we judge whether calling a certain condition pathological is valid or not." (Jefferson, 2014, p. 2).

While not directly responding to Stier (2014), Banner (2013) makes points relevant to the task at hand. She argues that mental disorders cannot be completely reduced to brain disorders on the basis that if a brain abnormality does not lead to a problem at the mental/behavioral scale then it is not a mental pathology. Rather, the label 'mental disorder' indicates a problem at the level of the person functioning in their environment[62]. While some mental disorders have been found to correlate with abnormalities at a brain level, Banner correctly points out that it is the dysfunction at the level of the person that makes it pathological, not its (partial) instantiation in the brain. While the general thrust of Banner's position is parallel to Stier's claims around irreducibility, there are two elements of Banner's construal that are particularly interesting. Firstly, she highlights the role of the social and environmental context in shaping what counts as disordered. Secondly, she defines mental disorder as specifically concerned with deviation from the individual's *functional norms*; those norms that support the functioning of persons within their context.

What can be learnt here?

Out of this discussion, and most clearly implied by Muders (2014), two key requirements emerge[63]: An evaluative concept of mental disorder must be clear about, 1) what kind of norms are at issue (i.e. whose norms are we talking about), and 2) where they come from (i.e. what are these norms and what gives them their normative status).

Regarding the first requirement, it should be apparent that the most contentious question is whether or not socio-cultural norms have a role to play in demarcating mental disorders from benign conditions. Classically speaking, if they do then this

[62] While not part of the sampled debate, Frisch (2014) makes very similar points based on an exploration of the ideas of Kurt Goldstein, one of the founders of clinical-neuropsychology. He suggests convincingly that Goldstein's ideas were remarkably similar to what I express later in this chapter.

[63] These requirements are not meant to be comprehensive – they are simply a vehicle for discussion.

would constitute a *strong* evaluativist position. If they do not, and the norms in question are simply those of the individual, then this would constitute a *weak* evaluativist position. However, one thing we can learn from the above debate is that this way of discriminating positions represents unduly dichotomous thinking. In particular, Banner's (2013) position seems to move beyond the strong vs weak evaluativist divide. To explain this, let's cycle back and briefly summarize some of the positions of the papers explored above while considering whose norms they place at issue.

Firstly, Stier (2013) suggests that psychiatry is currently acting on an implicit strong evaluativism, but he does not really comment as to whether this is justified. Jefferson (2014) in contrast, correctly points out that incorporating socio-cultural norms into the concept of diagnosis leaves us in an uncomfortably relativistic position whereby disordered status may completely change with shifts in cultural contexts. This would position Jefferson as weakly evaluative. Banner (2013) however does something quite different. She emphasizes the functionality of the individual within their social and environmental context and defines mental disorder upon the breaking of the norms that support this functionality. At a first glance, this may seem to be a strongly evaluative position because, given that the socio-cultural environment plays a large role in deciding whether an action is functional or not, it leads to a situation where what counts as disordered changes with cultural and situational context[64]. On the basis of this, it appears that Banner's position is strongly relativistic.

However, on further inspection Banner's (2013) position is a lot more nuanced than this, and indeed, more nuanced than she explicitly recognizes in her original paper. In being based on *functional norms*, her move allows only those socio-cultural norms that are crucial for the continued adaptive functioning of the individual within their context, while excluding those norms that merely serve the group or are merely statistical. While not explicitly stated, some socio-cultural norms are let in, and some are not, based on whether or not they contribute to the functioning of the individual.

[64] And indeed it does seem to, for example, Fulford and Jackson (1997)describe three cases of people who exhibit psychotic phenomena, the experience of which actually helped them in times of crisis. They demonstrate how the only successful way of demarcating such benign cases from pathological psychosis is by reference to the values and beliefs of the individual – these being obviously culturally influenced factors. For further examples of culture's pervasive influence on phenomena often seen as indicative of psychopathology, see: Larøi et al. (2014), NiaNia, Bush, and Epston (2016).

This then begins to move beyond the classic dichotomy between weak and strong evaluativist positions. Further, this positioning seems optimal in a sense, as it leaves the act of diagnosis as justifiable purely by reference to supporting the individual, as opposed to judging people by the standards of their society. This position thereby circumnavigates Foucauldian type claims, without ignoring the role of culture and context. It is for this reason that I think Banner's position is very successful. When explicating an embodied enactive concept, I will demarcate norms of relevance to mental disorder in a similar way.

If we consider the second requirement on a successful normative construal – that it needs to be clear regarding the source of the normativity at play – we can see where Banner's (2014) construal starts to fall down. What exactly does it mean to function well? And relatedly, what really is a functional norm? Banner's paper assumes that we have answers to these questions rather than providing answers to them. To be clear, Banner's paper was not really trying to achieve this; it was primarily a response to the idea that mental disorders are brain disorders. However, the problem remains, if we want to develop a position such as Banner's into a conceptual model of mental disorder that can compete with the models explored in chapter two – essentially a richer and less objectionable form of functionalism – then we need answers to these questions. Here again we can see the call for a richer framework of human functioning.

Groundwork for an Embodied Enactive Approach

I will now shift gears and explore how embodied enactivism can answer these questions. As a reminder these questions were: What does it mean to function well? And, what is a functional norm? To begin answering the first question, I briefly revisit the DCT and show how this idea is in part an understanding of how normativity can arise in a natural world. Following this, I overview two very similar systems of thought regarding the natural origins of normativity in complex autonomous systems (such as organisms) in order to clarify what I mean by the concept of a functional/natural norm; thus, answering the second question. Note that these two systems are external to embodied enactivism but are very much parallel to the DCT. Finally, I argue that embodied enactivist thinking, in particular something I refer to as the constitutional view of culture [CVC], allows for an extension of natural normativity to encompass

higher human values. Throughout this section I highlight two key concepts – namely, *self-maintenance* and *adaption* – as they play a key role in the following section where I pull together an understanding of what counts as mental disorder under embodied enactivism.

'Functioning well' under embodied enactivism and the DCT.

The DCT is the idea that that the origins of mind arise from the same process structures that support and define life (Kirchhoff & Froese, 2017; Thompson, 2007). Under the DCT, meaning arises from an organism's needful relation with its environment; to self-maintain requires the acquisition of energy from the world and avoidance of threats to the self (Thompson, 2007). At the cellular level this process is referred to as *autopoiesis* (Thompson, 2011; Thompson & Stapleton, 2009). As introduced in chapter four, for the enactivist this needs-based relationship changes the environment to one of *meaning* and *valence* for the organism, making the mind and our relation to the world inescapably affective in nature[65], whilst still thoroughly embodied (Colombetti, 2014; Colombetti & Thompson, 2008; Maiese, 2016). This does not mean that basic life forms, or plants, are conscious in a self-aware or reflective sense (this would be seen to come later, with the evolutionary development of a nervous system or some equivalent). Rather, according to the DCT, all life forms are viewed as having a non-conscious subjectivity or 'zero-point', and a non-conscious embodied 'concern' (i.e. a self-perpetuating structure) for the continuation of the self (*self-maintenance*) in the face of changing and precarious environmental conditions (*adaption*) (Di Paolo, 2005; Thompson, 2007; Thompson & Stapleton, 2009).

An enactivist primarily works with the DCT as a way to understand the emergence of *meaning* and *experience* for living beings. However, it is also an understanding of how *normativity* can emerge in a world of facts (Hume, 1978/1739). Indeed, in many ways 'meaning' is basically the subjective experience of normativity – normativity *for me*. Insofar as an organism *should* act to maintain its own life, there are

[65] It is fascinating to note that Okrent (2017) also arguably touches on this point. In chapter 2 he states, "for an organism to *perceive* its world is to perceive what is instrumentally important to the organism…" (p.31). Thus, he ties perception to the enaction of meaning via the pragmatic needs of a living organism, in a very similar way to the enactive authors cited here.

states, actions, and processes that the organism *should* be in or perform[66]. These states, actions, and processes change in accordance with the current needs of the organism and the constraints of the environment.

To begin answering the first question then. For an embodied enactivist 'functioning well' most primarily refers to an ability to *self-maintain* and *act adaptively* (to act in a way that supports self-maintenance within the constraints of a dynamic environment)[67]. These two concepts form the fundamental 'good' in the enactive view of normativity. Note though that embodied enactivism does offer a further layer of normativity (above and beyond self-maintenance and adaption) that emerges in socialized and self-understanding creatures – I will explore this in the 'Cultural Embeddment and Normativity' section below. For reasons I will explain I don't think this 'higher' form of normativity can be used to define mental disorder. For now, I need to shift to the second question and specify what exactly I mean by functional norm.

What is a functional/natural norm?

The ideas explored in this section have been developed by Wayne Christensen and Mark Okrent in separate works on the origins of normativity (Christensen, 2012; Christensen & Bickhard, 2002; Okrent, 2017). While these authors do not cite each other, they express remarkably similar ideas: that norms are inherent in self-maintaining and adaptive systems such as life forms and arise in service to those systems continued adaptive functioning within their environment. More specifically, norms are seen as supporting the *organizational autonomy*[68] of a system. Organizationally autonomous systems on Christenson's view are those that "…possess a process organization that, in interaction with the environment, performs work to guide energy into the processes of the system itself." (Christensen & Bickhard, 2002, p. 3). In

[66] De Haan (in press) uses the term 'relational realities' to refer to this dynamic.

[67] Originally self-maintenance was seen to be the central concept by itself, but Di Paolo (2005) noted this produces a dichotomous understanding of what is good for the organism (you either continue to live or you do not). With the addition of 'adaption' a graded (and much more useful) view emerges.

[68] I have added the descriptor 'organisational' to differentiate it from personal autonomy, a related but separate concept. For clarity I have, throughout this chapter, tended to refer to 'self-maintenance and/or adaption' so as not to introduce confusion with personal autonomy – valuation of which varies across cultures.

other words, they are thermo-dynamically open but self-maintaining systems[69]. To use the example of life forms, organisms are very much in a far-from-equilibrium state when contrasted with the wider environment within which they are embedded; it's very easy for life forms to die, but hard for them to keep living. The persistence of an organism relies on a set of balanced conducive states and processes (*self-maintenance*), but also that these states and processes change in response to alterations in the environment in a way that serves self-maintenance (*adaption)*. These states and processes occur both within the individual (e.g. blood pressure and circulation), and within the environment (e.g. sufficient oxygen). These states and processes are the *functional norms* of the organism. Importantly for our purposes, behaviors of the whole system, so long as they serve the continued function of the organism, can also be seen as functional norms (e.g. seeking food and shelter);

> "…for an entity to be alive is in part for it to interact with its surroundings in ways that are instrumental to its continuing life, given the kind of thing it is, from the 'standpoint' of the living thing there is a right and a wrong way to carry out that interaction." (Okrent, 2017, p. 28).

Parallel to the DCT then, these accounts view "… normativity as inherent in the organization or form of living systems…" (Christensen, 2012, p. 104).

The largest point of demarcation between these two authors is that Christensen is oriented to a systems perspective, and Okrent to one of organisms and agents. Both view norms as arising from the teleological purposiveness of *self-maintenance* and *adaptivity*. For Okrent this is grounded in the nature of being an organism, and whether other kinds of things can give rise to such norms is an open question[70]. Christensen is not bound to organisms as the only known sources of normativity in this way. Christensen's view makes it more explicit that ecosystems, social institutions, and other autonomous systems may conceivably have their own non-derivative functional norms

[69] The connection to embodied enactivism is clear here, e.g. Varela, Thompson, and Rosch (2017), but links can also be made to Free Energy Principle theory; see Kirchhoff (2016).

[70] Okrent does note that "Whether or not it is also the case that norms only arise in the context of life remains to be seen." (Okrent, 2017, p.28).

(Christensen, 2012; Christensen & Bickhard, 2002). For example, these norms might relate to levels of predation in an ecosystem, or availability of coffee in a busy office.

This overview hopefully elucidates what I mean by functional/natural norms. To further clarify however, I will briefly cover two types of norms that do not count as functional norms. Firstly, a functional norm is very different to norms based on typicality. Norms based on typicality are those that aren't functionally important and are simply based on deviation from the usual distribution – e.g. having a non-problematic benign growth or having purple hair; neither is typical, but neither is either a problem. These are often referred to as 'statistical' norms, and they are not seen as prescriptive (Okrent, 2017). As such, statistical norms cannot be of direct use for defining dysfunction or disorder, a point implicitly supported by Banner's (2013) construal and noted by Jefferson (2014) in the discussion earlier; "A statistical notion of dysfunction and pathology is too thin to be useful for medical practice." (Jefferson, 2014, p. 2).

Secondly, it is quite common in the literature to see norms of human functioning as derived from a component's apparent evolutionary function (Troisi & McGuire, 2002; Wakefield, 1992). Norms based on purported evolutionary function are much more similar to the account at hand than statistical norms, in that they are prescriptive rather than merely statistical in nature. However, as reviewed in chapter two, construing norms as natural based on their apparent evolutionary function faces a knowledge problem: we cannot know for certain that we have the evolutionary story correct, nor that other unknown functions aren't being simultaneously served by the state or process in question (Christensen & Bickhard, 2002). Furthermore, such an account does not leave room for adaptive deviations from the evolutionary norm (Christensen & Bickhard, 2002; Troisi & McGuire, 2002). This is hugely problematic given the importance of adaptive phenotypic variation for evolutionary theory. As such, I do not see evolutionary theory as providing a rich enough account of human functioning to support an understanding of disorder, at least within the mental realm (however, for a good attempt at such a construal, see Troisi & McGuire, 2002).

As a fictional example to flesh these differentiations between functional, statistical, and evolutionary norms I use the example of Jim. Jim has three arms, his third arm sits underneath his right. Importantly, Jim's third arm does not get in the way

of his functioning, in fact Jim is better at many tasks than plain old two-armed people. Jim's arm therefore breaks norms of typicality (most people don't have three arms), and etiological/evolutionarily derived norms (we did not evolve to incorporate a third arm). However, Jim's third arm does not break the functional norms of 'Jim the complex autonomous system' because it does not get in the way of Jim's ability to meet his needs relevant to self-maintenance, nor impact his ability to adapt to environmental changes. On my view then, Jim should not be seen as disordered.

As a point of clarification, I am not saying here that the existence of functional norms cannot sometimes be *inferred* from statistical comparisons across individuals. Taking the example of blood pressure: we know what sorts of parameters are medically acceptable based on research studies, and that certain blood pressure thresholds are associated with harmful outcomes such as fainting, heart attacks, strokes, etc. This sort of inference seems reasonable, at least at the physiological scale where the states and processes that constitute functional norms are somewhat more stable across individuals, and deviations from norms often have more obvious effects (e.g. blood pressure is clearly definable and measurable, similar levels count as too high or too low across individuals, and deviation from the norm can result in outcomes that immediately challenge the self-maintenance of the individual). The inference from typicality and associated risk across the population to a normative claim about an individual's blood pressure therefore seems reasonable. For reasons we will return to later, whether the same sort of inference can be made when considering behaviors of an organism that do not seem to directly serve some obvious biological norm remains to be seen. I will argue that they cannot. Before doing so, I need to first demonstrate how embodied enactivism has extended these ideas of natural normativity to inform an understanding of the normativity of complex human behavior.

Cultural embedment and normativity.

While sharing the same root understanding of a functional norm, embodied enactivism offers an important extension of this account of normativity. Two key 3e concepts are important here. Firstly, that of embedment described earlier, where interactions with the environment are necessary for the development of the mind. This refers to both a physical and, especially in humans, a socio-cultural environment. The

second key idea relevant to our purposes here is the constitutional view of culture (CVC). Most succinctly espoused in the introduction of Durt et al. (2017), the CVC is in many ways an elaboration of embedment. According to the CVC, groups of individuals and the artifacts they produce constitute a cultural ontology; a collaboratively generated shared world of significance and meaning that facilitates intra-group behavior and the transmission of tools, knowledge, and ways of knowing (Durt et al., 2017; Kirmayer & Ramstead, 2017). This shared world, or *habitus* (in the sociological sense), is embodied within the habits and practices of the group which are passed on to and developed by younger generations because they represent adaptive ways of understanding, managing, and altering the environment (Henrich, 2015; Heyes, 2018). Interestingly, such a perspective can even be shown to encompass so-called higher-level cognitive practices such as mathematics and reasoning about the minds of others (Gallagher, 2017; Heyes, 2018). Significantly, this shared world, while being co-generated by the group, also represents a major reshaping of the environment within which individuals reside, thereby molding the ontogenetic and phylogenetic development of individuals in ways that the group has found to be adaptive (Durt et al., 2017).

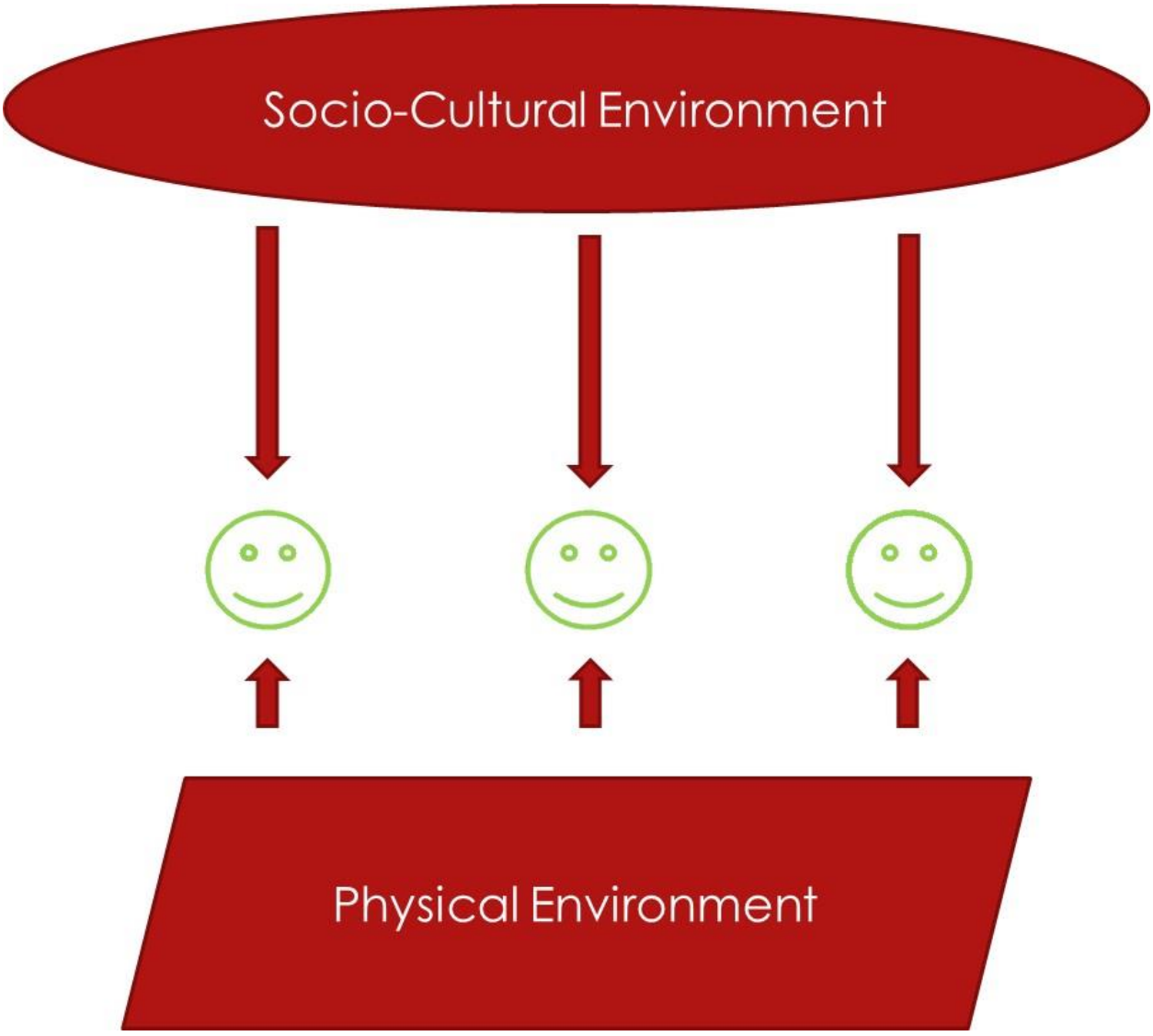

Figure 2. The Constitutional View of Culture [CVC]. The arrows here represent constraint on the development of individual's mode of functioning by their environments; both physical, and socio-cultural (which the individuals as a group constitute). This process is occurring across both life-span and evolutionary time-scales.

These two ideas have allowed authors such as Maiese (2016) and Di Paolo (2005) to describe how, in conceptually and socially sophisticated animals such as ourselves, more complex tendencies in behavior can develop, embodied within the dynamics of the

organism system[71]. Building upward from the enactive core of meaning and normativity rooted in the needful relation between organism and environment, Maiese and Di Paolo demonstrate how irreducible higher-order socio-culturally mediated functional norms can emerge. Over evolutionary and life-span time scales, behavioral/evaluative tendencies are selected for and developed*, as they allow the organism to flourish in accordance with the constraints of the socio-cultural environment* (which they as a group constitute and as individuals reside within). These behaviors are therefore irreducible *functional norms*, serving the flourishing and by extension the self-maintenance and adaption of the organism system within the socio-cultural environment. As individuals *and cultures* then, humans generate their own values/meaning. I refer to these socio-culturally generated functional norms as *interpersonal prudential norms*[72]; examples may include mastery, patience, personal autonomy, honor, and social connectedness.

Functional norms then, as used within our framework, are not simply inherent in those biological states, processes, and basic behaviors of the organism that immediately support them (e.g. seeking food and shelter), but are also evident in more complex behaviors that indirectly serve the continued functioning and maintenance of the organism via reciprocal relations with the socio-cultural environment. Norms of behaviour that facilitate the *fairing-well* of the individual within their socio-cultural environment may be less directly tied to biological functioning but are still, indirectly, functional-norms. Maiese (2016) offers the example of being a good driver: we wish to be good drivers not simply so that we can avoid crashing, but to demonstrate our mastery which has positive social implications for us. We feel proud of our ability to master such a complex skill.

---

[71] De Haan (in press-b, in press-a) develops a slightly different approach here, based on 'reflexivity'; our ability to see ourselves in the world and reflect on *how* we want to live. I will return to de Haan's work in chapter eight.

[72] In my first paper during this PhD I referred to these as values rather than norms, but have since shifted to the use of 'functional norms' for the sake of clarity (Nielsen & Ward, 2018). I found that use of 'values' tended to confuse people (e.g. people with anorexia value being thin, does this make it not disordered?). The term 'functional norm' highlights that the normativity of behaviour is more complicated, that norms can conflict etc. The term 'values' of the other hand is much more loaded and seems to encourage confusion because of the different understandings of 'values' that people hold. De Haan (in press-b) does use the notion of 'values' within her enactive framework and I discuss the term further in relation to this within the final chapter.

While biological norms are similar across individuals, interpersonal prudential norms vary in the degree to which they are endorsed across cultures. This is because culture constitutes a significant variation in the environment, thereby placing differing constraints on how individuals can best achieve their needs. Endorsement will also vary across individuals due to dispositional differences (whether learnt or genetic). This has implications for the process by which we can gain knowledge of norms. As discussed earlier regarding the norms of bodily processes such as blood pressure, the inference from typicality and associated risk at a population level to a normative claim about an individual seems reasonable. However, things get murkier when we shift to functional norms of behavior. For example, the degree of personal autonomy required to support functioning will vary across contexts, cultures, and individuals. The inference from typicality to functional norm is a lot more tenuous within the domain of behavior than it is when considering physiology or the like (Fulford, 2002). This is because there are many different ways for individuals, groups, cultures, societies, and ecosystems to meet the needs required for their self-maintenance. In other words, these higher-level systems have a larger set of functional states. In contrast, the human circulatory system and other such internal bodily systems, have a much smaller set of functional states – e.g. not much needs to change about the circulatory system to result in the death of an organism. In practice, this means that a clinical psychologist or psychiatrist will always be asking the question 'is this a problem for this individual within their context?' Whereas, for a medical doctor, answers to this question will be easier to arrive at[73].

What Then Counts as Mental Disorder?

Drawing together this groundwork and weaving it in with the previous discussion on normativity in the concept of mental disorder, a view emerges similar to but more developed than that argued for by Banner (2013) (also see Frisch, 2016). One that

[73] It is due to this complexity regarding the functionality of behaviour that I maintain (in the next section) that a link must be made to the more fundamental processes of self-maintenance and adaption if we are to label someone's behaviour as disordered (rather than basing this distinction on someone's ability to 'flourish' in a partially socio-culturally defined sense). This way we can be much more confident that it is the individual's functional norms that are at issue. The fact that everyone else in a culture does something one way and a client is doing them differently – that they are not living their culturally informed 'good life' – cannot be indicative of disorder (I see this as being connected to mental health in a positive sense, rather than to disorder). At the same time, it is vital to understand the client's culture so that you can understand how they learned to function.

maintains the intuitive appeal of functionalist positions such as the HD Analysis or the BST (see chapter two), but addresses the weaknesses of these positions by its subscription to a richer understanding of human functioning.

What counts as mentally dysfunctional on this view is any set of behaviors (inclusive of cognition, perception – anything the organism does) performed by an organism that significantly violates its own functional norms, in that it is acting counter to its own *self-maintenance* and *adaption* needs. The persistence of this pattern of behavior thereby threatens the organism's *organizational autonomy* and as such, should be considered disordered. I focus on the two processes of *self-maintenance* and *adaption* within this definition because, under an embodied enactive conception of human functioning, these are fundamental processes; they are the ultimate ends of all human action. Other values/functional norms – such as the aforementioned *interpersonal prudential norms* which support adaption by facilitating the individuals 'fairing-well' in their socio-cultural environment – are relevant in so far as they are functional norms of the individual rather than society (I will return to this in the 'A Possible Objection' section). A reasonable link needs to be made back to these fundamental processes of self-maintenance and adaption if a diagnostic label is going to be ethically applied. To not demonstrate such a link risks pathologizing individual or cultural variances in modes of functioning.

The current framework then is positioned in a similar way to Banner's (2013) construal in that it is functionally defined, thereby positioning it beyond the false dichotomy of weak versus strong evaluativism. As argued earlier, this is a significant strength as the act of diagnosis is then justifiable by reference to individuals and their needs; staving off Foucauldian type claims (unlike strong evaluativism), while also not ignoring how culture shapes many of those needs in the first place (as per weak evaluativism). However, being situated within the broader framework of embodied enactivism offers advantages over Banner's brand of functionalism. This framework brings greater conceptual specificity, provides justification for the use of functionality as the crux of the definition, encourages ecological considerations including socio-cultural elements, and offers a rich and coherent system for conceptualizing relevant factors such as mind and culture. I will now continue to develop this construal, first by

highlighting some key strengths, and then by exploring a foreseeable counter-argument to which I reply.

