The eccentric manners of explanatory models: towards an account of perspectival mosaic unity in psychiatry

Josephine Lenssen
The eccentric manners of explanatory models: 
towards an account of perspectival mosaic unity in psychiatry

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy 
an de Vrije Universiteit Amsterdam, 
op gezag van de rector magnificus 
prof.dr. V. Subramaniam, 
in het openbaar te verdedigen 
ten overstaan van de promotiecommissie 
van de Faculteit der Geesteswetenschappen 
op donderdag 16 mei 2019 om 13.45 uur 
in de aula van de universiteit, 
De Boelelaan 1105

door

Josephine Maxine Lessen
geboren te Aberdeen, Verenigd Koninkrijk
Promotoren: prof.dr. G. Glas
prof.dr. H. W. de Regt

Mede mogelijk gemaakt door:
Table of Contents

1 Introduction 10

2 A review of the current literature on apathy 18
   2.1 Introduction 18
   2.2 Schizophrenia 18
   2.3 Definition of apathy and related terms 19
   2.4 Apathy as a symptom of schizophrenia 21
   2.5 Literature overview 24
   2.6 Preliminary conclusions 42

3 The biomedical model and its inherent reductionism 46
   3.1 The history of the biomedical model 46
   3.2 Limitations of the Biomedical Model 49

4 The Biopsychosocial Model 60
   4.1 Enter Engel: The Biopsychosocial Model 60
   4.2 Background and development of the biopsychosocial model 60
   4.3 The Biopsychosocial Model 64
   4.4 Application of the biopsychosocial model 64
   4.5 General systems theory 67
   4.6 The notion of science in the biopsychosocial model 71
   4.7 Concluding remarks: Biomedical versus Biopsychosocial 73

5 A critique of the biopsychosocial model 76
   5.1 Introduction 76
   5.2 Four levels of analysis 77
   5.3 Back to Engel on the aim of the biopsychosocial model 82
   5.4 The false dichotomy 85
   5.5 Intermezzo: Evidence Based Medicine 87
   5.6 How would Engel implement a general systems approach? 88
   5.7 Answer to the first central question of the thesis 90
   5.8 Definition of disease 91
   5.9 A missing dimension 92

6 Philosophical Models of Explanation 98
   6.1 Introduction 98
   6.2 How does an explanation explain? 99
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td>Constraints for a good explanation</td>
<td>100</td>
</tr>
<tr>
<td>6.4</td>
<td>Three models of scientific explanation</td>
<td>103</td>
</tr>
<tr>
<td>6.5</td>
<td>Covering Law Model</td>
<td>104</td>
</tr>
<tr>
<td>6.6</td>
<td>Representation Model</td>
<td>106</td>
</tr>
<tr>
<td>6.7</td>
<td>Unification Model</td>
<td>109</td>
</tr>
<tr>
<td>6.8</td>
<td>Summary</td>
<td>111</td>
</tr>
<tr>
<td>7</td>
<td>Mechanistic Explanation</td>
<td>116</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>116</td>
</tr>
<tr>
<td>7.2</td>
<td>A Little on Levels</td>
<td>117</td>
</tr>
<tr>
<td>7.3</td>
<td>Levels of science versus levels of organisation in Engel</td>
<td>118</td>
</tr>
<tr>
<td>7.4</td>
<td>Levels in the literature on apathy</td>
<td>120</td>
</tr>
<tr>
<td>7.5</td>
<td>Levels of mechanism</td>
<td>122</td>
</tr>
<tr>
<td>7.6</td>
<td>Mechanisms and mechanistic explanations</td>
<td>123</td>
</tr>
<tr>
<td>7.7</td>
<td>Comparison of different definitions of mechanism</td>
<td>125</td>
</tr>
<tr>
<td>7.8</td>
<td>Entities and activities that make up a mechanism</td>
<td>125</td>
</tr>
<tr>
<td>7.9</td>
<td>Dualist and non-dualist accounts of mechanism</td>
<td>129</td>
</tr>
<tr>
<td>7.10</td>
<td>Ontic and Epistemic implications</td>
<td>130</td>
</tr>
<tr>
<td>7.11</td>
<td>Mechanistic approach in current apathy research</td>
<td>135</td>
</tr>
<tr>
<td>7.12</td>
<td>The Mechanistic Model in psychiatry: testing Craver’s additional norms</td>
<td>139</td>
</tr>
<tr>
<td>7.13</td>
<td>Mental Mechanisms</td>
<td>145</td>
</tr>
<tr>
<td>7.14</td>
<td>How Now? Where we stand</td>
<td>148</td>
</tr>
<tr>
<td>8</td>
<td>Network Analysis</td>
<td>152</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>152</td>
</tr>
<tr>
<td>8.2</td>
<td>What is Network Analysis?</td>
<td>153</td>
</tr>
<tr>
<td>8.3</td>
<td>Psychometrics</td>
<td>156</td>
</tr>
<tr>
<td>8.4</td>
<td>Types of networks</td>
<td>163</td>
</tr>
<tr>
<td>8.5</td>
<td>Advantages and Disadvantages of using Network Analysis</td>
<td>167</td>
</tr>
<tr>
<td>8.6</td>
<td>Network analysis and Craver’s causal constraints</td>
<td>170</td>
</tr>
<tr>
<td>8.7</td>
<td>Discussion: causality and network analysis in the science of psychiatry</td>
<td>178</td>
</tr>
<tr>
<td>8.8</td>
<td>Do Network Analysis and Mechanistic Explanation converge?</td>
<td>180</td>
</tr>
<tr>
<td>8.9</td>
<td>Mechanistic Property Clusters: a hybrid approach?</td>
<td>181</td>
</tr>
<tr>
<td>8.10</td>
<td>Conclusion</td>
<td>186</td>
</tr>
<tr>
<td>9</td>
<td>Perspectival Mosaic Unity</td>
<td>190</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>190</td>
</tr>
<tr>
<td>9.2</td>
<td>Mosaic Unity</td>
<td>191</td>
</tr>
<tr>
<td>9.3</td>
<td>Six fundamental assumptions</td>
<td>194</td>
</tr>
<tr>
<td>9.4</td>
<td>The ontic-epistemic divide</td>
<td>201</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>9.5 Teleology in mechanistic explanation</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>9.6 Network Analysis and mechanistic explanation as complementary methods</td>
<td>214</td>
<td></td>
</tr>
<tr>
<td>9.7 Reformulating the project’s second central question</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>9.8 Answering the central questions that fuelled this project</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>9.9 Conclusion</td>
<td>228</td>
<td></td>
</tr>
<tr>
<td>Epilogue</td>
<td>232</td>
<td></td>
</tr>
<tr>
<td>Bibliography</td>
<td>238</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>252</td>
<td></td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>258</td>
<td></td>
</tr>
<tr>
<td>Curriculum Vitae</td>
<td>262</td>
<td></td>
</tr>
</tbody>
</table>
Chapter I

Introduction
I Introduction

When I commenced my philosophy studies during my third year of medical school, an aspect of philosophy that appealed to me in particular was that it made me more aware of my assumptions and offered me tools to question them. Many of the things that I had so far taken for granted turned out to be open to further questioning. Although philosophy opens the way to new questions, this does not entail that these questions have any straightforward answers. The foundations of much that I had taken for a fact turned out to be crumbly and susceptible to multiple different interpretations, each with potentially very different implications. Learning to put up with this newfound vagueness and bearing the inherent inconclusiveness has been a bewildering yet ultimately enriching process.

In many ways, studying philosophy helped to prepare me for becoming a psychiatrist. Developing a tolerance for “knowing that I don't know” especially, and not letting this get in the way of my ability to be there for my patients, has proved invaluable. Psychiatry, perhaps more so than other medical fields, is riddled with grey areas (which many psychiatrists of a reductionist-biological persuasion would like to explain in terms of grey matter!) and explanatory gaps. We have formulated labels for collections of co-occurring symptoms based on controversial classificatory structures. These labels are linked to protocols for the state-of-the-art treatment of what we take to be mental disorders. The mechanisms through which these treatments achieve their (often frustratingly limited) effects are largely a mystery to us.

When presented with someone who is suffering from paranoia, who has lost loved ones due to their substance abuse, who has been forced to abandon their studies and career path due to an inability to concentrate, it can be a great challenge to offer guidance despite the lack of knowledge of the causes of this hardship on the part of the clinician. As a clinician, I am inclined to feel responsible for reducing my patients’ torment, for instance through diminishing their symptoms and aiding their recovery in various life domains. Yet, the tools available to do so are few: many pills have disappointing results and substantial side-effects; psychotherapy requires a massive investment on the part of both patient and clinician. Nonetheless, it is extraordinarily important to provide people with a positive perspective. I think that offering a positive perspective requires not just technical medical skills and some understanding of the mechanisms that underlie symptoms. Instead, I propose that developing a rapport with each individual patient (and, where possible, with their loved ones too), is essential to my role as a medical professional. It affords me insight into the individual and his/her context. Such a working allegiance means we can not only sketch out a realistic perspective but also
work towards this in a manner that suits the patient’s needs and abilities.

As I noted above, my philosophical studies have helped me be more at ease with my lack of an understanding of psychiatric disorders. I feel this has benefitted my ability to juggle the responsibility to help people despite having few tools and nonetheless not let this hinder the authenticity that is required to develop a working allegiance with my patients.

In my roles as a medical scientist, a clinician and a philosopher, I am fascinated by the challenges of achieving a sense of unity or at least some level of coherence in dealing with the many facets of psychiatric suffering. At root, this is what fuelled my choice of a PhD topic. I found the toolkits with which I have been attempting to equip myself during my medical and philosophical studies to be complementary. However, I also noticed that few researchers in psychiatry draw on philosophical tools and that few philosophers have hands-on experience with psychiatric practice. I therefore set out to examine whether there are ways to combine these toolkits so that they can be mutually informative and hopefully allow us to further our knowledge and understanding of psychiatric disorders so that we can more effectively support those suffering from them. For these purposes, I chose to focus on the notion of explanatory models in psychiatry.

Explanatory models are abstract descriptions that propose how and why a phenomenon occurs. For phenomena in psychiatry, I have noticed that there are many different possible models for a single phenomenon. In an ideal situation, these models are complementary: they clarify different aspects of a phenomenon. Rather more often, the models are conflicting. They offer radically different mechanisms and it may be exceedingly difficult to determine which is the most useful in a particular context. I have often been entangled in the jungle of different explanatory models and felt ill-equipped to choose an appropriate path along which to proceed. I thus made it my project to examine different types of explanatory model, from both medicine and philosophy in more detail to see if this could provide any guidance.

Below, I outline the central questions which I set out to address and the contents of the chapters that follow. As I discuss in Chapter 9, my findings during the project lead me to reformulate the second of the central questions. The central questions with which I began, and to which I refer in the chapters to come, are:

I.) What is the utility of the biopsychosocial model at the level of a) philosophy, b) science and c) clinical practice?
II.) How can we understand the use of the biological, psychological and social levels of explanation in science and account for the coherence and possibilities for integration of these different levels of explanation?

III.) What are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy?

Chapter 2: in order to stay as close to psychiatric practice as possible, I survey a literature sample. Because a large proportion of the patients I encounter meet the diagnostic criteria for schizophrenia and because the negative symptoms are a huge source of misery, the literature sample zooms in on apathy in schizophrenia. I offer short introductions of schizophrenia, the negative symptom cluster and apathy in particular before outlining my search strategy. Ten articles that meet my search criteria are then briefly summarized with a view to detecting possible philosophical assumptions at play and evidence of biomedical or biopsychosocial thinking. I show how evidence of specific medical explanatory models is indirect and meagre at best.

In Chapter 3, the biomedical model is introduced. It is a model with origins in the 19th century and with even earlier roots that date back to Cartesian thinking. The biomedical model’s main limitation is its adherence to reductionist methods: the model seeks to explain diseases solely in terms of pathophysiological mechanisms and thus neglects other facets such as interpersonal and psychological aspects. The problems of reductionism are the theme of the remainder of the chapter. Different contexts in which the term features are examined and various types of reductionism are distinguished. I advocate continuing the search for a model that avoids reductionism and thus does not seek to explain by reducing all phenomena to occurrences at a particular level of organisation such as the biological level.

Chapter 4 introduces the biopsychosocial model. The model was developed by Engel and first presented in his 1977 article. Engel expands the model in many of his later publications and this literature is examined. I discuss how the biopsychosocial model aims to avoid the problems of biomedical thinking. The model is inspired by the doctrine of dynamical systems thinking and Engel himself refers to the writings of von Bertalanffy. I introduce this doctrine and some different variants of systems theory. The degree to which systems theory is incorporated in the biopsychosocial model to help integrate the different domains turns out to be somewhat limited. The chapter concludes with a short comparison of the biomedical and biopsychosocial models.

The fifth chapter delves into the shortcomings of the biopsychosocial model, which,
as I note is the model that is commonly taught to medical students. The first line of critique is that Engel does not differentiate between clinical and scientific practice – in fact, he deems this a false dichotomy. I instead embrace the notion of different roles and perspectives and show that differentiating between them is a heuristic tool that can help avoid confusion. On Engel's account, it is unclear whether the model is intended as a clinical or scientific model. A second objection is that Engel alludes to, but does not further develop, the ideas of systems theory in the biopsychosocial model. He offers little guidance on how to bring together the biological, psychological and social domains. Third, the model fails to help provide a definition of what disease is in the first place and risks inviting an over-inclusive interpretation of “disease”. Finally, I address the question of whether the biopsychosocial model should not include other dimensions, such as an existential dimension. In the concluding section, I address the first central question of the thesis, proposing that the biopsychosocial model is best regarded as a guide for clinical practice rather than a scientific model.

Having discussed the two most common medical models of explanation, Chapter 6 turns to more abstract, philosophical models of explanation. Key facets of explanation, such as description, prediction and intervention are discussed. I borrow heavily from the philosophers Wesley Salmon and Carl Craver by adopting their five constraints on a good explanation. Put simply, the five constraints are aimed at detecting causal relevance. Armed with these constraints, I discuss three philosophical models of scientific explanation: the Covering Law model, the Representation model and the Unification model. I offer examples from the field of psychiatry and show how these models fail to meet the five constraints. There are of course many more models but I focus on these three because a more exhaustive treatment is beyond the scope of my project. In the following two chapters, I discuss two further models in greater detail as they hold more promise for the field of psychiatry: mechanistic explanation and a statistical method called Network Analysis.

Chapter 7 offers an in depth exploration of mechanistic explanation. The chapter zooms in on the notion of levels, their appearance in the psychiatric literature and also on Craver's idea of “levels of mechanism”. Furthermore, I investigate what a mechanism is, showing that it consists of parts or entities and their interactions. The accounts of different mechanistic thinkers (e.g. Glennan, Machamer, Darden, Bechtel, Abrahamsen, and Richardson) are compared and the consequences of the differences are discussed for the field of psychiatry specifically. I revisit the apathy literature sample, searching for evidence of mechanistic thinking, and provide examples of proto-mechanisms. I take issue with some inconsistencies within Craver's account and also with the aspects of the accounts of mechanistic explanation that hinder the model's applicability in the
field of psychiatry. The most important problem for mechanistic explanation is the lack of robustly detectable, physiologically plausible parts in mechanisms for psychiatric phenomena. I suggest that there is an inherent fuzziness in psychiatry.

Network analysis, the topic of Chapter 8, is a statistical approach that entails using mathematical analyses to identify relations between phenomena. However, before any data can be represented in terms of a network, it must be collected. Psychometrics is the field that studies how mental phenomena may be measured. I argue for the importance of sound psychometrics, highlighting downfalls such as statisticism, the measurement problem and the idea of the latent variable. Next, three types of network that are most commonly used to model psychiatric phenomena are distinguished. In order to determine whether Network Analysis can be regarded as a model of scientific explanation, I investigate if the three types of network can meet the constraints for a sound explanation introduced in Chapter 6. I conclude that none of them can fully meet the requirements imposed by Craver and adopted by me. I include a discussion on Mechanistic Property Clusters, a hybrid between mechanistic thinking and statistical approaches. I then proceed to further develop the idea of a hybrid model for explanation that draws upon both Mechanistic Explanation and Network Analysis to explain a psychiatric phenomenon.

Chapter 9, titled “Perspectival Mosaic Unity” serves to bring together the findings from the previous chapters and attempts to answer the remaining questions fueling this project. I begin by offering an enactivism-inspired set of basic assumptions upon which I base my own attempt to offer a hybrid account of explanation that is suitable for psychiatry. Many dichotomies that are pervasive in traditional philosophy and also in scientific practice, are afforded the status of heuristic tools rather than absolute categories (for instance the ontic-epistemic divide and the subject-object divide). I argue that the perspectival approach, based on the inescapable reciprocity between self and world at which I thus arrive, runs deeper than the dichotomy-dependent notion embraced by Craver. I finally return to the central questions of the project. Given my findings, I adapt the phrasing of the second central question: How can we account for the integration of different perspectives on a target phenomenon? In answer to the question, I explore the idea of perspectival integration as a means to develop a more unified scientific landscape. I then move on to the third central question: what are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy? I draw on an example from the scientific literature to illustrate the implications of my refutal of biomedical and biopsychosocial thinking. I instead offer a post-dichotomous account of mechanistic explanation and supplement it with Network Analysis which is more suitable for the thorny issues presented by the field
of psychiatry. In this account, I suggest that the use of statistical methods can help determine which parts are relevant to a mechanism in the first place by using statistical methods to identify possible causal relations.

Finally, I conclude the chapter and my thesis with reflections on how my findings have affected my clinical practice and suggest that an adapted account of mechanistic thinking can be of help in formulating a descriptive diagnosis and making individualised treatment plans with patients.
Chapter 2

A review of the current literature on apathy
2 A review of the current literature on apathy

2.1 Introduction

This chapter provides an overview of some of the scientific literature on apathy in patients suffering from schizophrenia. Although apathy is a symptom that is common to various clinical syndromes, I choose to focus on apathy in schizophrenia to narrow down my treatment of the phenomenon and because it is such a common and as yet poorly understood symptom of schizophrenia. Attempts to further our understanding of this phenomenon span many different levels. The study of apathy will therefore serve as a basis for the treatment of the two main explanatory models that are pervasive in medicine: the biomedical and biopsychosocial models, and in later chapters, as a case study to test the validity of other promising models of explanation (Chapter 7 & 8).