Evaluating this position.

A strength of this framework is that it is in many ways congruent with a medical understanding of physical illness, while also highlighting the differences between the bio-medical and psychological domains and their respective conceptual needs. A significant violation of the functional norms of an organism system at a biological scale essentially constitutes an injury or medical condition. Similarly, on my view a significant and continued violation of functional norms of the organism system at a behavioral or psychological scale *is* a psychological disorder. As explored above, one key difference between these domains is that in the former it is generally safer to infer the existence of a functional norm from a statistical one. Functionality of behavior and psychology is, in contrast, diverse in that there are many ways to be functional – as exampled by cultural variation (Fulford, 2002). It is therefore ethically questionable to infer that a norm derived from typicality is a functional one within the psychological domain, because whether it counts as a functional norm is going to be much more individually and contextually specific. This framework therefore prescribes great attention to the role of the context in shaping an individual's mode of functioning.

At all scales of analysis, the embodied enactive framework highlights that an organism is *attempting* to act in accordance with its inherent purpose – to adapt (Di Paolo, 2005) and self-maintain (Thompson, 2007). Just as getting a cold reflects a faltering of the immune system to adapt to the challenge of a pathogen, mental or behavioral disorders often reflects a faltering of the organism attempting to adapt to the challenge of a changing environment[74]. 'Faltering' is here used because outright failure is inappropriate; the organism is still alive. An example of this would be a child growing up in a difficult family context where cycles of coercion have negatively reinforced his escalating of aggressive behavior (Granic & Patterson, 2006; Smith et al., 2014). We know this will not serve the child well in other contexts, and may disrupt other norms of development (Erskine et al., 2016). However, the aggression has developed due to the

[74] I realize this is not a perfect analogy – many symptoms of a cold may actually be seen as a functional and typical response to the presence of the pathogen.

constraints of the family system and the child's adaption to this environment. A further example would be a refugee from a war zone whose previously adaptive bias towards interpreting others' actions as aggressive is now dysfunctional within their present, largely peaceful, context. Both examples highlight the need for consideration of context over time rather than just the role of the current environment. The 3e perspective allows for, and indeed encourages, recognition of both of these sides; that this behavioral pattern is an adaption to the environment, but that it is also very likely to be maladaptive in other contexts and is maintaining a family dynamic that is problematic for both other family members and the continuing development of the child. The current framework then, encourages the dynamic consideration of context. The question being: in what way is the behavior attempting to serve the person's needs within their context (past or present), and are there other ways for these needs to be met that would represent a more balanced normative equation?

This brings us to a further strength of thinking about disorder in this way. An individual's functional norms do not necessarily all point to a single prescribed action (and if they do, these tend to be areas in our lives in which decisions as to which action to take are clear and easy). Instead, functional norms often compete, and compromise is required. For example, it's ideal to sleep 6-8 hours a night, but sometimes we have some approaching deadline and need to compromise on this; staying up late to finish some important project. One can act in accordance with one norm, while violating another. When it comes to norms, compromise is the norm! If, however, I stay up late to complete work regularly, perhaps for less and less important projects and resulting in chronic tiredness, then the normative equation begins to look unbalanced. In other words, this pattern of behavior starts to look dysfunctional.

This idea of an *unbalanced normative equation* is worth fleshing out with a clinical example. Imagine a client where a behavior (e.g., self-cutting) is serving some function (e.g., emotional regulation). To use normative language, the cutting is serving the norm of emotional stability. However, in the process, other norms are adversely impacted (e.g., having unbroken skin). Two elements are of importance here. Firstly, while the cutting is serving a norm/function, this does not mean that it is on the whole 'functional'. Other norms are being violated by this action (having unbroken skin), and

there is risk of breaking even more vital norms (e.g., being infection free, undamaged arteries/veins). It is this element that is important when considering whether the equation is reasonably balanced or not; whether the pattern of behavior and its consequences are on the whole functional (ranging from ideal to roughly functional) or dysfunctional (the individual's functional norms are being or are at significant risk of being significantly impacted). The second element to consider in this example is whether there are clearly ways in which the function performed by the cutting behavior may be achieved in a significantly less normatively imbalanced way (e.g., emotional regulation strategies). Insofar as there is a less negatively impactful way to achieve some norm, and that the compromising of other collateral norms is significant, we are justified in offering assistance. When the functional norm breaking behavior takes a recognized causal and constitutional form, labeling with a diagnosis to facilitate communication and treatment across organizations is our society's way of achieving and providing this assistance.

A possible objection.

Many readers at this point will be concerned that I have ignored an obvious counter example. This would be a situation where the social context is placing unjustified constraints on someone, and where defiance of these constraints appears somewhat 'dysfunctional'. Examples would include acts of rebellion in a totalitarian society, and gay/queer people expressing their sexuality in a homophobic society. At first glance it may seem that according to this view these are instances of mental disorder, because both acts are seemingly not adaptive within the social context given the risk they bring to the individual. This is obviously a problematic conclusion. This issue seems to underlie the intuitive need for some sort of recognized 'dysfunction' or lesion alongside the normatively defined 'harm', as in the harmful dysfunction analysis (Wakefield, 1992, 1997). The argument seems to be that this requirement allows for an easy response to such counter examples; the ontic distinction from typicality at some sub-personal level makes the disorder seem more 'real'. However, I will argue that, with an addendum justified by the broader embodied enactive framework, the current functional construal can exclude such cases. It is therefore more parsimonious than two-

part models and does not unduly privilege the sub-personal. First however we must explore the issue in a little more depth.

In general terms the violation of norms of the socio-cultural systems (functional, legal, civil, or otherwise), do not represent mental disorder under our framework (they may however represent a crime, immoral act, or social faux pas). Rather we are specifically concerned with the *functional norms of the individual*. This is what separates my claim regarding the normativity of mental disorder from a Foucauldian type view, under which disorders are defined by the violation of socio-cultural norms (and are therefore not justifiable if the labeling of disorder is truly intended to be in the interest of the individual).

Unfortunately, things are rarely this simple. Under the CVC, one may note that there is a complex two-way relationship between the norms of an individual and the norms of a culture or society. While the norms of the culture serve the continued survival and functioning of the collective, the collective itself is of course constituted by the individuals and therefore the functional norms of the culture will, largely and for the most part, serve the majority of the individuals' survival. Going in the other direction, the norms of the group are a dominant constraint on how the developing individuals within that group context *learn to function*. Large parts of the intra-dependent set of functional norms operating on an individual are therefore shaped by their cultural context across development. Someone who grew up in urban Japan will have a different mode of functioning than someone who grew up in bible-belt USA, and so on. If culture is the ways of knowing, being in, functioning, and making sense of the world, shared across a particular group (Durt et al., 2017), then consideration of culture when asking normative questions is always going to be relevant.

This means that a discussion of individual normativity must explore the role of culture but, more practically, also makes teasing apart the functional norms of an individual from the norms of the culture in which they reside challenging. This is especially true when someone is part of a cultural minority or of a culture that is less recognized in the mainstream, as such individuals are effectively living between two worlds and exposed to contrasting ways of functioning. One particularly interesting example, that highlights the importance of interplay of individuals and culture in

shaping the functionality or disorder of a behavior, is how experiences that from a western viewpoint would certainly be classified as hallucinations are interpreted much less pathologically in many cultures. For individuals embedded within such cultures, the consequences for their functioning are much less severe, sometimes even positive (Fulford & Jackson, 1997; Larøi et al., 2014; NiaNia et al., 2016).

This gets us to the problem. In recognizing that social context is a huge part of the individual's environment, socio-cultural norms can sometimes be imported as derivatively functional for the individual. It therefore seems that such cases as rebellion in totalitarian society, or expression of queer sexuality in a homophobic society, must be counted as disordered under a functional construal. However, the 3e orientation of our framework can help us in navigating this situation. Embodied enactive thinking places the anchor point of consideration at the level of the individual; as the experiencing agent, for which meaning exists. In light of this, it seems very odd to refer to a norm as functional *for* an individual if it stems from a socio-cultural norm that does no work for, or in fact is running counter to, the self-maintenance and adaption of the individual in question. I therefore suggest the following addendum that helps clarify why such examples do not count as disorder under our framework:

> *A norm, even if apparently functional, should not be used to define disorder if it is derived from (secondary to) either:*
>
> a) *A non-functional norm of a higher-order system, or*
> b) *A functional yet arbitrary norm of a higher-order system that is impinging on the self-maintenance and/or adaption of the lower order entity.*

I will now explain and justify this addendum through the exploration of the problematic cases. Firstly, the expression of homosexual orientation in a homophobic society. As explored above, an argument could be made that this is not functional for the individual because it risks persecution. However, the socio-cultural norm of homophobia is a statistical/religious/erroneous moral norm, not a functional one. We now know that allowing honest expression of sexuality with our societies does not result

in societal collapse. Therefore, the constraint placed on the individual by the homophobic norm is not justified; the problem is with society and with its norms not working for the individual, not with the individual themselves[75]. Accordingly, the addendum above specifies that while it may, in a homophobic context, be somewhat functional to hide one's sexuality, continued expressions against this particular functional norm should not be seen to constitute mental disorder. This is because the dysfunctionality of honest expression of one's sexuality is derived from a non-functional socio-cultural norm. Insofar as, from a CVC view, it is society's role to serve its constituents, the dysfunction is with the homophobic society, not with the homosexual individual.

Secondly, concerning rebellion and other risky political acts. Once again, an argument can be made for such behaviors being dysfunctional because they risk the self-maintenance and/or adaption of the individual. This is a slightly more complicated situation because, despite moral qualms, it may be argued that the overly restrictive norms of a totalitarian society are functional in that they are helping to maintain the stability of the society in question[76]. However, there is a sense in which the functionality of such norms is arbitrary; we know that other societies exist that do not rely on totalitarian norms for their continuation. Assuming again that the purpose of a society is to serve its constituents (as per the CVC), the fact that this society is impinging on its member's self-maintenance and/or adaption to survive suggests that the dysfunction is at a societal level, not with the rebellious individual (and indeed this seems to go some way in justifying their action for change). In accordance with this reasoning, the above addendum rules out basing the labeling of mental disorder on seemingly functional norms derived from functional yet arbitrary socio-cultural norms.

Having questioned societal norms in light of individual norms, it is interesting to question individual norms in light of the social. There are certainly cases where the social trumps the individual. Even when some action is functional for the individual,

---

[75] As a parallel point it is also very difficult to see within this example how a norm that is so constraining on the autonomy of the individual can really be said to be 'functional' for that individual.

[76] Once again, it is difficult to see how such overly restrictive norms are in any true sense 'functional' for the individuals being constrained. However, given the context it becomes in a sense 'functional' to abide by it.

should it cross certain social norms then this would seem to constitute a social faux pas, crime, or immoral act. The framework presented here does not excuse such actions (although I would hope it would encourage compassion in seeking to explain them). In cases where patterns of such action become a learned way of functioning for an individual though, two interesting categories seem to emerge. The simplest of these is non-pathological; those that achieve their own self-maintenance and adaption in disregard of social/legal/moral norms. This category would range from selfish people to career criminals. The second case is more interesting for our discussion here; those whose patterns of social norm violations actually work against their own self-maintenance and adaption within their social environment, and are therefore pathological in the sense defined here – i.e. personality disorders. Under the current framework personality disorders do seem to count as disorder, but the harm to the individual is mediated by the breaking of social norms rather than by the crossing of individual norms directly. These constructs then are different in kind to both 'regular' psychopathology where individual functional norms are directly impinged, and social deviancy where social norms are directly violated.

In opposition to criminality, it is also interesting to consider altruism here. When an individual acts in the interest of their group, in contradiction to their own interests, then there may be a concern that our framework labels such a behavior dysfunctional, and its persistence disordered – some sort of 'altruistic personality disorder' if you will. This is an issue that needs further thought, but my sense is that the framework is not individualistically biased in this way. Within the timeframe of the act, altruistic behaviors seem to reflect an under-emphasis on individual and biologically immediate norms relative to socio-culturally generated norms. However, because these norms benefit the individual at other times, then as clarified by the addendum, these altruistic norms should not be used to define dysfunction. Thus, someone may, to a certain degree, *temporarily* act against their own self-maintenance and adaption in a non-dysfunctional way. The limiting factor is that, largely and for the most part, the norms they are following during the act must benefit them at other times.

Before closing this section, it is worthwhile briefly considering some current personality disorders as they represent complex normative cases. On the current

construal some personality disorders appear to be more valid disorders than others. For example, it is hard to imagine a social group that serves the interests of its constituent members well where interpersonal styles such as those seen in some personality disorders are encouraged. Such cases would include narcissistic personality disorder, borderline personality disorder, or anti-social personality disorder. The socio-cultural norms of relevance in these cases then, seem both functional and non-arbitrary, with the outcome being problematic interpersonal functioning (although in the case of narcissism there would be a genuine argument to be made that the problems that arise primarily concern the functional impact on others rather than the individual being diagnosed – in other words this may be a moral category). In other cases, however, such as schizotypal personality disorder, the socio-cultural norms being broken seem to be predominantly statistical, making this a very questionable diagnostic category under the current framework. Essentially this category describes those who are weird/odd to the point that other people treat them in a way that makes their social functioning difficult. Finally, schizoid personality is very interesting to consider. This category describes people with asocial (as oppose to anti-social) tendencies. These people simply care less about the social/interpersonal domain and would happily live by themselves. While statistical norms are certainly being broken in such cases, it is hard to understand how this disorder represents a functional problem for the individual concerned. This 'disorder' therefore seems more likely to be simply a different mode of functioning.

Conclusions and Summary

In chapter four I argued that, structurally speaking, mental disorders are likely constitutionally and causally complex phenomena, situated across multiple scales of analysis. Ultimately though, a diagnosis is a claim that something is *wrong* with a person's functioning. A diagnosis is therefore a *normative* claim. Overviewing a brief sample of literature in this area, the most pertinent question seemed to be 'which norms are relevant when demarcating disordered from benign conditions?' Stier (2013) described current practice as including socio-cultural norms within this distinction, while Jefferson (2014) suggested this is unjustified and risks unsustainable relativism. I suggested that the most viable move was exemplified by Banner (2013), who starts to move beyond strong versus weak evaluativism; instead defining disorder by the

functionality of behavior. In accordance with the requirements of a satisfactory mental disorder concept implied by Muders (2014), I have here attempted to develop such a functional view into a fully-fledged and coherent position within this debate that makes clear what norms are at issue and where they are seen to come from.

Reconciling this functional view with science's naturalized view of the universe required an account of how purposiveness and normativity can arise, in order for there to be purpose and norms against which functioning is contrasted. I argued that embodied enactivism offers such an account which we have here explored, alongside consilient views of normativity, whereby norms arise in life-forms due to their organismic self-maintaining process structure and their adaption to the constraints of their environment (Christensen, 2012; Christensen & Bickhard, 2002; Okrent, 2017; Thompson, 2007). From this position, mental disorder is a pattern of behavior (inclusive of *all* actions of the organism, such as thought and perception) that runs counter to its functional norms to a significant or atypical degree. Functional norms are those norms that support the organisms continued self-maintenance and adaption, and by extension, their ability to fare-well in their communities (Di Paolo, 2005; Maiese, 2016). What exactly it means to 'fare-well' for any individual will subtly change as a function of the individual, and will co-vary with the culture in which they learned to function. This welcomes intersection with cross-cultural psychology and psychiatry (Kirmayer & Crafa, 2014; Kirmayer & Ramstead, 2017).

Teasing apart the norms that serve the individual from those that serve the group is a complicated exercise. I have here argued that this distinction must rest on whether the norms of society are working for the individual, or put more technically, whether the norms in question support the individual's self-maintenance and adaption. A 3e orientation therefore prescribes strong consideration of context and culture over time, while also focusing on the individual and their needs. An embodied enactive perspective on mental disorder, in that it subscribes to embedment, must recognize the role of culture in shaping the way that an individual functions. The functionality of a behavior, even those which we may dismiss as inherently pathological from our received point of view, is often contingent on the social environment, as well as the culturally informed manner of functioning and definition of 'flourishing' that the individual subscribes to.

The embodied enactive perspective encourages us to consider such rich variation and, through its basis in the organism's strive to survive as a basic predicate of all life, provides a structure on which to begin to tease apart the disordered from the functional at the level of the individual. I have further specified that, within this framework, functional norms of individuals that are derived from non-functional or arbitrary socio-cultural norms should not play a role in demarcating disorder. Despite their apparent functionality, such norms seem to represent a disorder of society rather than disorder of the individual.

Understanding the normative nature of diagnosis is vital for the purposes of being able to ethically justify our practices as psychologists, psychiatrists, and researchers. An embodied enactive conception holds potential in this regard given its ability to bridge the natural and the normative, and I hope that my work here represents a step towards developing this perspective. As an upshot of the normative focus that the embodied enactive position brings, we must question the nature of the norms imposed by society. Institutions such as psychiatry and clinical psychology – in being the arbiters of such strong normative labels as diagnoses are, and advocates for those in or in need of our care – have a responsibility to be critical of the norms of society when they touch on our domain of expertise. Importantly, this includes reflecting on our own institutional and personal norms of practice. In the following chapter I collide this normative picture with the structural considerations from chapter four, and attempt to describe the fuller concept that arises when we consider the normative and structural together.

Chapter 6: A Concept of Mental Disorder and Two Challenges

Over the last two chapters have I have considered the nature of mental disorder through an embodied enactive lens while separating the structural and normative 'dimensions'. From an embodied enactive view however, the normative and the structural are not orthogonal as dimensions are supposed to be. As explored in the previous chapter, under embodied enactivism the normative naturalistically emerges *for* self-maintaining and adaptive structures in the world (i.e., life forms). In other words, we have closed the normative gap. Because of this, the normative and structural are better seen as continuous[77]. In turn, for my purposes, it is better to think of the normative and structural as different domains or parts of the same conceptual model of mental disorder (as opposed to separable dimensions). In this chapter, I will therefore discuss the normative and structural considerations together and describe the fuller conceptual model of mental disorder that emerges. First, I will briefly sum up the last two chapters, before sketching the emerging concept and briefly applying it to the example of anxiety disorders. I then evaluate the concept using Zachar and Kendler's (2007) conceptual taxonomy that I used to outline RDoC's conceptual position in chapter three. Following this I briefly make comparison to the most relevant of the

[77] On the embodied enactive view, (direct) meaning is literally the relevance of the world for an organism's survival and adaption, given its history and mode of functioning. De Haan (in press-b) refers to this direct meaning as the 'relational reality'. The *experience of* meaning therefore, is the experience/recognition of this relevance. Organisms develop the capacity to experience and respond to meaning across evolution and development because it facilitates action that accords with functional norms, thus encouraging the survival and reproduction of the organism. A functional norm in my sense of the term, is conceptually related to the direct meaning/'relational reality', in that functional norms are ways of acting/thinking/feeling that align with the relational reality of a situation and thus do work for the self-maintenance/adaption of the organism. On my view we do not engage with relational reality in an unbiased way, rather our sense of meaning is a leaned, evolved, and therefore imperfect mode of seeing the relevance of the world for us given our needs and histories. This must be the case because otherwise acting in accordance with our feelings/sense of meaning would always be functional. This is related to Okrent's (2017) distinction between following a functional norm (e.g., not walking under ladders because it is dangerous) and merely acting in accordance with it (e.g., not walking under ladders for superstitious reasons). Both of these actions accord with the functional norm and are therefore functional ways of understanding and acting in the world, but typically, recognising and following the norm more directly affords greater functionality in the long run. For example, lots of people have a fearful relationship with snakes. Such a reaction to snakes makes sense because many snakes are dangerous and a tendency to avoid them has therefore facilitated our ancestor's survival in the past. However, not all snakes are dangerous. If I can learn to distinguish between dangerous and non-dangerous snakes, then I can shift my experience of meaning and my related behaviours to be closer aligned to the relational reality (i.e. only being fearful of/avoiding snakes that are actually dangerous). I have learned about the world in finer detail and this affords me a greater range of ways to function, e.g., earning money as a snake handler.

conceptual models explored in chapter two. Finally, I comment on two challenges we face if we intend to put an embodied enactive concept of mental disorder to work within the later tasks of psychopathology.

Integrating into a Fuller Concept

In chapter four I argued that in a structural sense, the embodied enactive view reveals mental disorders as repetitive patterns of/tendencies in behaviour (inclusive of all actions of the organism, such as thoughts, emotions, sense-making in general, actions and perceptions), with causal structures best thought of as stable dynamic patterns across the brain-body-environment system. I suggested this pattern can be thought of as an MPC-kind structure, spanning the brain, body, and environment. In chapter five I showed that embodied enactivism subscribes to a view of normativity as emergent from self-maintaining complex systems, and thus features the tools required to develop a sophisticated systems-functionalism as a basis for the labeling of certain behaviors as dysfunctional or disordered. On this view mental disorders are patterns of behavior that run counter to the organism's own functional norms to a significant or atypical degree. I argued that significance in this context should be thought of as the negative implication of a person's self-maintenance or adaption processes, and that – while useful to consider – 'higher-order' layers of normativity/meaning such as socio-culturally derived values that do not support the individual's self-maintenance and adaption should not be used to define disorder. If we do, we risk pathologising individual variance.

Bringing these ideas from the structural and normative domains together, mental disorders can be seen as *dysfunctions in the behavioral and experiential processes of striving organisms*. These dysfunctions are constituted by relatively stable dynamic patterns (/networks of phenomena) within the brain-body-environment system of individuals, supporting behaviors[78] – themselves a key part of this pattern – that run significantly counter to the persons functional norms[79]. These dynamic patterns then

[78] Again, by 'behaviour' I refer to a wide range of phenomena such as actions, emotions, thoughts, even perception and attentional processes.

[79] Taking a perspectivist approach (Chang, 2020), where different modes of description can be seen as complimentary models of the same aspect of reality, it is also possible to approach this understanding of mental disorder using the more dialectical enactivist language of Di Paolo, Cuffari, and

are dysfunctional process-structures in the adaptive processes of agents, distinguishable by their negative functional effects and inflexibility[80].

Briefly applying this embodied enactive conceptual model to anxiety disorders as an illustrative example only, I would argue that it provokes a much richer understanding of anxiety than current approaches. Anxiety disorders are traditionally defined as levels of vigilance and/or fear, disproportional to perceived threats, to a degree that is atypical and produces significant harm or impaired functioning (American Psychiatric Association, 2013a). Rather than assuming this pattern of behaviour is caused by an underlying brain lesion (i.e., biological essentialism) or an error or difference in cognition (i.e., psychological essentialism), an embodied enactive view would consider an anxiety disorder as *a network of phenomena within the brain-body-environment system, that in sum represents a dysfunction in the behavioral and experiential processes of the striving agent*.

Some of the behavioral and experiential phenomena that together constitute the most obvious aspects of this pattern within anxiety may include: perceptual biases towards potential threat, the affective experiences of worry and fear, repeated checking behavior, fatigue, sleep disturbance, irritability, etc. This aspect of the network at the scale of behaviour and experience is not taken to be the complete picture however. Each

---

De Jaegher (2018): Mental disorders are parasitic partial 'autonomies' within the process structures of human functioning (i.e., mental disorders themselves partially 'self-maintain' within the context of the brain-body-environment system). This autonomy is dependent on – but in tension with – the biological, sensorimotor, and other adaptive autonomies of the host organism, and conflicts with the normative structures that arise from these, to the detriment of the organism's adaptive agency and likelihood of survival. Such a description has value because it emphasizes the partial entitativity of mental disorder. Mental disorders are processes within the agent-world system yet can be distinguished by their dysfunctional effects. While behaviour in general has a tendency to flow towards adaption and self-maintenance, mental disorders are process-structures that flow in the opposite, dysfunctional, direction.

[80] As an imperfect metaphor to try to capture this concept: If we take a river to represent the processes of human behaviour, then the ocean seems to represent the striven for state of self-maintenance/organizational autonomy. Stagnancy therefore represents death, and the general tendency of the river to flow towards the ocean – and to carve its own path through the landscape –represents adaption. Occasionally there are bends and rapids that represent challenges to the rivers flow; the trials and tribulations of life. Along the way, in interaction with these obstacles, eddies often form. These are normal back-flows in the fluid-dynamics of the river, representing normal but less-than-ideal behaviours; such as eating too much chocolate or staying up too late. In such eddies the behaviour is non-adaptive, but the flow is largely unimpeded (i.e. it is not dysfunctional/disordered). Within this image, mental disorder may start as an eddy, but gets larger and more persistent. Carried by the force of its own adaptive momentum and shaped in interaction with the dynamics of the landscape it flows through, the water cycles back around on itself, wearing its way into the bank until a pond is formed. The water still flows to the ocean, but its progress is significantly slowed-down; it risks stagnation.

of these component behavioral and experiential phenomena are themselves necessarily embodied, and therefore the current view immediately provokes questions as to how the observed behavioral and experiential phenomena are themselves composed at the biological scale. In anxiety these factors likely include but are not limited to: genetic polymorphisms, epigenetic factors, neurotransmitter levels, hormone levels, gut microbiota balances (Foster & Neufeld, 2013), neuronal structures, and the activity and structure of neural circuits and anatomical systems such as the amygdala and HPA axis. Still however, this is not taken to be the complete picture. On the embodied enactive view these behavioral and experiential phenomena are components of the agent's mode of functioning (even if together they are constituting a dysfunctional mode), and functioning is always embedded. Thus, the embodied enactive view immediately demands that we consider how the pattern of behavior and experience has been constrained by the physical and sociocultural environment. This will include direct causal links from environmental factors to the behavior, but also indirect causal links via the constituent biological factors. Examples include but are not limited to: childhood history, the actual threat level of previous environments, modeling of anxious behaviors, the amount of food available and which nutrients and vitamins this food contains, exposure to drugs including licit ones such as caffeine and alcohol, relationship history (including parental relationships) and whether these relationships supported the development of self-efficacy, gender norms concerning management of distress, the culturally mediated understanding of what it means to be anxious, etc.[81].

Getting More Precise

To add further detail and precision to this sketch, I will evaluate the concept against the six-factor conceptual taxonomy presented in Zachar and Kendler (2007) that I applied to RDoC in chapter three.

Causalism/descriptivism. The embodied enactive perspective developed here conceptualizes disorders as relatively stable dynamic causal patterns within the brain-body-environment system supporting the continued engagement with significantly dysfunctional behaviour. Given the sheer complexity of this system, such patterns will

[81] For further example and comparison to extant conceptual models see table two in chapter eight.

have multiple causal components and will differ across individuals. This view therefore highlights that categorizing mental disorders based on their causes is always going to be a challenging endeavor. Ultimately though, this concept of mental disorder is still a causal one. This position aligns with *causalism,* but stresses the complex nature of the causes at play.

The stable dynamic pattern view of disorder ultimately begs two questions: what is it that makes some individuals more likely to fall into this pattern in the first place (etiological mechanisms)? And what is it that makes the pattern relatively stable (maintenance mechanisms)? Maintenance mechanisms seem more likely to be common across different manifestations of a disorder, and also more relevant to treatment. For this reason, maintenance mechanisms may be better suited for a role in demarcating diagnostic entities. It is likely that during the developmental process of a classification system, causal knowledge in the form of empirical and theoretical science will continue to develop. Our understanding then, of causal and maintenance mechanisms, will in time shift from quite general to more specific until an optimal level for the pragmatic purpose of classification is reached. As a note, the view described here is open to the possibility of transdiagnostic etiological and maintenance mechanisms.

Essentialism/nominalism. Interestingly, the conception of mental disorder expressed in this thesis leans slightly more towards essentialism than one might think. If a pattern of recurrent dysfunctional behavior, with a similar MPC causal structure, is seen to be occurring with some regularity across individuals, then this suggests that there is some tendency within the dynamics of the human (in a nomothetic sense) brain-body-environment system, to fall into such a pattern; much like an attractor basin in dynamic systems theory. This fact makes disorders 'real' (as kinds/phenomena, rather than idiosyncratic instances of human suffering). They are recurring phenomena *discovered* in the world, rather than being concepts invented for practical reasons and/or capturing divergent occurrences that don't belong together in a meaningful sense. To be clear, I am not advocating a total essentialist or even discrete kind view here (realism and essentialism are often conflated). On the current view, psychiatric disorders are bound together by similarities in their causal network rather than by sharing some stable essence. They are much more like a biological species than an

atomic element in this regard, however they have no causal lineage as a species does. The two pools of water example I gave in chapter two seems a much better analogy then. This type of multiply realized kind is referred to as a "*type-causal"* kind (Magnus, 2014a). The core concept I am describing here could therefore be described as a fuzzy type-causal MPC kind. Note how this brings a certain flexibility; some mental disorders may have tighter hubs of causal relations within the brain (such as, arguably: ADHD, schizophrenia), and some may be more diffuse (such as, arguably: alcohol dependence, depression)[82]. All of these cases can be described as fuzzy type-causal MPC kinds with causal structures spanning multiple scales, but the distribution of causal influences across these scales likely differs (Kendler, 2012b).

The essentialism/nominalism continuum has particular relevance for the task of classification. In turning to the task of classification, we must remember that a classification system is a practical human endeavor, and will therefore always be influenced by pragmatic and concerns and its own historicity (Zachar, 2018). Furthermore, given that the view of mental disorder presented in this thesis acknowledges the complexity and multi-scale nature of causal structures supporting dysfunctional behavior, it seems very unlikely that a classification system is going to accurately 'carve nature at its (fuzzy) joints' any time soon. Because of these reasons it seems important to make a clear distinction between 'the reference' and 'the referenced' when thinking about classification systems. Even though the current view holds mental disorders (the referenced) to be real, diagnostic entities (the reference) should not be viewed as completely 'real'. They will likely always, or at least for a very long time, remain imperfect representations of the mental disorders that they are trying to capture. Thus, the current view makes a distinction between mental disorders in nature (seen as type-causal MPC kinds), and mental disorders as diagnostic concepts (which are probably best described by what Zachar and Kendler call *moderate nominalism*; they

[82] Some disorders may even have dense hubs of connection in the (socio-cultural) environment and thus in a sense be 'top-down' disorders; this seems to be the case with dissociative identity disorder (i.e. 'multiple personality disorder'), or other 'culturally bound' syndromes. I put quotes around 'top-down' as I do not wish to imply a hierarchy here, nor to fail to recognise the relational quality of such disorders. (they are not 'caused' by society so much as concern the relation between the individual and their community/socio-cultural environment).

are always partially shaped by our purposes, needs, socio-cultural values, and historical decisions).

Objectivism/evaluativism. At first pass, the conceptual framework presented here comes down clearly on the side of *evaluativism*. I mean this in the sense that, from an embodied enactive view, values and norms are a vital component in the conceptual fabric of mental disorder. However, I want to partially follow de Haan (in press-b) here, in noting that there is a conflict between the very nature of this proposed 'objectivism/evaluativism' continuum and the central tenants of enactivism. As explored in chapter five, embodied enactivism sees normativity as continuous with the natural world rather than as something that must be expunged to reach some 'objective view'. Norms and values are a part of the natural world when viewed through a non-reductionistic lens, and so it makes no sense to oppose evaluativism with objectivism as Zachar and Kendler do in their taxonomy (Zachar & Kendler, 2007). Instead, the DCT offers us a way to collapse the normative gap and see mental disorders as both objective things in the world *and* as strongly dependent on the negative implication of the sufferer's functional norms. Under this view, norms/values are seen as ubiquitous, and therefore necessary for a comprehensive understanding of human behavior and functioning. When I label the embodied enactive view as *evaluative* it is this thorough-going role of values that I am attempting to highlight, rather than that a behaviour being disordered is somehow not objective or less real.

Emphasizing the evaluative (yet still objective/real) nature of mental disorder, brings certain advantages. For example, in chapter five I explored how, while physiological norms of functioning will be the same across individuals, 'higher level' norms/values of both social and prudential kinds differ across individuals given different genetic and epigenetic predispositions, learning histories, and socio-cultural contexts. This highlights the need, from an embodied enactive viewpoint, to consistently consider cultural values in both practice and research within this domain. Open recognition of the ubiquity of norms and values in practices such as clinical psychology and psychiatry, as well as science in general, is viewed here as essential for supporting ethical decision making (Douglas, 2009). The fact that an embodied enactive view supports such a thorough-going role of norms and values seems best represented by the

label of evaluativism, but this, again, is not to imply that they are any less real for their normativity-ridden nature.

Internalism/externalism. It should hopefully be fairly obvious that the view presented here holds that both internal and external factors, as well as the interactions between them, are vital for a complete understanding of behavior and disorder (as per embedment). The current concept would therefore, at a minimum, fall under what Zachar and Kendler refer to as *moderate externalism*. This position would be flanked on one side by 'internalism' which basically refers to the idea that everything important is happening inside the organism; quite a reductionist view[83]. On the other extreme we would find a 'total' or 'radical' externalism which might hold that mental disorder is always caused by socio-cultural factors such as the stresses of capitalism; quite a radical view. The positioning of the current framework on such a 'continuum of externalism' can be further specified using Roberts, Krueger, and Glackin's (in press) taxonomy of externalist positions regarding mental disorder. This taxonomy separates between different classes/kinds of externalism regarding mental disorder. Under Roberts, et al.'s taxonomy the current view is at a minimum within the class of 'causal externalism' – essentially equivalent to moderate externalism here defined. Further, the position here espoused likely qualifies as what Roberts et al. refer to as 'relational externalism'. This is a position that holds mental life – and psychopathology – to be relationally constituted and therefore inherently interactive.

The reason I hesitate with the application of this 'relational externalism' label is complex; let me briefly expand. At its simplest the current perspective holds that mental disorder is a repetitive behaviour (/tendency in behaviour) that has the normative status of being significantly dysfunctional for the individual. Through the concept of embedment, we can see that the environment (both physical and socio-cultural) plays a vital and likely non-linear causal role in shaping and occasioning behaviour. It also plays a large role in determining the viability and therefore the normative status of the

[83] Internalism could be further separated into those that think everything important is happening a holistic physiological level, the level of 'neuro-circuitry (such as RDoC), at the level of brain-chemistry, at the genetic level, etc. An embodied enactive view rejects all such views by its commitment to embedment (the recognition of the contextually dependant nature of behaviour), and the taking of the whole brain-body-environment system as its focus of analysis.

behaviour (i.e., different behaviors work in different environments). Both the behaviour itself and its normative status are therefore contingent upon organism-environment relations. As explored in chapter four, the current view also accepts the relational nature of mental life – meaning and experience are *for* the organism, enacted in accordance with its needs in relation to the world. However, I do not see environmental factors as a *constitutive* part of the behaviour itself (i.e., my commitment to embedment rather than extension within this thesis). Ultimately, I see the causal/constitutive distinction as primarily an epistemological tool to be used as needed[84] (for relevant discussion see; Kirchhoff, 2015). While I am sympathetic to relational externalism of mind, I avoid constitutive expansion of the mind (i.e., 4e; see chapter four for my reasons). I am unsure whether this excludes the current perspective from being considered a true example of relational externalism in Roberts et al.'s intended sense.

Entities/agents. Zachar and Kendler (2007) describe how the entity position generally views "…individuals as *vehicles* for pathological syndromes…", while the agentic position holds that "…each psychiatric disorder as manifest in an individual patient is relatively unique." (p. 559). The current framework would certainly view each manifestation of dysfunctional behavior as unique in important ways. Moreover, the very reason a cluster of phenomena should be seen as a *disorder* at all is because it will ultimately run counter to the functional norms of the agent. This concept of disorder is therefore inextricable from a purposive and agential view. The embodied enactive framework here developed would therefore, at first pass, be best described as sitting under the *agential* position.

However, this is not to say that meaningful regularities across agents (e.g., disorders) cannot be extrapolated (i.e. classification is presumed to be a fruitful exercise, with fuzzy categories discernable based on the similarity of the causal/process structures across instances of disorder). Nor is it to say that, under the current view, mental disorders cannot be considered to be conceptually isolatable processes,

[84] I think also it makes pragmatic/communitive sense to utilize the skin boundary in this way – it seems meaningful and it's how the general public speak. That said – I am aware of an interesting current discussion concerning the use of 'Markov blankets' to determine (potentially pluralistic/nested) constitutive boundaries (Markov blankets are a way to calculate optimal system boundaries). See Ramstead et al., (2019) and Ramstead (2019).

demarcated by their negative functional effects. What I am effectively challenging here then is the idea that an agential view is necessarily opposed to an entitative one. Rather than a single continuum, entitativity and agentiality are better seen as two separate continua (albeit with a potential relation in that a fully entitative/disease model concept seems incompatible with an agential view). The description of mental disorder from an embodied enactive view at the start of this chapter described mental disorders as process-structures within the agent-world system, discernable by their significantly dysfunctional effects and their similarity to other cases of the disorder in question. This description highlights both the ultimate dependency of the disorder process on the striving agent, as well as the conceptual separability of disorder from the agent. This separability is based on the functional outcomes that the process-structure has for the agent (i.e., the fact that it is flowing in opposition to the striving agents' adaptive processes), and the fact that it is a pattern we can see in others. Overall this is certainly an agential view, but also recognizes a partial entitativity[85].

Categories/continua. The position sketched out in this thesis views mental disorders as dysfunctional patterns or tendencies in an agent's striving, constituted by many interrelating causal factors across the brain, body, environment system. Many of these compositional factors will themselves be continuous in nature, and therefore, the constituted patterns of behavior seem very unlikely to be definable in a clearly categorical manner. There may also be compositional overlap between individual disorders, despite their being important differences between them. For example, on this view two different kinds (or perhaps sub-types) of depression may hypothetically be isolated on the basis of the presence or absence of some important mechanism, while still sharing other important mechanisms and features. A good degree of fuzziness therefore seems to be predicted by the embodied enactive view, as can be seen by the description of mental disorders as fuzzy type-causal MPC kinds when discussing essentialism/nominalism. The precise degree of fuzziness however, will be different across different disorders and is really an empirical question. A blanket statement committing the embodied enactive view to either a continuous or reasonably categorical

[85] De Haan (in press-b) has extended my thinking on this issue since publication of Nielsen and Ward (2018).

view would therefore be inappropriate. Instead, as argued by Kendler, Zachar and Craver (2011), viewing the constitutional structure of mental disorders as fuzzy MPC kinds affords a large degree of flexibility – some disorder may turn out to be reasonably discrete, and some may turn out to be nigh on continuous. To stress however, this not to say that the differences between different kinds of mental disorder, nor the distinction between non-ideal behavior and dysfunctional behavior, will be meaningless. Rather these distinctions are based on the degree of (functional) norm-crossing, and empirical regularities across kinds respectively[86]. These boundaries are certainly fuzzy, but far from arbitrary.

Comparing Conceptual Models

Now that the central concept has been explored it seems pertinent to make comparison to some of the more relevant extant conceptual models explored in chapter two. Not all models will be compared in-depth, rather focus is given to popular views, and those that provide an interesting contrast. Following this, I will shift to noting two significant challenges that the current framework faces for it continuing development.

Structural models.

Within the structural dimension, the embodied enactive model of mental disorder developed here aligns best with a fuzzy/MPC kind view. It is explicitly recognized that, as a pattern of behaviour, the causal structure of mental disorders likely spans brain, body, and environment. The embodied enactive view however, also reminds us that this pattern is not occurring in a vacuum, or as some entity that can ever really be completely abstracted out from the agent concerned. Instead, mental disorder is considered to be residing within the process structures of the striving individual (themselves in context), and any model of the 'disorder process' is seen as necessarily an idealization (although potentially a useful one for explanation and the development of

[86] As argued by Zachar (2014), this genuine fuzziness invites pragmatic decision making in the development of diagnostic systems. The problem is of course the divergent purposes that these diagnostic systems are meant to serve. In the service of different purposes (e.g., explanatory efforts, the treatment of individuals, the development of talk therapies, the development of pharmacological treatments, diagnosis as relevant for legal decisions) different degrees of abstraction may be pertinent. How those performing the task of classification should respond to these different needs is an entirely different thesis.

treatments). The embodied enactive view therefore, is more strongly agential than the MPC view.

The most interesting structural model to contrast the embodied enactive view with however, is the likely the essentialist notion of mental disorder. In particular, I want to focus on two places of near similarity between the essentialist and embodied enactive views, so that it can be seen more clearly what separates them.

Firstly, on the current view and as predicted by essentialism, some disorders may, in time, be discovered to feature an 'underlying' hub of tight causal connections that are central to the disorder process. This causal hub of activity may potentially (but not necessarily) be 'in the brain'. For example, we may discover that some causal sub-type of depression reliably involves some alteration in some neural network 'X'. However, on the current view, to conclude from such a discovery that such a hub represents the 'essence' of a disorder would likely be mistaken. The discovery of such a hub – and coming to understand its mechanistic relation to the wider pattern of dysfunctional behaviour – would obviously be hugely useful (hence my argument in the following chapter that reductionistic strategies such as RDoC are likely to be fruitful). However, coming to see such a hub as 'the-disorder-proper' would likely represent a gross decontextualization under the developed view. Dysfunction lies in the wider pattern of behaviour and its (lack of) adaptivity for the agent in their environment; it exists in the relation between the organism and their environment. To abstract away from this complexity and instead focus on some apparent 'essence' risks reifying the mental disorder in question as a 'disease' or brain pathology, leading to a hyper-focus on the apparent disease processes at the expense of coming to understand the person and their context. To continue the above example, depression is so much more than simply a brain disease; coming to recognize some important brain mechanism at play should not change our recognition of this. The exception to this is if it can be shown that the disorder in question really is better thought of as a disease entity – that the proposed 'essence' really is the one key constitutional factor at play and can truly be said to cause the dysfunction. In this case however, the disorder in question seems to look more like a brain pathology with behavioral symptoms (e.g. Parkinson's), rather than being a

mental disorder[87]. This divide – between a brain pathology with behavioral symptoms, and mental disorder – is better seen as continuous rather than categorical. Recognizing this continuity doesn't fit with the essentialist view. By taking a fuzzy/MPC kind view, the current position allows for recognition of this continuity because, as mentioned in chapter two, the notion of an MPC is flexible; able to capture more, or less, heterogenous clusters.

Secondly, as touched on in chapters four and six, the embodied enactive view developed here sees mental disorders as real patterns to be found in the world; it is a mode of realism about mental disorder[88]. We can draw a parallel to essentialism here because the realist commitments of both positions mean that a causalist classification system is seen as a genuine possibility. Contrary to an essentialist view however – where types of disorder would be clustered due to a shared essence – the current view would prescribe classification based on the *similarity* of the causal patterns supporting disorder. Again, we can see similarity to the MPC kind view.

Normative models.

Turning to the normatively focused conceptual models, it should by now be apparent that the embodied enactive view presents a much richer and justifiable variety of 'functionalism' than either the statistical or evolutionary positions. By its situation within a richer framework of human functioning, the embodied enactive concept moves beyond considering what the individual 'should' be able to do according to either evolutionary or statistical norms at a species or reference-class level – positions which we saw in chapter two face significant limitations. Instead, the embodied enactive view recognizes that the assessment of somebodies functioning is always, in a certain sense, evaluative.

---

[87] Note the similarity to a Szaszian position here in making a distinction between a medical disease and mental disorder. Contra Szasz, this distinction is seen as fuzzy, and the current position also carves out a distinct conceptual space for mental disorder in a way that Szasz did not.

[88] Realism and essentialism – along with internalism – are often conflated. See Hartner and Theurer (2018) for discussion (although note that I disagree with their ultimate conclusion as it seems to rest on a the assumption that normativity cannot be part of the natural world – hence ruling out mental disorder as a fruitful target for mechanistic explanation).

Those well-versed in the philosophy of psychiatry may take opposition here, for evaluativism is traditionally seen as the antithesis of functionalism. Evaluativism sees mental disorders as irreducibly value-laden, while functionalism attempts to find an objective demarcation between the disordered and the benign. Pervasive in western thought is the idea that values/norms and objectivity don't mix; thus, the observed tension between evaluativism and functionalism. As foreshadowed by Thornton (2000) however, embodied enactivism allows us to plug the normative gap; to collapse this dichotomy and move to a synthesis view. This is because embodied enactivism recognizes that – assuming naturalism – if values and norms exist then they must simply be part of the natural world. Through its commitment to the DCT, embodied enactivism offers an account of norms and values as arising for particular organizational structures in the material world; i.e., purposive and precarious systems, striving to self-maintain and adapt (Di Paolo, 2005; Maiese, 2016). Hence, the current concept of mental disorder is evaluativist, yet no less real because of it, for it is tied to the real functional norms of the individual diagnosed; to the adaptive fit between the behaviour of the organism and its environment. To again put this most simply; whether the behaviour is working for the individual.

This move will hopefully go a long way in satisfying the evaluativist because it recognizes a role for norms and values in the concept of disorder; namely those norms that support the individual's self-maintenance and adaption. However, on the other side this move also avoids Foucauldian-style critique, because it is not the norms of society that are seen as at issue, only those norms that support the adaption and self-maintenance of the individual. Hence the current concept cannot be seen as a socially constructed label for deviance, rather it provides a conceptual route to offering diagnosis in the interest of the client. Observations such as those of Stier (2013) – that wider normative features such as the values of the clinician often play a role in diagnostic decisions – would therefore be seen as erroneous influences rather than reflective of what a diagnosis should represent; i.e., that the client is acting against their own best interests.

Unfortunately, this positioning beyond the oft-assumed dichotomy between objectivism and evaluativism – responsible for the concepts ability to navigate many of

the criticisms plaguing other models – is also what leads to the first of the key challenges I will soon explore. How should we go about operationalizing adaption so that we can make diagnostic decisions in line with the prescriptions of this concept, and avoid erroneous influences such as those observed by Stier (2013)? Until we can do so, the current position is open to a charge of offering only an ideal with very little practical value (although, this is not to say the concept itself cannot be useful, indeed it could be argued that an ideal is exactly what a conceptual model should represent!). This is one of the biggest challenges that the current view faces and I will return to it in the next section.

Briefly comparing to pragmatism before moving on, as explored when discussing essentialism vs. nominalism earlier, the practical and political realities of generating a classification system mean that diagnostic entities are never going to be perfect representations of the patterns of dysfunction people experience. Mental disorders as represented in our classification systems will always be biased and distorted by the needs and values of the groups generating those systems, as well as the practical limitations placed upon such groups (Zachar, 2018). I am therefore not opposed to the observations of moderate pragmatists who recognize this degree of nominalism regarding diagnostic entities. Through its commitment to realism however, the current position is in direct conflict with the radical pragmatism/near total nominalism of the likes of O'Connor (2017). The rejection of nominalism, and commitment to an ideal concept of what mental disorder is, above and beyond how it is used, means that the current concept offers exponentially more guidance than a radially pragmatic position.

Two Challenges

Embodied enactivism, with its roots in dynamical systems theory and its call for understanding the organism as a whole – in context and across time – seems well situated to comprehend the complexity of human behaviour. It has the potential to facilitate the convergence of psychological, neuroscientific, and phenomenological perspectives around a central conception of mental disorder, without undervaluing any one of these approaches. If we think back to the analysis of the conceptual foundations of both the DSM and RDoC (reviewed in chapter three), I would hope it is becoming clear that, in comparison, embodied enactivism seems to offer a superior basis for

understanding human behavior and mental disorders. Comparing to the conceptual models overviewed in chapter two also, this developing embodied enactive perspective is clearly a potentially fruitful perspective to explore. There appear to be certain primary strengths that support it in this regard: it prescribes a more comprehensive view (Fuchs, 2009), it brings many explanatory tools such as emergence and constitution to the table (see chapter three), and it also features a naturalistic account of functional norms/values (Di Paolo, 2005; Di Paolo et al., 2010; Maiese, 2016). Considering RDoC in particular, these strengths exist where RDoC is weakest. The comprehensive viewpoint and the availability of theoretical tools such as constitution stand in direct contrast to RDoC's neurocentricism, and the naturalistic account of norms/values – as seen in chapter five – offers a source of conceptual validity unavailable to RDoC, bound as it is in its implicit statistical functionalism. In closing this chapter however, rather than just summarizing the strengths I see in the embodied enactive approach, I want to point out two challenges that will need to be overcome if the potential of an embodied enactive psychopathology is to be actualized.

Operationalizing adaption.

I have argued that, in order to ethically label someone as having a disorder, we need show that their behaviour is not working *for them* (see chapter five). From an embodied enactive view this amounts to demonstrating significant impact on their processes of adaption and/or self-maintenance (again, this includes socio-culturally informed interpersonal-prudential norms, but not socio-culturally informed norms that do not do work for the individual). Conceptually, this is a very different approach – and a more justifiable approach – compared to simply measuring people's behaviour against the statistical norm, against the standards of their society, or against some concept of an evolutionary 'design' (see chapter two). But this also gets us to the first challenge with this view. Self-maintenance is easy to measure (if you aren't doing it, you are dead). But, how exactly do we measure adaption? The very definition of adaption is outcome-based as opposed to means-based; the more supportive of your survival a behaviour is, then the more adaptive it is seen to be. The embodied enactivist, with their understanding of behaviour as emergent from factors across the brain-body-environment system, centered around the organisms striving agency, and with their understanding of how

our cultural embedment shapes our mode of functioning, feels very uncomfortable being prescriptive here. There is no *one way* of being functional; *you lay your own path as you walk it*. There is a tension between this non-prescriptive stance and the needs of psychiatry and clinical psychology as institutions which demand operationalizable standards. In practice we need a more standardized way of assessing whether a behaviour is dysfunctional or not. The framework provided here does not currently provide this.

Firstly, I want to say that it is *not* my intention to meet this challenge in this thesis. For the time being this will remain a significant limitation of the embodied enactive view developed here. What I do want to do, briefly, is suggest a *possible* way forward. This would be to maintain (conceptually) that it is ultimately adaption and self-maintenance that the concept of mental disorder is concerned with, but import a framework or set of measures with a greater degree of specificity and prescription in practice. This would offer guidance in the evaluation of individuals as to the functionality of their behaviour. Importantly however, this framework/set of measures would be understood as a *heuristic*; an understanding of the ways that *most* people adapt. To restate a point from chapter five, an ethically minded psychiatrist/clinical psychologist will always be asking the question 'is this behaviour a problem for this person in their context?'

There are some extant frameworks/measures that may be able to serve this role as a heuristic guide to the assessment of functionality. One (more evaluative) approach may be to use something like 'The Good Lives Model' (T. Ward & Maruna, 2007) from forensic psychology, which lists a set of common domains of achievement in a typical 'good life' against which individuals could be evaluated. Another (more statistical functionalist) approach – similar to current standard practice in clinical psychology – would be to utilize a suite of relevant standardized symptom measures which attempt to approximate functioning via contrast to statistical norms (e.g. standard measures of depression, etc.). These two approaches have different strengths and weaknesses. Using a framework such as The Good Lives Model offers flexibility, and as such is likely more applicable across cultures, however using this framework would also require a lot of in-depth work by both the diagnostician and client. The embodied enactive framework

developed here would prescribe this flexible framework approach in a diagnostic setting over the standardized measure approach, in the hope that it would encourage a more client-centric and critical perspective. In research settings however, practical limitations may require the standardized approach. Developing and/or incorporating a heuristic framework to serve this role is an area of future development for the current embodied enactive conceptual framework[89].

Managing holism.

Embodied enactivism requires a comprehensive multi-scale and constitutionally minded view, consisting of brain, body, and environment. By taking such a view, there seems a danger that we may get 'lost in the infinite', and this presents us with the second challenge. Much like we saw with the biopsychosocial movement in chapter two, there is a real concern that the embodied enactive approach may risk unsustainable holism (Ghaemi, 2009). How should we go about investigating dysfunctional human behaviour and developing explanations thereof, when we are sitting within a conceptual framework that forces us to recognize the sheer complexity of the subject matter? A parallel issue here is that, given its grounding in dynamic systems theory, an embodied enactivist perspective can sometimes discourage researchers from taking an interest in causal mechanisms, instead encouraging them to map the observable dynamics of the system (Bechtel, 1998). This tendency is observable in the previous attempts to develop embodied enactive explanations of mental disorders that I reviewed in chapter four which seemed to lack explanatory purchase. Moreover, as mentioned in chapter five, taking an embodied enactivist stance means that we assume 'down-ward' causation, and this denies us access to the tool of theory reductionism (which many find useful in the face of such complexity).

[89] Another possibility seems to be implied here by Ramstead (2019). While my understanding of the mathematics involved is (very) limited, Ramstead's thesis implies a possibility of formalizing adaption through the notions of active inference and the free energy principle (FEP): "Systems that obey the FEP via adaptive action are said to engage in *active inference*; since it will look as if such systems are inferring the causes of their sensory states, via the selection of adaptive action policies, i.e., those associated with the least free-energy" (Ramstead, 2019, p. 27, emphasis in original). While the information-theoretic approach seems potentially at odds with some of the other theoretical commitments in this thesis, the ability to mathematically specify adaption – or perhaps 'approximate' it (i.e. using active inference theory as an epistemological model) – is an intriguing possibility. Another possibility, as I will mention in chapter 8, will be to utilize something like de Haan's (in press-b) four general characteristics of pathological sense-making as a descriptive guide.