This chapter begins with a brief introduction to schizophrenia and apathy. I quickly zoom in on apathy as a symptom of schizophrenia tracing it from its origins in the late nineteenth century to present. Next, the search strategy for my literature search is presented and the chapter continues with a treatment of each of the selected articles. In the literature, I search for evidence of biomedical and biopsychosocial thinking (the topics of the following two chapters) and point out commonly-used metaphors that carry implicit philosophical assumptions. Note that although there are many more philosophically complete models of explanation (which are discussed in Chapter 6), I choose to zoom in on the biomedical and biopsychosocial models only here. This is because these are the only two models of explanation with which most medical practitioners and researchers are familiar. Searching for evidence of other, perhaps more complete (traditional) philosophical models of explanation is not the goal of this chapter. However, in discussing new developments in the literature on models of explanation in biology and the neurosciences, I frequently draw on the literature discussed in this chapter.

2.2 Schizophrenia

Schizophrenia is a debilitating disorder that is thought to have a world-wide prevalence of around 1% and an incidence of around 1,5 per 10 000 people (McGrath, 2008). Symptoms can be divided into three major groups: psychotic or positive symptoms, negative symptoms and cognitive impairment (Mueser & McGurk, 2004; Foussias et al., 2015). Psychotic symptoms include hallucinations that can be tactile, gustatory, olfactory, visual and auditory. Patients may also have various types of delusions such as delusions of grandeur, somatic delusions, paranoid delusions or delusions of persecution.
A review of the current literature on apathy

and delusions of reference that lead patients to believe that things in their environment are communicating with them. Bizarre behaviour and loss of contact with reality are also psychotic symptoms. Negative symptoms include lack of emotion and interest in life, apathy, affective flattening, poverty of speech and social isolation, which is related to the inability to make or keep friends. Examples of cognitive symptoms are disorganized thinking, inability to understand, poor memory and concentration and difficulties expressing and integrating thoughts and feelings. Patients suffering from schizophrenia are often unable to hold down a job, attend school or see to their personal everyday needs. Close relationships with friends and relatives and parenting too, are frequently problematic.

2.3 Definition of apathy and related terms

The word *apathy* stems from the Greek *apatheia*, and is comprised of *a-* which means “without”, and “*pathos*” which is “suffering” or “feeling”. Whereas 2000 years ago, being apathetic was considered a virtue because extreme emotions were taken to inhibit clear rational thinking, nowadays, apathy is considered a deficit (Starkstein & Leentjens, 2008). It indicates a lack of interest, enthusiasm or concern. The neurological and neuropsychiatric literature frequently use terms such as “indifference” and “flat affect” as synonyms. Foussias et al. (2014), in their work on the negative symptoms of schizophrenia regard “apathy” as synonymous with “amotivation” and “avolition” (p. 696).

Whereas some authors refer to apathy as a symptom (for instance of schizophrenia), others refer to it as a syndrome in itself. Two frequently cited definitions of apathy are those by Fisher (1983) and Marin (1990). Fisher (1983) states that apathy is a symptom of a syndrome which he calls “abulia”: “abulia is suggested for the specific neurological syndrome comprising slowness, decreased responsiveness, apathy, etc.” (Fisher, 1983, p. 9). In contrast, Marin, in various publications from the early 1990s suggests that it should be regarded as a distinguishable neuropsychiatric syndrome with its own set of diagnostic criteria and methods of validation. Marin (1991) expands the definition of apathy stated earlier for clinical purposes, proposing that it entails “diminished motivation not attributable to a diminished level of consciousness, cognitive impairment, or emotional distress. In its differential diagnosis, abulia, akinesia and akinetic mutism, depression, dementia, delirium, despair, and demoralization must be ruled out” (p. 22).

Starkstein (2000) and in a later publication with Leentjens (Starkstein & Leentjens, 2008) indicate that apathy is a deficit state, which is associated with a lack of motivation, interest, and enthusiasm. It is characterized by a lack of initiative, a reduced will to work, and a decreased desire to maintain social relationships. The neurological and neuropsychiatric literature frequently use terms such as “indifference” and “flat affect” as synonyms. Foussias et al. (2014), in their work on the negative symptoms of schizophrenia regard “apathy” as synonymous with “amotivation” and “avolition” (p. 696).

1 http://www.oxforddictionaries.com/definition/english/apathy
2008) embraced the project of regarding apathy as more than just a symptom and further expanded the diagnostic criteria for apathy as a syndrome, suggesting A, B, C and D criteria to produce a set of criteria organized in a manner reminiscent of other DSM classifications:

“(A) Lack of motivation relative to the patient’s previous level of functioning or the standards of his or her age and culture as indicated either by subjective account or observation by others.

(B) Presence for at least 4 weeks during most of the day, of at least 1 symptom belonging to each of the following three domains:
   Diminished goal directed behaviour
   1. Lack of effort or energy to perform everyday activities.
   2. Dependency on prompts from others to structure everyday activities.

   Diminished goal directed cognition
   1. Lack of interest in learning new things, or in new experiences.
   2. Lack of concern about one’s personal problems.

   Diminished concomitants of goal directed behaviour
   1. Unchanging or flat affect
   2. Lack of emotional responsivity to positive or negative events.

(C) The symptoms cause clinically significant distress or impairment in social, occupational or other important areas of functioning.

(D) The symptoms are not due to diminished level of consciousness or the direct physiological effects of a substance.”
   (Starkstein & Leentjens 2008, adapted from Starkstein, 2000)

Stuss, van Reekum and Murphy (2000), argue against a definition of apathy in terms of amotivation. They offer two reasons for this proposal. First, motivation or “an absence of will” or “lack of an inner urge” is extraordinarily difficult to evaluate, for it requires third-person assessment of behaviour and affect. The “inner state” of the patient is then inferred from these observations. Second, the group of disorders of which apathy is taken to be a symptom is so heterogeneous and multi-facetted that it is very difficult to distil apathy from the broad range of other concurrent symptoms. The authors therefore suggest a narrow and behavioural definition of apathy in order to support the hypothesis that there are distinct kinds of apathy (to be categorized in terms of neuropsychological
A review of the current literature on apathy

The negative symptoms of schizophrenia as a psychiatric disorder are an example of “the common causes of apathy states” (2000, p. 343).

Furthermore, apathy features in the DSM-IV-TR as a symptom of many different diagnoses, including schizophrenia, schizophreniform disorder, major depressive disorder, dysthymic disorder, mental disorder due to a general medical condition, schizoid and schizotypal personality disorder, autistic disorder and post-traumatic stress disorder. It is frequently part of the A-criteria and is found amongst other symptoms such as affective flattening, alogia, avolition, apathetic immobility, social withdrawal, loss of pleasure and interest, and decreased activity, effectiveness or productivity (Clark et al., 2011). However, the DSM-IV-TR itself does not provide a clear definition of “apathy”. In sum, there are various definitions of “apathy” at play in the scientific literature. Some definitions refer to apathy as a symptom, others refer to it as a syndrome. Some definitions allow for interchangeability of “apathy” and terms such as “amotivation”, “flat affect” and “indifference”. Thus “apathy” is employed either as an umbrella-term or, alternatively, “apathy” is itself tucked under another umbrella-term, which renders it a circularly defined phenomenon. For Fisher, apathy is a symptom of abulia (i.e. under the umbrella) whereas for Marin, apathy and abulia are distinct (cf. Stuss, van Reekum & Murphy, 2000, p. 341) and amotivation in cognitive, affective and behavioural domains is a key feature of apathy as a syndrome (i.e. apathy is itself an umbrella-term) (Marin, 1996). In order to narrow down the study of apathy, I move swiftly on to apathy as a symptom of schizophrenia.

2.4 Apathy as a symptom of schizophrenia

The history of the use of the term “apathy” in the context of what we today call schizophrenia, can be traced back to the late nineteenth century. In this period, physicians strove to establish the field of psychiatry as a scientific discipline in order to free it of its association with spiritualism and mesmerism (Porter, 2002, pp. 183-185). Nowadays, we conceive of schizophrenia as entailing both “positive” and “negative” symptoms. The term “apathy” is frequently found in conjunction with other terms comprising these negative symptoms. The positive-negative symptom dichotomy was first introduced in the late nineteenth century by John Russell Reynolds, a neurologist who suggested a distinction between symptoms with a “loss of vital properties” as opposed to the positive symptoms which reflected an excess of “vital properties” (cf. Pearce, 2004). John Hughlings Jackson, a neurologist renowned for his work on epilepsy, further developed this distinction in the late 19th century in his writings on neurological disorder (1883). Hughlings Jackson proposed an anatomical and physiological hierarchy in which higher
centres such as the cortex, regulated lower centres. Negative symptoms, on his account, result from loss of the regulatory cortex function, whereas positive symptoms are due to disinhibition of lower centres (York & Steinberg, 2011).

At the very end of the nineteenth century, Emil Kraepelin, a German psychiatrist, popularized “dementia praecox”, a term coined by Arnold Pick, a Czech psychiatrist and neurologist in 1891. Like many of his contemporaries, Kraepelin worked within a predominantly biomedical paradigm, in which physicians and scientists sought to explain clinical phenomena in terms of biological causes. Kraepelin's methods included a longitudinal approach which involved describing his patients' life histories in great detail and carefully documenting the outcomes of their psychiatric disorders: he was more interested in the course of their symptoms than in focusing on all the symptoms at one moment in time only. He also pioneered the use of psychological tests on his patients. The Kraepelinian dichotomy refers to Kraepelin's distinction between dementia praecox, which involved “atrophy of the emotions” and “vitiation of the will” on one hand and manic-depressive psychoses (i.e. psychoses with a clear affective component) on the other (cf. Porter, 2002, p. 185). Dementia praecox is often regarded as the conceptual precursor to schizophrenia.

In the early 1900s, the Swiss psychiatrist Bleuler coined the term “schizophrenia”. He further developed the concept in his 1911 writings: “Dementia Praecox, oder Gruppe der Schizophrenien” (Dementia Praecox, or the Group of Schizophrenias). In his work, Bleuler brought together a biomedical approach and a Freudian approach by applying Freud's ideas to Kraepelin's work on dementia praecox. He sought to help complete the work of Kraepelin rather than refute it. Bleuler regarded the concept of dementia praecox valid but found its reference to dementia misleading. In coining “schizophrenia” Kraepelin aimed to emphasize his idea that a “split of personality” or “destruction of the inner coherence of the personality” as he called it, is the main symptom (Stotz-Ingenlath, 2000).

Soon, the terminology of positive and negative symptoms came to be applied to schizophrenia, in which, as mentioned earlier, amotivation, blunted or flattened affect, avolition, alogia, attentional deficit and anhedonia are regarded as negative symptoms (e.g. Andreasen & Olsen, 1982; Foussias & Remington, 2010; Limosin, 2014). Positive symptoms include hallucinations, delusions, some types of formal thought disorder, and bizarre behaviour (e.g. Andreasen & Olsen, 1982). The distinction is sometimes referred to as that between the florid (positive) and defective (negative) symptoms. Andreasen and Olsen (1982) conceive of “negative schizophrenia” as follows:

1. At least two of the following are present to a marked degree.
A review of the current literature on apathy

a. Alogia (e.g., marked poverty of speech, poverty of content of speech)
b. Affective flattening
c. Anhedonia-asociality (e.g., inability to experience pleasure or to feel intimacy, few social contacts)
d. Avolition-apathy (e.g., anergia, impersistence at work or school)
e. Attentional impairment

2. None of the following dominates the clinical picture or is present to a marked degree.
   a. Hallucinations
   b. Delusions
   c. Positive formal thought disorder
   d. Bizarre behaviour

   (Andreasen & Olsen, 1982, p. 790)

In more recent research, further attempts have been made to disentangle the various aspects of schizophrenia. The negative symptoms have thus been grouped into two subsets with apathy on one hand and diminished expression on the other (Blanchard & Cohen, 2006; Strauss et al., 2013; Kirkpatrick, 2014; Foussias et al., 2015; Hartman et al., 2015). Strauss et al. (2013) emphasize the “issue of heterogeneity of symptom presentation in schizophrenia” (p. 784) which is a drawback of factor analytic studies. Factor analytic studies are limited to analysing the clustering of symptom sub-domains and do not provide insight into how individual people group together based upon symptom profiles.

Note that the attempts to disentangle symptom sub-domains within the group of negative symptoms of schizophrenia coincides with a transition from a clinically driven approach to a data-driven approach. Whereas much of the above outline of apathy is primarily fuelled by the observation of clinical phenomena followed by further analysis of these phenomena, i.e. a top-down approach, the data-driven approach uses factor analytic studies in an attempt to demarcate separate sub-syndromes within the group of schizophrenia sufferers rendering it a bottom-up approach. This transition seems to express the point made by Strauss et al. (2013), that we do not (yet) have an adequate explanation for the heterogeneity of the group of people who meet the diagnostic criteria for schizophrenia. Factor analytic approaches provide a way to tackle the task of discerning sub-domain clusters. The next question is then whether these sub-domain clusters correspond to clinical subtypes or whether they provide insights that are relevant to prognosis, treatment etc.
In sum, the aim of this sub-section on apathy specifically as a symptom of schizophrenia was to provide a sketch of how these terms came into use. The section followed the evolution of the diagnosis “dementia praecox” into “schizophrenia” and traced the roots of the positive and negative symptom dichotomy and the place of apathy within that dichotomy. At a more abstract level, the section also briefly touched upon the changing contexts in which this terminology emerged, without at all delving into further explanations of how the context affected the way the terminology developed. This brought to light an oscillation between a predominantly biomedical model in the late nineteenth century and attempts to form a hybrid model from psychoanalytic and biomedical roots in the early 1900s and later back to a more biomedical (and perhaps later still, a biopsychosocial) model in the latter part of the twentieth century. Finally, I noted a transition from nosology derived from clinical observations to nosology derived from statistically derived symptom clusters. All in all, however, the section has not yielded a concise definition of “apathy” in the context of schizophrenia. Researchers continue to approach the phenomenon from different angles and have highlighted or excluded different aspects, including emotional, behavioural and cognitive aspects.

Having addressed apathy as a theoretical construct in the current section, the following section of this chapter is an overview of some of the scientific literature on apathy in schizophrenia as a clinical phenomenon. The research summarized below spans the biological (such as neurotransmitters and neuroanatomy), psychological (for instance performance on neuro-psychiatric tests) and social domains (including social functioning).

### 2.5 Literature overview

The goal of this section is twofold. As mentioned above, I seek to outline the current scientific literature on apathy and will provide short summaries of various (review) articles in which I attempt to stay true to the structure of the article summarized. This understanding of apathy will serve as a basis for the chapters to follow. Meanwhile, at a meta-level, I also reflect on the explanatory models within these articles. In order to get a sense of the often implicit explanatory models at play, I study the way in which the research hypotheses are embedded in the researchers’ current take on the subject matter, how they are stated, in which manner the methodology used to test hypotheses reflects the researchers’ philosophical assumptions and the way in which the empirical results are presented. For instance, examples of metaphors used to suggest a “mapping” (also a metaphor!) of the domain of neurobiology onto behaviour are pointed out. Spatial metaphors are amongst the most common metaphors in this context: phenomena at different levels of explanation are presented as related in a spatial manner. In fact,
the phrase “levels of explanation” itself harbours a spatial organisation, for a level is a horizontal plane in which something is situated. I return to the specifics of levels of explanation in a later chapter.

Furthermore, because this is very much an interpretative process in which particular statements in the articles may spark a range of associations in readers working in different paradigms, I provide ample citations to support my own interpretations. Although this discussion may at times bring philosophical inconsistencies to light, it is not my intention to single out and criticize particular articles. The overall aim here is constructive: to gain an understanding of the implicit explanatory models (and I limit the discussion here to two main models in medicine: the biomedical and biopsychosocial models) and more specifically the way in which different levels of explanation are related to each other by researchers at the cutting edge of science. The section is structured as an article by article discussion. This is a departure from the usual style of empirical reviews, in which articles are grouped together in terms of the themes or aspects of the phenomenon under study. Each article summary will be followed by a meta-analysis as discussed above.

2.5.1 Pubmed search strategy
A PubMed search on 18-02-2015 with the term “apathy” yielded 3326 hits whereas the search for “schizophrenia AND apathy” yielded 227 results, of which 31 were clinical trials, and 42 were labelled as reviews. A total of 117 of the articles have been published in the past 10 years, with 75 in the past 5 years. 58 articles had the search terms in either the title or the abstract. To narrow down the search for my current purposes, I began with the articles that were about humans and that had the search terms in either the title or the abstract. Many of the articles found through this strategy were not specifically about apathy in schizophrenia. Therefore, I instead found further relevant articles through the literature cited in the search results. Broadly speaking, the articles on apathy and schizophrenia, are articles that have one (or more) of the following goals:

1) Clarify the concept of apathy (within the context of negative symptoms of schizophrenia) in terms of aetiology and/or validation of methods for measuring it.
2) Treatment of apathy in schizophrenia based on clinical observations, patient experience or neurobiological hypotheses.
3) Compare patient experiences of apathy with clinical observations (1st-3rd person).
4) Compare clinical observations with brain imaging data (behavioural-functional anatomy).
5) Compare psychological test outcomes with brain imaging data (psychological-functional anatomy).
6) Compare performances on various psychological tests/tasks with clinical observations
(psychological-behavioural).

7) Compare patient and control groups in terms of brain-imaging data or performance on psychological tests (pathological-control groups on levels: psychological and biological).

Below, examples of these different types of articles are outlined briefly and implicit assumptions concerning the way in which the above-mentioned research domains are related to each other are highlighted.