Being mindful of this challenge, the question I am asking here is 'if this is what mental disorders are, how should we seek to explain them?' We need a way to parse the system and reduce complexity, while still maintaining an awareness of the complex whole, and that any reduction is always going to represent an idealization. I propose that a *pragmatically inspired mechanistic*[90] *reductionism* may suit this task[91]. This would be an approach to the task of explanation where we attempt to break down mental disorders into component 'parts', while explicitly recognizing that we are working with an idealized model (of a complex disorder process within an even more complex brain-body-environment system). In the following chapter I turn away from the task of conceptualization and shift to considering the task of explanation in order to propose one possible solution to this problem. I first overview some current approaches to the explanation of mental disorders, before attempting to develop a meta-methodological framework for explanation that makes sense in light of the concept developed across the last three chapters.

---

[90] The sense of 'mechanism' that I refer to here is different to that used in chapter two. Here I mean it in a minimal sense with no implication that the mechanism in question is functional/purposive. For example, this understanding of 'mechanism' could apply to an explanatory model of how a geyser shoots water into the air at regular intervals, just as well as it could be applied to an explanation of how bee's regulate the temperature of their hive (Illari & Glennan, 2017).

[91] A concern was raised by one of the reviewers of my first paper that this talk of mechanisms may introduce the very dualism that enactive perspectives are trying to overcome. I appreciate this concern but believe that enactive thinking is compatible with a mechanistic view, so long as the phenomenological and systemic views are not under-valued as a consequence. By explicitly recognizing that we are *modeling* and thus *idealizing* the complex realities of mental disorder, I don't see conflict here. What I propose in the next chapter is an epistemological method rather than a metaphysical commitment. Similar calls for a synthesis between dynamic and mechanistic approaches has been made before, see Bechtel and Abrahamsen (2010), Kaplan (2015), Fagan (2015).

Chapter 7: The RAP Approach to Explanation

> The so-called forms of illness in their present-day delimitation have turned out to be too large, on the other hand the elementary symptoms, because they represent single phenomena, are less useful for distinguishing between the various conditions. Between these two ranges of phenomena would be the *symptom complexes*.
>
> -Alfred Hoche (p. 342, 1991/1912, emphasis added)

To treat and manage mental disorders more effectively it is first necessary to develop good explanations of them. There is now growing recognition that the *status quo* approach of launching research expeditions in and around current DSM constructs has not resulted in sufficient progress (L. A. Clark et al., 2017; Insel et al., 2010; Lilienfeld & Treadway, 2016; Wakefield, 2015; Whooley, 2014). Recent responses to this challenge have been made in the form of several proposals regarding what mental disorders are and how we should go about explaining them. Key examples include the RDoC and Symptom Network Modeling ([SNWM]; Borsboom et al., 2018). These methods vary as to what they see as the most appropriate *explanandum* (i.e. the thing-to-be-explained; alternatively labeled 'target of explanation'). In the first part of this chapter I briefly examine these approaches, focusing on how their explananda are conceptualized and considering the degree to which such targeting will support the timely development of good explanations in psychopathology. Due to space constraints it is not my intention to review these approaches, merely to demonstrate the space they leave for the complimentary approach I subsequently present.

I suggest that the key weakness of current DSM style syndrome-based approaches is that they are focused on explanatory targets that are too complex and heterogenous. Such explananda tend to lack reliability and validity. I therefore see RDoC as a shift in the right direction, given that it prescribes the targeting of smaller, more reliable explananda. However, I suggest that RDoC (which is predominantly focused on sub-personal and/or single-level *abnormalities* in human functioning) goes

too far; setting its sights too small and arguably losing sight of the larger purpose of explanation in psychopathology research. The RDoC instead seems to be serving a separate important task: that of uncovering potential 'ingredients' to serve within our boarder explanations. SNWM approaches meanwhile, model the larger picture that the RDoC overlooks, and has many strengths in this regard. As targets of explanation however, SNWMs have weaknesses in their replicability and their thin symptom-level focus.

In the second half of this chapter I present my own method for developing explanations of psychopathology which I label the *Relational Analysis of Phenomena* (RAP) approach. This approach is inspired by the work in previous chapters. As mentioned at the end of the last chapter it is intended to be congruent with an embodied enactive understanding of mental disorders, but blend in a pragmatically inspired mechanistic reductionism in order to address the challenge of balancing holism with cognitive manageability on the part of the explainer. I also wanted the RAP to be accessible to researchers with diverse theoretical commitments and so throughout this chapter I have generally avoided embodied enactive terminology. I should also note at the outset that the RAP is really designed to produce an understanding of the *maintenance* of disorder rather than necessarily its etiology.

Rather than targeting large heterogenous syndromes or only focusing on single phenomena, the RAP approach conceives of its explananda as small *idealized* systems of phenomena I label *phenomena complexes* (PCs). Such systems (/models) are much smaller and simpler than current SNWM-style networks or DSM-style syndromes, thus improving reliability across individuals. They are composed of small sets of clinical phenomena and their apparent causal interactions. This allows explanatory focus to be given to the *relationships between phenomena*, a novel focus compared to the other methods reviewed. It is important to stress that I adopt a pluralistic perspective to the explanation of psychopathology, and as such view the RAP approach as providing an additional explanatory method, rather than being necessarily the 'right way'.

Explananda in Current Approaches

The role of explanandum has two major competing requirements. Firstly, a good explanandum should be a robust and reoccurring phenomenon (Haig, 2014). This means that the thing we are trying to explain needs to be reasonably similar (both in appearance and constitution) across different instances/persons – i.e. it needs to demonstrate *construct stability* (Sullivan, 2014). At the same time, explanations have a motivating context within which they are sought, and this context forms the pragmatic landscape; rightfully influencing many of our decisions during the explanatory process (Potochnik, 2010, 2016, 2017; Thagard, 2017). Within science, this often takes the form of a research problem that motivates the enquiry process (Haig, 2014). The second major requirement is that a good explanandum must maintain its *relevance* to this reason for seeking an explanation. We primarily seek explanations of psychopathological phenomena because they bring about harm and dysfunction in people's lives. We want to know how to alleviate this harm as effectively as possible, and to do so quickly. By 'relevance' within this context then, I mean the degree to which an explanandum is related to the harm and dysfunction that mental disorder represents. It is, after all, this impact on the lives of individuals that motivates our explanatory efforts.

Targeting phenomena in our explanations that balance the two requirements of *stability* and *relevance* should result in better explanations and encourage explanatory progress. In this section I briefly overview three current approaches to selecting and describing the explananda of psychopathology research.

DSM based approaches.

Historically and currently, DSM syndromes are commonly used to define the explanandum in psychopathology research (Berenbaum, 2013). There are many recognized problems with the DSM which were summarized in chapter three. The issue most relevant here is that of heterogeneity (Lilienfeld, 2014). This is where individuals with the same diagnoses often have differing patterns of symptoms with differing levels of severity (i.e. *symptomatic heterogeneity*), suggesting that the diagnostic label in question may be capturing more than one underlying causal process (i.e. *etiological heterogeneity*). This concern is well evidenced; prototypical disorders such as post-

traumatic stress disorder [PTSD], eating disorders, schizophrenia, and depression have all been shown to capture large and heterogenous populations (Contractor et al., 2017; Dickinson et al., 2017; Galatzer-Levy & Bryant, 2013; Hawkins-Elder & Ward, in press; Monroe & Anderson, 2015). Ultimately this suggests that there are good reasons to doubt the etiopathological validity of the DSM's diagnostic constructs – i.e. that they pick out similarly constituted entities with common causal processes/structures. For someone seeking etiopathologically and constitutionally valid, rather than simply descriptively valid, entities, DSM-style syndromes seem artifactual and not the sorts of stable and relevant things we should seek to explain.

Research domain criteria (RDoC) based approaches.

In response to the problems with the DSM mentioned above, the US National Institute of Mental Health (NIMH) launched the RDoC; a research framework with the goal of shifting attention from signs and symptoms to the underlying casual processes that generate them. In doing so it assumes mental disorders to be disorders of 'brain circuitry' (Insel et al., 2010; Morris & Cuthbert, 2012).

RDoC adopts an organizational matrix with a horizontal axis containing seven 'units of analysis' (which specify structural 'levels' of enquiry), and a vertical axis listing basic psychological functions (Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Lilienfeld & Treadway, 2016; Morris & Cuthbert, 2012). The explanatory aim is to study how a phenomenon observed at a particular unit/level (e.g. higher levels of striatal dopamine, lower dendritic spine density in brain area X) affects the degree to which the basic functions are achieved (e.g. response to acute threat, approach motivation). The hope is that this process will uncover transdiagnostic mechanisms relevant to current diagnostic labels (Cuthbert & Insel, 2013; Hoffman & Zachar, 2017). Under RDoC 'transdiagnostic mechanisms' refer to neural circuit abnormalities that negatively affect the specified functional domains.

While many have concerns surrounding the potential neurocentricism and reductionism of the RDoC movement (Berenbaum, 2013; Hershenberg & Goldfried, 2015; Kirmayer & Crafa, 2014; Lilienfeld, 2014; Lilienfeld & Treadway, 2016; Nielsen & Ward, 2018; Wakefield, 2014a), the shift to focusing on transdiagnostic mechanisms

and their relation to specified functions represents a shrinking of explanatory targets towards more stable phenomena. This move seems an advisable response to the heterogeneity plaguing DSM-defined targets. Regarding this move however, Hoffman and Zachar point out a concern that we share:

> "[t]he worry is that in order to achieve the fineness of grain needed for elucidation of causal mechanisms, we risk losing connection to the "coarse" clinical phenomena of interest." (Hoffman & Zachar, 2017, p. 68).

This relates to the aforementioned requirement that an explanandum maintain its relevance to the reason for seeking an explanation in the first place. The concept of mental disorder is inherently normative, yet outside the specified basic functional domains there is no broader normative element within RDoC with which to give RDoC's findings meaningful conceptual validity (Nielsen & Ward, 2018; Wakefield, 2014b).

Essentially then, there seems to be a possibility that RDoC represents an overcorrection in the grain size of the explanatory targets in psychopathology – in which targets do not maintain their relevance to the wider dysfunction and suffering that motivates our enquiries. Ultimately, this concern is probably outweighed by the sheer amount of basic research that RDoC will facilitate. But we need to be clear about what RDoC is doing. Research within the RDoC framework searches for (largely sub-personal) abnormalities that likely play constitutional and/or causal roles as *components of* psychopathology. This is vital work, as it discovers and confirms phenomena that can then be used to weave together an explanation – but such phenomena do not themselves constitute explanations of psychopathology.

RDoC grants greater freedom to researchers, in that under the RDoC framework they no longer have to justify their research interests by linking them to some particular and established problem (i.e. DSM syndromes). This freedom will be good for psychopathology as a complete scientific endeavor (Casey et al., 2013), but the component task of *developing explanations* of psychopathology has different requirements to the larger science within which it sits. As discussed, ideal explananda balance stability and relevance to the larger disorder space. By targeting largely sub-personal abnormities and investigating their potential role as transdiagnostic

mechanisms, RDoC seems to prioritize the prior at the expense of the later. In doing so, RDoC seems be performing a different task to that of picking out ideal targets and explaining them. Rather, it seems to be providing the sub-personal ingredients for our explanations.

Symptom network modeling [SNWM] based approaches

SNWM is presented by its advocates as a new model of mental disorder that rejects the search for underlying cause/s of psychopathology. Instead, SNWM assumes that many mental disorders are better understood as *networks of symptoms,* which can then be statistically modeled. Symptoms within these networks are hypothesized to cause each other, with recursive feedback resulting in the relative stability of the network over time (Borsboom et al., 2018; Cramer et al., 2010; McNally, 2016). Recent years have seen a significant increase in SNWM research, with many examples of it being used successfully in empirical studies (Fried et al., 2017).

There is considerable debate over whether SNWM is really a new way of thinking about mental disorder, or simply a new and promising measurement tool (Bringmann & Eronen, 2018; Epskamp et al., 2017; Fried & Cramer, 2017; Haig & Vertue, 2010; Humphry & McGrane, 2010; Molenaar, 2010). Following Ward and Fischer (2019), I consider SNWMs to be best thought of as *phenomenal models*[92]. Phenomenal models *describe* the explanandum, particularly as it changes over time (Hochstein, 2012, 2013, 2016). Such models do not do explanatory work themselves, but are vital for the task of explanation, especially when the focus of enquiry is complex.

The key strength that SNWMs bring as phenomenal models is specificity. We can ask more specific questions that we can with DSM syndromes such as: why particular symptoms predict others, or why certain networks of disorder predict the 'activation' of other networks. I refer to this specificity regarding associations as *horizontal detail*. One potentially useful element of horizontal detail is the ability to measure the *centrality* of a symptom within a network, effectively mapping the strength of its associations with other relevant symptoms (Fried et al., 2017). Centrality then, may potentially act as a

[92] Rather than being a novel conceptual model, hence why I have not reviewed SNWM in earlier chapters.

guide as to where we should most efficiently focus our exploratory and explanatory work[93] (Fried et al., 2016). There are recognized ways that horizontal detail could be improved in SNWM approaches, namely: shifting to the use of directed networks (these incorporate longitudinal data), focusing on individual rather than group abstracted networks, and increasing the sample rate to produce greater temporal resolution/dynamicism (Bringmann et al., 2013; Bringmann & Eronen, 2018; Fried et al., 2017; McNally, 2016; Molenaar, 2010; Wichers et al., 2017). Methods to implement such improvements are being developed (Booij et al., 2018; Cramer et al., 2016).

Another potential strength SNWM features is relevance. By this I mean that the collection of symptoms and their associations being described by a SNWM generally seem to represent genuine problems for the people in which these symptom dynamics are embodied. This is a strength relative to the more microscopic view of the RDoC[94].

SNWMs also have significant weaknesses as phenomenal models. SNWMs are often generated from group-level data with large sample sizes, and currently there are no established 'goodness-of-fit' measures that asses the reliability with which the group level abstracted network matches the pattern of associations within individuals (Beard et al., 2016). This is problematic because patterns that emerge at a group-level are not always present at an individual-level, yet ultimately it is the individual that we are most interested in when developing explanations and treatments (Barlow & Nock, 2009; Beltz et al., 2016; Blampied, 2017). The lack of an appropriate measure of group-level to individual-level reliability brings into question SNWMs ability to meet the stability requirement we have outlined. The suggestion of shifting to the measurements of symptom networks *in* individuals *across* time mentioned earlier may go a long way in addressing this issue (e.g., see Fisher et al., 2017).

---

[93] There are some concerns surrounding the interpretation of network centrality in this way; conceptual overlap between recognized symptoms may artificially inflate measures of centrality and central symptoms do not necessarily reflect causal importance (Bringmann & Eronen, 2018; Dablander & Hinne, 2018; Fried & Cramer, 2017).

[94] That said, one criticism not mentioned above is that the reason that a given network of symptoms should be labelled a 'mental disorder' is left unclear (Cooper, 2013b; Zachar, 2010). This is of concern if one views SNWM as a *concept* of mental disorder, but even viewing SNWMs as phenomenal models as I do this is still potentially problematic. The association with dysfunction, harm, or distress is the reason for seeking explanations and therefore seems like something that should be captured in the description of the explanandum.

The biggest weakness of SNWMs as phenomenal models however, is that they lack *vertical detail*. By this I am referring to the fact that SNWMs operate purely at the 'symptom level' (T. Ward & Clack, 2019a). This may seem an odd criticism to make – they are *symptom* networks after all. However, Borsboom et al. (2018) claim that SNWM as an idea is inspired by, or at least conceptually related to, the concept of a *mechanistic property cluster* (MPC); "A research program that has put…[mental disorders as MPCs]… to work is the network approach to mental disorders." (p. 12). As seen earlier, under the MPC view mental disorders are *explicitly multi-scale* clusters of mutually-reinforcing causal mechanisms. Given this grounding in the MPC concept, it is unclear why SNWMs should be limited to the symptomatic scale. For SNWMs to act as phenomenal models of disorders that facilitate mechanistic insight, they need to map rich *multi-scale* detail, including the constitution of the symptoms themselves (Ward & Clack, 2019).

Summary

To summarize, DSM defined targets seem too unstable to serve as explananda productively. SNWM and the RDoC meanwhile, both perform complementary but distinct roles relative to the method presented in the latter half of this chapter. SNWMs are useful for *describing* functional relationships between symptoms, and the RDoC will help uncover (largely sub-personal) differences in those who experience disorder; phenomena which may then play a role in our explanations. The RAP method I will present is designed to focus explanatory attention on the *relationships between* symptoms (or rather *clinical phenomena*). I will now overview why we think such a focus will be productive before presenting the method itself.

Groundwork for an Alternative Proposal

Recent conceptual models concerning the structure of mental disorders highlight the possibility of *emergent stability* playing a key role in their maintenance (Borsboom et al., 2018; Fuchs, 2009, 2017; Kendler et al., 2011). Emergent stability refers to the idea that something (e.g. a mental disorder) may persist due to the causal relationships between its parts cycling back and resulting in a stable pattern or state. Primary among these views is the MPC view, which as I mentioned earlier underlies the currently

popular SNWM approach. Again, this view holds that many mental disorders may be constituted by *mutually reinforcing* causal mechanisms that cross scales of analysis (Kendler et al., 2011). The embodied enactive conception of mental disorder developed over the last few chapters also takes such a view (although it sees this stable pattern as existing within the wider brain-body-environment system, compromising the persons processes of adaption and/or self-maintenance).

The common thread to these views is the *highly circular process structure* of the mechanisms seen to constitute disorders (Fuchs, 2009, 2017). In effect, this circularity can be seen as a basic form of *self-maintenance* (see chapter six, description two). For example, there is now converging evidence that non-suicidal self-injury (NSSI) such as cutting, scratching, punching objects etc., self-perpetuates due to it serving an emotion regulation function. Short-term, engagement with NSSI has been shown to alleviate emotional distress. In the long-term, it fails to relieve distress and discourages the use of alternative regulation strategies. This then seems to lead to continued engagement with NSSI despite its significant risks[95] (Chapman et al., 2006; Robinson et al., 2018).

To optimize *relevance* then, what we need to understand about mental disorders is this self-maintaining dynamic. As reviewed, the RDoC is targeted at a grain-size that is likely inappropriate for the purposes of capturing this circular causality and the ensuant maintenance of dysfunctional behaviour. SNWM is of an appropriate grain-size for this purpose but is focused on *describing* the relationships between the parts (symptoms) of disorder rather than explaining them. How then should we best seek to understand the relational structure –the diachronic constitution – of mental disorders? I see room here for a method that focuses on the relationships between the parts of a disorder, as a way of developing an understanding of how mental disorders self-maintain. One way that this could be achieved is to repeatedly select out small

[95] Other examples are: negative reinforcement in substance dependency, reinforcement of anxiety in parent-child dyads. Intuitively, schizophrenia is an example of a disorder that does not seem to feature this self-maintaining process structure. However, in terms of schizophrenia producing distress and dysfunction, phenomenological analyses highlight the role of psychosocial alienation feeding back to produce feelings of distress and 'un-worlding' (de Haan & Fuchs, 2010; Fuchs & Röhricht, 2017; Maiese, 2016). While it is not clear whether a circular cause underlies hallucinations and delusions, it does seem to play a role in maintaining and moderating the distress and dysfunction that arises from these symptoms (and thereby its disordered status).

systems/models of interacting parts from the wider disorder for closer analysis. This is the core idea of the *Relational Analysis of Phenomena* (RAP) approach.

The Relational Analysis of Phenomena (RAP) Approach

The RAP framework is designed for use by researchers when attempting to develop explanations in the field of psychopathology, by research teams planning multi-disciplinary investigatory projects, and may also be useful for clinicians reflecting on their explanatory methods (although it is not intended to be applicable to clinical practice wholesale). It is particularly focused on the development of explanations of the *maintenance* of disorder, in that it is primarily designed to produce constitutional explanations of the dysfunctional behavioral pattern (e.g. why the components hang together/continue to be engaged in) rather than etiological ones (e.g. what led to the development of the disorder in the first place). From an embodied enactive perspective this method is seeking to develop an idealized mechanistic model of the stable dynamic pattern that constitutes the disorder process within the brain-body-environment system.

According to the RAP, it is not the objects of our classification system that we seek to explain. Classification systems are simultaneously ontological lists, diagnostic tools, and socio-political documents (Zachar, 2018). Each of these tasks bring their own biases and constraints. This issue requires much deeper analysis than space allows but, suffice to say that classification systems in psychopathology will always be subject to such competing purposes. Instead of recommending the complete separation of psychopathology research from DSM categories, as per the RDoC, the RAP allows DSM syndromes to point out potential areas of exploration, while not allowing for DSM syndromes to *define* either the local explanandum nor the wider disorder one is trying to understand. I henceforth refer to the wider disorder as the *problem space* for the purposes of clarity, and to highlight this decoupling.

The three phases of the RAP are: Phase one – *List and map*; Phase 2 – *Focus and enrich*; Phase 3 – *Explain and evaluate*. As I will describe, these phases are designed to allow investigators to go back and improve their explanations and processes over time (see figure 2). The RAP is also iterative in a larger sense. Cycling back and seeking to

explain different overlapping PC structures should produce an understanding of the wider problem space; an understanding of the constitutional structure of the disorder process.

Phase 1: List and map.

List (1a). The key task at this stage is to develop a list of reliable phenomena within the problem space. A comprehensive literature review is called for, identifying and evaluating the reliability of possible phenomena (e.g. checking for replicability, multi-method triangulation, lack of conceptual overlap with other phenomena). The objective is a listing of *clinical phenomena*.

I have used the term 'phenomena' throughout this thesis, but it is worth specifying my meaning here. The concept of central importance during this stage of investigation is the data/phenomena distinction. This distinction is made by Bogen and Woodward (1988) and discussed further in Haig's (2014) Abductive Theory of Method. According to this distinction, *data* are observable things such as recordings or reports about the state of the world. Unfortunately, data is inherently noisy and often biased. A *phenomenon* meanwhile is an apparent fact about the world, inferred from the data based on reliable patterns therein. A reliable phenomenon will be inferred from multiple and replicable sources of data (e.g. self-report, observation, behavioral tests). On this view scientific theories do not explain *the data*; rather they explain *the phenomena*. This explanation is achieved through the postulation of causal or constitutional mechanisms. Generally speaking, phenomena can take many forms such as objects, states, processes, events, and effects (Haig, 2014).

When I speak of *clinical* phenomena, I refer to phenomena that are relatively specific to the target population compared to the wider population, or that otherwise seem to be playing a role in the problem space. Within the RAP approach I explicitly use 'clinical phenomena' to refer to *behavioral and phenomenological* instances found reliably within the problem space. This is not to say that these phenomena only exist at these scales/perspectives (phenomena are usually observable at multiple scales), only that within this method, phase 1 should be limited to detecting phenomena within these domains. This is done to help reduce the complexity of the task at hand and keep things

manageable. The choice to limit phase 1a in this way will likely influence the overall image of a disorder that the RAP will produce; anchoring our understanding to the level of behaviour and experience. While this represents a bias/idealization away from the likely complexity of actual disorder processes, this makes sense from an embodied enactive view which highlights the important role of phenomenology and action.

At phase 1a the focus is on states (e.g. moods and emotions, levels of awareness), events/actions (e.g. self-harm, outbursts of anger, bodily sensations or other perceptual experiences), and tendencies/dispositions (e.g. thought-action fusion, apparent perceptual biases, anhedonia, paranoia). Effect and process type phenomena are of interest but are incorporated in phase 1b. (e.g. that purging often follows binging, that anhedonia often increases with chronic stress).

Phenomena that occur within other disorders *should* be included on this list. While it may be tempting at this stage to simply import the DSM criteria, which are often taken to describe the recognized problem, this will not be a fruitful approach. We want to eventually explain the disorder as it *actually* occurs, not as it is idealized in our diagnostic manuals which have been heavily biased for diagnostic reliability and other purposes (Zachar, 2018). Phenomena measured by psychometric tests may also be unexhaustive and should be supplemented by comprehensive literature trawling and observation.

Map (1b). The key question at this stage is 'what are the known/apparent relationships between the clinical phenomena?' Technically these relationships (when reliably detected) are themselves phenomena. For clarity I therefore refer to them as *relational* phenomena. Investigators should seek to map the clinical phenomena listed in phase 1a into a network of relations. Here I am drawing on the SNWM approach. This can be done using directed symptom network modelling (dSNWM), some other form of dynamical modeling, or (in lieu of such tools) a time-sensitive conceptual sketch. The relational phenomena that emerge should themselves be subject to the requirements of replicability and multi-method triangulation to ensure their status as phenomena.

At this stage, awareness needs to be drawn to the fact that relational phenomena in psychopathology exist at varying time scales. For example, panic disorder is defined

by the presence of panic attacks leading to persistent worry and/or maladaptive behavioral changes in response to the panic attacks (commonly taking the form of agoraphobia). The development of fear and avoidance strategies, and the possible resulting low moods and other secondary impacts, occur on a time scale ranging from days to months. Compare this to panic attacks themselves; a collection of physiological/experiential phenomena which occur over a timescale of minutes. In managing this temporal complexity, it may be necessary to produce multiple maps of associations between phenomena at different time scales. The tighter frame-rate associations may then be nested into the wider time-frame network as a composite phenomenon.

A tangential but important task at this stage is to perform a validity check of sorts. Investigators should consider whether the behaviors understand are genuinely problematic. The question here is, 'what is it that makes this a problem for individuals within their physical and cultural context?'. Creating a list of the functional norms typically being impinged by this network of behavioral and phenomenological phenomena may be helpful, and – alongside the centrality of clinical phenomena – can also be used to guide the targeting process in phase two.

Phase 2: Focus and enrich.

Focus (2a). The key task at this stage is to select a cluster of two to four clinical phenomena and their relations from within the now mapped problem space, and to model them as a small system of interacting phenomena (i.e. temporarily ignoring their relation to clinical phenomena outside this selected system). This idealized model is referred to as the *phenomena complex* (PC). At least two phenomena are obviously needed so that there is a relation to explain. The suggested upper-limit of four phenomena within the PC is chosen simply to support manageability on the part of the explainer. Once richly described (in phase 2b), this PC will take the role of explanandum.