2.5.2 Article I: Chue and Lalonde (2014). “Addressing the Unmet Needs of Patients with Persistent Negative Symptoms of Schizophrenia: Emerging Pharmacological Treatment Options”

Chue and Lalonde (2014) provide a review of pharmacological treatments of the negative symptoms of schizophrenia, which include apathy. Various classes of agents such as those that modulate glutamergic and cholinergic systems (through targeting NMDA receptor function, metabotropic glutameric receptor 2 function and nicotinic acetylcholine receptor function), psychostimulants, and anti-inflammatory drugs were reviewed as possible treatments for negative symptoms. They summarize the mixed, sometimes promising preliminary results of trials with these agents. Nicotinic acetylcholine receptor agonists appear to show the most potential based on the studies reviewed in their article. In their introduction on the “neurobiology of negative symptoms” the authors state: “The neurobiological underpinnings of negative symptoms are complex and continue to be the subject of discussion and research. While the basis of positive symptoms may be explained by dopaminergic models, the explanatory hypotheses of negative symptoms encompass a broader array of candidate neurotransmitters and associated pathways. There is no shared common understanding of the neurobiology of negative symptoms, which may eventually be revealed as too heterogeneous a concept to unite around a single biological hypothesis, as there is for positive symptoms. Although the cellular, molecular, and circuit alterations underlying these symptoms are not clear, studies focusing on motivation, reward, working memory, and goal-oriented behavior may help shed light on the circuitry that is altered in negative symptoms, [Dowd & Barch 2010; Gold et al. 2008] especially as these physiological functions correspond to the clinical dimension elucidated by factor analyses of negative symptoms discussed later.” (Chue and Lalonde, 2014, p. 778)

This excerpt shows that the authors take negative symptoms of schizophrenia to be expressions of processes at the neurobiological level. This neurobiological level and its “circuitry” “underpins” and “underlies” the clinical level and the authors lament that negative symptoms may be “too heterogeneous a concept to unite around a single
biological hypothesis” (Ibid., p. 778, my italics). This terminology suggests a spatial organization of levels in which the neurobiological level is below and somehow affects the level of clinically observable phenomena. In the final sentence of the excerpt, the authors switch to the term “correspond” to address the relationship between findings in the neurobiological and clinical domains. In other passages in the article, terms such as “correlate”, “contribute” “associate” or “involve” are used to bring together findings from domains addressed in their review, including neurobiology, (psychosocial) functioning and subjective variables. This “bringing together” can imply a causal relation (as suggested by the use of “contribute” and “involve”) or a more descriptive relation (as suggested by “correlate” and “associate”). Interestingly, these terms, in contrast to terms such as “underpin” and “underlie” do not imply a spatial organisation of levels of explanation. Although the authors do not explicitly specify the type of explanatory approach underlying their research, it would appear that they are working primarily from a biomedical paradigm, according to which clinical and subjective phenomena should all be reduced to neurobiological phenomena.

2.5.3 Article 2: Kirkpatrick (2014). “Recognizing Primary Vs Secondary Negative Symptoms and Apathy s Expression Domains”

Kirkpatrick (2014) gives an overview of the current definition of negative symptoms of schizophrenia. He states that the rating scales such as the “Scale for the Assessment of Negative Symptoms” (SANS) and “Positive and Negative Syndrome Scale” (PANSS) do not help clinicians to determine the cause of the negative symptoms, i.e. whether they are primary or secondary. Primary negative symptoms are those that form “an integral part” of schizophrenia for which there are not as yet any effective treatments; secondary negative symptoms are caused by factors such as positive symptoms, medication side effects, depression and substance abuse (Ibid., p. 4) and are “usually responsive to treatment of the underlying cause” (Ibid., p. 3). Furthermore, the rating scales fail to take into account more recent research in which a distinction between apathy and diminished expression has been made. This distinction may provide new research directions as the emergence of two groups could mean that these groups have separate risk factors, pathophysiology, prognosis and treatment targets. Kirkpatrick goes on to discuss early trial results from phase 2 trials in which the treatment response to galantamine (an acetylcholinesterase inhibitor), anabaseine (a partial nicotinic agonist) and bitopertin (a selective glycine reuptake inhibitor) differs between the two sub-domains of negative symptoms. Adjunctive galantamine improved symptoms on the alogia subscale of the SANS but did not differ from placebo for the total score

2 Phase II trial: the drug or treatment is given to a larger group of people to see if it is effective and to further evaluate its safety. (http://www.nlm.nih.gov/services/ctphases.html)
on the SANS. Anabaseine showed significant improvement on SANS total scores and improved scores on the alogia and anhedonia subscales. Bitopertin improved apathy and social deficits (with a lesser effect on expressive deficits) (Ibid., p.5). Kirkpatrick underlines the need for further research and replication of these results before drawing further conclusions from them.

Kirkpatrick's article is a publication that is part of a “Continuing Medical Education” (CME) activity and readers can obtain credits by reading it and answering subsequent questions. It argues that distinguishing between symptom domains in schizophrenia can aid the search for new treatment targets: “[t]o address this treatment gap, research into new treatment targets, such as the glutamate NMDA receptor, is currently underway. Clinicians need education in the recognition, identification, and assessment of the different symptom domains in schizophrenia to better promote optimal outcomes for their patients” (Ibid., p. 2). In this excerpt, the author implies a link between treatment targets at the neurobiological level and differentiation at the clinical level of symptom domains. The author does not offer any suggestions as to the nature of the relation between these two levels.

Later, the author describes the distinction between primary and secondary negative symptoms, but does not link this to other levels of explanation. He states that primary symptoms are an “integral part of schizophrenia”. But, given that schizophrenia is a descriptive diagnosis, defined by a set of clinically observable criteria, Kirkpatrick’s definition of primary negative symptoms as part of schizophrenia does not shed any light on what they are. Interestingly, the definition provided by Kirkpatrick is somewhat circular: primary negative symptoms are defined in terms of schizophrenia and schizophrenia is defined (in part) in terms of primary negative symptoms; i.e. they are primary in that they are not attributable to any other known cause such as substance abuse or depression as stated in de DSM-IV-TR criteria (D and E criteria) (cf. DSM-IV-TR (2000) 4th ed., text rev.).

In the article, Kirkpatrick makes no explicit claims about the way in which the levels (or types) of explanation on which he focuses are joined. He does imply that distinguishing between sub-groups of patients can aid the development of new drugs. He suggests exploring whether patients in the subgroups respond differently to pharmacotherapy and thus assumes a correlation between aetiological pathway (and using this as a target for new drugs) on one hand and clinical presentation on the other.

In sum, due to the brief and descriptive nature of the publication, it is not possible to extract much concrete information about the explanatory models that are at play.
Kirkpatrick refrains from committing, yet indirectly, he seems to adhere to a biomedical model which harbours the potential to display a one-to-one mapping of clinical symptoms onto processes in the brain.

2.5.4 Article 3: Limosin (2014). “Neurodevelopmental and Environmental Hypotheses of Negative Symptoms of Schizophrenia”

Limosin (2014) offers an epidemiological review of environmental and neurodevelopmental factors thought to play a role in (the development of) the negative symptoms of schizophrenia. In his treatment of the differences between positive and negative symptoms, the author posits that “negative symptoms should not be considered only as a unitary construct” (Limosin, 2014, p. 3). In the introduction, he states that environmental factors and neurodevelopmental factors cannot be considered in isolation. Limosin draws upon “integrated vulnerability models” and distinguishes between causal factors and vulnerability factors which “are considered to be predisposing disease factors (…) whose influence alone would be insufficient to trigger the onset of the disorder” (Limosin, 2014, p. 2). Furthermore, he adopts a broad notion of “environmental” which entails both psychosocial factors such as immigration, urban environment and disturbances in early parent/child relationships and biological factors, for instance toxicity from psychoactive agents such as cannabis, amphetamines and cocaine and infectious agents such as influenza and finally obstetrical complications such as pre-eclampsia, gestational diabetes and delivery complications.

Neurodevelopmental factors associated with negative symptoms are mainly extrapolated from animal models. Mice subjected to prenatal inflammatory responses in the final gestation period are more likely to show behavioural abnormalities and learning disabilities, social and working memory deficits and also anhedonic behaviour compared to controls. Male mice exposed to prenatal inflammatory responses in particular, lacked cognitive and behavioural flexibility (Limosin, 2014, p. 4). Maternal deprivation in rats has also been associated with a decreased ability to regulate the response to sensory information and cognitive disorders with impaired learning abilities.

The authors discuss evidence supporting the traumagenic neurodevelopmental model. The model is based on 1) the high rate of childhood abuse in patients with schizophrenia and 2) similar effects on brain development of trauma from abuse and some of the “biological abnormalities” found in patients with schizophrenia (e.g. hyperreactivity of the HPA axis, cerebral atrophy, ventricular enlargement, reversed cerebral asymmetry and hippocampal lesions). Whereas some victims of abuse show dissociative and positive symptoms, others show hypervigilance, which is thought to trigger pruning (a process which entails a decrease in the number of synapses and neuronal loss) (Ibid.,
This latter response to trauma may be linked to a dysfunction of corticostriatal glutameric transmission due to which excessive dopamine release following stress may induce cerebral damage (excitotoxicity and oxidative stress) (Ibid., p. 6).

The author concludes by highlighting the importance of understanding “pathogenic mechanisms that underpin the occurrence of negative symptoms” (Ibid., p. 6) in order to identify therapeutic targets. He proposes that the neurodevelopmental approach is “particularly relevant” and also calls for the consideration of psychological factors as viable treatment targets.

In this BioMedCentral open access article, Limosin provides a carefully formulated overview of the neurodevelopmental aspects of negative symptoms. He argues that “negative symptoms should not be considered only as a unitary construct” (Ibid., p. 3). This excerpt shows both that he regards “negative symptoms” as a construct and takes care not to reify them and that he recognises that they are not unitary, but rather multifaceted (consisting of “motivation, pleasure, and emotion expression factors” (p. 3). The neurodevelopmental approach, which seeks to understand “pathogenic mechanisms” spans multiple levels of explanation, as is shown in the summary, where the author moves from brain abnormalities to behavioural dysfunction.

Finally, the use of “vulnerability” in this article warrants further exploration. It is interesting because it is used in such a broad range of contexts and spanning different levels of explanation. It is mentioned:

1) at the level of (negative) symptoms
2) at the level of biology (e.g. hyperreactivity of the hypothalamic-pituitary-adrenal axis and dysfunction of corticostriatal glutameric transmission” (Ibid., p. 1)
3) as something to be approached from a neurodevelopmental perspective (Ibid., p. 1)
4) in the context of statistics: “genetic factors accounted for 64% of the vulnerability to develop schizophrenia, compared to 36% of environmental factors” (Ibid., p. 2)
5) as “non-specific” factors (…) related to the environment, whether of a psychosocial or biological nature, whether occurring early or later (…)” (Ibid., p. 1)

The above summary of the ways in which the notion of vulnerability appears in the article shows that it is difficult to pinpoint precisely what “vulnerability” entails. The author concedes this in calling the notion of vulnerability “complex” (Ibid., p.2) and calling for “integrated vulnerability models”. These models should take into account first that environmental factors exert effects both on occurrence and progression of disorder, second, that they are highly varied and third, that they cannot be considered in...
isolation due to the interactions between factors (Ibid., p. 2).

All in all, Limosin’s call for “integrated vulnerability models” rather than attempting to “consider environmental factors in isolation” and his appeal to regard psychological factors as therapeutic targets rather than focussing solely on the brain, taken together with his commitment to the neurodevelopmental approach suggest that he is working from a biopsychosocial model of explanation.

2.5.5 Article 4: Foussias, Agid, Fervaha and Remington (2014). “Negative Symptoms of Schizophrenia: Clinical Features, Relevance to Real World Functioning and Specificity Versus Other CNS Disorders”

The authors review developments regarding our understanding of the phenomenology of negative symptoms in schizophrenia. They zoom in on the literature that offers evidence for “separate, yet inter-related, sub-domains” (Ibid., p. 695) of negative symptoms which include amotivation on one hand and diminished expression on the other and set these apart from cognitive and depressive symptoms in schizophrenia. They advocate the removal of attentional impairment, inappropriate affect and poverty of content of speech from the list of negative symptoms (as specified by Andreasen, 1982) because factor analyses, usually based on the Scale for the Assessment of Negative Symptoms (SANS), have shown that they fit better within the disorganized symptom cluster (Foussias et al., 2014, p. 694). Furthermore, a study by Kelley et al. (1999) found that these subdomains hold, regardless of the patients’ medication status (i.e. regardless of whether the negative symptoms are primary or secondary).

The question “whether the negative symptoms are best viewed as a categorical construct characterizing a distinct subgroup of individuals with schizophrenia, or dimensionally across a continuum of severity” (Foussias et al., 2014, p.694), is raised. The idea, developed in the 1980s, that negative symptoms (or deficit syndrome), are best viewed as a categorical construct, has found support in subsequent research. Yet, research on negative symptoms in other disorders or on negative symptoms found in an attenuated form in relatives of patients suffering from schizophrenia have provided support for a dimensional rather than a categorical approach. On both dimensional and categorical approaches, the negative symptom construct has predictive value for outcome.

The authors zoom in on “anhedonia”, one of the negative symptoms. They distinguish between subjective rating instruments to measure hedonic experience and objective experimental paradigms to measure responses to emotion-eliciting stimuli. Overall, both types of research show that the hedonic capacity in patients with schizophrenia is generally preserved: the intensity of experiences does not differ between individuals
with schizophrenia and healthy controls and is also independent of disease stage and medication status.

Recent studies, which attempt to uncover the “discrete components of motivation in schizophrenia” (Ibid., p. 697) are discussed and Foussias et al. propose that defeatist beliefs about performance and biases around low expectancies for success in patients with schizophrenia may contribute to amotivation (Ibid., p. 697). The authors next expand their examination of negative symptoms to other neuropsychiatric illnesses and explore the differences and their relevance to functional outcome. In discussing negative symptoms in relation to depression, they state: “[c]ross-sectional and longitudinal examinations of the relationship between negative symptoms and depression have also found that these two domains exhibit, at most, modest and inconsistent overlap” (Ibid., p. 698), i.e. negative symptoms and depression are distinct.

Neurocognitive dysfunction and negative symptoms are distinct yet have a complex interrelationship: “[m]otivational deficits in schizophrenia have also demonstrated cross-sectional relationships with discrete cognitive domains including executive function, visuomotor sequencing, working memory, and visual learning and memory deficits, but no consistent longitudinal relationship” (Ibid., p. 698). The authors add that motivational deficits (which they take to be synonymous with apathy) and also defeatist performance beliefs may be “particularly important in explaining the relationship between both neurocognitive and social cognitive dysfunction and functional outcomes in schizophrenia” (Ibid., p.699); they later refer to amotivation as a “mediator” (Ibid., p. 699 and again on p. 703). As yet, little is known about the brain structures related to specific negative symptom subdomains and differential responsiveness to pharmacologic agents has not yet become apparent in trials. This contrasts the finding by Grant et al. (2012) that amotivation in particular improves during cognitive behavioural therapy (Foussias et al., 2014, p. 703).

The authors outline research on the course of negative symptoms and find evidence for “a discrete negative symptom domain early in the developmental course of schizophrenia” (Ibid., p.700), i.e. in the prodromal phase. Although individuals with schizophrenia experience enduring negative symptoms more frequently than those with major depressive disorder, negative symptoms alone (taken in a cross-sectional context) cannot serve as a valuable discriminator precisely because they can be a feature of various other non-schizophrenic disorders such as major depressive disorder, affective disorders, personality disorders and neuroses (Ibid.). In addition, factor analyses have shown that depressive and negative symptoms form distinct phenomenological symptom clusters (Ibid., p.702).
A review of the current literature on apathy

In line with their other publications, these authors have remained relatively theory-neutral with regard to the way they bring together the levels of explanation that feature in their literature. They do not, for instance, anywhere use the word “cause” and other synonyms are sparse. They thus abstain from suggesting a direction of causality. Instead, they draw on “relationship”, “has a role”, “has influence on” and “substrate” to describe the connection between phenomena at different levels of explanation. Interestingly, they mention “complex interrelationships” and introduce the notion of amotivation as a “mediator” in the interaction between cognitive dysfunction and outcome. Although the authors do not articulate this, these terms do allow for a causal reciprocity, in line with the biopsychosocial model in medicine. Their use of “construct” to refer to for instance negative symptoms, suggests that the authors want to take care not to reify a descriptive concept.

Moreover, the authors often highlight theoretical difficulties explicitly. An example is their discussion of categorical versus dimensional description of negative symptoms. The authors note that both approaches are valid, making it difficult to choose the most fruitful one. Another example is their mention of the difficulties surrounding the gap between subjective and objective ratings of hedonic experience. All in all, the findings suggest that the authors are grappling with the challenge of finding accurate terminology to describe the relation between the levels of explanation.

2.5.6 Article 5: Fervaha, Foussias, Agid and Remington (2013). “Neural Substrates Underlying Effort Computation in Schizophrenia”

Fervaha, Foussias, Agid and Remington (2013) provide an extensive overview of available methods for assessing effort and motivation. In order to maintain focus, I will zoom in on those aspects of their review that are most relevant to understanding apathy in schizophrenia. They first explain that the reliance on clinical rating scales based on patient recall is very common, yet it is crude - more rigorous methods for measuring this “construct” are needed. Rating scales such as the PANSS and SANS were not designed with a view to assessing amotivation in schizophrenia “rather they include specific items that tap into the construct” (Ibid., p. 2650). And even if rating scales designed specifically to measure amotivation were developed, it still remains to be seen whether they will help to further our understanding of the phenomenon. Marin’s “Apathy Evaluation Scale” (AES) is an example of a scale designed to measure apathy based on patients’ behaviour and “internal motives” and has proven useful for predicting functional outcomes (Ibid., pp. 2650-2651).