The selection of phenomena at this stage is not arbitrary, but at the same time PCs are not intended to be 'real' things in the sense of being naturally separable parts of disorders. Instead, they are pragmatically defined abstractions that try to balance

relevance, fertility, and manageability[96] (Potochnik, 2017). Accordingly, there are certain considerations that should inform the selection. Firstly, early in the project investigators should prioritize core phenomena and relations that appear to be doing a lot of work in the network produced during phase 1b (SNWM measures such as centrality may be useful here), or phenomena that seem important because of their particularly negative impact on people's lives. Secondly, PCs are primarily identified pragmatically, however targeting of apparent natural clusters within the mapped problem space is a good option. The limiting factor is that the PC should be limited in size as to keep the task of explanation manageable.

Finally, *ideal* PCs will feature a circular organizational structure. This circular process structure is conceptually what allows for the self-maintenance of the dysfunctional behavioral pattern. Selecting circular structures as PCs then, effectively balances the two key explananda requirements of stability and relevance. The competing need to keep the task of explanation cognitively manageable however should not be under-valued. If capturing this circularity is not possible while keep the number of constituent phenomena low, then ignoring the possibility of capturing circularity within the current PC and focusing on the selected relational phenomena is perfectly valid. As we will show later, the iterative nature of RAPs design allows for some exploration here – there is no one correct selection of phenomena.

These PCs then are small systems of two to four clinical phenomena. The relations between the constituent clinical phenomena are seen to be *potentially* causal in that there is good evidence for a causal link between the constituent phenomena, but the exact mechanism is unknown. By conceptualizing the explanandum in this way, explicit attention is drawn to the process structure of the disorder space and how this supports the organisms continued engagement with dysfunctional behaviour. What we seek to explain (in phase 3), is the nature of the relationships between the constituent phenomena.

Prototypical examples of ideal PCs already exist, such as the binge-purge cycle or self-starvation spiral in eating disorders (Hawkins-Elder & Ward, in press), experiential

[96] It is for this reason that I do not believe that PC selection represents a carving error (Franklin-Hall, 2016)

avoidance cycle in OCD, or escalation cycles in the families of children with conduct problems. Note that these are ideal examples in that they are all highly circular structures, metaphorically acting as 'engines of distress'. To restate, PCs do not always have to feature this circular organization. Readers may protest at this point that these examples seem to be theories rather than descriptions, and they would be in-part correct. These examples seem to foster a degree of understanding as to why individuals continue to binge and purge, starve themselves, perform bizarre rituals, or consistently misbehave respectively. But as 'theories' they are remarkably thin. They rely on intentional and empathetic inferences on the part of the person using them to understand someone's behaviour. Beyond this, the mechanisms remain largely unknown. It therefore seems more accurate to consider them as phenomenal models (Hochstein, 2012, 2013, 2016), or as cyclical mechanism sketches waiting to be filled out (Bechtel, 2011; Piccinini & Craver, 2011).

Enrich (2b). The task at phase 2b is to develop constitutional descriptions of each constituent clinical phenomenon. The constitution of the selected clinical phenomena must be described across scales of analysis both below *and above* the behavioral and phenomenological.

Here we draw on the ideas of Ward and Clack (2019a) and Hochstein (2016), in that the constitution of each clinical phenomena should be described via a set of *friendly* models at varying scales of analysis. This method is required given the constitutional complexity of clinical phenomena. The term 'friendly' refers to the fact that the descriptive models should be reasonably coherent, but not necessarily integrated or reducible to each other. The reason for this use of pluralism is that explanations at different scales make different *idealizations*, i.e. different models of the same phenomenon or mechanism are designed to abstract away from certain elements and to focus on different elements (Hochstein, 2016). Consider the phenomenon of anhedonia. One popular neurological model of anhedonia, postulated by Ferenczi et al. (2016), focuses on activations/modulations of different brain areas/neuro-chemical systems and their ensuant effects of reward seeking behaviour. In doing so it abstracts away from individual differences and contexts, and indeed genetic factors that may be playing a role in the wider phenomenon. Compare this model to behavioral models that may

focus on wider contextual factors (e.g. stress) and map the behaviour in finer detail; phenomenological models that attempt to richly describe the difficulties with feeling pleasure from a first-person perspective[97]; or cultural models that try to capture how different kinds of positive emotions may be more important across different cultures, thus changing the impact of anhedonia in different contexts. A truly rich understanding of the constitution of a clinical phenomenon requires description through a plurality of models across scales of analysis; see Ward and Clack (2019a, 2019b), and Hochstein (2016) for further discussion of this pluralistic method of description.

Incorporating externalism? As a guide for structuring pluralistic description, Ward and Clack (2019) suggest the possibility of utilizing the RDoC units of analysis. The use of an organizing structure in which to nest the set of descriptions is a useful one, however using the RDoC units does risk importing its neurocentricism. The mechanisms supporting behavior are not necessarily within the individual, but often span the environment. When seeking to understand complex systems such as humans are, we must – in Bechtel's (2009b) words – look not just down, but also up and around.

Consequently, I support Ward and Clack's (2019a) suggestion that the RDoC units may provide a helpful structure to support multi-scale description, but strongly suggest that investors add to this heuristic structure in a manner that prompts consideration of the situational, developmental, historical, and cultural contexts in which the phenomenon occurs, fails to occur, or occurs in a different form. Table 1 gives a hypothetical example of pluralistic multi-scale modeling of the clinical phenomenon of hyper-vigilance.

[97] 1st and 2nd person narrative accounts may be useful. See Fuchs (2017) on dual aspectivity.

Table 1. Hypothetical example of multi-scale description looking at the phenomenon of hyper-vigilance.

Scale of Analysis	**Description/Model**
Neurological/Physiological	Increased amygdala response to threat; reduced activation of ACC – associated with the regulation of emotional responses (Garfinkel & Liberzon, 2009; Liberzon & Martis, 2006); Hyper-sensitive sympathetic arousal (i.e. increased heart rate, sweating)
Behavioural	Persistent checking of environment for threat; Increases during times of stress; Hyper-reactive anger/fear response; Avoidance of novel situations
Phenomenological	Vivid awareness of escape routes in all situations; Constant feelings of being 'on edge'; Lowered, awareness of the current social context due to monitoring of environment for threat; Difficulty feeling that others are trustworthy
Social	Fearful or distrustful response by others, or frustration at inattentiveness; Employment difficulties

Note: In actual practice this description would be a lot more detailed, and based on detailed literature review.

Endpoint of phase 2b. Once the multi-scale models for each constituent clinical phenomena have been collected, the PC is seen to be complete and ready to serve as a pragmatically defined explanandum in the explanatory phase. By this stage PCs should:

- Be composed of parts that reliably correlate, or better yet, parts that have a longitudinally or experimentally evidenced directional relationship
- Be more (mereologically) simple than current diagnostic constructs, yet much more richly described (in terms of constitution)
- Be thought to play a role within one *or more* recognized disorders (not necessarily DSM recognized)

Further to these requirements, *ideal* PCs may be highly circular in their organization (Fuchs, 2017) – e.g. they already have a simple form of self-maintenance. It is this causal structure that represents work against the self-maintenance and adaptivity of the individual (see chapters five and six). This causal structure may not be present at the PC level, but instead may emerge as investigators cycle back and develop explanations of other PC structures in the problem space. Figure one offers a visualization of a PC structure.

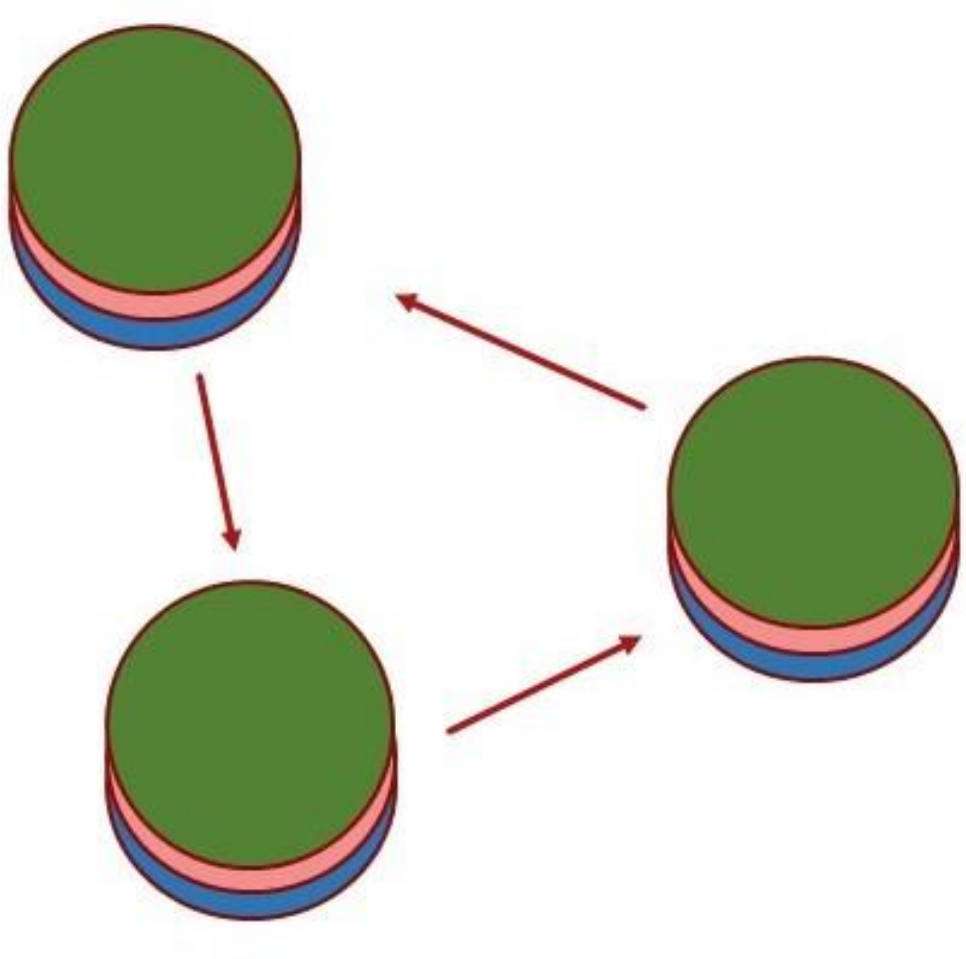

Figure 3. Visualization of a Phenomena Complex [PC]. Each constituent phenomenon is described at multiple levels using friendly models (represented here by the different colored circles). The relational phenomena (red arrows) will become the focus of explanation in phase 3.

Phase 3: Explain and evaluate.

Explain (3a). The task at this point is to make an inference to a constitutional explanation of the PC. Effectively the aim is to utilize the rich understanding of the constitution of the clinical phenomena (developed in phase 2) to infer explanations of the relational phenomena, thus explaining the internal structure of the PC.

The inference here is an *abductive* one. At its simplest, abductive inference is the postulation/recognition of some state of the world that serves to make another state of

the world (the explanandum) less surprising (Haig, 2014). A vital point here is that investigators should be looking for both potential *causal* links between the constituent clinical phenomena, but also be looking for potential constitutional overlap (analogous to a latent variable approach). Within the explanations generated no scale of analysis should be given *a priori* preference.

As a simple example we will look at two hypothetical relationships and their possible explanation. Let's say we have mapped out the problem space of 'depression' (phase one), and from this mapping we have isolated out a PC containing three phenomena: high stress, sudden waking during sleep, and weight gain (phase 2a). At phase 2b we have described these three phenomena at multiple scales. When describing 'high stress' and 'night-time waking' at a biological scale, we may notice these both of these phenomena commonly involve some kind of dysregulation of the cortisol system. Investigating this link further – though literature review or empirical investigation – we may discover that this issue with the cortisol system is plausibly underlying both phenomena. This then is an explanation of the relationship between stress and night-time waking by noting *constitutive overlap* – i.e. both phenomena are underpinned by the same mechanism, rather than the relationship being causal. Comparatively, when looking at our collected descriptions of weight-gain and stress we may note that a common reaction to stress is 'stress-snacking' which seems to provide temporary relief but also weight-gain. Weight-gain in turn is often associated with fear of negative social evaluation, plausibly increasing stress. In explaining this relation then we may propose a mutually reinforcing *causal* relationship. In actual practice explanations would be more detailed and rigorous than in this example.

Evaluate (3b). The task at this final stage is to evaluate the explanations generated at phase 3a. There may well be multiple possible explanations for each relational phenomenon, so the job here is to choose the best ones. This selection should be made on the basis of the competing explanations epistemic values. Epistemic values are qualities of explanations that we value because they make the explanations more likely to be accurate (Haig, 2014). Epistemic values include: *external coherence* (whether the explanation fits well with our other systems of knowledge, e.g. biological plausibility), *internal coherence* (that the explanations postulated don't conflict with

each other or their own internal postulations), and *parsimony* (which can be thought of as simplicity divided by the scope of the explanation).

As part of the evaluation process investigators are free to cycle back to an earlier phase (see the blue arrows on figure 4). By making different choices along the way, and then comparing resulting problem spaces, descriptions, or explanations, this allows for continually refinement of the outputs at each stage. Returning to Phase 1 allows for refining of the problem space. This may include the removal or addition of clinical phenomena, the merging or splitting of phenomena, or even the splitting or lumping of entire problem spaces as evidence emerges. For example, as evidence is uncovered investigators may decide that it is better to split the classic depression rated phenomenon of 'anhedonia' into separate phenomena (e.g. avolition towards pleasurable activities, diminished experience of pleasure when the activities are engaged in, reduced focus on pleasure when remembering activities that were enjoyed at the time). Alternatively, if initially splitting the phenomena up this way, investigators may decide these phenomena occur so regularly together that it is better to think of them as one phenomenon; see Ward and Clack (2019a, 2019b) for further discussion of this example.

Returning to phase 2b or 3a will produce different explanations of the same PC which can then be compared. Repeatedly cycling back to 2a and selecting a different set of phenomena will (over time, and different research groups) produce a network of models that explain overlapping PCs. An important idea here is that these overlapping explanations will eventually populate the problem space with a rich and distributed understanding of the relationships between the clinical phenomena that constitute the disorder under investigation.

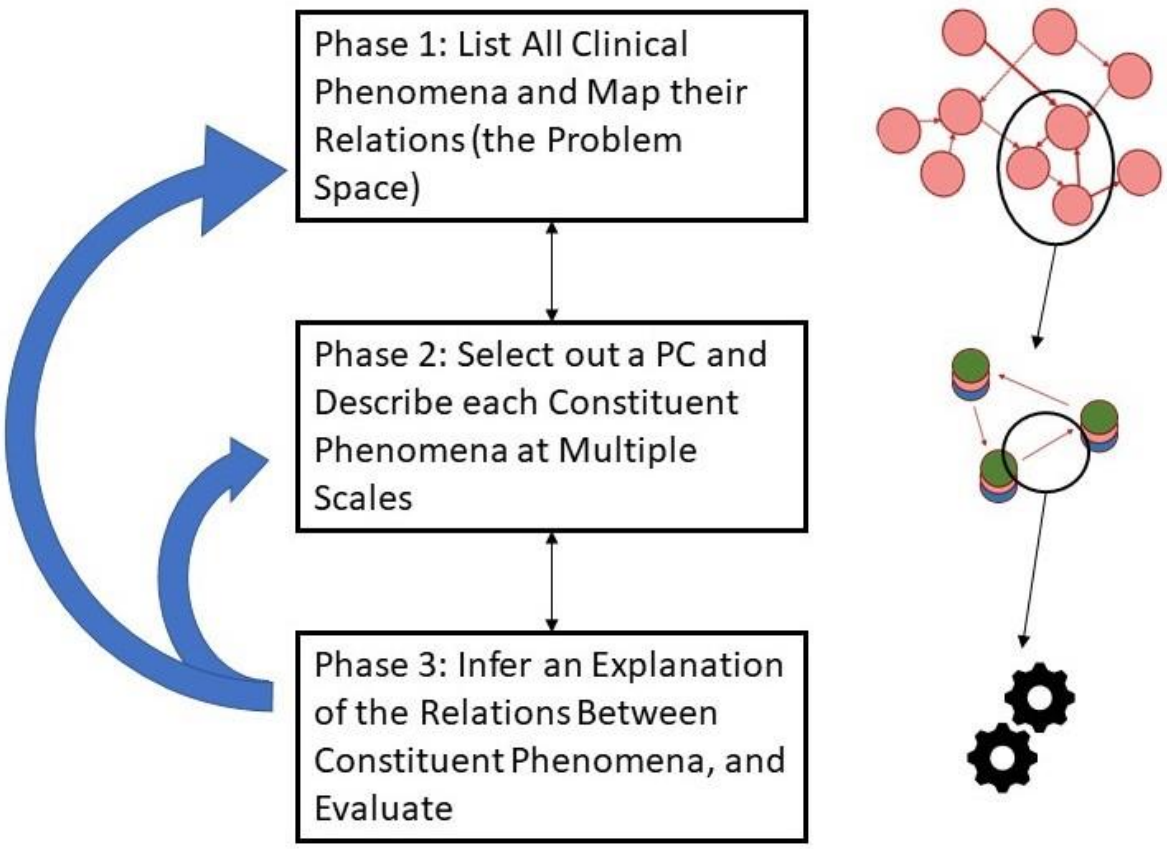

Figure 4. Schematic Representation of the RAP Process. The images on the right represent the output at each phase. Note the blue arrows, representing how RAP allows for refinement of the output at each phase (e.g. remapping the problem space, lumping or splitting of phenomena, adding or removing descriptive models of each phenomena, choosing a different set of phenomena within the PC). Returning to select different phenomena to form the PC should over time produce and understanding of the constitutive structure of the problem space.

Summary.

To summarize the RAP, there are three over all phases (see figure 2). First investigators list all the behavioral and phenomenological (i.e. clinical) phenomena present within the problem space and map out their relations. Second investigators artificially select two to four phenomena and idealize them as a small system which I have referred to as a *phenomena complex* (PC). Each constituent clinical phenomenon

is then richly described using a selection of friendly models from across different scales of analysis. Thirdly, investigators infer the nature of the relations observed between the constituent phenomena (i.e. the relational phenomena), before evaluating these explanations. At any stage necessary, investigators are free to return to an earlier phase to improve and refine their earlier explanations, descriptions, or mapping of the problem space. As a corpus of explanations of PC structures develop, this knowledge will represent greater understanding of the internal causal structure of the problem space – i.e. a distributed model of the stable dynamic pattern in the brain-body-environment system that supports continued engagement with the relevant dysfunctional behaviors.

Limitations and counter-arguments.

Generalizability. The RAP is grounded in a particular view of what mental disorders are (i.e. complex multi-scale entities with fuzzy boundaries and circular process structures). This aligns with the view argued for in previous chapters but may also represent a potential limitation of the approach. Just as we have different types of mental disorders (depression, anxiety, schizophrenia etc.), there are likely different kinds of mental disorders – e.g. organic diseases, socially constructed disorders, dysfunctional extremes on a normally distributed trait, etc. (Haslam, 2002, 2014; Kendler et al., 2011). All of the conceptual models presented across this thesis, including the embodied enactive model developed, are likely more or less relevant for different kinds of mental disorder. It is therefore likely that the RAP is a 'better fit' for explaining certain kinds of mental disorders than others. In particular, one may have a concern that the RAP is not a good fit for explaining disorders that turn out to be organic diseases of the brain (for example schizophrenia is often assumed to be such a 'disease'). The concern is that the causal work seems to be predominantly occurring inside the brain and that this may not sit as well with the RAP approach, because the descriptive phases are oriented around the behavioral and symptomatic scales.

I have, however, presented the RAP as an idea worth pursuing across psychopathology. I have done this for multiple reasons. Firstly, even the so-called 'organic' disorders such as schizophrenia are best seen as having a 'dappled nature', with some causal factors meaningfully clustered in the brain but others spanning across the brain-body-environment system (Kendler, 2012b). Secondly, in regard to the association

with dysfunction at the level of the person in their environment – which, on the current view, is essential to the disordered status – some degree of this circular process structure seems very likely to be present. Because of this, even if analysis at the scale of the RAP does not assist with uncovering the etiological causes of these more 'organic' disorders, it may still be useful in analyzing how these brain differences are associated with distress and dysfunction. Such a focus will thereby still be fruitful for the development of management strategies.

Is this pluralism sustainable? The externalism and multi-scale pluralism prescribed at phase 2b presents a potential challenge. It is simply not an achievable task to model a phenomenon at every conceivable scale from chemical through anatomical, to economic and cultural. The use of pluralistic modeling in this phase then, seems to require the instantiation of reasonable limits to make it sustainable. Without such limits it could be claimed that investigators will not know when to stop describing the constituent phenomena and when to move on to inferring an explanation of the PC. But which scales should investigators restrict themselves to? This is a reasonable concern and providing a definitive solution to this issue is an area of potential future development for the RAP framework.

Put simply the issue here is 'how much detail is required before moving to phase 3'? When presenting his distributed theory of mechanistic explanation from which we – as well as Ward and Clack (2019a) – draw this pluralistic model of phenomena description, Hochstein (2016) presents the limits of *non-redundancy* and *relevance*. Effectively, if an investigator is considering adding a model to her set of descriptors, she should make sure it makes a meaningful contribution to her understanding of the phenomenon; tracking novel difference makers to the occurrence or form of the explanandum (see Craver and Kaplan (2018) for further development of these limits).

Further to these limits, pragmatic guidance can be found in the work of Potochnik (2017), who highlights need to consider the wider purpose for seeking an explanation. Within the RAP the purpose of richly describing the constituent phenomena is to inspire the creative exercise of abductive inference to an explanation of the PC (phase 3a). The number of models required to do so, and which particular scales of analysis will be fruitful, will change with every investigation and investigator. While

the RAP represents the shift from phase 2b to phase 3 as a definitive step, the reality is that in practice this shift in the investigatory process will be iterative and gradual. Investigators are free to go backwards and forwards between phases 2b and 3 while exploring their theoretical speculations, adding to and removing descriptive models as required. The general suggestion then is to start with three or four different models, going back and adding more or removing them as required.

Conclusions and Summary

A vital element in any explanatory endeavor is the selection and depiction of the thing you are trying to explain. DSM syndromes are likely too heterogenous to serve this role. SNWMs have great potential in their ability to map the wider problem space that a disorder reflects, but seem too unstable and thin to act as good models of the explananda by themselves. Regarding the RDoC, I suggested that it will support the discovery of relevant (largely sub-personal) phenomena rather than explanations of disorder *in toto*. While knowledge of these phenomena will be vital, they are too distant to the disorder-as-a-whole to serve as ideal targets of explanation.

I have proposed the RAP as a *complimentary* method to these approaches. With its focus is on developing deep explanations of the relationships between clinical phenomena, the RAP is designed to fulfil a separate purpose to either the SNWM approach or the RDoC. The RAP is a meta-methodological framework that conceives of its targets as *phenomena complexes* (PCs), so named in reference to Hoche's (1991/1912) work within classification that argues for a similar shift to this middling level of complexity.

By focusing on PC structures, the RAP isolates explananda that are more manageable (and likely less heterogenous) than DSM syndromes or SNWMs, yet more directly relevant to the perpetuation of dysfunctional behaviour than RDoC derived targets. I suggested that RDoC is largely focused on uncovering dysfunctional neural mechanisms[98]. Comparatively, the RAP is focused on uncovering the wider mechanisms of disorder in people's lives. This makes the RAP a useful tool if researchers are

[98] There is some word play on mechanism here. In this sentence I am returning to the more restricted sense of (evolved/normal) mechanism. In the sentence following I am using mechanism in the more open minimal sense (Illari & Glennan, 2017).

interested in understanding the self-maintaining process structures of a mental disorder. If interest is in a different facet of mental disorder, then a different tool may be better suited.

Further to developing a method of explanation coherent with – and complimentary to the potential weaknesses of – the concept of mental disorder developed across earlier chapters, I have developed the RAP with an intent for it to be used by individual theoreticians, inter-disciplinary research teams, and potentially as a framework for encouraging co-ordination within the wider sciences of psychopathology. My hope is that, further to its use, my presenting of this framework will contribute to a dialogue concerning how to co-ordinate our investigatory and explanatory efforts (Sullivan, 2017). Targeting at this level should facilitate the timely development of explanations as to the maintenance of mental disorder, and with some luck, more efficacious treatments.

Chapter 8: Summing Up and Moving Forward

In this final chapter I will firstly explore another embodied enactive framework of mental disorder recently developed by Sanneke de Haan (in press-b, in press-a, 2017). Through comparing our positions, I demonstrate how the two frameworks are performing different work and point to areas of future development for an embodied enactive psychopathology. Following this I summarize the development of ideas across this thesis. Finally, I highlight some of the implications of the current framework for the tasks of classification and explanation, explore some further limitations to consider, and draw the thesis to a close.

Another Enactive Perspective

Toward the end of writing this thesis I became aware of the work of the philosopher Sanneke de Haan (in press-b, in press-a, 2017). De Haan has been working in the same area as myself; considering the nature of mental disorder through an embodied, embedded, and enactive lens. Her book 'Enactive Psychiatry' – an adaption of her PhD thesis – is due for release early 2020[99]. Her work represents an alternative – but seemingly not opposing – perspective on what mental disorders are through an embodied enactive lens[100]. In this section, I will briefly summarize her position through a comparison of our work. I will give some discussion to the similarities and differences between our views, as well as the differing merits of our perspectives. The intention in this section is to highlight points of difference between our views, in contribution to the ongoing development of an embodied enactive psychopathology.

In terms of her wider argument, de Haan (in press-b, in press-a) recognizes many of the same strengths in embodied enactivism as I have done. Specifically, she argues that embodied enactivism has huge potential in helping to address 'the integration

[99] I became aware of de Haan's work after writing the first three papers I published during the time-frame of this thesis, the first two of which were already accepted for publication. At the time of writing this (mid-2019) her 2017 paper in 'Mental Health, Religion & Culture' is the only published paper I am aware of that is in the area of developing an enactive framework of psychopathology. This 2017 paper is particularly focused on de Haan's notion of the 'existential aspect' of mental disorder, but does introduce her description of mental disorder as a biased pattern of sense-making. I regret not finding this paper sooner, however it is interesting to consider the similarities and differences between the views we have developed in parallel.

[100] Thank you very much to Dr. de Haan for providing an advanced draft of her upcoming book.

problem' in psychiatry. By 'the integration problem' de Haan refers most basically to the fact that we know of so many different causal factors at play in psychopathology – from social stressors to genetics – but have no clear way to consider how these factors come together to produce the disorders we recognize. De Haan's formulation of this 'integration problem' construes mental disorder as being composed of four dimensions. The first three of these – the experiential, physiological, and socio-cultural – basically relate to the three dimensions of the biopsychosocial model. The fourth dimension that de Haan considers she labels as the 'existential dimension' (2017). This existential dimension is important to de Haan's work and I will return to it later as her emphasis on this existential dimension represents an important difference between our perspectives. De Haan's (in press-b, in press-a) overarching argument is that an embodied enactive approach can integrate all four of these dimensions by viewing them as different aspects of the same complex whole, i.e., the organism standing in relation to its environment[101]. I see this argument as parallel to the central claim of this thesis – that embodied enactivism represents a useful framework of human functioning from which to consider mental disorder.

There are three key differences between de Haan's work and the framework presented in this thesis. Here I have used these differences to structure the following summary and discussion of de Haan's framework. Note that by 'differences' I do not necessarily mean points of conflict between the frameworks, rather it seems that these differences predominantly reflect differences in emphasis and choice of approach. These

[101] De Haan (in press-a) also reviews extant overarching frameworks and shows that all either fail to appropriately consider one of these four dimensions or fail to show how these dimensions may be integrated. Neuro-reductionist approaches arguably 'integrate' all four dimensions through the assumption that they are all ultimately caused by activity in the brain. De Haan argues against neuro-reductionism as such an assumption is ultimately un-evidenced and opens-up ethical concerns by de-emphasizing the psychological, socio-cultural, and existential dimensions. Evaluativist positions are shown to have a relative strength in that they recognize a role for values, however, they ultimately rest on a dualist worldview. De Haan points out that by assuming such a chasm between the natural world and the world of values, such an approach will always fail to offer a fully integrated view (this is similar to arguments made by Thornton (2000) that I reviewed in chapter two). The biopsychosocial model is also reviewed. As mentioned in chapter two this approach is explicitly integrative but is quite light on how exactly this integration should be achieved. De Haan points out the same concerns, and also notes that this approach fails to recognize her notion of the existential dimension. Finally, de Haan also considers the SNWM approach. Similar to my assessment in chapter seven, she sees SNWM as a tool for measuring the association between symptoms rather than presenting a novel concept/framework. Relatedly she criticizes SNWM for not providing a principled way of deciding what should be included in the network model and what shouldn't be.

differences are: 1) the central description of mental disorder as biases in sense making; 2) the normative conception of MD as Roschian rather than functional; 3) the notions of the existential dimension and existential (non-metabolic) values.

Biases in sense-making.

De Hann's (in press-b, in press-a) central claim is that mental disorders can be understood as *biases in sense-making*:

> "…[Mental disorders] refer to cases in which the evaluative interactions of a person with her world go astray. These interactions may include the person's thoughts, feelings, and/or behaviour – towards the world and/or to herself." (de Haan, in press-b, p. 234).

Remembering back to chapter four, 'sense-making' refers to an organism responding differentially to features of its world in accordance with the organism's purpose – to an organism enacting meaning in the world. For the enactivist, sense-making is therefore the defining act of cognition itself; to be a 'cognitive' system *is to be* a sense-making system. De Haan further specifies that the observed bias in sense-making has to be stable:

> "…a single instance of inadequate sense-making does not yet amount to a disorder. Psychiatric disorders refer to a more or less stable pattern in how someone's sense-making goes astray over time" (de Haan, in press-b, p. 234).

Structurally then, de Haan conceptualizes mental disorder as when a person's understanding of the world is significantly "…*biased in a specific direction:* the world appears overly threatening, or meaningless, or meaningful, or chaotic"(in press-b, p. 234).

This structural description of mental disorders in terms of sense-making is parsimonious and intuitively accurate from the enactivist position; to be 'minded' is to engage in sense-making, so to engage in dysfunctional/disordered sense-making is for the mind to be dysfunctional/disordered. This description, in a structural sense, also seems reasonably congruent to the concept of mental disorder described in this thesis. By this I mean that de Hann's description of mental disorders as *biases in sense-making*

could reasonably be placed alongside my description of mental disorder in chapter six. Remembering back, the description I gave was of mental disorders as dysfunctions in the behavioral and experiential processes of striving organisms, constituted by relatively stable dynamic patterns (/networks of phenomena) within the brain-body-environment system. This description highlights the disorder as conceptually separable from the agent, and as being composed of constituent and potentially isolatable phenomena/'parts'. De Hann's description in contrast, tries to capture mental disorder from the perspective of the sufferer as a holistic agent, an aspect that my description in retrospect did not accentuate. There seems to be no direct conflict however, between the descriptions in chapter six and the idea that mental disorders are biases in sense-making. Rather, the difference appears to be in the languages used and the emphasis these languages offer.

This is not to say that these differences in emphasis are not important. By describing disorder through the lens of the sense-making process, de Hann's description reminds us that disorder is occurring in the context of the person-as-a-whole and is, in a sense, inseparable from how the person understands the world and acts in it. In other words, by taking more of a 1st person perspective, de Hann's description adds greater emphasis to the *holistic*, *agential*, and *experiential* elements of mental disorder. This will be helpful as it encourages an empathetic stance – to richly consider the sufferers experience. In contrast, while still emphasizing the *holistic* and *agential,* the description I used in chapter six emphasizes the *mechanistic* and the partial *entitativity* of disorder; accentuating how we might disentangle pathological processes from the wider brain-body-environment system[102]. This emphasis provides more potential for the purposes of explanation; i.e., for managing the issue of unsustainable holism. The pragmatic mechanistic reductionism of the RAP for example, would seem an odd approach to take after hearing de Hann's description, but I hope makes sense following my descriptions. To put this in other words, the level of description used in the current framework seems to be placed at a more fertile level of abstraction for the task of explanation.

[102] These differences in emphasis may well reflect the differing interests of our professions. As a theoretical psychopathologist I am interested in 'breaking-down' and explaining mental disorders and my more mechanistic approach makes sense in light of this. As a phenomenologically informed philosopher de Haan was likely more motivated to capture the experiential aspects.

Ultimately, while these descriptions highlight different aspects of mental disorder through an embodied enactive lens, and thereby bring differing strengths, they appear to be different modes of description on a very similar thing. Continuing the enactive tradition of modeling life with life, let's return to the metaphor of a bacterium in its environment to demonstrate this similarity. Remember this environment contains sugar (the bacterium's food) and toxins (which degrade the bacterium in some way). Through an embodied mechanism, bacteria have a tendency to motivate toward the sugar and away from many toxins. If however, a bacterium started heading toward the toxins and away from the sugar then, under both de Haan's description and the view presented in this thesis, such action would represent a good analogy to mental disorder. De Haan's description of mental disorder as a bias in sense-making would prompt consideration of the bacterium as a whole system – how it's change in behaviour reflects a change in its relation to the world. Briefly scaling up to a human client, capturing this aspect is important because it will help inspire phenomenological consideration – how the clients very experience of the world is altered. The view espoused in this thesis however – while still recognizing that the bacterium is an irreducible whole – seems more likely to prompt consideration of which parts of the system seem particularly relevant to the dysfunctional change in behaviour, and to sit more comfortably alongside the idea that we may be able to develop an idealized model of the mechanisms underlying it. Our two perspectives then, while similar, perform different work. There are however, important differences in *why* this action might be considered disordered under each of our views and I will now shift to discussing these.

Demarcating pathology in sense-making.

Regarding the question of what counts as *sufficiently* biased sense-making so as to count as disorder, de Hann utilizes a Roschian/Wittgensteinian formulation where four general characteristics of pathological sense-making are listed but none are necessary or sufficient (in press-b). These four characteristics are that: 1) Pathological sense-making is often 'inappropriate' in the context ('appropriateness' is described as being assessable in contrast to current socio-cultural norms – i.e., does it conflict with 'common' sense?); 2) Pathological sense-making is often 'inflexible', i.e., the person acts the same way even in contexts when the action is not adaptive; 3) Pathological sense-

making often involves inflexible stance-taking, i.e., the person finds it very difficult to see/imagine things another way; 4) Pathological sense-making often results in suffering.

This Roschian-style approach, where not all of these characteristics are required for something to count as an instance of disordered sense-making, is obviously very different to the enriched systems-functionalism that I espoused in chapter five. The difficulty with the Roschian approach is that it does not paint a coherent picture of *what it means* for somebody's sense-making to be disordered. Rather, it only describes some of the characteristics common to those whose sense-making is disordered. Contrasting this Roschian approach with the approach espoused in this thesis, the current view provides a *central ideal* – a concept of what it means for someone's thoughts, behaviors, and/or emotions to be disordered versus not. On the enactive view, the mind is synonymous with the adaptive striving of the organism. The mind *is* the process of recognizing and responding to meaning in the world (i.e., sense-making), 'meaning' that ultimately derives from the purpose of the organism to self-maintain and adapt. Therefore, for the mind-organism to consistently act against this purpose *is* for it to be dysfunctional. A Roschian approach meanwhile provides no central concept like this, and it is left unclear why these characteristics should be privileged over other common features of disorder. At the same time, this discussion points to a strength of de Haan's (in press-b) Roschian approach within the normative domain. This is that her approach is a lot more practical than the current framework in its present state. By assessing for the presence of de Haan's four characteristics a clinician can make a reasonable estimate as to whether mental disorder is present or not. As explored at the end of chapter six, a limitation of the current framework concerns its ability to operationalize functionality in this way.

When discussing possible solutions regarding the challenge of operationalizing adaptivity, I suggested that one possibility would be to develop a *heuristic framework* that assesses functionality using proximal measures while acknowledging that it is self-maintenance and adaption that are being estimated. It is possible to understand de Haan's four characteristics of pathological sense-making as serving this role as a heuristic framework. On de Haan's current formulation it is unclear whether these characteristics are intended to play this descriptive/operationalized role (i.e., to provide

a way to approximately demarcate pathology), or are intended to be more intrinsic to the concept (i.e., to capture *what it means* for sense-making to be disordered). If they are intended to play the later more conceptual role – to define what it means for sense-making to be disordered – then we face the question 'why these four characteristics?'. If they are intended to play the more descriptive/operationalized role, we face a different question, that of what the characteristics are describing. In other words, how do we know what constitutes an improvement to these characteristics or not? Considering de Haan's view in isolation, it does not seem that this question is currently answered. On the view espoused in this thesis, the answer to this question concerns the degree to which the listed characteristics approximate the impact on self-maintenance and adaption. In other words, we return to the central question 'is this working for this person?'.

As our two frameworks currently stand therefore, they feature complimentary strengths within the normative domain. De Haan's framework offers more pragmatic guidance, while the current framework presents a clearer central ideal. Moving forward, the continued development of an embodied enactive approach to psychopathology seems to call for a synthesis of the strengths of these two approaches – i.e., continued refinement of the ideal concept, as well the continued refinement of something like de Haan's four characteristics in light of this central ideal.

An existential transformation vs. cultural embeddedness.

A final difference between the view developed in this thesis and de Haan's perspective concerns her underlying formulation of enactivism, particularly the role of biological functionality and how central this is to understanding human behaviour and structures of meaning. In her book, de Haan develops an 'enriched' or 'existentialised' version of enactivism (in press-b). De Haan sees the development of this existentialised enactivism as necessary because she argues that a more standard understanding of enactivism will struggle to capture the 'existential dimension' of mental disorder (and human experience more broadly). In this section I will unpack this notion of the 'existential dimension', as well as related concepts such as the 'existential/reflexive stance' and the role de Haan grants to 'existential values'. I will then briefly critique some aspects of these ideas and compare them to the relevant aspects of the current

framework. Without further dialogue it is difficult to know whether our differences in this domain are variances in emphasis, or whether they reflect more fundamental differences between our frameworks.

By the 'existential/reflexive stance', de Haan refers to the human capacity to recognize and take an evaluative stance upon the self:

> "The 'existential dimension' refers to the dimension that opens up due to the capacity to relate to our experiences. That is, we do not just experience things but we can also take stances on these experiences, on our ourselves and on our situation." (de Haan, 2017, p. 528).

For example, I can not only have a meaningful friendship with someone (a first-order enaction of meaning) but I can also consider whether I am a good friend, thus taking an evaluative stance upon myself and the meaning I experience (in a sense, a second-order enaction of meaning). The 'existential dimension of psychiatry' therefore refers to the way that sufferers of mental disorder understand and perceive themselves – how they consider their own existence. De Haan offers examples of this 'reflexive' stance operating in mental disorder, such as the fear-of-having-a-panic-attack that defines panic disorder, or the secondary feelings of guilt often associated with having depression (in press-a).

For de Haan (in press-b), this existential dimension is a fundamental component, not just of mental disorder, but of enactivism and the origin of human values. She makes a distinction between directly *metabolic values* (e.g., warmth, water, sociality) and *existential values*. By existential values de Haan refers "…to what motivates certain actions: actions that are not motivated by the drive to stay alive, but rather have to do with living a good, meaningful, or dignified life." (in press-b, p. 193). There is a division forged here between directly functional or metabolic values, and those that involve the existential stance – i.e., between what is good for you as an organism, versus reflecting upon one's self and how one wants to be. For de Haan, the ability to take this existential stance and thus develop existential values constitutes a qualitative shift from *organism* to *person*: "…we witness a *transformation* of the whole system from an organism-environment to a person-world system" (in press-b, p. 228). While de Haan emphasizes

that this 'transformation' is continuous and gradual, involving a complex interaction between individual and their socio-cultural world across development, she maintains that "existential beings do present a different form of life" (in press-b, p. 230).

There is much I agree with in this view. Existential values as de Haan describes them do seem to be importantly different to directly metabolic values. Existential values seem much more complex and this complexity does seem to, in part, concern the involvement of the human reflexive capacity. De Haan gives many examples of the existential stance at play in mental disorder such as the secondary effects of a diagnosis feeding back to alter a person's understanding of themselves (e.g., self-stigma and guilt), how existential values interact with choices concerning treatment (e.g. "I am not someone who takes psychoactive medication", "I am autonomous and don't need professional help"), or finally how most therapies are actually using our reflexive abilities to foster behavioral change (e.g., CBT encourages the client to consider alternative stances on the self and the situation, Mindfulness encourages a non-judgmental stance on one's own thoughts, ACT explicitly has clients consider what their values are and align their actions with them). These examples show that understanding the role of the existential stance in mental disorder and its treatment is important. Further, it is something that the view developed in this thesis does not explicitly work with. The fact that de Haan's framework incorporates consideration of these kinds of values is an absolute strength, as well as a strength relative compared to the current framework. The notion of an existential value will be an important tool going forward. There does however, seem to be room for improvement in this area.

For de Haan, the existential stance allows us to nigh on transcend biological functionality. Essentially, our capacity to take a stance – not just on the world but on ourselves and our relation to the world – is seen through its reflexive structure to open-up a level of autonomy that supersedes biological functioning. This also 'folds back', changing the nature of the entire system so that even our basic relation to the world (i.e., our sense-making) becomes 'existentialised':

> "Once you have become conscious of yourself as being visible to others, of the fact that others can see you and have a perspective on you and can evaluate you, there is no going back to oblivion" (de Haan, in press-b, p. 159).

For example, if my cat loves to eat my lasagna off the bench, then this concerns the direct relation of meaning between my cat and this food source (assuming here that my cat is not particularly reflexive or existential). If *I* love lasagna however, then the meaning relation is more complicated. My love of lasagna has ramifications for who I am; through loving lasagna I become 'someone who loves lasagna'. While this is a slightly silly example, it shows that there is a degree of face validity to this concept of 'existentialised' sense-making. At the very least, humans do seem very good at generating (sometimes overly-)complex structures of meaning.

However, I disagree with the idea of *transformation* within this space. The idea that there is a distinction of kind between basic sense-making and sense-making that involves an existential or reflexive stance seems somewhat in conflict with the central tenant of enactivism – that meaning is built upon precariousness and thus at least distantly rooted in biological functionality. Rather, I would sooner emphasize the continuity between these two 'forms' of sense-making by seeing the difference as one of increased complexity of the meaning-structure enacted and decreased immediacy of the relationship with biological functionality. De Haan is careful not to overstate her case here, she maintains for example that despite the level of autonomy introduced, existentialised sense-making remains thoroughly embodied (in press-b). She also rejects the label of total 'transcendence' from biological functionality, instead preferring to refer to a 'transformation of the system'. None-the-less, as described above, the autonomy that arises from the reflexive stance is seen to result in a qualitatively different kind of life form with a qualitatively different experiential life.

Through committing to the idea of an existential transformation, de Hann occasionally comes close to what could be labeled as *human exceptionalism*. For example, she states that:

> "[o]nly organisms capable of stance-taking, of being self-conscious, of relating to past and future, of evaluating themselves and others, of making moral judgements, of living a good life are vulnerable to psychiatric disorders" (de Haan, in press-b, p. 163).

While I agree that our human tendency to over-complicate things might be considered a vulnerability to certain mental illnesses, this stronger claim does not seem accurate. One only has to visit an animal shelter and observe the effects of trauma and abuse on an animal's mental wellbeing to question this assumption. If, through years of abuse, an animal has learned that human-beings are unpredictable and likely to bring it pain, then this learning represents a (previously adaptive) bias in its sense-making. Unlearning this response to humans is challenging (i.e., the sense-making is inflexible) and it maintains behaviors that push away people who are trying to help (e.g. biting a vet who is trying to treat an injury, thereby resulting in further suffering). Under both de Haan's formulation and my own, such an animal seems to qualify as having a mental disorder. Certainly, the existential stance has a greater role to play in human forms of mental distress, because thinking existentially is something that humans are very good at[103]. It is however not clear at all that existential thinking is a *requirement* for experiencing mental disorder. This tinge of human exceptionalism seems to stem from de Haan's underlying 'existential transformation' formulation of values.

An alternative to de Haan's existential transformation formulation, would be to view the emergence of more complex meaning structures – including existential values – as emerging from the complex relationship between individuals and a culture over time. In other words, to explain the emergence of complex human values through cultural embeddedness. I explored such an account in chapter five (see figure two). On this view, values are not at their core 'motivations' that emerge from our own understandings of ourselves as beings in the world, but rather are fundamentally *tendencies in action and experience* across a culture. Through basic associative learning and modeling processes, some tendencies in action and sense-making become part of the habitus of a culture over time, partially because they facilitate survival for the group and the self-maintenance and adaption of individuals within their environment (including their *fairing-well* in the socio-cultural context that said culture constitutes). There is presumably also a large degree of randomness and contingency at play here, a

[103] Arguably we are good at this because it is part of the niche we have constructed as a species – thinking existentially seems deeply and bi-directionally connected to our sociality. This is part of the reason I don't like the idea of an 'existential transformation' – it seems to separate us out from/raise us above other animals on the basis of something that we value because of the niche we fill.

sort of 'cultural drift' that arises from a tendency for people to model the behaviour of those around them, akin to the emergence of accents and dialects (although this modeling tendency itself seems likely functional). Over time, the culture recognizes some of these tendencies as something they value because it contributes to the cohesiveness/survival of the group and/or the success of individuals within that context, and so it is labeled as an explicit 'value' (e.g. being 'respectful', 'confident', 'fair'). This then begins a reification process by which the tendencies in action are explicitly taught and become imbued and entangled with idiosyncratic cultural markers (e.g., the idea that being 'respectful' involves standing up straight or having a firm handshake). Some people are labelled as more or less respectful, kind, honorable, or courageous. Seemingly, only once the pattern of behaviour is represented within the language of that culture can the 'value' itself become a motivation, because it is then recognized as something for individuals to aspire too. Before it is labeled, the value in question is merely an aspect of the implicit socio-cultural habitus – part of the culture's way-of-being and shared structures of meaning. As such, preceding their reification these tendencies in action and experience are certainly evaluative, but they are also habitual and immediate rather than necessarily concerning reflexivity.

This alternative formulation emphasizes the continuity between biological functionality and more complex socio-cultural values. While the capacity to take a reflexive stance is seen to play a role when value-labels act as motivations for human action, this occurs after the more implicit emergence of behavioral tendencies indicative of individual or cultural value/meaning systems that are more directly tied to the functionality of the individual within the socio-cultural environment. While there is much contingency and likely a degree of stochasticity underlying the emergence of tendencies in action within a particular habitus, a dominant 'selective pressure' on these tendencies in behaviour is that they do functional work for most people most of the time. On this view then, there is less transcendence and more continuity between complex meaning structures and biological functionality[104]. This seems to sit more

[104] There is a possibility here that the differences between de Haan's (in press-b)position and my own are again to do with the emphasis that our interests and professions bring. I am interested in 'explaining' the values and understanding their role in behaviour, I therefore have no major issue with evolutionary functionalist accounts of values. De Haan however discounts an evolutionary functionalist

comfortably within the wider embodied enactive perspective and avoid the potential for human exceptionalism[105].

One thing to stress for the purposes of clarity here is that, much like the current framework, de Haan (in press-b, in press-a) does not utilize values to answer the normative question of what counts as disorder, instead using her four Roschian characteristics for this purpose. While it may seem counter-intuitive, values cannot be used to demarcate pathology and de Haan's notion of existential values highlights this very well. Many values, despite their often-positive affective pull in terms of their phenomenology, may be tendencies in behaviour that are orthogonal to functioning, that once supported but no longer support functioning, or are even contrary to the functioning of the individual. Consider the valuing of the thin ideal in anorexia, being 'courageous' enough to die for your country, or the fuzzy boundary between valuing achievement and perfectionism. Simply put, values are not always good for you. Any particular value (e.g. materialism) may have been fostered within someone through a complex socio-cultural process (e.g. growing up in a capitalist society) and, much like the above examples, may not actually support the functioning of the individual in question. Values do not therefore seem like a tool with which we can answer the normative question 'what counts as mental disorder?'. It is for this reason that within the current framework mental disorder is defined using *functional norms* rather than values.

So, to summarize this third area of difference between our frameworks, de Haan (in press-b, in press-a) see's existential values as a vital part of the sense-making processes at play in mental disorders. In other words, she sees existential values having a role in the *structure* of disorder in that it is an existentialised sense-making process

---

account of values on the basis that it 'explains values away' and therefore fails to capture their felt importance/centrality of existential values in our lives. She also finds strength in an objectivist account of values on the basis that it aligns with this felt centrality. I think an enactive-functionalist account can satisfy both of these requirements by seeing 'meaning/value' as the proximal cause of action, without denying that the enaction of said meaning (and hence the action) may have its historical roots in biological functionality. On this account functionality doesn't seem to 'explain away' meaning or experience – values can still be 'relational realities' to use de Haan's phrase.

[105] This said there are current movements in the enactive field do explicitly make room for these 'higher levels' of autonomy; see Di Paolo, Cuffari, & De Jaegher (2018) on the idea of linguistic autonomy, or in their terminology how we are 'linguistic bodies'.

that is biased/altered. The current framework does not seem in conflict with this but does place less emphasis on existential values and reflexivity – seeing them as *important* but *not necessary* aspects of mental disorder. Underlying this difference is a deeper distinction between our understanding of enactivism and the emergence of values and meaning. De Haan's view emphasizes the human reflexive capacity and suggests this constitutes an existential transformation of the system – one from organism to person. My understanding instead emphasizes the complex relationship between individual and culture over time and sees the distinction between basic sense-making and complex and functionally-distal meaning structures as a difference of degree rather than kind. As stated at the start of this section, it is hard to know if these differences really reflect fundamental differences in our positions or whether they are differences in emphasis and wording. Either way, De Haan's concept of an existential value will likely be a very useful concept for considering the structure of mental disorders. Her observations concerning the role of reflexivity in mental disorder highlight an important aspect of mental disorder that is not captured well under the current framework. This represents a potential area of future development.

The value of different perspectives.

De Haan's (in press-b, in press-a) framework is an exciting step in the development of an embodied enactive concept of mental disorder. Her work shows, in concordance with the arguments of this thesis, that embodied enactivism has huge potential as a perspective from which to consider the nature of mental disorder. The three areas of difference highlighted here hopefully represent fruitful areas for continued development of an embodied enactive approach to psychopathology. A relative strength of de Haan's framework is the emphasis on the experiential aspects of mental disorder, including the relevance of existential values. A relative strength of the current approach concerns fertility for the task of explanation. By emphasizing the *mechanistic* and the partial *entitativity* of disorder the current framework is more suggestive of how we might disentangle pathological processes from the wider brain-body-environment system.

It is interesting to consider the degree to which the differences explored in this chapter reflect the differing interests of our professions. As a theoretical

psychopathologist I am interested in finding a justifiable way to 'break-down' and explain mental disorders, and my more mechanistic approach makes sense in light of this. As a phenomenologically informed philosopher on the other-hand, de Haan was likely more motivated to capture the experiential aspects of disorder. Her description of mental disorders as biases in sense-making and her emphasis on existential thought makes sense in light of this. In sum, our two frameworks appear to do different work, and in relation to this they emphasize different elements of mental disorder. Despite this the two frameworks seem to be largely compatible. Moving forward there is great potential for dialogue and debate and the continued refinement of our frameworks.

The Thesis in Brief

In chapter one I introduced the idea that the study of psychopathology can be broken down into component tasks and situated this thesis predominantly within the conceptual phase. In chapter two I overviewed a selection of extant conceptual models, breaking these down into those that make structural claims and those that propose a particular normative basis for using the label of disorder/dysfunction. I made two observations here. Firstly, that peoples' understanding of dysfunction/disorder is conceptually related to their understanding of human functioning. This raised the question of whether assuming a well-suited framework of human functioning might facilitate the development of a novel and fruitful conceptual model of mental disorder. The second observation, following Thornton (2000), was that using a non-reductionistic framework for this purpose may allow us to move beyond the evaluative-objectivist divide. In chapter three I reviewed both the DSM and RDoC, suggesting that both have issues regarding their underlying frameworks of human functioning. While both are well intentioned, the DSM was seen to be thin and non-committal in its descriptivism, and the RDoC was seen to be overly neurocentric/reductionistic, producing problems with conceptual validity and explanatory potential.

Beginning in chapter four I began to introduce embodied enactivism, a non-reductionistic understanding of human functioning that includes a naturalistic account of values and normativity. I suggested that, considering the structure of mental disorders from this perspective, mental disorders can be seen as stable dynamic patterns of causal relations within the brain-body-environment system. These causal structures

sustain repetitive patterns of behaviour, or tendencies in behaviour. But what makes such patterns disordered? In chapter five I took up this normative question, exploring how embodied enactivism contains, at its core, an understanding of normativity as arising from self-maintaining and adaptive systems. I showed that this understanding can be used to produce a richer form of functionalism where behaviour is considered disordered if it is working against the self-maintenance and adaption of the individual (including the fairing well of the individual in their socio-cultural context). I also explored here how embodied enactivism encourages consideration of how an individual's mode of functioning is socio-culturally informed, and how it offers a way to distinguish between *functional* norms and *statistical* socio-cultural norms. Relying on the former to demarcate disorder seems justifiable, as it means that diagnosis is offered in the interest of the individual. Relying on the later opens us up to anti-psychiatric critique by expanding the concept of mental disorder to include those whose mode of functioning differs from the standards of their society.