The authors move on to discuss laboratory-based assessments of goal-directed behaviour. These include the use of actigraphic monitors to quantify movements, quantification of
self-initiated purposeful behaviour and exploratory behaviour. These studies have shown that there is no difference in exploratory behaviour in patients with schizophrenia versus controls, that patients with schizophrenia walk more and that their pattern of movement is less predictable than that of healthy controls. Other laboratory assessments include the self-reported ratings of pleasure/arousal in response to images compared to the amount of effort exerted to seek or avoid future exposure to these images. The results show that schizophrenia patients have deficits when it comes to coupling behaviour and rating compared to healthy controls (Ibid., p. 2651). The authors cite a study by Tremou et al. (2010) which showed that there was no difference between patients and controls in terms of experienced pleasure or self-reported motivation. The authors conclude that these studies suggest a “disconnect between motivation and affective response (…) possibly impacting volition” (Fervaha et al., 2013, p. 2651).

In the next section of their article, the authors note that effort is not a singular construct, but rather one that involves numerous processes such as the evaluation of the reward, determination to expend effort and the action plan (Ibid., p. 2652). They cite studies on rodents which show that dopamine depletion after administration of haloperidol to the nucleus accumbens decreases their willingness to exert effort to obtain a reward. Further studies have shown that dopamine depletion however, does not affect hedonic response to or preference for reward. In contrast, glutamergic and serotonergic alterations do not lead to a reduction in effort (Ibid., p. 2653). Increasing dopamine levels through administration of amphetamine increases motivated behaviour. There is evidence to suggest involvement of the anterior cingulate cortex (lesions to this area decrease willingness to expend effort for reward) and amygdala (linked to effort-based decision-making) in motivated behaviour. The nucleus accumbens and anterior cingulate cortex appear in some (not all) studies to be part of a circuit involved in “processing” the decision-making. fMRI studies have shown increased blood-oxygen level dependent (BOLD) signals in the nucleus accumbens and ventral pallidum in anticipation of high effort. Patients with auto-activation deficit resulting from lesions to the striato-pallidal area are unable to modulate their behaviour despite valuing greater rewards more (Ibid., p. 2655). They conclude the section: “[t]aken together, studies with healthy human subjects show that [anterior cingulate cortex] and [nucleus accumbens] activity are important for effort-based decision making” (Ibid., p. 2655).

In subjects with schizophrenia, willingness to expend effort was lower in patients with more severe negative symptoms but did not correlate specifically with amotivation as measured with clinical rating scales (Ibid., p. 2655). The authors underscore again that motivation is a very multi-faceted process. They cite an abundance of studies to explain that schizophrenia has been associated with abnormalities in self-reported intrinsic
motivation, apathy and recall of experience of pleasure. Schizophrenia is also associated with deficits in motivated action, the learning, prediction, anticipation, and exploration of reward, and also the neural response to reward, value representations, neural encoding of goal-directed behaviour, tolerability of temporal delays in reward receipt, impulsivity, consistent decision-making, strategic decision-making, decision-making biases, cognitive-affective integration, prospection, planning or executive functions, contextual processing, and working memory. Finally they observe that “[c]onspicuously, hedonic experience derived from the consummation of reward seems to be intact in individuals with schizophrenia” (Ibid., p. 2656).

This article discusses a wealth of research. Perhaps due to the nature of a review, the authors are not explicit about their own philosophical assumptions and they do not go into great detail when treating the various articles they outline. They use terms such as “(neural/neurobiological) substrate”, “are important for”, “vary with” and “have a role in” to tie together processes at explanatory levels such as the level of clinical observation, performance on psychological tasks and neuroimaging data. This terminology seems to reflect the care with which the authors refrain from committing to a specific type of relation between these levels or to a preferred level of explanation (such as the biological level in adherents of the biomedical model).

At times, they use words suggesting a hierarchy of levels, such as “underpinning” and “underlying” and of course “substrate” (from the Latin word “substratum” or under-layer). In these cases, the neural processes “underlie” behaviours, or as in the title of the article “[n]eural substrates underlying effort computation in schizophrenia” (my italics); they are examining the “neural underpinnings of an animal’s willingness to expend effort for a reward” (Ibid., p. 2649, my italics), or the “neural underpinnings of value guided choice” (Ibid., p. 2652, my italics).

Furthermore, in discussing findings at the neurobiological level of explanation, they add subheadings such as “neurochemical basis of effort” and “neuroanatomical basis of effort”. This, along with statements such as “(…) the reduction in effort seen following dopamine antagonism is not seen after glutamatergic (Floresco et al., 2008) or serotonergic alterations (Denk et al., 2005; Izquierdo et al., 2012), but is seen following manipulations of other neurochemical systems” (Fervaha et al., 2013, p. 2653) suggests a specific, temporally defined direction of causality in which a neurochemical event precipitates a change at the behavioural level. If taken alone, this would suggest a biomedical model. Yet, the authors also note causal relations running in the opposite direction and thus do not seem to prioritize the biological level as their preferred level of explanation. For instance when they write: “[i]t should be noted that the relationship between these
processes [at the neurobiological and behavioural levels] is not unidirectional; rather, processing during one step may affect any subsequent or previous step, thereby causing continual updating” (Ibid., p. 2652).

All in all, these findings seem to fit with a biopsychosocial model, in which the biological level is the bottom layer, the psychological level lays on top, followed by the social level. The authors allow for reciprocity between the causal effects of events at these levels.


Clarke et al., provide an overview of the definitions of apathy and discuss the validity of the most commonly used questionnaires to measure the phenomenon in a wide range of disorders, including Parkinson’s disease (PD), other dementias, following traumatic brain injury (TBI) or a cerebral vascular accident (CVA), and in schizophrenia. In the introductory section, treatment options for apathy are briefly reviewed. The authors state that there is evidence for the effect of atypical antipsychotic medications “particularly in patients with schizophrenia, whose negative symptoms strongly resemble apathy” (2011, p.74).

Next, definitions (or rather, the lack of consensus concerning them), are treated. The authors note that apathy can be regarded as either a symptom or a syndrome. Examples of apathy as a phenomenon that occurs independently of other “neurological/psychiatric or medical conditions” include the “why try” attitude of prisoners, institutionalised elderly persons and immigrants (Ibid., p. 75). Sometimes, apathy is a “maladaptive consequence” of a situation or disorder. Recognition of these cases can help further our understanding of the “psychological and neurological mechanisms that mediate this loss of goal-directedness/interest or motivation” (Ibid., p. 75).

Differences in the definitions posited by central researchers are outlined. For Marin, apathy is a loss of motivation; for Stuss, van Reekum and Murphy (2000) it is an absence of responsiveness to stimuli characterized by a lack of self-initiated action. It seems that Clarke et al. are inclined to regard apathy as a “neurobehavioural” syndrome. They note that the Stuss, van Reekum and Murphy (2000) highlight a further quandary that requires research: i.e. that rather than being a single syndrome, apathy may be “separable states depending on which neural system is involved and which (…) response is affected” (Clarke et al., 2011, p. 75).

Finally, the authors review the rating scales used to measure apathy and comment on
both their reliability and validity. I will give a very brief summary of the rating scales outlined by Clarke et al. (2011) which have been found to be valid and reliable for assessing apathy in schizophrenia. The “Apathy Evaluation Scale” (AES), which consists of 18 items to quantify the behavioural, affective and cognitive domains, was found to be both reliable and valid for the measurement of apathy in schizophrenia. The Brief Psychiatric Rating Scale (BPRS), Scale for the Assessment of Negative Symptoms (SANS) and Positive and Negative Symptom Scale (PANSS) too, have been examined as tools for the evaluation of apathy in schizophrenia and there is evidence that all three are both valid and reliable.

In their conclusion, the authors highlight the importance of rating scales for lack of a golden standard to guide further research. Apathy is not yet a precisely defined “construct” which fuels terminological difficulties and the potential for biases (e.g. due to the debate about whether apathy is a symptom or a syndrome) in rating scale development. The authors lament that for the above reasons, at present, researchers will have to make do with inferior validation methods such as demonstrating discriminant, convergent and predictive validity of the measures.

The terminology used in this article suggests that the researchers are struggling to bring together the levels of explanation that are at stake. They describe apathy in various different terms, for instance as a “construct”, as a “neurobehavioural syndrome” and later they note that the phenomenon can also be a “maladaptive consequence of psychiatric, medical, and neurological disorders” (Ibid., p. 75). I discuss three points about this terminology. First, the use of “construct” to refer to apathy is very common in the literature and seems to reflect attempts not to reify the phenomenon given the current shortcomings in our knowledge of its aetiology. Second, the latter quotation is interesting because it discloses the way in which the researchers place psychiatric disorders (and neurological disorders) outside of the realm of medical disorders. Third, the two latter terms (syndrome versus maladaptive consequence) used to refer to apathy are somewhat mutually exclusive: on one hand, the authors embrace regarding apathy as a (neuropsychiatric) syndrome in itself, yet this makes it difficult to allow for the idea that apathy can also present itself as a “maladaptive consequence of other disorders”. All in all, this seems unparsimonious, for it would suggest that apathy is both a symptom of “psychiatric, medical and neurological disorders” and a syndrome in itself. Although the authors note the conceptual difficulties they do not treat possible solutions beyond advocating the use of rating scales for want of a more accurate measuring method. On the same page of the article, they express the need to understand the “psychological and neurological mechanisms that mediate this loss of goal-directedness/interest or motivation” in apathy as a maladaptive consequence of other disorders (Ibid., p.
75). Note here the use of “mechanism” and “mediate”, which suggests an assembly of dynamic parts, the workings of which “mediate” or lead to apathy as an end-product.

In sum, it is difficult to distil the type of model (biomedical or biopsychosocial) at play in their approach, due to the authors’ focus on the validity of psychometric models precisely because understanding of mediating mechanisms is lacking in the first place.

2.5.8 Article 7: Selten, Wiersma and van den Bosch (2000). “Distress Attributed to Negative Symptoms in Schizophrenia”

Selten et al. (2000) describe their own phenomenological study in which they employ the Scale for Assessment of Negative Symptoms (SANS) and the Subjective Experience of Negative Symptoms (SENS) (Selten et al., 1993) to examine to which of the negative symptoms hospitalized patients suffering from schizophrenia attribute distress. Furthermore, they searched for correlations between distress on one hand and a wide range of variables, such as positive and negative symptoms, psychiatric disability, depression, anxiety, legal status of stay, type and dosage of medication, length of illness, severity of illness, age at first admission and length of admission, on the other hand. The apathy and avolition subscale, which included lack of energy and impersistence, were found to be regarded as the most bothersome symptoms in the interviewees. Patients with comorbid depression and insight into positive symptoms were most likely to have high distress scores. Interestingly, they found that the severity of symptoms and distress attributed to them was not correlated.

This research is an example of a quantitative study that brings together clinical observations of symptoms (i.e. the third person account) quantified through attributing scores to answers with first person accounts of (quantified) distress attributed to those symptoms. The results are framed in terms of statistical correlations. This use of statistics is an approach that allows researchers to remain agnostic about causal relations or the precise mechanisms that may (or may not) connect the phenomena under study. The use of statistics and formation of networks of related symptoms will be treated in more detail in the following chapters. Because their study focuses on the psychological level (comparing first and third person findings) and does not span multiple levels of explanation, it is not possible to discern whether the authors are operating from a biomedical or biopsychosocial model. It is however, clear that their project is not reductionist in nature: they compare rather than aim to reduce first person experiences to third person, clinical observations.
2.5.9 Article 8: Liemburg et al. (2015), “Neural Correlates of Planning Performance in Patients with Schizophrenia – Relationship with Apathy”

In their study published in 2015, Liemburg et al. explore the neural correlates of apathy in patients diagnosed with schizophrenia. Based on studies by Levy and Dubois (2006), Faerden et al. (2010), Foussias and Remington (2010) and Roth et al. (2004), they infer that apathy may arise from planning deficits and set out to find neural correlates for this association. They used the Positive and Negative Syndrome Scale (PANSS) to provide a measure of apathy at the behavioural level and used fMRI data of patients performing the Tower of London (ToL) task (a task that is taken to require higher order planning processes) in order to look for correlations between a) PANSS score, b) performance on the Tower of London task and c) fMRI data. They conclude that:

“Higher levels of apathy were associated with less task-related activation within the inferior parietal lobule precuneus and thalamus. Compared to controls, patients showed lower activation in lateral prefrontal regions, parietal and motor areas, and a higher activation of medial frontal areas.

Apathy was related to abnormal activation in thalamus and parietal regions during the ToL task. This supports the hypothesis that impaired function of brain regions involved in planning and goal-directed behavior may underlie apathy in schizophrenia. Moreover, impaired lateral prefrontal activation in schizophrenia patients compared to controls is consistent with the hypofrontality model of schizophrenia.” (Liemburg et al., 2015, p. 367)

The quantitative study outlined above is of interest to this project because, like many other studies on apathy, it spans multiple levels of explanation. Results from the PANSS pertaining to the phenomenon “apathy” are compared to the results of fMRI data and the above citation of the results show that the authors assume that there is a hierarchy of levels in stating that “impaired function of brain regions involved in planning and goal-directed behavior may underlie apathy in schizophrenia” (Ibid., my italics). Another example: “the parietal cortex, besides the prefrontal cortex, appears to have an important role in the mental processes that underlie apathy” (Ibid., p. 372, my italics).

Interestingly, the authors sometimes use “activate” in an active sense for instance in the following three examples: “healthy control subjects activated mostly parietal brain areas during the planning condition” (Ibid., my italics), “healthy subjects deactivated temporal regions in response to higher task complexity” (Ibid., p. 372) and “[p]atients indeed failed to activate prefrontal areas in response to increased task complexity” (Ibid., p.
Elsewhere the authors use “activate” passively, that is, without ascribing agency to anything or anyone in particular: “[s]triatal areas were not activated (…)” (Ibid., p. 372, my italics) and “thalamus, medial frontal and motor areas were more strongly activated in the count balls condition” (Ibid., my italics).

In the discussion section, terms such as “related to”, “have a role in”, “associated with” and “involved in” are used to describe the relation between fMRI findings and PANSS scores. These terms suggest that the authors are being careful not to imply a specific direction of causality or implicate a preferred level of explanation. Yet, as stated above, they do commit to a hierarchy of levels, as is shown by their use of the verb “to underlie”. In sum, it seems that the authors are agnostic about the way in which the levels of explanation can be brought together or related to one another, and they do not attempt to reduce findings at the psychological level to findings at the biological level. Therefore, they seem to be working from a predominantly biopsychosocial model rather than a biomedical, reductionist model.

2.5.10 Article 9: Roth et al. (2004). “Apathy in Schizophrenia: Reduced Frontal Lobe Volume and Neuropsychological Deficits”

Roth et al. take the following as their point of departure: “[t]he neurobiological underpinnings of apathy in schizophrenia have not been systematically investigated, although frontal lobe abnormalities have been shown in both structural and functional imaging studies of deficit syndrome schizophrenia.” (Roth et al., 2004, p. 157). Their goal is to explore the correlation between neuropsychological performance, as measured by various scales such as the SANS, and the Schedule for Assessment of Positive Symptoms (SAPS), and compare these results with MRI data on temporal lobe size. They find that schizophrenic patients with high levels of apathy also show lower performance IQ, poorer visuomotor sequencing, verbal and learning memory deficits and also smaller bilateral frontal lobe volumes than schizophrenia patients with low levels of apathy and controls. In the discussion, the authors propose that:

“[i]mpaired sequencing ability in our high apathy group is consistent with prefrontal-thalamus-basal ganglia circuitry involvement in sequencing, the finding of reduced frontal lobe volume in this group, and neuroimaging evidence of disruption of this circuitry in patients with schizophrenia with prominent negative symptoms (…). The specific cognitive processes through which frontal-subcortical or right hemisphere dysfunction may lead to apathy in schizophrenia remain to be elucidated.” (Roth et al., 2004, p. 159)

Note that, as in the previous example, these researchers too, implicitly seem to adhere to
A review of the current literature on apathy

A hierarchy of explanatory levels in which the neurobiological level appears to underpin the behavioural level. The terminology that is used to bring together the domains under study: “associated with”, “is consistent with”, “involvement” and finally “may lead to apathy”. The use of terms such as “involvement” and “is consistent with” again suggests careful agnosticism on the part of the authors with regard to the nature of the relationship found between the behavioural and neural domains. “May lead to” appears to imply the assumption of a causal relationship in which the direction of causality runs from the neural to the behavioural level of explanation. In all, it is not entirely clear whether these authors are working in a predominantly biomedical or biopsychosocial paradigm because on one hand they are agnostic about the relation between the levels of explanation, whereas on the other hand, some terminology suggesting a direction of causation and a preferred level of explanation has made its way into their writings.


Foussias et al. (2015), writing in the Journal of Psychopharmacology, provide a conceptual review of the negative symptoms of schizophrenia and draw on these in order to discuss treatment options. They focus on the two subdomains of negative symptoms: diminished expression and amotivation and give an overview of ways in which these theoretical constructs can be assessed clinically in both humans and non-humans. For various symptoms of the amotivation subdomain, i.e. hedonic capacity, reward prediction/learning, reward valuation, effort valuation, goal-directed decision-making and social motivation, the authors offer “translational paradigms” for their evaluation. For instance, for the symptom “effort valuation” they suggest using the “effort-cost decision making task” and/or the “energy expenditure for rewards task” (EEfRT) (Ibid., pp.118-119). The overview shows that most of the methods for evaluating the amotivation subdomain are task-based, except for the method of “laboratory-based assessments of emotional experience in response to emotion-eliciting stimuli” (Ibid., p. 119) which can be used to evaluate hedonic capacity according to the authors.

Like many other authors, Foussias et al. write about the “negative symptom construct” (cf. Ibid., p. 123) and they deem the “negative symptom construct” and the “motivational system” a feature of this construct that is “multi-facetted” rather than unitary (Ibid.). Foussias et al. recognize the heterogeneity in clinical presentation, and write about “the clinical expression of symptoms”. The latter phrase suggests that the clinical presentation is something that results from expression of a symptom. Their statement that the negative symptoms “exhibit a moderate degree of interrelatedness, underscoring the possibility that they may have some shared, as well as some distinct, neurobiological origins” (Ibid., p. 117) moreover suggests a specific direction of causation, where neurobiological
processes lead to symptoms. They mention the “neurobiological underpinnings” of negative symptoms and clinical presentation, which, as is very common in the scientific literature, suggests a hierarchy of levels. However, it is not clear how the authors seek to bring together these facets in an integrated explanation.

Interestingly, Foussias et al. begin their introduction by stating that “[s]chizophrenia is a chronic brain disorder (…)” (Ibid., p. 116). Their definition of schizophrenia as a brain disorder and the use of the expression “neurobiological origins” suggests that the authors regard the neurobiological level as the preferred level of explanation for negative symptoms. Finally, based on these observations, it would seem that the authors are working in a predominantly biomedical paradigm.

2.6 Preliminary conclusions

The summary of ten articles in the scientific literature provided above serves a twofold purpose. First, the review articles have given some insight into what is known about apathy in schizophrenia. The research articles were added in order to give readers a sense of the ways in which the phenomenon can be investigated experimentally. The literature summarized was chosen because it spans a variety of themes, including conceptual aspects of apathy, measurement of apathy, and of course psychological, social and biological aspects of apathy. Second, each summary was followed by an interpretative section in which I discussed how the explanatory levels at play in the articles were brought together and whether this seemed to fit with either a predominantly biomedical or rather a biopsychosocial model of scientific explanation. Although there are many different explanatory models, I chose to focus on the biomedical and biopsychosocial models because these are the most well-known models in medicine and because a more philosophical discussion of other models and the way in which they tie in with the biomedical and biopsychosocial models will follow in later chapters.

We have seen that none of the authors summarized have explicitly named a scientific model to which they adhere. It was therefore necessary to attempt to distil this information from the way in which the authors worded their articles. In doing so, I have taken some expressions to point to a biomedical paradigm. For instance, when the (neuro)biological level is the preferred level of explanation and when processes at the (neurobiological) level of explanation are thought to give rise to processes at those higher levels. Tell-tale signs of a biopsychosocial paradigm are that the authors do not commit to a single preferred level of explanation for all phenomena under study and that they employ terminology which does not commit to a unidirectional causal relation between biological processes and phenomena at the psychological and social levels. Both
paradigms have in common that there is a hierarchy of levels of explanation that we encountered very frequently in the literature discussed above, in which the biological is followed by the psychological and social levels.

In some of the articles it was difficult to detect any predominant paradigm, for instance because the research discussed pertained to one level of explanation only or because the authors at times use expressions that suggest a biomedical model and at times imply a biopsychosocial model. Furthermore, there is a large degree of overlap of both content and authors, but it seems that not all articles on which one author has worked (even if they were published within the space of five years) draw upon the same scientific model of explanation. This finding suggests that many authors are perhaps not aware of their own philosophical assumptions or of the implications of certain ways of expressing their findings.

Taken together, these literature-based findings highlight the confusion amongst scientists investigating apathy when it comes to the challenge of integrating their findings from different domains of study. And, although many researchers may be only marginally aware of them, these problems may in fact be obstructing further progress when it comes to understanding apathy because philosophical assumptions within a research paradigm inescapably shape the way experiments are designed, executed and described. The hypothesis that confusion about the integration of levels of explanation may be obstructing scientific progress will be explored in more detail later.

---

3 One of my basic assumptions throughout the interpretative section has been that when the authors mention “levels” (for instance the neurobiological level), they are referring to levels of explanation. I will discuss different types of levels in detail in a later section.
Chapter 3

The biomedical model and its inherent reductionism
3 The biomedical model and its inherent reductionism

This chapter first briefly follows the development of the biomedical model which has guided clinical and scientific practice in both the past and the present, and then zooms in on the problems of reductionism that are inherent in this model of explanation in particular.

3.1 The history of the biomedical model

The biomedical model has its origins in the 19th century. It is frequently stated to be rooted in Cartesian mind-body dualism (Engel, 1977; Ghaemi, 2010, p. 93) and is also often associated with Rudolf Carl Virchow (e.g. DeWalt & Pincus, 2003; Wade & Halligan, 2004; Brown & Fee, 2006) and other early microbiologists such as Louis Pasteur working in the 19th century. Romano, a professor of psychiatry with whom Engel collaborated extensively, suggested that the dualistic approach in psychiatry followed that in medicine and was fuelled by the need to bridge the gap between the physical and biological sciences on one hand and the social sciences on the other. He added that “psychiatry is attempting to bridge this gap in a setting which, far from being contemplative, requires that the needs of the desperately sick be met, not tomorrow, but now and today” (1949, p. xii).

The mind-body dualism inherent in the biomedical model can be traced back to the 17th century philosopher and mathematician, René Descartes (1596-1650). Descartes was a “substance dualist”. He argued for the metaphysical position that mind and body are different “substances” and he was puzzled about the nature of their relationship. On his account, the body is material: it is spatially extended and is therefore referred to as the “res extensa” which is divisible (Descartes, 1996, p. 59) and governed by the laws of mechanics. Descartes’ account is reductionist because he regarded the body as a machine, definable in terms of parts and their mechanistic interactions. However, he found the workings of the human to be too different from the workings of other things in the world. He regarded the ability to think as the distinguishing factor and therefore introduced the “res cogitans”. The res cogitans is immaterial and indivisible (i.e. it is “quite single and complete”, (Ibid., p. 59). Furthermore, the res cogitans cannot be explained in terms of shape, size and the motions of interacting parts, its essential property is that it thinks.

The following excerpt demonstrates Descartes’ dualistic reasoning:

(…) I can infer correctly that my essence consists solely in the fact that I am a thinking thing. It is true that I may have (…) a body that is very closely joined to
me. But nevertheless, on the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and on the other hand I have a distinct idea of body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it. (Ibid., p. 54)

Thus, the question of how the two could interact arose: "(…) I am very closely joined and, as it were, intermingled with [my body], so that I and the body form a unit" (Ibid., p. 56). In his attempt to determine how something immaterial can influence the body, Descartes posited that it was the pineal gland in the brain that had this unifying function. The reason for choosing the pineal gland, appears to have been its central location within the anatomy of the otherwise bilateral brain (Smith, 1998). The biomedical model itself does not offer any suggestions about how mind and body can interact, but it is clear from the above that the dualism inherent in the biomedical model stems from Descartes: the biomedical model harbours an implicit divide between the experience of symptoms (taken to be mental) and the “underlying” physical deficit (in the body).

Some two hundred years later, the story on the origins of the biomedical model again gathered pace with the advent of microbiology. Virchow was born in Pomerania (former Eastern Prussia) in 1821 and received his medical training in Berlin, Germany. Besides his clinical work he was also an avid researcher. He advocated a departure from the then pervasive humoral principles, proposing a focus on cellular principles instead. His contributions to the field of cellular medicine included research on the pathophysiology of leukocytosis, leukemia, thrombosis, pulmonary embolism, mycoses, amyloid and trichinosis (Wade & Halligan, 2004). At the age of 26, he founded a new journal called “Archives for Pathological Anatomy and Physiology and Clinical Medicine” (Brown & Fee, 2006). Although Virchow is known mainly for his focus on the biomedical aspects of medicine, he later dabbled in politics and in fact promoted medicine as a social science. Nonetheless, the central idea in the biomedical model, that diseases result from cellular pathology, is frequently attributed to Virchow (Wade & Halligan, 2004). This central idea is reductionist: disease results from abnormalities at the biological level and can therefore be fully explained in terms of biological processes.

The French chemist and microbiologist, Louis Pasteur, was a contemporary of Virchow’s. Born in 1822, he is widely regarded as one of the fathers of germ theory. The process of heating foods to reduce the number of microbes and prevent spoilage is named after him: pasteurisation. Together with Ferdinand Cohn and Robert Koch, Pasteur was amongst the earliest bacteriologists. He pinpointed microorganisms as a cause of disease and
developed a variety of vaccines. He disproved the then pervasive idea of “spontaneous generation” (this is the idea that micro-organisms could arise from inanimate matter) and thus helped sow the seeds for a biomedical approach to disease.

3.1.1 The Biomedical Model

According to the biomedical model, medical disease can be ascribed to a bodily defect or malfunctioning at the biological level (for example at the level of molecules, cells, tissues, or organs) rather than to processes at for instance the psychological and social levels. Unlike the biopsychosocial model, which was coined and described by Engel (as discussed in more detail below), the biomedical model is not as clearly outlined in a single document by a specific thinker, as is demonstrated by the outline of its history in the previous section. The biomedical model is primarily a physicalist reductionist model: disease or the absence of health is explained in terms of pathogenesis at the biological level. The biomedical model of mental disorder has as its central claim that mental disorders are brain diseases.

The biomedical model has proved very successful in furthering our understanding of a great number of diseases. Take for example cystic fibrosis: we now know that it is an autosomal recessive genetic disorder in which abnormal sodium and chloride transport across epithelium, resulting primarily in pulmonary and gastrointestinal symptoms, is caused by a mutation in the gene for the protein cystic fibrosis transmembrane conductance regulator. Cystic fibrosis and many other predominantly “somatic” diseases are candidates for explanation on the biomedical model because there is an identifiable biological substrate, such as a single gene mutation in the case of cystic fibrosis. Other examples of disorders that lend themselves well to a biomedical approach are those researched by Virchow.

Both reductionism and mind-body dualism are inherent in the biomedical model. These two basic assumptions have clearly helped direct medical research (i.e. to the biological level) and have served to provide a limited, but at least coherent (within that limited scope,) framework from within which to devise and test hypotheses empirically.

As demonstrated in Chapter 2, the biomedical model is still pervasive in the recent literature on apathy in schizophrenia. The basic tenet of the biomedical model as applied in psychiatry is that mental disorders are brain diseases (e.g. Deacon, 2013). The reduction inherent in the biomedical approach to psychiatry involves explaining entities at the psychological, social and experiential levels of explanation in terms pertaining to the biological level of explanation.

One example of this view in practice is provided by Andreasen (2001) who, in the
introductory pages of her book “Brave New Brain”, makes the strongly reductionist claim that mental illnesses “affect the brain and its product, the mind” (p.6). The scientific literature on apathy in schizophrenia is packed with more examples. For instance, in Chapter 2, Foussias et al. (2015) were quoted stating that “[s]chizophrenia is a chronic brain disorder (…)” (p. 116). In the following section, I zoom in on why reductionism and mind-body dualism are problematic.

3.2 Limitations of the Biomedical Model

Unfortunately, there are many conditions or phenomena pertaining to biomedically explicable diseases for which the biomedical model cannot provide an explanation. For instance, even disorders such as cystic fibrosis are associated with psychological effects that have so far proven irreducible to the biological level of explanation. The genetic defect that is associated with the pulmonary and gastrointestinal symptoms of cystic fibrosis cannot directly account for the existential pain of having a chronic disease. In Chapter 2, various examples of a biomedical approach to researching apathy in schizophrenia were outlined and discussed. Despite advances in imaging techniques such as fMRI and the rise of (epi)genetics, the mist shrouding many aspects such as the etiology and pathophysiology of this debilitating phenomenon, has not yet cleared. So far, it has not been possible to pinpoint with any certainty a biological defect that can account for the symptoms of what we call schizophrenia. The aim of the following section therefore, will be to address at a more abstract level the two major, and closely related, basic philosophical assumptions that render the biomedical model obsolete: mind-body dualism and reductionism.

3.2.1 Reductionism: a many-faced beast?

First of all, a word on the difference between reduction and reductionism is warranted. Both concern intertheoretical connection. Reduction refers to a particular, single situation in which one phenomenon is explained in terms of a different, usually “lower” level of explanation. Reductionism, in contrast, is a more all-encompassing view, which holds that the method of reduction is preferable to a holistic approach. The refutation of reductionism therefore does not render reduction itself obsolete, but rather suggests that reduction alone is insufficient when it comes to explaining phenomena (e.g. Klein, 2009, p. 52; Fang, 2011).

The term “reduction” is used in many contexts. It originates from the Latin “reducere” which means “to lead back to”. Its use differs depending on the context in which it is used: everyday language, science or philosophy. Within science, the term “reduction” implies that one entity (be it a thing or a theory) can be brought back to another, usually
lower-level entity. The notion of scientific reduction is a central theme in the philosophy of science. Here, reduction can take on strong and weak meanings and apply to ontology or epistemology. It can entail that one entity is more basic than another or that if one entity is reducible to another, that the former is nothing over and above the latter.

Table 3.1 Contexts in which “reduction” features

<table>
<thead>
<tr>
<th>Contexts for the use of the word “reduction”</th>
<th>Meaning and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyday language</td>
<td>For instance: making something smaller, diminishing its numbers</td>
</tr>
<tr>
<td>Science</td>
<td>Technical uses within sciences, for example: in <strong>physics</strong>: a decrease in positive valence or increase in negative valence by gaining electrons. <strong>Chemistry</strong>: a reaction in which hydrogen is combined with a compound or in which oxygen is removed from a compound. <strong>Biology</strong>: a reduction division is the first meiotic division of a cell in which the chromosome number is halved. More abstract use in science: one entity (be it a thing, a process or a theory) can be <em>brought back</em> to another, usually lower-level entity</td>
</tr>
<tr>
<td>Philosophy (of science)</td>
<td>Reduction makes ontological, epistemological and/or methodological claims about the relation between scientific domains. For instance the questions: can genetics be reduced to molecular biology? Can the mind be reduced to the brain?</td>
</tr>
</tbody>
</table>

Reductionism raises philosophical questions that are best not conflated. For instance, does a particular model of reduction demonstrate the structure of an explanation? Or does it rather describe a research strategy? Does reduction apply to the domain of epistemology (i.e. is it about reducing theories, laws or facts to other, more fundamental theories, laws or facts) or ontology (i.e. is it about showing that entities are composed of other entities)?

In order to navigate the rocky terrain of reductionism in the philosophical domain, three types of reduction can be defined: methodological (or “explanatory”), epistemological (or “theory”), and ontological (or “constitutive”) reduction (Sarkar, 1992). In explanatory reduction, “(...) the reduced entity is explained by the reducing entity no matter whether these entities are theories, laws, empirical generalizations or even individual observation reports” (Sarkar, 1992, p. 170). Explanatory reduction holds that the relation between reduced and reducing entities is one of explanation. On Sarkar’s account, these entities are epistemological entities such as theories, laws, and even individual observations (Sarkar, 1992, p. 170).

Theory reduction refrains from making claims about ontology. The reducing theory
is derived from the reduced theory, for instance, the theory of special relativity can be reduced, and thereby simplified, to classical mechanics if the speed of light is taken to approach infinity. However, calculations using equations in classical mechanical theory only hold for large objects not travelling at speeds close to the speed of light. Explanation is often involved, but this is not necessary for this type of reduction. Thus, theory and explanatory reduction frequently overlap. Sarkar nevertheless makes the distinction because, as he argues, it is particularly relevant for “actual scientific reductions”, which are “messy” because they combine fragments of incomplete theories and mechanisms (*Ibid.*, pp. 175-176).

Constitutive, or “ontological” reduction holds that upper level systems are composed of lower level systems and that they conform to the laws governing these lower level systems. In other words, constitutive reduction is about explaining wholes in terms of their parts (*Ibid.*, p. 171).

Note that the three types of reduction outlined above are defined in terms of relations between entities. This form of reduction is termed “synchronic reduction”. Synchronic reduction contrasts diachronic reduction in which reduction is regarded as an activity (it is “what scientists do”). Furthermore, note that Sarkar’s three types of (synchronic) reduction are related to a degree that corresponds to one’s own intuitions and beliefs concerning a.) the relation between epistemology and ontology and b.) whether or not reduction necessarily involves explanation. Concern a.) falls beyond the scope of this project and does not affect the account provided in this thesis, a point I will return to later. How one deals with concern b.) depends on how “explanation” is defined. The notion of explanation is treated in more detail in Chapter 6.

Table 3.2 Types of reduction

<table>
<thead>
<tr>
<th>Types of reduction in philosophy (Sarkar 1992)</th>
<th>Methodological (explanatory) reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The reducing entity explains or describes the reduced entity</td>
</tr>
<tr>
<td>Epistemological (theory) reduction</td>
<td>A relation between theories: the reducing theory is derived from the reduced theory</td>
</tr>
<tr>
<td>Ontological (constitutive) reduction</td>
<td>A relation between levels of organization: upper-level systems are composed of lower level systems</td>
</tr>
</tbody>
</table>

3.2.1 Subtypes of intertheoretical reduction

Schaffner (1967) provides a helpful overview of four approaches to intertheoretical (i.e. epistemological or theory) reduction. He distinguishes between 1) the Nagel, Woodger and Quine (NWQ) model in which the terms of one theory are related to terms of another theory. 2) The Kemeny and Oppenheim model (KO) which takes a
more indirect approach to reduction, stating that the same observable predictions are obtained by both theories. 3) The Popper, Feyerabend and Kuhn (PFK) model holds that one theory explains why another works and corrects that theory. 4) The Suppes paradigm holds that isomorphism of theories is insufficient for reduction. Schaffner argues that in fact the Suppes paradigm is a weaker version of the NWQ model. In his later writings, including his 1969 article which focuses on reductionism in the Watson-Crick Model of DNA, Schaffner further develops his own model of reductionism in the field of biology. Schaffner is optimistic about the possibility of reducing biology to chemistry: “ultimately, biological organisms are nothing but chemical systems” (1969, p. 346, italics in original). However, he argues that analyses of reductionism in physics do not lend themselves for a straightforward translation to the field of biology (Ibid., p. 325). One important reason for this is the problem of empirical support: in biology, it is much more challenging to devise experiments that demonstrate the relation between the reduced and reducing theories. This is largely due to difficulties surrounding the characterization of the entities, which must have similar or relatable properties, that feature in the reduced and reducing theories. Also, Schaffner draws upon the importance of organisational features (Ibid., p.343). Biological organisms are biological in virtue of the structure or organisation of the chemical complexes of which they are composed. It is this particular organisation of parts, and not just any random organisation of the chemical components, that interests biologists. In parallel, chemists are interested in specific organisations of atoms and therefore treat their subject matter at the level of atoms (as a physicist might) rather than their component articles. Schaffner is also careful to state that the reducibility of biology to chemistry for him does not mean that “living organisms can only be fruitfully studied as chemical systems” (Ibid., p. 346, italics in original). In order to take into account the organisational aspect, Schaffner posits that both the parts themselves and their interactions need to be addressed (Ibid., p.345). Finally, besides pinpointing organisation as a reason for inter-science differences in reduction, Schaffner also adds that in biology, the organisation is already given, raising the added question of how these organisms came to be (Ibid., p.345).