In chapter six I collapsed the normative and structural considerations explored in the previous two chapters and attempted to describe a more complete concept of mental disorder from an embodied enactive view. I suggested that mental disorders are dysfunctions in the behavioral and experiential processes of striving organisms, constituted by relatively stable dynamic patterns (/networks of phenomena) within the brain-body-environment system. I then specified these descriptions by rating this concept on Zachar and Kendler's (2007) conceptual taxonomy, highlighting the causal (but complex), agential (but still real), evaluative (but still objective), and externalist nature of the concept. At the end of this chapter I noted two challenges that will need to be overcome if such an embodied enactive concept of mental disorder is to realize its potential.

The first of these challenges was how to operationalize adaptivity. From an embodied enactive view, the diagnostic process is a lot more complicated than simply placing someone in a category based on their symptoms. Rather, diagnosis is seen as an evaluative process where the diagnostician is first asking 'is this pattern of behaviour working for this individual?'. I suggested that in the future development of these ideas, it may be useful to import or develop a framework that can scaffold this process. The

second challenge I noted was that of managing the holistic viewpoint that an embodied enactive perspective entails. The issue here is how to balance this holism with the need to understand the mechanisms at play in mental disorder. In chapter seven I attempted to provide a possible solution to this challenge in the form of the RAP, a meta-methodological framework for guiding the task of explanation. The RAP takes a pragmatic and mechanistically reductionistic approach, trying to keep in sight the system-wide and agential view that an embodied enactive perspective entails, while still providing a path to breaking-down and studying disorders as idealized entities. The ultimate aim of the RAP is to gradually reveal common mechanisms of dysfunction in people's lives.

Finally in the current chapter I have explored the work of Sanneke de Haan (in press-b, in press-a, 2017). Similarities and differences between our views were explored and some relative strengths of both views suggested. Through its greater emphasis on the mechanistic and partially entitative aspects of mental disorder the current framework was suggested to have more fertility for the task of explanation, as exampled by the RAP in chapter seven.

Disordered Eating as a Summary Example

Actual application of the ideas in this thesis to the development of an explanatory model is well beyond the constraints of this thesis. However, as a summary exemplar only, it is useful to consider how the embodied enactive approach here developed might reconfigure our conceptualization of disordered eating in comparison to other positions explored. Of particular interest is the relation between these conceptualizations and the kind of explanatory attempts we might take when holding these positions. A selection of conceptual positions are explored in this way within table 2.

Table 2. How various conceptual positions may relate to conceptual and explanatory approuches in the study of anorexia nervosa.

Conceptual Position	**Conceptualisation of Anorexia**	**Congruent Explanatory Strategy**
Biological Essentialism	Anorexia is a lesion/difference in someones brain/biology that produces a pattern of self-starvation. No nessecary commitment to why this is a disorder (but often coupled with statistical or evolutionary functionalism).	Study the brain. Compare the brains of those who do and do not self-starve in an attempt to locate the lesion. Seek to understand how the lesion produces self-starvation. RDoC seems a viable approuch.
Psychological Essentialism	Anorexia is a lesion/difference in cognition that produces a pattern of self-starvation. No nessecary commitment to why this is a disorder.	Study people's thinking. Compare the cognitive processes of those who do and do not self-starve in an attempt to locate the lesion. Seek to understand how the lesion produces self-starvation.
Socio-cultural Realism (Natural or Discrete Kind)	Anorexia is a distinct pattern of self-starvation caused by the pressures of society (e.g., media representations, the thin-ideal). No nessecary commitment to why this is a disorder.	Study the social locations of those who self-starve. Compare to the social locations of those who don't. Infer the social pressures of relevance and seek to understand how they produce self-starvation.
Social Constructionism (Deflationary)	Anorexia is an unduly pathologising label given by society to those that self-starve.	There is potentially nothing to explain. Instead we need to question the institutions that

	Those captured by the label may not represent a meaningful group, may be expressing normal distress, or may be responding to problems in society.	are pathologising people, and offer non-medical support for those in distress.
Fuzzy MPC Kind	Anorexia is a pattern of self-starvation behaviour brought about and maintained by a fuzzy network of causal mechanisms, potentially spanning the brain, body, and enviroment. No nessecary commitment to why this is a disorder.	Multiple approuches needed (i.e., methodological pluralism). Intention is to identify and understand the causal mechanisms supporting the pattern of behaviour. The RAP seems a viable approuch.
Embodied Enactive View	Anorexia is an alteration in the adaptive and sense-making processes of the agent in the world, supporting a pattern of self-starvation behaviour. This alteration is constituted by and can be modeled as a fuzzy network of causal mechanisms, very likely spanning brain, body, and enviroment. This is a disorder because, by self-starving, the person is acting against their own functional norms as a self-maintaining system.	Multiple approuches needed (i.e., methodological pluralism). Intention is to identify and understand the causal mechanisms supporting dysfunction at the level of the agent adapting to their enviroment. First person and third person persectives will be useful (i.e., phenomenonlogy will play an important role). The RAP seems a viable approuch.

Key Implications

Across this thesis many implications of the view developed have been mentioned, but before closing it is worth explicitly highlighting those that are most important. As mentioned in chapter one, the most immediate implications pertain to the tasks of classification and explanation.

Classification.

The current view supports a gradual shift towards a causalist diagnostic system as our explanations of mental disorder develop. Mental disorders are causal structures in the world which we can come to understand and categorize on the basis of similarity. At the same time, under the current framework the task of classification is seen as a partially pragmatic endeavor. This is due to the current paucity of causal understanding regarding mental disorders, the presumed complexity of mental disorders, and the fact that – as per Zachar (2018) – diagnostic systems are inherently political documents. Diagnostic systems will always be subject to social needs and pressures in accordance with this, and as such will likely never represent a 'perfect' natural ontology within their domain (even if such a thing is possible within an ever-changing socio-cultural and technological environment). Researchers and practitioners who wish to take an embodied enactive approach should therefore be ever critical of the nosological system of their time, although this is hardly particular to the embodied enactive view.

What does pertain to a novel aspect of the embodied enactive view here developed is the exclusive use of functionality to answer the normative question (i.e., 'what counts as mental disorder?'). This formulation entails an exclusion of *alternative modes of functioning* from the category of 'mental disorder'. This is taken to be ethically advantageous. On the current view the label 'mental disorder' is not about being different, nor about failing to adhere to societies standards and values, rather it signifies that the person is displaying a pattern of behaviour that is not working for them as an organism in their context. As an ideal at least, this concept thus avoids Foucauldian critique which holds mental disorder as a mode of social control/punishment for those that diverge from the norm. For example, as mentioned in chapter five, someone who displays a schizoid phenotype is not necessarily disordered under the developed

framework, rather we need to consider the possibility that such an individual is just functioning differently.

More importantly than this however, this exclusively functional formulation has flow on positive ethical effects for *cultural responsivity*. The normative formulation of the current framework, in being tied to functionality alone, removes reliance on statistical norms. This is advantageous because reliance on statistical norms can result in a pathologising of modes of functioning that develop within cultures underrepresented in the sample from which the norms are gathered. For example, NiaNia et al. (2016) report multiple case-studies in which young Māori who display 'symptoms' usually taken as indicative of psychosis are successfully 'treated' in a way that understands their 'symptoms' as culturally specific phenomena rather than as indicative of a recognized mental disorder. The framework developed in this thesis would perceive these cases as falling into two categories demarcated by functionality: culturally specific disorder, and non-pathological culturally specific phenomena/experiences which are causing distress. For both categories the framework developed here would suggest the possibility of finding a mode of functionality that works for the individual and/or reduces distress. By decoupling from the received view of 'mental disorder', NiaNia et al. take a very similar approach – finding paths to functionality and/or alleviation through a collaboration between traditional and psychiatric approaches. Such an approach would seem inconsistent with an underlying concept of mental disorder based on contrast to statistical normality, our limited understanding of evolutionary normality, or the values and norms of wider society, as such concepts would in practice entail a blindness to the unique socio-cultural milieu in which these young people's development was embedded.

Explanation.

Many of the implications of the developed framework for the task of explanation are formalized in the RAP (chapter seven). In summarizing the thesis however, it seems worthwhile briefly highlighting some of them here.

Two key points of difference between the embodied enactive concept developed and *status quo* approaches are the open commitments to moderate externalism and

anti-essentialism (Roberts et al., in press; Zachar & Kendler, 2007). On the current view, mental disorder pertains to the functional status of the relationship between the action of the organism and its environment. Further, the causal structures that support continued engagement with said dysfunctional action are seen to span brain, body, and environment. Mental disorders are therefore not merely 'in people' but between people (understood constitutionally) and the world they are embedded in. The ramifications for the task of explanation here are hard to overstate because, compared to an often assumed biological or psychological essentialist type view, the very nature of what we are seeking to explain is changed. There is an anchoring of the explananda to the scale of the individual acting in their environment, and the network of causal factors maintaining the dysfunctional behaviour is presumed to be disperse and complex. Further, the relationship between a token/ideographic instance of disorder and the kind/nomothetic disorder is presumed to be variable – kindship of an instance of mental disorder to a type of mental disorder (such as depression) is defined by similarity rather than sameness or causal lineage (i.e. structurally mental disorders are seen as fuzzy type-casual MPCs). This dramatically changes what the task of explanation will look like.

For a start, on this view we aren't just looking for one 'nugget of truth' which will explain a mental disorder. Rather than a moment of discovery like the myth of Newton and his apple, we would expect a more gradual process of knowledge gathering. This would be a process where researchers from across the globe slowly work to reveal the network of mechanisms that constitute the causal structure of a mental disorder. Instead of one paradigm defining discovery, coming to understand a mental disorder will probably be much more like a team of paleontologists slowly brushing away dirt to reveal a set of fossils, and developing theories about how all these bones fit together to form a complete dinosaur. Relatedly, instead of developing a single theory – e.g., the X theory of depression – we will likely need multiple explanations that each focus on different mechanisms in the network and how they operate. Rather than somebody developing a successful explanation of depression as a whole, we would instead expect smaller scale explanations to be developed mechanism by mechanism. As hypothetical examples, we might see theories emerging at a neurological-level that concern how

difficulties experiencing pleasure relate to difficulties sleeping, or at a psychological/ecological-level how changes that depressed people make to their environments may actually contribute to the perpetuation of their low mood[106]. Moreover, not all mechanisms uncovered may be relevant for all cases of a disorder. Due to the fuzzy nature of the kind concept, certain mechanisms may be playing a greater role in some instances of disorder than in others. The kindship relation maintains – and does practical work for the diagnostician seeking to understand an individual's dysfunction – due to the meaningful similarity between instances of a disorder.

This all leads to a rejection of neurocentric approaches such as RDoC and non-mechanistic approaches such as SNWM. While both approaches will likely serve distinct and important roles in the process of coming to understand mental disorders, neither seems appropriately targeted to produce an understanding of the mechanisms of dysfunction in people lives. Achieving this will likely require the coordination of multiple explainers using multiple methods, a task for which the RAP was explicitly designed.

Further Limitations

Further to the two key challenges already highlighted, there are limitations to the work done that bear considering as the thesis is drawn to a close. These limitations concern appropriate use of the model developed, how we will be able to falsify/evaluate the model, and how applicable the model is to the range of currently accepted mental disorders.

Appropriate use.

The work of this thesis is conceptual, and not explanatory. This was clearly stated at the outset but bears remembering. The concern here is that the framework developed may foster a sense of understanding when considering a particular mental disorder, even though we do not yet have a quality explanation for it. Consider the example of anxiety given in chapter six. By offering a description of pathological anxiety that incorporates many of the known causal factors into a model of the agent in the world,

[106] This last example is inspired by Krueger & Colombetti (2018)

the description may spark a feeling of understanding, in that the pattern of anxiety no longer seems surprising when all of these potential causal factors are highlighted. While there is an interesting question to be asked about whether causally heterogenous descriptions of mental disorders can do limited explanatory work – by highlighting *potential* causes and *ruling out* other causes, for discussion see Maung (2016) – we must remember that the 'causes' mentioned are only potential, insufficiently evidenced, and will certainly not apply in every token case of anxiety. Explicit explanatory models which postulate, on the basis of good evidence, how exactly constituent phenomena within particular disorders are related is the *next* step in the development of an embodied enactive psychopathology. The claim made in this thesis is that the conceptual framework here developed will be helpful for this development of explanatory models, not that the conceptual framework itself or examples given perform any meaningful explanatory work.

Falsifiability/explanatory value.

Given the conceptual nature of the work in this thesis, it essentially makes no empirically testable claims. This is potentially problematic because good science is falsifiable science. Thought must therefore be given to what would it mean for the framework developed in this thesis to be incorrect or useless. The answer to this problem is that, while the framework is not *directly* testable, it is *indirectly* falsifiable and open to evaluation through the explanatory models it fosters. If the framework helps produce valuable explanatory models of disorder and/or the relation between clinical phenomena within a disorder (i.e., models that stand up to empirical testing, that are parsimonious, that point to successful treatments) then this will constitute reason to believe that the wider framework is mapping on to reality in a useful way. If the framework instead facilitates the production of explanatory models that consistently fail to generate accurate predictions, that are overly complex, or that otherwise don't seem useful, then this will constitute reason to revise or disregard the framework.

Applicability.

The framework developed in this thesis presents a particular understanding of what mental disorder is. While it is argued that this perspective can provide an

interesting and useful way to think about all mental disorders, I do not want to make the dogmatic claim that this will be the *best* perspective to take for all mental disorders currently recognized. Rather, the way of thinking about mental disorder developed in this thesis may turn out to be more useful for some conditions (and purposes) than for others. For example, consider neuropsychological conditions such as Attention Deficit Hyper-Activity Disorder (ADHD). Perhaps ADHD is best seen as a difference in people's behavioral and attentional processes supported by a network of causal factors that span brain, body, and environment, all across development. On the other hand, if I am a neuroscientist trying to understand how the brains of those that experience ADHD differ from others, then it may actually be more useful to me if I approach ADHD as simply a brain disorder. Utilizing a simpler conceptual model may allow me to ignore environmental and developmental factors as extraneous and doing so may well facilitate my discovery of reliable differences (i.e., lesions) in the brains of those with ADHD or subtypes thereof. This would represent very useful information in our understanding of the condition. While the framework developed in this thesis seems likely to be useful across the study of most mental disorders, this does not mean that the framework will be the best option for all disorders and investigatory tasks.

Returning to our Starting Questions

At the closing of this thesis it is interesting to revisit the three general questions I listed in chapter one.

Are mental disorders something you get or something you do? In highlighting the agential nature of disorder, the embodied enactive view developed here sides with the idea that mental disorders are something people do. This of course not to suggest that mental disorders are entirely volitional, only that they concern a person acting in the world, the functionality of this action, and the flexibility with which action is altered when it is not serving the self-maintenance, adaption, and faring well of the organism within its environment.

Are mental disorders defined by brute facts or by social norms and values? On the developed view mental disorders are factual, yet still normative in a functional

sense. Deciding whether someone's behaviour constitutes a mental disorder concerns a thoroughgoing evaluation of the behaviour – of whether it is working for them.

Finally, does a mental disorder exist inside someone's brain or is it dispersed across their brain, body and environment? On the embodied enactive view the answer to this question is complex. Structurally mental disorders are complex networks of causal factors interacting within the brain-body-environment system. These structures are then defined as mental disorders on the basis of the functional/dysfunctional nature of the relation between the organism's environment and the organism's pattern of behaviour over time.

Conclusions

This thesis took as its central question 'what exactly is mental disorder?'. It was proposed that answers to this question depend on our fundamental assumptions about human functioning. As a set of assumptions seemingly fit for purpose, the position of embodied enactivism was explored and the nature of mental disorder from this perspective considered. The embodied enactive approach developed allows for the convergence of psychological, neuroscientific, and phenomenological perspectives around a central conception of mental disorder, without prejudice. The view presented: moves beyond the internalist bias of many current conceptual models, defines an ethically and scientifically justifiable role for normativity within the nature of disorder, encourages consideration of cultural and individual variance, does not unduly prioritize brain-level explanations of human behaviour, and can sit comfortably within a wholly natural world view.

There is of course much work to be done moving forward. Within development of the framework itself there is an unresolved limitation concerning the operationalization of adaption. As suggested, some form of heuristic framework to assess the impact of a pattern of behaviour on a person's adaptive fit to their environment is called for here. Outside of the conceptual work proper, an obvious next task in the development of these ideas is to try to use the RAP to develop an explanation of a phenomena complex, isolating and unpacking a selection of mechanisms within a recognized disorder. With

any luck this embodied enactive perspective on mental disorder will face critique on multiple fronts, fostering its continued conceptual development.

Just like any science, the study of psychopathology necessitates a partnership between explanatory theory and empirical investigation. But new explanations don't just come from nowhere. Rather, explanations emerge from a complex relationship between extant theory, discovery, and the conceptual framework in which the science is being done. This thesis was situated at this later conceptual level, considering the base assumptions at play within the sciences of psychopathology and attempting to reformulate them in line with the naturalistic principles of embodiment, embedment, and enaction. The conceptual framework developed represents one plausible alternative answer to the question 'what is mental disorder?'. If the sciences of psychopathology are to progress, we need to keep asking this question and refining our answers.

References

Agerbo, E., Sullivan, P. F., Vilhjálmsson, B. J., Pedersen, C. B., Mors, O., Børglum, A. D., Hougaard, D. M., Hollegaard, M. V., Meier, S., & Mattheisen, M. (2015). Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: A Danish population-based study and meta-analysis. *JAMA Psychiatry*, *72*(7), 635–641.

Albert, P. R., Benkelfat, C., & Descarries, L. (2012). *The neurobiology of depression—Revisiting the serotonin hypothesis. I. Cellular and molecular mechanisms*.

American Psychiatric Association. (1980). *Diagnostic and statistical manual, 3rd edn (DSM-III)* (3rd ed.).

American Psychiatric Association. (2013a). Anxiety Disorders. In *Diagnostic and Statistical Manual of Mental Disorders*. https://doi.org/10.1176/appi.books.9780890425596.dsm05

American Psychiatric Association. (2013b). *Diagnostic and Statistical Manual of Mental Disorders, 5th edn (DSM-5)* (5th ed.).

Andersen, H. (2014a). A field guide to mechanisms: Part I. *Philosophy Compass*, *9*(4), 274–283.

Andersen, H. (2014b). A field guide to mechanisms: Part II. *Philosophy Compass*, *9*(4), 284–293.

Andersen, H. (2016). BIOMEDICAL SCIENCES. *The Routledge Companion to Philosophy of Medicine*, 81.

Andrews, G., Slade, T., & Issakidis, C. (2002). Deconstructing current comorbidity: Data from the Australian National Survey of Mental Health and Well-being. *The British Journal of Psychiatry*, *181*(4), 306–314.

Banner, N. F. (2013). Mental disorders are not brain disorders. *Journal of Evaluation in Clinical Practice*, *19*(3), 509–513.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*(1), 19–21.

Beard, C., Millner, A. J., Forgeard, M. J., Fried, E. I., Hsu, K. J., Treadway, M., Leonard, C. V., Kertz, S., & Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine*, *46*(16), 3359–3369.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*(3), 295–318.

Bechtel, W. (2009a). Explanation: Mechanism, modularity, and situated cognition. *The Cambridge Handbook of Situated Cognition*, 155–170.

Bechtel, W. (2009b). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543–564.

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, *78*(4), 533–557.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A*, *41*(3), 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003

Beck, A. T., & Bredemeier, K. (2016). A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clinical Psychological Science*, *4*(4), 596–619.

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment*, *23*(4), 447–458.

Berenbaum, H. (2013). Classification and psychopathology research. *Journal of Abnormal Psychology*, *122*(3), 894.

Bergner, R. M. (1997). What is psychopathology? And so what? *Clinical Psychology: Science and Practice*, *4*(3), 235–248.

Bergner, R. M. (2004). An integrative framework for psychopathology and psychotherapy. *New Ideas in Psychology*, *22*(2), 127–141.

Bergner, R. M., & Bunford, N. (2017). Mental disorder is a disability concept, not a behavioral one. *Philosophy, Psychiatry, & Psychology*, *24*(1), 25–40.

Bhavsar, V., Boydell, J., Murray, R., & Power, P. (2014). Identifying aspects of neighbourhood deprivation associated with increased incidence of schizophrenia. *Schizophrenia Research*, *156*(1), 115–121.

Bingham, R., & Banner, N. (2014). The definition of mental disorder: Evolving but dysfunctional? *Journal of Medical Ethics*, *40*(8), 537–542.

Bird, A., & Tobin, E. (2018). *Natural Kinds*. The Stanford Encyclopedia of Philosophy.

Blampied, N. M. (2017). Analyzing therapeutic change using modified Brinley plots: History, construction, and interpretation. *Behavior Therapy*, *48*(1), 115–127.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303–352.

Booij, S. H., Wichers, M., De Jonge, P., Sytema, S., Van Os, J., Wunderink, L., & Wigman, J. T. (2018). Study protocol for a prospective cohort study examining the predictive potential of dynamic symptom networks for the onset and

progression of psychosis: The Mapping Individual Routes of Risk and Resilience (Mirorr) study. *BMJ Open*, *8*(1), e019059.

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

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science*, *44*(4), 542–573.

Boorse, C. (2014). A second rebuttal on health. *Journal of Medicine and Philosophy*, *39*(6), 683–724.

Borrell-Carrió, F., Suchman, A. L., & Epstein, R. M. (2004). The biopsychosocial model 25 years later: Principles, practice, and scientific inquiry. *The Annals of Family Medicine*, *2*(6), 576–582.

Borsboom, D., Cramer, A., & Kalis, A. (2018). Brain disorders? Not really… Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 1–54.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, *61*(1–2), 127–148.

Brigandt, I. (2013). Explanation in biology: Reduction, pluralism, and explanatory aims. *Science & Education*, *22*(1), 69–91.

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PloS One*, *8*(4), e60188.

Buchanan, J. (1995). Social support and schizophrenia: A review of the literature. *Archives of Psychiatric Nursing*, *9*(2), 68–76. https://doi.org/10.1016/S0883-9417(95)80003-4

Casey, B., Craddock, N., Cuthbert, B. N., Hyman, S. E., Lee, F. S., & Ressler, K. J. (2013). DSM-5 and RDoC: progress in psychiatry research? *Nature Reviews Neuroscience*, *14*(11), 810.

Chang, H. (2020). Pragmatism, Perspectivism, and the Historicity of Science. In M. Massimi & C. D. McCoy (Eds.), *Understanding Perspectivism* (pp. 10–27). Taylor & Francis.

Chang, H., & Shaw, D. S. (2016). The emergence of parent–child coercive processes in toddlerhood. *Child Psychiatry & Human Development*, *47*(2), 226–235.

Chapman, A. L., Gratz, K. L., & Brown, M. Z. (2006). Solving the puzzle of deliberate self-harm: The experiential avoidance model. *Behaviour Research and Therapy*, *44*(3), 371–394.

Christensen, W. D. (2012). Natural sources of normativity. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 104–112.

Christensen, W. D., & Bickhard, M. H. (2002). The process dynamics of normative function. *The Monist*, *85*(1), 3–28.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, *18*(2), 72–145.

Colombetti, G. (2014). The feeling body. *Cambridge (Mass.)*.

Contractor, A. A., Roley-Roberts, M. E., Lagdon, S., & Armour, C. (2017). Heterogeneity in patterns of DSM-5 posttraumatic stress disorder and depression symptoms: Latent profile analyses. *Journal of Affective Disorders*, *212*, 17–24.

Cooper, R. (2013a). Avoiding false positives: Zones of rarity, the threshold problem, and the DSM clinical significance criterion. *The Canadian Journal of Psychiatry*, *58*(11), 606–611.

Cooper, R. (2013b). *What's special about mental health and disorder?*

Cramer, A. O., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS One*, *11*(12), e0167490.

Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Complex realities require complex theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences*, *33*(2–3), 178–193.

Craver, C., & Kaplan, D. M. (2018). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *The British Journal for the Philosophy of Science*.

Cuthbert, B. N. (2014). The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, *13*(1), 28–35.

Cuthbert, B. N., & Insel, T. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 126.

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology*, *122*, 928–937.

Dablander, F., & Hinne, M. (2018). Centrality measures as a proxy for causal influence? A cautionary tale [Preprint]. *PsyArXiv*. *November 7th*. https://doi.org/10.31234/osf.io/nue4z

de Haan, S. (in press-a). An enactive approach to psychiatry. *Philosophy, Psychiatry and Psychology*.

de Haan, S. (in press-b). *Enactive Psychiatry* (draft). Cambridge University Press.

de Haan, S. (2017). The existential dimension in psychiatry: An enactive framework. *Mental Health, Religion & Culture*, *20*(6), 528–535.

de Haan, S., & Fuchs, T. (2010). The ghost in the machine: Disembodiment in schizophrenia–two case studies. *Psychopathology*, *43*(5), 327–333.

De Jaegher, H. (2013). Embodiment and sense-making in autism. *Frontiers in Integrative Neuroscience*, *7*.

Di Paolo, E. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452. https://doi.org/10.1007/s11097-005-9002-y

Di Paolo, E., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic Bodies: The Continuity between Life and Language*. MIT Press.

Di Paolo, E., Rohde, M., & De Jaegher, H. (2010). Horizons for the enactive mind: Values, social interaction, and play. *Enaction: Towards a New Paradigm for Cognitive Science*.

Dickinson, D., Pratt, D. N., Giangrande, E. J., Grunnagle, M., Orel, J., Weinberger, D. R., Callicott, J. H., & Berman, K. F. (2017). Attacking heterogeneity in schizophrenia by deriving clinical subgroups from widely available symptom data. *Schizophrenia Bulletin*, *44*(1), 101–113.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press. https://books.google.co.nz/books?id=LcFvKeOJRmgC

Doust, J., Walker, M. J., & Rogers, W. A. (2017). Current dilemmas in defining the boundaries of disease. *Journal of Medicine and Philosophy*, *42*(4), 350–366.

Drayson, Z. (2009). Embodied cognitive science and its implications for psychopathology. *Philosophy, Psychiatry, & Psychology*, *16*(4), 329–340.

Durt, C., Fuchs, T., & Tewes, C. (2017). *Embodiment, Enaction, and Culture: Investigating the Constitution of the Shared World*. MIT Press. https://books.google.co.nz/books?id=OJakDgAAQBAJ

Ebisch, S. J., & Gallese, V. (2015). A neuroscientific perspective on the nature of altered self-other relationships in schizophrenia. *Journal of Consciousness Studies*, *22*(1–2), 220–240.