3.2.2 Reductionism in the biomedical model
This section introduces three tiers to structure my treatment of discerning which of Sarkar’s types of reductionism are at stake in the biomedical model. First is the overarching problem of projecting a philosophical distinction onto scientific practice in general. Second is the inter-field argument: following Schaffner, I argue that an analysis of reductionism in one scientific discipline need not necessarily hold for other disciplines. Third is the intra-field argument based on the heterogeneity of scientific practice even within a discipline: there is no reason to assume that an analysis of reductionism that holds for one subset of scientific practices will necessarily hold for other practices,
The biomedical model and its inherent reductionism

even within one discipline or field. In other words, the difficulty pertains to the intra-disciplinary methodological heterogeneity which entails that a one-size-fits-all approach to projecting one type of reduction onto a sub-discipline may be misguided. This is because the methods employed are contextually determined, where “context” is taken in the broadest possible sense. This hearkens back to Heisenberg’s contention that “what we observe is not nature itself but nature exposed to our method of questioning” (1958, p. 58).

In order to flesh out the first tier, I return to the central question of this section. When stating that the biomedical model is reductionist, which of these three types of reductionism borrowed from Sarkar is (or are) at stake? Answering this question is not straightforward, as is evidenced by the considerable body of literature on the nature of the reduction inherent in microbiology (e.g. Wimsatt, 1976; Rosenberg, 1985; Nagel, 1949 & 1961; Kitcher, 1982 & 1984; Sarkar, 1989; Hull, 1972 & 1976; Schaffner, 1969 & 1993). In tackling this task, philosophers such as those quoted above thus proceed by reviewing the way scientists form their hypotheses and then attempt to deduce from these analyses the way in which reduction takes place.

In the case of the scientific literature on apathy, this would entail trying to find examples of each of these types of reduction. In doing so, one of the sources of confusion is the distinction between ontological and epistemological reductionism. As mentioned earlier, where one draws the line between ontological and epistemological reduction (a philosophical distinction) is not as clear as the above definitions may suggest. In the case of reduction in psychiatry research, it would depend on whether the mind is regarded as a thing in the world, i.e. an ontological entity or instead as a theoretical (or even linguistic) construct.

Furthermore, determining which type of reduction is at stake in the biomedical model in medicine in general and in psychiatric research in particular, requires projecting the philosophical distinction between theory, explanatory and constitutive reduction onto scientific practice. Following the table of contexts for the operationalisation of “reduction”, this involves moving from philosophical to scientific operationalisation of the term. This is problematic: as the table highlights, the meaning of the term is highly context-dependent. A philosopher and a scientist conversing about reduction are likely to be speaking different languages, operating within frameworks that are only vaguely related. The projection of a philosophical technical term onto scientific practice post hoc is therefore nothing more than an interpretative exercise on the part of the philosopher. Many nuances are likely lost in translation. Although an interesting exercise (as shown in Chapter 2), it mainly serves to uncover philosophical intuitions of researchers through
what they write. And, as it turns out, these philosophical intuitions are frequently incomplete and internally inconsistent in the eyes of philosophers. Interestingly, the lack of a coherent philosophical framework in the work of highly respected and experienced researchers and clinicians operating at the cutting edge of science or psychiatric practice does not seem to affect their status as successful scientists and clinicians (e.g. Ralston, 2013). Thus, for the first tier the problem of projecting philosophical distinctions between types of reduction onto scientific practice emerges. In this vein, following Schaffner (1969, p. 325), I advocate a strategy of mutual illumination rather than one-way projection.

The second tier zooms in on the philosophical analysis of reduction in different sciences such as physics, chemistry, biology and medicine. As argued by Schaffner, an analysis of reduction that holds for physics is not necessarily accurate for reduction in other sciences. There are inter-science differences, for instance in the way hypotheses are empirically tested, and in the organisational features which affect the characterization of the relationship between reduced and reducing theories, processes and entities. Take for instance the recent empirical evidence for gravitational waves. The existence of gravitational waves was anticipated, theoretically, by Einstein a century ago. In early 2016 the first journal articles emerged, showing empirical evidence to support the theory. This evidence is based on experiments conducted with a laser beam, which is split and reflected back to a set point by a series of mirrors. Gravitational perturbations lead to disturbances of the space in which the laser beam operates and thus affects the way in which the recombined beam hits the photodetector. The perturbation that sparked the recent articles on gravitational waves was traced back to a warping of space-time due to gravitational waves caused by the collision of two black holes. In this example, in order to detract evidence for gravitational waves from the mind-boggling amount of “noise” detected by the apparatus was comprised of multiple steps. First, predictions about the existence of measurable gravitational waves sparked the development of models to determine patterns (calculated by running models on supercomputers) in the data that could be regarded as empirical evidence for gravitational waves. Next, the experimental setup had to be designed to mechanically cancel out as much noise as possible. Remaining noise (due to for instance the waves crashing into our coastlines, earthquakes and human activity) then had to be identified and methods were developed to statistically cancel this out in the data generated by the apparatus.

Contrast the above example in which theory and practice meet, to an example taken from medicine, where the relation between reduced and reducing entities, processes and theories is nowhere near as clear-cut as it can be in physics. One major difference is the availability in physics of a more secure theoretical framework where theory precedes
empirical testing. In medicine, the hypotheses that are to be empirically tested are rarely seated in as firm or unified a theoretical framework. Furthermore, in medical research it is extremely difficult to correct for the huge role that a person's wider context may play in the manifestation of disease or efficacy of a drug. Like physicists, medical researchers too, rely on statistics (gathering large numbers of measurements for mathematical manipulation) to approximate (or correct for) individual contexts. For instance, the life expectancy of a human suffering from cystic fibrosis depends on the type of health care to which they have access, their adherence to available therapy, the ability of their immune system to fight infection etc. In this case, there are so many variables that it is very challenging to take all of them into account. Even in vitro experiments and animal experiments, in which researchers are able to fix a number of contextual variables, still call for statistical analysis in order to render their results useful. In contrast to the problem of cancelling out noise in the example taken from physics, the “noise” caused by contextual factors in the medical example is crucial to understanding the phenomenon at hand. Thus in other words, in physics, reducing data to support an already available theory is an entirely different process than in medicine, where theory development is more data-driven in the first place (accumulating data and discerning relevant factors contributes to theory formation). The difference outlined here thus runs parallel with the difference between a top-down and bottom up approach.

The third tier further extrapolates the line of argument expounded in the second tier: even within a scientific discipline or sub-discipline, there may be organisational principles and limitations in the possibilities for empirically testing hypotheses which make it difficult to outline anything more than a very abstract notion of reduction that holds for all relationships between theories, processes and entities within that field. This somewhat contrasts Schaffner’s claim that he had devised “a biological principle of reduction” that also holds for other biological theories (1969, p.345). He specifies that:

“[t]he principle can be stated as follows: given an organism composed out of chemical constituents, the present behaviour of that organism is a function of the components as they are characterisable in isolation plus the topological causal infrastructure of the chemical constituents”. (*Ibid.*, pp. 345-346)

My response to his claim is that the principle is indeed outlined in very vague terms which make it both more widely applicable, but also less informative when it comes to the mission of describing the type of reduction that is at play in specific instances.

Another way of expressing the difficulties in categorizing the reductionism inherent in the biomedical model as summarized above, is stating that an analysis of reductionism
in medicine and in the use of the biomedical model in psychiatry in particular, involves navigating a transition from reduction as a philosophical notion that describes relations between entities (for instance between psychological and biological phenomena or theories) to reduction as a scientific activity performed by scientists and clinicians. In other words, this entails a switch from synchronic to diachronic reduction. Diachronic reduction implies temporality and even, as some would argue, is required for scientific progress in time: one theory is replaced by, and improved upon, by its successor.

I now move on to why reductionism in the biomedical model is problematic and will show that the arguments why reductionism is problematic can hold for all three types of reductionism and therefore singling out examples of each type in specific instances of scientific practice will prove unnecessary to support my claim that reductionism (and also mind-body dualism) is what renders the biomedical model obsolete.

3.2.3 So why is reductionism a problem?
It is the exclusiveness of reductionism which forms a problem. As mentioned earlier, reduction in itself need not cause problems, whereas the view that reduction and reduction only can further scientific understanding is problematic. First, reductionism entails choosing a level of explanation as the essential or preferred level of explanation. The finding that some phenomena are more easily understood or described or explained in terms at one level of explanation, does not necessarily entail that this holds for all phenomena in all contexts. In the biomedical model, the biological level of explanation is privileged; it is chosen as the level on which phenomena ought to be examined and understood. This choice seems relatively arbitrary and subject to change, as is illustrated by the outline of the history of the biomedical model earlier in this chapter. In sum, there is no compelling reason to absolutely prefer one level over others for the explanation of all medical phenomena.

Second, reduction entails simplification and abstraction which often detracts too much from the character of the phenomenon to be explained. Take for instance a patient suffering from auditory hallucinations: on a daily basis, he hears angry voices talking about him in strong language, saying that he is a worthless human and making fun of the way he goes about his (already limited) everyday activities. The patient feels frightened and frustrated. He wonders whether the voices perhaps have a valid point and his self-confidence plummets. In a desperate attempt to make sense of his unsettling experiences and to avoid providing the voices with further grist to their mill, he becomes increasingly isolated and tapers down his activities. Trying to reduce this patient's deeply troubling experiences to processes located in the brain, leaves no room for consideration of what these complaints mean for this individual patient in his day to day life. Pointing
to fMRI studies on brain activation during hallucinations, genetic predisposition and the dopamine hypothesis of psychosis will do little to further our understanding of the effects of the hallucinations on this particular patient’s life. Reducing a psychiatric complaint to a brain defect thus dehumanizes psychiatric practice. It detracts meaning from the experiential and psychological domains of psychiatric phenomena by privileging the biological level of explanation. Similarly, reducing a psychiatric complaint to a behavioural defect does away with biomedical and existential aspects too.

The next chapter offers an examination of Engel’s biopsychosocial model which he posited in an attempt to integrate different levels of explanation rather than attempting to choose a single preferred level of explanation in the reductionist “one-size-fits-all” vein of the biomedical model.
Chapter 4

The Biopsychosocial Model
4 The Biopsychosocial Model

4.1 Enter Engel: The Biopsychosocial Model

In this chapter, the biopsychosocial model as described by Engel, is introduced. I begin by offering a very brief biography of George Engel, to illustrate the context in which the model was developed. Next, I provide an outline of Engel’s writings on the biopsychosocial model including textual evidence and schematic figures to identify key concepts in a largely chronological order. I zoom in on the work of the biologist von Bertalanffy. Engel refers to von Bertalanffy’s work on general systems theory. Engel suggests it may be used to integrate the biological, psychological and social aspects that are important in medicine. These introductory sections on the biopsychosocial model then serve as the basis for discussing various lines of critique and for identifying and salvaging the strongest aspects of the model in the following chapter. The exploration of these lines of critique in Chapter 5 makes it possible to answer the first central question of this thesis: what is the utility of the biopsychosocial model at the level of a) clinical practice b) science and c) philosophy? The answer to this question prepares the way to the treatment of the second central question of this thesis: how can we understand the use of the biological, psychological and social levels of explanation in science and account for the coherence and possibilities for integration of these different levels of explanation? This second question will be the topic of later chapters.

4.2 Background and development of the biopsychosocial model

George Libman Engel was born in New York in 1913 and completed his medical degree at Johns Hopkins University in Baltimore in 1938. On graduating, Engel was mainly interested in physical explanations of disease and was sceptical of psychosomatics and psychoanalysis. However, during his early career he began collaborating with the psychiatrist John Romano, first at the University of Cincinnati (around 1942) and later at the University of Rochester Medical Center (in 1946), and became increasingly interested in psychosomatics. Engel was appointed at both the psychiatry and medicine departments at the University of Rochester Medical Center and helped to establish a medical psychiatric liaison service. Engel advocated psychiatric training in the medical school curriculum and commenced his own psychoanalytical training. The titles of his publications follow this transition from a more medical focus to a deep interest in psychosomatics: his earliest publications are on nerve metabolism and decompression sickness. Later on, he focusses on physiological and psychological aspects of fainting (Engel, 1950) and psychogenic pain (Engel, 1959). Throughout the 1950s, he continues to develop his ideas on the concepts of health and disease (e.g. Engel, 1953 & 1960).
The short excerpt below from his 1960 article “A Unified Concept of Health and Disease” demonstrates the impressive scope of Engel’s thought: for his concept of disease he differentiates between types of explanation (those that explain how versus those that explain why) and the effects of these two approaches. For instance, he relates this to the normativity inherent in the disease concept and highlights that one of the effects of favouring how explanations over why explanations affects the patient’s relation to his disease, reducing him to a “helpless victim” of disease.

Regardless of whether disease is equated with a physiological, an anatomical, or a biochemical lesion, the psychic distortion in the observer involves the idea of a discrete “thing” inside the body, an entity having an existence of its own, apart from the patient, who is the helpless victim. Still another variation on this theme is the mechanistic concept, according to which the body is seen as a machine and disease as a condition due to a defective part. In modern medicine this takes the form of the mechanism (the “how”) being used as the explanation of the disease state. From this perspective, peptic ulcer is considered to be due to overactivity of the vagal outflow. The “why” is ignored. A misbehaving organ or system again fulfills the psycho-economic requirement of the disease as a thing apart. Even the various “stress” theories currently popular are not immune to this psychic influence, “stress” being conceptualized as some kind of “bad” force of which the person is a helpless victim. (Engel, 1960, p. 462)

Engel’s 1960 article develops various lines of thought which later lead him to propose the biopsychosocial model. Below, I outline these themes to provide a deeper understanding of the soils from which the biopsychosocial model sprouted.

In his 1977 article “The Need for a New Medical Model: A Challenge for Biomedicine” Engel goes on to identify a crisis in medicine (p. 129). He ascribes the crisis to adherence to the biomedical model which is “no longer adequate for the scientific tasks and social responsibilities of either medicine or psychiatry” (Engel, 1977, p. 129). “Medicine’s crisis stems from the logical inference that since “disease” is defined in terms of somatic parameters, physicians need not be concerned with psychosocial issues which lie outside medicine’s responsibility and authority” (Ibid.). Engel discusses two common positions assumed by clinicians as a response to the crisis: the exclusionists and the reductionists. Exclusionists advocate excluding psychiatry from medicine. Szasz (1961) for instance, argues that mental illness is a “myth” because it does not fit the biomedical definition of disease. Reductionists advocate sticking to the biomedical model. Ludwig (1975) states that “sufficient deviation from normal represents disease, that the disease is due to known or unknown natural causes, and the elimination of those causes will result in
Chapter 4

Table 4.1 Key Themes from Engel (1960)

<table>
<thead>
<tr>
<th>Key themes in “A Unified Concept of Health and Disease” (Engel, 1960)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Normativity of the disease concept (p. 459)</td>
</tr>
<tr>
<td>• Disease concept not attuned to factors such as personality, bodily constitution and life style of patient (p. 460)</td>
</tr>
<tr>
<td>• Engel refutes the idea of disease as entity separate from its (helpless) victim, rather the “material of medicine is sick or disturbed man” (p. 460)</td>
</tr>
<tr>
<td>• The physician too cannot detach himself from his material (p. 460) and is drawn to the idea of disease as a separate entity which he can “attack” (p. 461)</td>
</tr>
<tr>
<td>• Explanations of how (as in the mechanistic tradition) versus why disease occurs, i.e. what function do the changes brought on by the disease have? (p. 462; pp. 474-475)</td>
</tr>
<tr>
<td>• The inhibiting influence of nosology: falsely implying certainty, drawing attention to only some aspects of the disease (p. 463)</td>
</tr>
<tr>
<td>• Diagnostic labels fail to fully define illness, but can have statistical and predictive value (p. 643)</td>
</tr>
<tr>
<td>• Complete physical recovery need not imply complete mental recovery (p. 468)</td>
</tr>
<tr>
<td>• Manifestations of disease are relative and system-dependent (p. 468)</td>
</tr>
<tr>
<td>• There is no sharp dividing line between health and disease (p. 469)</td>
</tr>
<tr>
<td>• Personal, social and cultural factors affect whether and how a complaint it communicated (p. 470)</td>
</tr>
<tr>
<td>• Absence of a complaint need not imply absence of disease (p. 470)</td>
</tr>
<tr>
<td>• Engel advocates viewing disease from a naturalistic rather than an institutional perspective. For, traditionally, what is categorized as disease is restricted to what the physician can understand/recognise and this stands in the way of recognizing disease as a natural phenomenon (p. 472)</td>
</tr>
<tr>
<td>• Etiologic factors can originate within the organism as well as in its environment and must be considered in relation to the system(s) upon which they are operating (p. 473)</td>
</tr>
<tr>
<td>• Factors that determine an organism’s capacity to “grow, survive and adapt” (p. 475): “genic” (purely biochemical) (p.475-6) and developmental factors (p. 476)</td>
</tr>
<tr>
<td>• Factors that strain current capacities of the organism (p. 478): physical or chemical, injurious factors (p. 478), deficiency states, microorganisms and parasites, and psychological stress (pp. 480-483)</td>
</tr>
</tbody>
</table>
cure or improvement in individual patients” (p. 603). Ludwig suggests simply limiting the field of psychiatry to those behavioural disorders that ostensibly or presumably arise from brain dysfunction or other “natural causes” (as opposed to “metapsychological, interpersonal or societal causes”) and therefore excluding disorders such as “problems of living, social adjustment reactions, character disorders, maladaptive learning patterns, dependency syndromes, existential depressions” (Ibid.). Engel finds both positions lacking.