Elbau, I. G., Binder, E. B., & Spoormaker, V. I. (2019). Symptoms are not the solution but the problem: Why psychiatric research should focus on processes rather than symptoms. *Behavioral and Brain Sciences*, *42*.

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, *196*(4286), 129–136.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927.

Fagan, M. B. (2015). Collaborative explanation and biological mechanisms. *Studies in History and Philosophy of Science Part A*, *52*, 67–78.

Falcon, A. (2015). Aristotle on Causality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015). Metaphysics Research Lab, Stanford University.

Ferenczi, E. A., Zalocusky, K. A., Liston, C., Grosenick, L., Warden, M. R., Amatya, D., Katovich, K., Mehta, H., Patenaude, B., & Ramakrishnan, C. (2016). Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior. *Science*, *351*(6268), aac9698.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology*, *126*(8), 1044.

Foster, J. A., & Neufeld, K.-A. M. (2013). Gut–brain axis: How the microbiome influences anxiety and depression. *Trends in Neurosciences*, *36*(5), 305–312.

Foucault, M. (2003). *Madness and civilization*. Routledge.

Franklin-Hall, L. R. (2016). New mechanistic explanation and the need for explanatory constraints. In *Scientific composition and metaphysical ground* (pp. 41–74). Springer.

Fried, E. I., & Cramer, A. O. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020.

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time… Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354.

Frisch, S. (2014). How cognitive neuroscience could be more biological—And what it might learn from clinical neuropsychology. *Frontiers in Human Neuroscience*, *8*, 541.

Fuchs, T. (2009). Embodied cognitive neuroscience and its consequences for psychiatry. *Poiesis & Praxis*, *6*(3–4), 219–233.

Fuchs, T. (2017). *Ecology of the Brain: The phenomenology and biology of the embodied mind*. Oxford University Press.

Fuchs, T., & Röhricht, F. (2017). Schizophrenia and intersubjectivity: An embodied and enactive approach to psychopathology and psychotherapy. *Philosophy, Psychiatry, & Psychology*, *24*(2), 127–142.

Fuchs, T., & Schlimme, J. E. (2009). Embodiment and psychopathology: A phenomenological perspective. *Current Opinion in Psychiatry*, *22*(6), 570–575.

Fulford, K. (2001). 'What is (mental) disease?': An open letter to Christopher Boorse. *Journal of Medical Ethics*, *27*(2), 80–85.

Fulford, K. (2002). Values in psychiatric diagnosis: Executive summary of a report to the chair of the ICD-12/DSM-VI Coordination Task Force (Dateline 2010). *Psychopathology*, *35*(2–3), 132–138.

Fulford, K., & Colombo, A. (2004). Six models of mental disorder: A study combining linguistic-analytic and empirical methods. *Philosophy, Psychiatry, & Psychology*, *11*(2), 129–144.

Fulford, K., Davies, M., Gipps, R., Graham, G., Sadler, J., Stanghellini, G., & Thornton, T. (2013). *The Oxford handbook of philosophy and psychiatry*. OUP Oxford.

Fulford, K., & Jackson, M. (1997). Spiritual experience and psychopathology. *Philosophy, Psychiatry, & Psychology*, *4*(1), 41–65.

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science*, *8*(6), 651–662.

Gallagher, S. (2006). *How the body shapes the mind*. Clarendon Press.

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford University Press. https://books.google.co.nz/books?id=Z28sDwAAQBAJ

Gardner, A., & Boles, R. G. (2011). Beyond the serotonin hypothesis: Mitochondria, inflammation and neurodegeneration in major depression and affective spectrum disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *35*(3), 730–743.

Garson, J. (2017). Mechanisms, phenomena, and functions. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 122–133). Routledge.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford Series in Cognitive Dev.

Ghaemi, S. N. (2009). The rise and fall of the biopsychosocial model. *The British Journal of Psychiatry*, *195*(1), 3–4.

Gibbs, Jr. R. W. (2005). *Embodiment and cognitive science*. Cambridge University Press.

Ginsborg, H. (2014). Kant's Aesthetics and Teleology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014). Metaphysics Research Lab, Stanford University.

Gleaves, D. H. (1996). The sociocognitive model of dissociative identity disorder: A reexamination of the evidence. *Psychological Bulletin*, *120*(1), 42.

Glennan, S., & Illari, P. (2017). *The Routledge handbook of mechanisms and mechanical philosophy*. Taylor & Francis.

Gould, S. J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues*, *47*(3), 43–65.

Graham, G. (2013). *The disordered mind: An introduction to philosophy of mind and mental illness*. Routledge.

Haig, B. D. (2014). *Investigating the Psychological World; Scientific Method in the Behavioural Sciences*. Massachusetts Institute of Technology.

Haig, B. D., & Vertue, F. M. (2010). Extending the network perspective on comorbidity. *Behavioral and Brain Sciences*, *33*(2–3), 158–158.

Hartner, D. F., & Theurer, K. L. (2018). Psychiatry should not seek mechanisms of disorder. *Journal of Theoretical and Philosophical Psychology*.

Harvey, M. I. (2015). Content in languaging: Why radical enactivism is incompatible with representational theories of language. *Language Sciences*, *48*, 90–129. https://doi.org/10.1016/j.langsci.2014.12.004

Haslam, N. (2002). Kinds of kinds: A conceptual taxonomy of psychiatric categories. *Philosophy, Psychiatry, & Psychology*, *9*(3), 203–217.

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*(1), 1–17.

Haslam, N., Holland, E., & Kuppens, P. (2012). Categories versus dimensions in personality and psychopathology: A quantitative review of taxometric research. *Psychological Medicine*, *42*(5), 903–920.

Hawkins-Elder, H., & Ward, T. (in press). Theory Construction in the Psychopathology Domain: A Multi-Phase Approuch. *Theory & Psychology*.

Hershenberg, R., & Goldfried, M. R. (2015). Implications of RDoC for the research and practice of psychotherapy. *Behavior Therapy*, *46*(2), 156–165.

Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.

Hoche, A. E. (1991). Die Bedeutung der Symptomenkomplexe in der Psychiatrie. *History of Psychiatry*, *2*(7), 334–343. https://doi.org/10.1177/0957154X9100200711

Hochstein, E. (2012). Minds, models and mechanisms: A new perspective on intentional psychology. *Journal of Experimental & Theoretical Artificial Intelligence*, *24*(4), 547–557.

Hochstein, E. (2013). Intentional models as essential scientific tools. *International Studies in the Philosophy of Science*, *27*(2), 199–217.

Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, *193*(5), 1387–1407.

Hochstein, E. (2019). How Metaphysical Commitments Shape the Study of Psychological Mechanisms. *Theory & Psychology*, *29*(5), 579–600. https://doi.org/10.1177/0959354319860591

Hoffman, G. A., & Zachar, P. (2017). RDoC's metaphysical assumptions: Problems and promises. *Extraordinary Science: Responding to the Crisis in Psychiatric Research*, 59–86.

Hucklenbroich, P. (2014). Medical criteria of pathologicity and their role in scientific psychiatry—Comments on the articles of Henrik Walter and Marco Stier. *Frontiers in Psychology*, *5*, 128.

Hume, D. (1978). A treatise of human nature [1739]. *British Moralists*, 1650–1800.

Humphry, S. M., & McGrane, J. A. (2010). Is there a contradiction between the network and latent variable perspectives? *Behavioral and Brain Sciences*, *33*(2–3), 160–161.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. Mit Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT Press.

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*, 155–179.

Illari, P., & Glennan, S. (2017). Varieties of mechanisms. In *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (pp. 109–121). Routledge.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). *Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders*.

Insel, T., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science*, *348*(6234), 499–500.

Jefferson, A. (2014). Mental disorders, brain disorders and values. *Frontiers in Psychology*, *5*, 130.

Johnstone, L., Boyle, M., Cromby, J., Dillon, J., Harper, D., & Kinderman, P. (2018). *The power threat meaning framework*. British Psychological Society.

Kaplan, B. J., Rucklidge, J. J., Romijn, A., & McLeod, K. (2015). The emerging field of nutritional mental health: Inflammation, the microbiome, oxidative stress, and mitochondrial function. *Clinical Psychological Science*, *3*(6), 964–980.

Kaplan, D. M. (2015). Moving parts: The natural alliance between dynamical and mechanistic modeling approaches. *Biology & Philosophy*, *30*(6), 757–786.

Karter, J. M., & Kamens, S. R. (2019). Toward Conceptual Competence in Psychiatric Diagnosis: An Ecological Model for Critiques of the DSM. In *Critical Psychiatry* (pp. 17–69). Springer.

Kendler, K. (2008). Explanatory models for psychiatric illness. *American Journal of Psychiatry*, *165*(6), 695–702.

Kendler, K. (2012a). Levels of explanation in psychiatric and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry*, *17*(1), 11.

Kendler, K. (2012b). The dappled nature of causes of psychiatric illness: Replacing the organic–functional/hardware–software dichotomy with empirically based pluralism. *Molecular Psychiatry*, *17*(4), 377.

Kendler, K. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5–12.

Kendler, K., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., Üstün, T. B., & Wang, P. S. (2009). The global burden of mental disorders: An update from the WHO World Mental Health (WMH) surveys. *Epidemiology and Psychiatric Sciences*, *18*(1), 23–33.

Khalidi, M. A. (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge University Press.

Kinderman, P. (2005). A psychological model of mental disorder. *Harvard Review of Psychiatry*, *13*(4), 206–217.

King, A. J., & Sumpter, D. J. T. (2012). Murmurations. *Current Biology*, *22*(4), 112–114.

Kingma, E. (2007). What is it to be healthy? *Analysis*, *67*(2), 128–133.

Kirchhoff, M. D. (2015). Extended Cognition & the Causal-Constitutive Fallacy: In Search for a Diachronic and Dynamical Conception of Constitution. *Philosophy and Phenomenological Research*, *90*(2), 320–360.

Kirk, S. A., Wakefield, J. C., Hsieh, D. K., & Pottick, K. J. (1999). Social context and social workers' judgment of mental disorder. *Social Service Review*, *73*(1), 82–104.

Kirmayer, L. J., & Crafa, D. (2014). What kind of science for psychiatry? *Frontiers in Human Neuroscience*, *8*, 435.

Krueger, J., & Colombetti, G. (2018). Affective affordances and psychopathology. In *Philosophical Perspectives on Affective Experience and Psychopathology: Vol. XXVIII–2* (pp. 221–247). Quodlibet.

Kvaale, E. P., Haslam, N., & Gottdiener, W. H. (2013). The 'side effects' of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clinical Psychology Review*, *33*(6), 782–794.

Kyselo, M. (2016). The enactive approach and disorders of the self-the case of schizophrenia. *Phenomenology and the Cognitive Sciences*, *15*(4), 591–616.

Labella, M. H., & Masten, A. S. (2018). Family influences on the development of aggression and violence. *Current Opinion in Psychology*, *19*, 11–16.

Larøi, F., Luhrmann, T. M., Bell, V., Christian Jr, W. A., Deshpande, S., Fernyhough, C., Jenkins, J., & Woods, A. (2014). Culture and hallucinations: Overview and future directions. *Schizophrenia Bulletin*, *40*(Suppl_4), S213–S220.

Lebowitz, M. S., & Appelbaum, P. S. (2019). Biomedical Explanations of Psychopathology and Their Implications for Attitudes and Beliefs About Mental Disorders. *Annual Review of Clinical Psychology*, *15*, 555–577.

Lee, H. (2012). *Biological Functionalism and Mental Disorder* [Dissertation]. Bowling Green State University.

Lilienfeld, S. O. (2014). The Research Domain Criteria (RDoC): An analysis of methodological and conceptual challenges. *Behaviour Research and Therapy*, *62*, 129–139.

Lilienfeld, S. O., & Marino, L. (1995). *Mental disorder as a Roschian concept: A critique of Wakefield's" harmful dysfunction" analysis*.

Lilienfeld, S. O., & Marino, L. (1999). *Essentialism revisited: Evolutionary theory and the concept of mental disorder*.

Lilienfeld, S. O., & Treadway, M. T. (2016). Clashing diagnostic approaches: DSM-ICD versus RDoC. *Annual Review of Clinical Psychology*, *12*, 435–463.

Lim, C., Barrio, C., Hernandez, M., Barragán, A., & Brekke, J. S. (2017). Recovery From Schizophrenia in Community-Based Psychosocial Rehabilitation Settings:Rates and Predictors. *Research on Social Work Practice*, *27*(5), 538–551. https://doi.org/10.1177/1049731515588597

Magnus, P. (2012). *Scientific enquiry and natural kinds: From planets to mallards*. Springer.

Magnus, P. (2014a). Epistemic categories and causal kinds. *Philosophy Faculty Scholarship*. https://doi.org/10.1016/j.shpsc.2014.10.001

Magnus, P. (2014b). NK≠ HPC. *The Philosophical Quarterly*, *64*(256), 471–477.

Maiese, M. (2016). *Embodied Selves and Divided Minds*. Oxford University Press. https://books.google.co.nz/books?id=w_quCgAAQBAJ

Maiese, M. (2017). Can the mind be embodied, enactive, affective, and extended? *Phenomenology and the Cognitive Sciences*, 1–19.

Mallon, R. (2016). *The construction of human kinds*. Oxford University Press.

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). *The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review*.

Maung, H. H. (2016). Diagnosis and causal explanation in psychiatry. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *60*, 15–24.

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104.

Megone, C. (1998). Aristotle's function argument and the concept of mental illness. *Philosophy, Psychiatry, & Psychology*, *5*(3), 187–201.

Molenaar, P. C. (2010). Latent variable models are network models. *Behavioral and Brain Sciences*, *33*(2–3), 166–166.

Monroe, S. M., & Anderson, S. F. (2015). Depression: The shroud of heterogeneity. *Current Directions in Psychological Science*, *24*(3), 227–231.

Morris, S. E., & Cuthbert, B. N. (2012). Research Domain Criteria: Cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience*, *14*(1), 29.

Muders, S. (2014). On the concept of the normative in the assessment of mental disorder. *Frontiers in Psychology*, *5*, 129.

Murphy, D. (2017). Can psychiatry refurbish the mind? *Philosophical Explorations*, *20*(2), 160–174.

Murphy, D., & Woolfolk, R. L. (2000). The harmful dysfunction analysis of mental disorder. *Philosophy, Psychiatry, & Psychology*, *7*(4), 241–252.

Nesse, R. M. (2001). On the difficulty of defining disease: A Darwinian perspective. *Medicine, Health Care and Philosophy*, *4*(1), 37–46.

NiaNia, W., Bush, A., & Epston, D. (2016). *Collaborative and Indigenous Mental Health Therapy: Tātaihono–Stories of Māori Healing and Psychiatry*. Taylor & Francis.

Nielsen, K., & Ward, T. (in press). Phenomena Complexes as Targets of Explanation in Psychopathology: The Relational Analysis of Phenomena (RAP) Approach. *Theory & Psychology*.

Nielsen, K., & Ward, T. (2018). Towards a New Conceptual Framework for Psychopathology: Embodiment, Enactivism and Embedment. *Theory & Psychology*, *8*(6), 800–822. https://doi.org/10.1177/0959354318808394

Nielsen, K., & Ward, T. (2019). Mental Disorder as both Natural and Normative: Developing the Normative Dimension of the 3e Conceptual Framework for Psychopathology. *Journal of Theoretical and Philosophical Psychology*. https://doi.org/10.1037/teo0000118

Nordenfelt, L. (2007). The concepts of health and illness revisited. *Medicine, Health Care and Philosophy*, *10*(1), 5.

O'Connor, B. (2017). Mental Disorder as a Practical Psychiatric Kind. *Philosophy, Psychiatry, & Psychology*, *24*(4), E-1-E-13.

Okrent, M. (2017). *Nature and Normativity: Biology, Teleology, and Meaning*. Routledge.

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology*, *123*(2), 452.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Ossorio, P. (1985). Pathology. In *Advances in Descriptive Psychology* (Vol. 4, pp. 151–202). JAI Press.

Parnas, J., & Sass, L. (2010). The spectrum of schizophrenia. *The Embodied Self*, 227–243.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Potochnik, A. (2010). Levels of explanation reconceived. *Philosophy of Science*, *77*(1), 59–72.

Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science*, *83*(5), 721–732.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science*, *79*(1), 120–140.

Price, J., Sloman, L., Gardner, R., Gilbert, P., & Rohde, P. (1994). The social competition hypothesis of depression. *The British Journal of Psychiatry*, *164*(3), 309–315.

Radden, J. (2006). *The philosophy of psychiatry: A companion*. Oxford University Press.

Ramstead, M. J. (2019). *Have we lost our minds? An approach to multiscale dynamics in the cognitive sciences* [PhD thesis]. McGill University.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: Beyond internalism and externalism. *Synthese*, 1–30.

Rietschel, M. (2014). Mental disorders are somatic disorders, a comment on M. Stier and T. Schramme. *Frontiers in Psychology*, *5*, 53.

Roberts, T., Krueger, J., & Glackin, S. (in press). Psychiatry beyond the brain: Externalism, mental health, and autistic spectrum disorder. *Philosophy, Psychiatry, & Psychology*.

Robinson, K., Garisch, J. A., Kingi, T., Brocklesby, M., O'Connell, A., Langlands, R. L., Russell, L., & Wilson, M. S. (2018). Reciprocal Risk: The Longitudinal Relationship between Emotion Regulation and Non-suicidal Self-Injury in Adolescents. *Journal of Abnormal Child Psychology*, 1–8.

Rucklidge, J. J., & Kaplan, B. J. (2013). Broad-spectrum micronutrient formulas for the treatment of psychiatric symptoms: A systematic review. *Expert Review of Neurotherapeutics*, *13*(1), 49–73.

Sadler, J. Z. (1999). *Horsefeathers: A commentary on" Evolutionary versus prototype analyses of the concept of disorder.".*

Sadler, J. Z. (2005). *Values and psychiatric diagnosis* (Vol. 2). Oxford University Press.

Sadler, J. Z., & Agich, G. J. (1995). Diseases, functions, values, and psychiatric classification. *Philosophy, Psychiatry, & Psychology*, *2*(3), 219–231.

Smith, J. D., Dishion, T. J., Shaw, D. S., Wilson, M. N., Winter, C. C., & Patterson, G. R. (2014). Coercive family process and early-onset conduct problems from age 2 to school entry. *Development and Psychopathology*, *26*(4pt1), 917–932.

Stein, D. J., Phillips, K. A., Bolton, D., Fulford, K., Sadler, J. Z., & Kendler, K. S. (2010). What is a mental/psychiatric disorder? From DSM-IV to DSM-V. *Psychological Medicine*, *40*(11), 1759–1765.

Stier, M. (2013). Normative preconditions for the assessment of mental disorder. *Frontiers in Psychology*, *4*, 611.

Sullivan, J. A. (2014). Stabilizing mental disorders: Prospects and problems. In J. A. Sullivan & H. Kincaid (Eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*. MIT Press.

Sullivan, J. A. (2017). Coordinated pluralism as a means to facilitate integrative taxonomies of cognition. *Philosophical Explorations*, *20*(2), 129–145.

Szasz, T. S. (1960). The myth of mental illness. *American Psychologist*, *15*(2), 113.

Szasz, T. S. (1963). *Law, liberty, and psychiatry: An inquiry into the social uses of mental health practices*. Syracuse University Press.

Szasz, T. S. (1974). *The myth of mental illness: Foundations of a theory of personal conduct, Rev*. Harper & Row.

Tabb, K. (2016). Philosophy of psychiatry after diagnostic kinds. *Synthese*, 1–19.

Tekin, S., & Bluhm, R. (2019). *The Bloomsbury Companion to Philosophy of Psychiatry*. Bloomsbury Publishing.

Telles-Correia, D., Saraiva, S., & Gonçalves, J. (2018). Mental Disorder—The Need for an Accurate Definition. *Frontiers in Psychiatry*, *9*, 64.

Thagard, P. (2017). *Natural philosophy: From Social Brains to Knowledge, Reality, Morality, and Beauty* (draft 3).

Thomas, J. G., & Sharp, P. B. (2019). Mechanistic science: A new approach to comprehensive psychopathology research that relates psychological and biological phenomena. *Clinical Psychological Science*, 2167702618810223.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press. https://books.google.co.nz/books?id=OVGna4ZEpWwC

Thompson, E., & Cosmelli, D. (2011). Brain in a vat or body in a world?: Brainbound versus enactive views of experience. *Philosophical Topics*, *39*(1), 163–180.

Thompson, E., & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, *28*(1), 23–30.

Thornton, T. (2000). Mental illness and reductionism: Can functions be naturalized? *Philosophy, Psychiatry, & Psychology*, *7*(1), 67–76.

Tsai, K.-Y., Chung, T.-C., Lee, C.-C., Chou, Y.-M., Su, C.-Y., Shen, S.-P., Lin, C.-H., & Chou, F. H.-C. (2014). Is low individual socioeconomic status (SES) in high-SES areas the same as low individual SES in low-SES areas: A 10-year follow-up schizophrenia study. *Social Psychiatry and Psychiatric Epidemiology*, *49*(1), 89–96.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind: Cognitive science and human experience*. MIT press.

Varga, S. (2011). Defining mental disorder. Exploring the'natural function'approach. *Philosophy, Ethics, and Humanities in Medicine*, *6*(1), 1.

Wakefield, J. C. (1992a). Disorder as harmful dysfunction: A conceptual critique of DSM-III-R's definition of mental disorder. *Psychological Review*, *99*(2), 232.

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

Wakefield, J. C. (1997a). Diagnosing DSM-IV—Part I: DSM-IV and the concept of disorder. *Behaviour Research and Therapy*, *35*(7), 633–649.

Wakefield, J. C. (1997b). Diagnosing DSM-IV—Part II: Eysenck (1986) and the essentialist fallacy. *Behaviour Research and Therapy*, *35*(7), 651–665.

Wakefield, J. C. (1997c). Normal inability versus pathological disability: Why Ossorio's definition of mental disorder is not sufficient. *Clinical Psychology: Science and Practice*, *4*(3), 249–258.

Wakefield, J. C. (1999a). Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology*, *108*(3), 374.

Wakefield, J. C. (1999b). *Mental disorder as a black box essentialist concept*.

Wakefield, J. C. (2000a). Aristotle as sociobiologist: The" function of a human being" argument, black box essentialism, and the concept of mental disorder. *Philosophy, Psychiatry, & Psychology*, *7*(1), 17–44.

Wakefield, J. C. (2000b). Spandrels, vestigial organs, and such: Reply to Murphy and Woolfolk's" The harmful dysfunction analysis of mental disorder". *Philosophy, Psychiatry, & Psychology*, *7*(4), 253–269.

Wakefield, J. C. (2007). The concept of mental disorder: Diagnostic implications of the harmful dysfunction analysis. *World Psychiatry*, *6*(3), 149.

Wakefield, J. C. (2013). The DSM-5 debate over the bereavement exclusion: Psychiatric diagnosis and the future of empirically supported treatment. *Clinical Psychology Review*, *33*(7), 825–845.

Wakefield, J. C. (2014a). Wittgenstein's nightmare: Why the RDoC grid needs a conceptual dimension. *World Psychiatry*, *13*(1), 38–40.

Wakefield, J. C. (2015). DSM-5, psychiatric epidemiology and the false positives problem. *Epidemiology and Psychiatric Sciences*, *24*(3), 188–196.

Wakefield, J. C. (2014b). *The biostatistical theory versus the harmful dysfunction analysis, part 1: Is part-dysfunction a sufficient condition for medical disorder? 39*, 648–682.

Walker, E. R., McGee, R. E., & Druss, B. G. (2015). Mortality in mental disorders and global disease burden implications: A systematic review and meta-analysis. *JAMA Psychiatry*, *72*(4), 334–341.

Walker, M. J., & Rogers, W. A. (2018). *A new approach to defining disease*. *43*, 402–420.

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, *36*(3), 365–375.

Ward, T., & Clack, S. (2019a). *From Symptom to Clinical Phenomena*. *54*, 40–49.

Ward, T., & Clack, S. (2019b). From symptoms of psychopathology to the explanation of clinical phenomena. *New Ideas in Psychology*, *54*, 40–49. https://doi.org/10.1016/j.newideapsych.2019.01.004

Ward, T., Clack, S., & Haig, B. D. (2016). The Abductive Theory of Method: Scientific Inquiry and Clinical Practice. *Behaviour Change*, *33*(4), 212–231.

Ward, T., & Fischer, R. (2019). Behavioral and Brain Sciences Commentary on Borsboom, Cramer and Kalis: Families of network structures—We need both phenomenal and explantory models. *Behavioral and Brain Sciences*, *42*(E31). https://doi.org/10.1017/S0140525X1800122X

Ward, T., & Maruna, S. (2007). *Rehabilitation*. Routledge.

Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V., & Vos, T. (2015). The global burden of mental, neurological and substance use disorders: An analysis from the Global Burden of Disease Study 2010. *PloS One*, *10*(2), e0116820.

Whooley, O. (2014). Nosological reflections: The failure of DSM-5, the emergence of RDoC, and the decontextualization of mental distress. *Society and Mental Health*, *4*(2), 92–110.

Wichers, M., Wigman, J. T., Bringmann, L. F., & de Jonge, P. (2017). Mental disorders as networks: Some cautionary reflections on a promising approach. *Social Psychiatry and Psychiatric Epidemiology*, *52*(2), 143–145.

World Health Organisation [WHO]. (2016). *Mental Disorders [Fact Sheet]*. http://www.who.int/mediacentre/factsheets/fs396/en/

Zachar, P. (2010). The abandonment of latent variables: Philosophical considerations. *Behavioral and Brain Sciences*, *33*(2–3), 177–178.

Zachar, P. (2014). *A metaphysics of psychopathology*. MIT Press.

Zachar, P. (2015). Psychiatric disorders: Natural kinds made by the world or practical kinds made by us? *World Psychiatry*, *14*(3), 288.

Zachar, P. (2018). Diagnostic Nomenclatures in the Mental Health Professions as Public Policy. *Journal of Humanistic Psychology*, 0022167818793002.

Zachar, P., & Kendler, K. S. (2007). Psychiatric disorders: A conceptual taxonomy. *American Journal of Psychiatry*, *164*(4), 557–565.

Zachar, P., & Kendler, K. S. (2017). The philosophy of nosology. *Annual Review of Clinical Psychology*, *13*, 49–71.

Zautra, N. (2015). Embodiment, interaction, and experience: Toward a comprehensive model in addiction science. *Philosophy of Science*, *82*(5), 1023–1034.