Engel notes that the biomedical model is primarily a scientific model. It was designed as a tool for scientific research, prescribing rules based on shared assumptions, but has become our dominant “folk” model or “cultural imperative” and even a “dogma” according to Engel (Engel, 1977, p. 130). He argues that the model is not suited to clinical practice and outlines specific consequences of adhering to the biomedical model for clinical practice (Ibid., pp. 131-132). These are outlined below:

I. Regarding identification of a biochemical defect as a criterion for diagnosis does not allow for situations in which there is a biochemical defect but the patient isn’t ill, or where the patient is ill despite the absence of an identifiable biochemical abnormality.

2. The biomedical model neglects the importance of the relationship between biochemical processes and clinical data (patient reports, the meaning of the symptoms for the patient).

3. The biomedical model fails to take into account the effects of “life and living” that influence the disease’s onset and course.

4. Whether and when a person regards himself as sick and whether and when a person seeks healthcare in the face of biochemical abnormalities depends on psychological and social factors too.

5. Removal of the biochemical abnormality does not necessarily amount to restoration of the patient’s health.

6. The biomedical model cannot account for the way in which the physician-patient relationship affects therapeutic outcome.

The consequences of biomedical thinking outlined above and Engel’s identification of a “crisis” in medicine call for a new notion of what a disease or illness is. This notion must include normative aspects so as to be sensitive to and adequate in the context in which it is operationalised. These normative aspects pertain to what it means for the individual to be ill within his historical and current context: each of the above consequences is an example of this. Engel designed the biopsychosocial model to try to allow for the normativity, which is, on my account, inherent to medicine both as a science and as
clinical practice. The question of whether he succeeds at this, is a topic to be discussed in the following chapters.

4.3 The Biopsychosocial Model

The goal of the biopsychosocial model is to take into account the missing dimensions of the biomedical model to accommodate the humanness and normativity of medicine by drawing attention to processes at levels other than the purely biological level of explanation in order to explain disease. For Engel, diseases are “natural phenomena” and “the major criteria for identification of disease have always been behavioural, psychological, and social in nature” (1977, p. 130). He adds that the boundaries between “sick” and “well” or between “illness” and “problems of living” will never become clear because they depend on the above-mentioned factors and cultural factors too (pp. 132-133). In observing this, Engel aims to broaden the scope of the physician to take into account personal and social factors, which he considered to be at least as important to understanding pathogenesis. Or in his words: “A biopsychosocial model which includes the patient as well as the illness would encompass both circumstances” (Ibid., p. 133). He leaves it up to the physician to “weight the contributions of social and psychological as well as of biological factors implicated” (Ibid.).

Towards the end of the paper, Engel proposes a general systems theory perspective as a conceptual framework to reconcile the biological, psychological and social domains. He draws upon von Bertalanffy suggesting that there are isomorphisms across the organisational levels from which basic laws and principles may be uncovered. Engel holds that there is a hierarchical connection between the biological, psychological and social levels of organisation which entails that a change at one level affects the other levels too. I return to von Bertalanffy’s general systems theory in more detail later and first delve a little further into Engel’s writings about the biopsychosocial model.

4.4 Application of the biopsychosocial model

In his 1980 article Engel focusses on the clinical application of the biopsychosocial model, which he deems a scientific model (Engel, 1980, p. 535). He does so by introducing us to the case of Mr Glover, a patient brought into the emergency department after suffering his second myocardial infarction. Engel walks the reader through the various factors at play, demonstrating the interplay between biological mechanisms (e.g. blockage of a coronary artery, damaging his myocardium), psychological (e.g. experience of discomfort, remembering his earlier myocardial infarction, denial) and social aspects (e.g. his behaviour towards colleagues and medical professionals) following
The Biopsychosocial Model

his myocardial infarction. Engel writes:

while changes are taking place at the levels of tissue, cell, molecule, organ, organ system, and nervous system, illness and patienthood do not become issues until the person level is implicated, that is, not until the person experiences something untoward or exhibits some behavior or appearance that is interpreted as indicating illness.

For Mr. Glover such changes began around 10 in the morning. While alone at his desk he began to experience general unease and discomfort and then during the next minutes growing “pressure” over his mid-anterior chest and an aching sensation down the left arm to the elbow. The similarity of these symptoms to those of his heart attack six months earlier immediately came to mind. Thus began the threat of disruption at the person level and with it still another wave of reverberations up and down the systems hierarchy. (Ibid., p. 539, italics in original)

For his article, Engel deliberately chooses an example of a patient with a somatic (rather than psychiatric) problem in order to emphasize that even for diseases that are relatively well understood from the biological perspective, the psychological and social domains require consideration if the disease is to be more fully understood.

He invokes general systems theory (referring to Weiss and von Bertalanffy) as his preferred conceptual framework in order to overcome the “centuries-old limitation”
(Engel, 1980, p. 536) of the biomedical model, i.e. that a factor-analytic approach is necessary for the scientifically valid manipulation of data required for furthering our understanding of pathophysiology. In doing so, Engel argues for a broader concept of science – one in which a holistic approach also qualifies as “scientific”. Engel suggests that there is a hierarchy which may “be represented schematically by a vertical stacking to emphasize the hierarchy and by a nest of squares to emphasize the continuum. (...) Each level in the hierarchy represents an organized dynamic whole, a system of sufficient persistence and identity to justify being named” (ibid.). He stresses that the hierarchy of levels of organisation, in which the central relationship is that of part to whole, is also a continuum (Ibid., p. 537). For instance, a neurotransmitter can only be a neurotransmitter within the context of the nervous system and a patient is only a patient in virtue of being embedded in a social system in which this role or relationship exists.

Not only does Engel argue for a new conception of what qualifies as science, his emphasis on a hierarchy of levels of organisation also raises questions as to how “disease” is defined. The aim of the biomedical model was to explain disorders at the biological level of organisation. However, as the case discussed in Engel’s 1980 article highlights, the explanation must be sought at various levels, i.e. the biological, psychological and social levels. For instance, as Engel’s account of Mr Glover shows, his sense of responsibility at work affected his willingness to take on the patient role after the onset of his symptoms, and the unsuccessful attempt at arterial puncture in the hospital which worsened his angina (and perhaps even brought on the ventricular fibrillation). The above examples show how events at the psychological and social levels resonate with his physiological functions. The physiological events cannot be fully explained if the psychological and social aspects are neglected nor can the psychological and social aspects be reduced to biological processes.

So, in short, Engel’s ideas, demonstrated through a case study on Mr Glover, can be stated as follows: 1) there are various levels of organisation besides the biological level and 2) each level warrants the use of different methods for its study and explanation and 3) these levels cannot simply be reduced to the biological level. Therefore, 4) exclusive focus on the biological level will not yield full understanding of (psychiatric) disorders. To remedy this, we need to extend our regard to the psychological and social levels of organisation and their specific methods of study for a more complete understanding of disorders. This lands Engel with a new problem: how can we interpret and integrate our findings from the various domains in a scientifically valid manner in light of the premise that the one cannot be reduced to the other? His proposed solution is to invoke general systems theory where the focus is not only on the parts of a system (studied at just one level of organisation) but importantly, also on the interrelations between them.
4.5 General systems theory

Engel takes inspiration from “general system(s) theory” as defined by von Bertalanffy and Weiss. He explains that whereas the scientist can single out parts of the human, the physician cannot do so (and focus only on body and disease) without risk of neglecting the patient as a whole. Engel laments the pervasive idea of an antithesis between science and humanity, which results (in part at least) from the idea that “the factor analytic approach of reductionism alone qualifies as scientific” (1980, p. 536). Engel turns to systems theory to overcome this antithesis. In systems theory, Engel sees a way to allow that the methods required for studying “what makes for the wholeness” (of the cell, organ, person etc.), differ from those required for study of any of its parts or subsystems.

Engel invokes a hierarchy of natural systems, with the biosphere at the top, followed by society-nation, culture-subsociety, community, family, two-person, person, nervous system, organs, tissues, cells, organelles, molecules, atoms, and subatomic particles respectively. In another diagram he represents these systems as a continuum: “[e]ach system is at the same time a component of higher systems” (Ibid., p. 537, italics removed). He adds that “[t]he designation “system” bespeaks the existence of a stable configuration in time and space, a configuration that is maintained not only by the coordination of component parts in some kind of internal dynamic network but also by the characteristics of the larger system of which it is a component part” (Ibid.). Thus, Engel allows for reciprocity between parts within a system, and also between the parts and whole of a system. Taking into account this reciprocity and the complementarity of the aspects is of importance for both scientist and physician: Engel characterises it as a “guide” (p.538). Unfortunately - and I return to this point later when addressing shortcomings of Engel’s account - Engel does not provide further details about the relation between the levels of systems, the role of potential isomorphies (he only states that different systems require different approaches for their study) or mathematical methods for their integration. Engel also does not make explicit which role philosophy could play in the integration of the different levels on a systems theory approach. In later chapters I will look more closely at this problem of integrating the levels and possible solutions. I turn first to some of von Bertalanffy’s works to which Engel refers for an account of systems theory and give an introduction here to set the stage for a more detailed discussion in the following chapters.

In his 1952 book, *Problems of Life: An Evaluation of Modern Biological Thought*, von Bertalanffy reflects on the unity of science. One of the common traits of the sciences, on his account, is the “basically statistical character of the laws of nature and the intrinsic contradictoriness of reality” (Ibid., p. 176). The contradictoriness is found in
complementary conceptions such as wholes versus parts, organisation versus disorder, and the “resolution of rigid structures into dynamics” (Ibid., pp. 177-179). In the field of biology, mechanistic thinking motivated attempts to understand life in terms of isolable parts and processes. A major limitation of this view was that it reduced living beings to mere machines. One of the responses to the limitation of viewing living beings as machines was vitalism, the view that living beings are governed by different principles (for instance by purposive agents) than inanimate things. The obvious problems with vitalism in turn instigated “organismic conceptions” (Ibid., p. 181). These organismic conceptions, which is how von Bertalanffy terms his own account, seek to “give a scientific meaning to wholeness” (Ibid.). To do so, an understanding of how the parts themselves work, must be augmented with an understanding of how they interact. Von Bertalanffy differentiates between the laws governing the parts and those governing the interactions between them: “[n]ew laws appear at each level of organization” (Ibid.). Furthermore, the subject matter of biology is not static. Instead, there is a “ceaseless stream of matter and energy, passing through the organism and forming it at the same time”. The idea of a ceaseless flow hearkens back to Heraclitus, one of the forefathers of philosophy, who also described the unity of opposites addressed above.

An organism is thus an open system with a “dynamic morphology” which means that it has “an ordered flow of processes” (Ibid., pp. 185-6). Von Bertalanffy highlights the importance of regarding organisms as having a dynamic morphology for the fields of medicine and psychology. The view helps to clarify the relation between parts and whole within the organism and between organism and environment. For instance, an intelligent sentence is more than the sum of the meanings of its individual words; a living being is more than a lump of interacting cells (Ibid., pp. 189-190). Von Bertalanffy offers the example of Gestalt theory, which is based on the idea of psychical wholes, that in turn call for a dynamic, systemic approach for their understanding. The gestalt of an object or phenomenon can stay the same despite changes in its parts; different words may convey the same message just as a melody can be played in different keys. As stated earlier, elucidating how this is possible requires moving beyond simple addition of parts of the perceptual process (“perception is not a mosaic of single sensations” (Ibid., p. 192)) and zooming in on their interrelations and broader context too. The organisation of the system is due to “dynamic regulation” between the parts of which it is comprised. Dynamic regulation contributes to maintaining or restoring an equilibrium. Von Bertalanffy labels “equilibrium” the fundamental principle of the general system theory approach (1952, p. 193 & p. 199).

At the close of the book, von Bertalanffy delineates levels for description of phenomena:
I. Analogies: “superficial similarities in phenomena that correspond neither in the

2. Homologies: Although causal factors may differ, phenomena are “governed by structurally identical laws” (*Ibid.*).

3. Explanation: “statement of the conditions and forces present in the individual case, and of the laws following therefrom” (*Ibid.*).

He suggests that general system theory may help to distinguish analogies, which are “scientifically worthless” from homologies. Uncovering of homologies renders it possible to transfer laws applicable to one phenomenon to another phenomenon (perhaps even in an entirely different scientific field). Von Bertalanffy compares general system theory to the theory of probability: from a logico-mathematical point of view, both are purely formal and are applicable to a vast array of fields. For philosophy, the author suggests that general systems theory may even serve to replace ontology with “an exact system of general principles” (*Ibid.*, p. 201). The physicist for instance does not answer the question of what an electron really “is”. Instead, he states the laws that characterise what we collectively call “an electron”. Interestingly, in the field of psychiatry, the question of what a mental disorder “is”, is a central theme (see for instance Philips et al., 2012). Perhaps this is because the field, due to its inherent complexity, does not present researchers with clearly demarcated phenomena (such as electrons) for which to find descriptive laws in the first place. Chapter 2 of this thesis, which begins with a sketch of the history of the concept “apathy” (especially in schizophrenia), further highlights this point: it is difficult to define the phenomenon under study in the first place. In fact, the search for governing laws and delineation of the phenomenon under study frequently appear to co-evolve in the field of psychiatry. This observation is on par with the way experts in the field have set about developing the various versions of the *Diagnostic and Statistical Manual of Mental Disorders* (for instance DSM 5). This fits with von Bertalanffy’s earlier statement, that with new levels of organisation, new laws too must emerge for the description and delineation of key phenomena. In fact, von Bertalanffy suggests a systems theory approach to psychology and psychiatry, noting that it appears to “provide a consistent framework for psychopathology” because “[m]ental disease is essentially a disturbance of system functions of the psychophysical organism” (1968, p. 218).

Later, in an article written in 1972, von Bertalanffy expands on the types of systems for which systems theory is of interest. Systems theory has scientific, mathematical, technological and philosophical aspects. “The first may be circumscribed as *systems science*, that is, scientific exploration and theory of “systems” in the various sciences (...) and general systems theory as the doctrine of principles applying to all (or defined subclasses of) systems” (1972, p. 414). He states that in order to understand the whole,
one needs an understanding of the parts and of the interrelations between these parts. One way of understanding these interrelations is through describing them in mathematical terms “because mathematics is the exact language permitting rigorous deductions and confirmation (or refusal) of theory” (Ibid., p.415). Von Bertalanffy identifies “dynamic system theory” as a subtype of systems theory which “is concerned with the changes of systems in time” (Ibid., p. 417). Von Bertalanffy stresses that a systems approach may lead to one system having multiple descriptions, but that this is a healthy sign rather than being indicative of a problem (Ibid., p. 415). Frankl makes this point beautifully too, drawing on stereoscopic vision as an analogy:

“There is a difference between the right and the left pictures that are offered to you. But it is precisely this difference that mediates the acquisition of a new wholeness, of an additional dimension, the third dimension of space. To be sure, the precondition is that we achieve a fusion between the picture on the right and on the left. And what holds for vision is also true of cognition: unless we obtain a fusion, confusion may be the result”. (Frankl, in Koestler & Smythies, 1972, p. 396)

Systems technology pertains to the technologies required to deal with the problems that arise in “modern technology and society, including both “hardware” (control technology, automation, computerization, etc.) and “software” (application of system concepts and theory in social, ecological, economical, etc., problems)” (Ibid., p. 420). Von Bertalanffy points out that traditional branches of technology are no longer equipped to accommodate the level of complexity of modern technology and society. Interdisciplinary, holistic and generalist approaches are called for, and this is what “systems technology” entails.

As for “systems philosophy”, von Bertalanffy states that “we must first find out the “nature of the beast”: what is meant by “system,” and how systems are realized at the various levels of the world of observation. This is “systems ontology” (Ibid., p. 421), which is contrasted with “systems epistemology” and “values”. Within systems ontology, “real systems” (entities perceived in or inferred from observation and existing independently of an observer” (Ibid.)) must be distinguished from “conceptual systems” (“such as logic or mathematics, which essentially are symbolic constructs” (Ibid., p. 422)) and “abstracted systems” (“conceptual systems corresponding with reality” (Ibid.)). On the systems epistemology view, knowledge is an “interaction between knower and known, and thus dependent on a multiplicity of factors of a biological, psychological cultural and linguistic nature” (Ibid., p. 423). This renders science but a “perspective of man” (Ibid.).

Von Bertalanffy’s treatment of different levels for the analysis of systems and
The Biopsychosocial Model

implementation of general system theory, resonates well with elements of the hierarchy (specifically, the distinction between the philosophical and scientific perspectives) which I introduced in an attenuated form in Chapter 3. In the following chapter, I discuss to the distinction between the everyday, clinical, scientific and philosophical levels in more detail. I have adapted the hierarchy slightly to provide a schematic illustration of von Bertalanffy’s variants of systems theory (see able 4.3).

Table 4.3 Contexts for “General System Theory”

<table>
<thead>
<tr>
<th>Contexts for general system Theory</th>
<th>Meaning and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systems science</strong></td>
<td>Scientific exploration and theory of systems in sciences such as physics, biology, psychology and social sciences. Simply studying parts of systems in each of these fields and then adding them together again does not yield an explanation of the system as a whole. Complex systems have many, but not infinitely many parts. From the laws governing specific systems in specific fields, general laws, (based on homology/isomorphism) applicable to many systems across an array of sciences can be uncovered (von Bertalanffy, 1972, pp. 414-415). Systems science concerns wholes and wholeness, pertains to the “interrelations between a superordinate “whole”” (Ibid., p. 416) and often requires an interdisciplinary approach. Mathematics provides the language for the description of the laws of systems, because “mathematics is the exact language permitting rigorous deductions and confirmation (or refusal) of theory” (Ibid., p. 415).</td>
</tr>
<tr>
<td><strong>Systems Technology</strong></td>
<td>Modern society is so complex, that new, systems-based approaches to technology, spanning various disciplines and dealing with large numbers of variables, are required (Ibid., p. 420-421). Regarding things in the world from a systems-theory approach (i.e. systems science), requires the application of a systems-based technology. Examples are control and information theory, game theory, decision theory, theory of circuits. Concepts and models that originated in one discipline, may turn out to have much broader significance, for instance feedback loops can be used to describe biological processes as well as electric circuits: “they are independent of their special realizations” (Ibid., p. 421).</td>
</tr>
<tr>
<td><strong>Systems philosophy</strong></td>
<td>Systems philosophy pertains to the “re-orientation of thought and world view following the introduction of “system” as a new scientific paradigm (1972, p. 421). “System” is thus the new Kuhnian paradigm and replaces the parts-focused paradigm. Systems philosophy in turn includes three branches: Systems ontology Systems epistemology Values</td>
</tr>
</tbody>
</table>

4.6 The notion of science in the biopsychosocial model

In Engel’s “From Biomedical to Biopsychosocial: Being Scientific in the Human Domain”, an article written in 1997, the author calls for more explicit attention to humanness in medicine and reflects on the complementarity of the physician’s own experiences on one hand and his observations on the other, which are required to define which is necessary for “being scientific in the human domain” (pp. 521-522). He posits that
“to accommodate the human domain, science and being scientific must be redefined” (Ibid., p. 523).

In this vein, Engel points out three typically human attributes: first, that the meaning of being sick is deeply related to our awareness of the inevitability of death. Second, that we suffer when interpersonal bonds are compromised and are comforted when they are repaired. Third, we can examine our own “inner life and experience” and communicate about it. Our capacity for language brings together these attributes: it allows the physician to combine his “inner” experiences with observations of the “outer world”, i.e. the patient’s account. On perceiving a threat to their well-being, such as sickness, which, according to Engel usually begins as a private experience, “not necessarily knowable to anyone else” (Ibid., p. 525), most patients depend on language to communicate this. The patient’s own account is thus a primary source of data for the physician. This is especially true in the field of psychiatry due to psychiatrists’ dependence on clinical observation and anamnesis and the lack of other disease markers (such as imaging and biochemical options). Next, Engel draws on Heisenberg’s famous quote: “what we observe is not nature itself, but nature exposed to our method of questioning” (W. Heisenberg, 1958, p. 58). This quote holds true not just for physics, but also for medicine. Engel uses anamnesis as an example: the information a physician gets from a patient depends on the questions posed and the manner in which they are posed. Much like the relation between observer and observed cannot be ignored in physics, we cannot afford to neglect the relationship between physician and patient in medical practice (Engel, 1997, p. 523).

On Engel’s account, it is through the process of anamnesis that the physician tries to access the patient’s private world and considers the reliability of his account. Following the Heisenberg quote, anamnesis entails a reciprocity between physician and patient, rather than a unidirectional process in which the patient is merely an object of study. Engel argues that it is important to allow that this process is scientific, despite his acknowledgment of the fallibility of human accounts. Below, I explore Engel’s writings in an attempt to understand how Engel intends to include humanness into the realm of science.

In the 1997 article Engel delves into the notorious topic of where the limits of science lie. For instance, he poses the question of whether concepts such as “intimacy” and “narration” fall within the scope of science due to the lack of consensus on their definitions (p. 525). Yet, he also dismisses the notion, which he deems an “a priori assumption” of the biomedical model, that “patient-derived data and the means of their acquisition are neither teachable nor subject to systematic study” (Ibid.). Scientific
practice requires that we “understand and respect the natural state of whatever domain we are concerned with” (Ibid., p. 526). He contends that patient accounts contribute to precisely this process of understanding and respecting the natural state of the human being and therefore defends accepting them as scientifically valid.

In sum, Engel labels the biopsychosocial model as a conceptual framework that is more complete and inclusive to “guide clinicians in their everyday work with patients” (Ibid., p. 522). He concludes that the biopsychosocial medical model introduces the human dimension (back) into medicine and calls for a more flexible interpretation of what counts as scientific by advocating the inclusion of patient accounts as an acceptable, even indispensable source of data for the physician. However, it is still not entirely clear from his writings whether, or to what degree Engel differentiates between medicine as clinical practice and medicine as a scientific practice. I return to this point in the following section where I expand on various lines of critique on the biopsychosocial model and refer to another of Engel’s writings.

4.7 Concluding remarks: Biomedical versus Biopsychosocial

The biopsychosocial model overcomes the problems of the reductionist biomedical model because it does not embrace a specific level of explanation or organisation as the preferred level to which to reduce phenomena under study. In the previous chapter, I outlined how the biomedical model is founded on mind-body dualism. The biopsychosocial model is not based on the assumption that mind and body are of a fundamentally different essence. Although many adherents of the biopsychosocial model still implicitly have dualist tendencies, the biopsychosocial model promotes the view that mind and body are a continuum. Another important difference between the models is that the biopsychosocial model allows for and in fact makes explicit the inescapable normativity of both scientific and clinical practice and the humanness of researchers, clinicians and patients alike. Taking this normativity into account furthermore allows for a more nuanced way of differentiating between “sick” and “well” on the biopsychosocial model. Despite its merits, the biopsychosocial model as presented by Engel harbours some thorny issues. These are discussed in detail in the following chapter with a view to answering some of the central questions of this thesis.
Chapter 5

A critique of the biopsychosocial model
5 A critique of the biopsychosocial model

5.1 Introduction

This chapter zooms in on various tensions in Engel’s account of the biopsychosocial model as presented in the previous chapter. Earlier, in Chapter 4, I outlined the development of the biopsychosocial model in response to the shortcomings of the biomedical model, which I argued in Chapter 3 to be due to its inherent reductionism and dualism. It emerged that, at face value, the biopsychosocial model is a clinically relevant improvement on the biomedical model. Furthermore, it resonates well with the systems theory approach. However, there are important shortcomings, which may in part account for the pervasiveness of its biomedical precursor in scientific research, as demonstrated in Chapter 2. In this chapter, various lines of critique of the biopsychosocial model are discussed. Although I draw on other thinkers’ commentaries on the biopsychosocial model and treat different lines of critique, the chapter is ultimately geared towards trying to answer the first of the three central questions of this thesis and prepare the way for treating the other questions too. To recap, these questions are:

I. What is the utility of the biopsychosocial model at the level of a) philosophy, b) science and c) clinical practice?

II. How can we understand the use of the biological, psychological and social levels of explanation in science and account for the coherence and possibilities for integration of these different levels of explanation?

III. What are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy?

Below, I set out by expanding upon my preferred account of how to conceptualise the difference between the everyday, clinical, scientific and philosophical levels of analysis or perspectives. An attempt to answer the first central question requires determining how Engel differentiates between clinical practice, science and philosophy and whether he intends the biopsychosocial model as a clinical, scientific or philosophical model. I tackle these queries by close reading Engel’s works, providing ample textual evidence along the way and pointing to tensions in how he conceives of the difference between clinical and scientific practice, which he deems a false dichotomy, in particular.

The second line of critique concerns the implementation of a general systems approach. Guidelines for how to do this are conspicuously absent in Engel’s writings, which means that a lot of work remains to be done in order to further bring together the biological, psychological and social levels of explanation and move beyond pseudo-holism.
Third, I offer a brief treatment of various other lines of critique which are important for a balanced account of the problems with the biopsychosocial model, that yet only indirectly contribute to answering the central questions of my thesis.

The lines of critique can thus be summarized as follows:

1. Does Engel intend the biopsychosocial model as a clinical or scientific model?
   - How does Engel conceive of the difference between clinical and scientific practice?
   - Critique of his suggestion of a false dichotomy between clinical and scientific practice.
2. How would Engel implement a general systems approach?
3. How can users of the biopsychosocial model avoid an inclusive definition of disease?

Pertaining to the first line of critique, which is closely related to the first of my three central questions I contend that, in Engel’s writings, it is not immediately clear for which context the biomedical model was designed: was it intended as a clinical or scientific model? I provide an analysis of the scope of and relations between concepts such as “science” and “humanness” in his account and point out inconsistencies. After providing textual evidence that Engel at least intended the biopsychosocial model as a clinical model, I address the question of whether it also qualifies as a scientific model. Next, I explore and refute his contention that there is a false dichotomy between clinical and scientific practice based on the argument that it doesn’t follow from his line of reasoning. Furthermore, I draw upon the afore-mentioned hierarchy of perspectives or levels of analysis to illustrate my own interpretation of the distinction between scientific and clinical practice.

5.2 Four levels of analysis

In order to set the stage for my treatment of this aspect of Engel’s presentation of the biopsychosocial model, I first expand upon the different perspectives or levels of analysis that were introduced previously in the context of reductionism:

1. The everyday level
2. The clinical context
3. The scientific context
4. The philosophical context
These related yet distinct contexts help to structure my analysis of various notions, such as “reduction” (as in the third chapter), “knowledge”, and “practice”. “Knowledge” and manners for organising and accumulating it are important themes because both the biomedical and biopsychosocial models can be regarded as models for the organisation and accumulation of knowledge in the medical domain. I take “practice” to refer to the application or operationalisation of knowledge. In the table below, the use of the term “knowledge” and “practice” are outlined for the everyday, clinical, scientific and philosophical contexts.

Table 5.1 Levels of analysis.

<table>
<thead>
<tr>
<th>Contexts for “knowledge” and “practice”</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyday language</td>
<td>The action of acknowledging or owning something; acknowledgement, confession; an instance of this; acknowledgement or recognition of a person’s position, title, etc. (from: Oxford English Dictionary⁴).</td>
</tr>
<tr>
<td>Clinical practice</td>
<td>Acquaintance with facts, principles and truths through study or general investigation, pertaining to diagnosis and/or treatment of a patient and his or her disposition as a human (e.g. Glas, 1991, p. 35). Knowledge in the clinical domain concerns the individual person, it is about understanding patients. This requires practical knowledge, or skills and sometimes the application of technology required to obtain relevant information from the patient. The clinician combines knowledge obtained from the patient, through anamnesis, observation of non-verbal behaviour, manner of presentation, and further tests with theoretical medical (scientific) knowledge in order to diagnose and treat the person. Thus clinical methods and decisions are based on much more than just the result of experiments in which data are controlled, measured and quantified (e.g. Malterud, 2001).</td>
</tr>
</tbody>
</table>

⁴ http://www.oed.com/view/Entry/104170?rskey=MuBPxL&result=1&isAdvanced=false#eid
Clinical knowledge can be regarded as distinct from scientific knowledge because it pertains to the person as a whole. This is in contrast to scientific knowledge, which zooms in on the pathophysiology of a specific part of the human, for instance an organ, tissue, cell, protein or gene (cf. Schaffner, 1986; Montgomery, 2006; Glas, 1992). Scientific knowledge in medicine is often described in terms of a mechanism and for many researchers and clinicians this is associated predominantly with biomedical knowledge. Malterud, writing in the Lancet, exemplifies this intuition, stating that “[a] patient’s life, history, and feelings are not easily translated to biomedical variables and statistics” (2001, p. 398). Yet, being biomedical is not a necessary condition for knowledge to be scientifically valid. This was shown in Chapter 2, where outcomes of psychological tests and clinical observations too, contributed to testing hypotheses in apathy research.

Clinical practice and scientific practice are separate in this model, despite the fact that many clinicians themselves refer to medicine as a science. Patel, Evans, and Groen (1989) note that the traditional view of the scientific method, which does not reflect how physicians work, holds that “scientific theories consist of generalizations, logically deduced from a set of basic premises, which permit one to generate hypotheses that can be tested empirically, resulting possibly in further refinements to the set of basic premises” (pp. 54-55). Montgomery (2006) argues that calling medicine a science in the traditional positivist sense is a “misdescription” (pp. 29-30). The role of a clinician is very different to the role of the scientist. She observes that, in terms of its methodology and rationality, medicine has more in common with the social sciences than with the natural sciences. This is because the social sciences are better equipped to deal with uncertainty and complexity and are more sympathetic to including subjective experience and incorporating (inherently fallible) historical reconstruction (Ibid., Chapter 2). Montgomery dismisses the oxymoronic notion of medicine as the “science of individuals” (Ibid., p. 32), and settles instead for medicine as “science-using” practice.

Montgomery points to the tension in medicine, between art and science: either medicine must be neither or both at once. Yet, how can it be both at once, because science and
“Art” is generally associated with characteristically un-scientific phenomena such as “intuition”, “subjective skill” and “tacit knowledge” (*Ibid.*, p. 30). “Science” in turn is supposed to involve the “replicable, invariant, universalizable description of material reality” (*Ibid.*, p.31). She construes medicine as “phronesis”, or practical reasoning, in which scientific knowledge and well-practiced clinical skills melt together (pp. 33-34). This tension between art and science is also pointed out by Engel, who attempts to overcome it by subsuming the art-aspects of clinical practice under scientific practice. I provide a more detailed account, and critique of Engel’s approach to this tension below.

Clinical knowledge is the basis for understanding the patient and explaining his or her affliction. It requires taking into account and effectively combining the general laws, principles and mechanisms associated with disease that are subsumed under the heading of scientific knowledge, and combining this with knowledge about the individual patient and the context in which his or her affliction is occurring, i.e. it is, to use Montgomery’s term, *science-using*, but not scientific in and of itself. Clinical knowledge entails the ability (or skill) to interact with a patient in order to distil clinically relevant information from anamnesis, observation, and further tests (which in turn entails the skills to determine when and how to use technology to obtain relevant information). “Distilling” here serves as a metaphor for the complex process of interpreting information provided actively and passively, verbally and non-verbally etc. by the patient and perhaps family members and other carers. And, as Malterud adds: “The medical research tradition lacks strategies for the study of interpretive action, its dynamics and its consequences” (Malterud, 2001, p. 397).

Furthermore, Malterud in the same article points out that: “[c]linical interaction requires the understanding of particulars to be integrated with the understanding of universals. When medical knowledge generated from groups is applied to individuals, careful negotiations with the specific patient and situation are essential for adequate understanding and management. A patient’s life, history, and feelings are not easily translated to biomedical variables and statistics” (p. 398). Although Malterud seems to lament the lack of methods to scientifically study the clinical interpretative process, her biomedical orientation, which Engel would undoubtedly refute, surfaces with her suggestion that we ought to try and translate biographical and emotional aspects into biomedical variables.

For the sake of thoroughness, I take Malterud’s contention a step further by suggesting that interpretation of information in the clinical setting is a selective process, guided not just by the patient’s manifestation and the clinician’s clinical experience, but also
A critique of the biopsychosocial model

by assumptions based on the clinician’s own context in a far broader sense. Groopman, in “How doctors Think” (2008) describes a number of errors which highlight just how challenging it is for the physician to transcend his context. He outlines cognitive processes such as “anchoring”, “affective error” and “availability” on the part of the physician which lead to confirmation bias in clinical practice. Confirmation bias entails “confirming what you expect to find by selectively accepting or ignoring information” (Ibid., p.65). “Anchoring” refers to settling on one possibility at an early stage rather than considering multiple possibilities. Affective error occurs when the physician’s feelings towards a patient, be they positive or negative, cloud their ability to gather and survey the available data. “Availability” refers to a physician’s tendency to opt for possibilities (concerning for instance diagnosis and/or treatment) that come readily to mind. Each of these errors can lead to confirmation bias if settling on and pursuing just one possibility occurs prematurely because the physician is not aware of his own psychological processes. To avoid these potential errors, Groopman prescribes reflection of the physician on his own role and context and defends the importance of a thorough and patient-centered anamnesis. Glas further broadens the “context” to include aspects from the physician’s personal life and biography which percolate into the way he fulfils his role as a physician (Glas, 1992; cf. Montgomery, 2006; Groopman, 2008). For instance, a general practitioner who, based on his religious background, is against abortion, will likely offer different advice to pregnant woman seeking advice about abortion than a clinician who is not fundamentally opposed to abortion. A second example: a psychiatrist who in his personal life is dealing with the grief of losing a loved one to suicide, is likely to be more sensitive and risk-averse in dealing with a patient struggling with suicidal tendencies. These examples serve to highlight that, like his patients, the clinician too is primarily a human being, situated not only in his professional role as a physician but also embedded into his broader life context. This humanness of the clinician entails that his own values percolate into the way he fulfils his professional role (and vice versa). Clinical practice is inescapably normative and therefore, clinical knowledge and practice are inherently value-laden (cf. Glas, 1992, Chapter 2 & 6).

The normative and individualistic aspects of clinical practice are, due to their complexity, notoriously difficult to render in terms universal enough to afford the label “scientific” in the sense adhered to by biomedical thinkers. I agree with Engel that this should not lead to the conclusion that rendering them in biomedical terms is the solution to making them “scientific”. I return to the problem of integrating non-biomedical aspects into the scientific domain and the potential of a general systems theory approach to this project in more detail later.
5.3 Back to Engel on the aim of the biopsychosocial model

Engel contends that the biopsychosocial model is designed to “guide clinicians in their everyday work with patients” (1997, p. 522). He contrasts this to the primarily scientific utility of the biomedical model (Ibid.). We can therefore safely conclude that he is positioning the biopsychosocial model as a model for organizing the knowledge relevant to clinical practice. However, as pointed out, he also wants to use the model as a basis for bringing humanness back to science. He links these two claims by calling the distinction between scientific and clinical practice a “false dichotomy”. With this move, he seems to want to legitimize his allusion to science in the context of clinical practice. Yet, his bold refutation of a dichotomy raises further questions pertaining to his underlying philosophical assumptions concerning the scope of science. Moreover, if there is no distinction between clinical and scientific practice, can we infer that the biopsychosocial model is intended to serve as a scientific model too?

To address the questions above, an examination of how Engel characterizes the relationship between science and “humanness” is called for. As noted, Engel deems the biomedical idea that “only what can be measured in the laboratory truly qualifies as scientific” (Engel, 1987, p. 108) overly narrow. Medical science, which in turn informs clinical practice, requires the inclusion of human data. Human nature is “no less accessible to scientific inquiry than any other natural phenomena” (Ibid.). For Engel, the consulting room and bedside should therefore be regarded as the medical domain’s laboratories: “meaning is preeminent to measurement” (Ibid., p. 110). Engel introduces the distinction between the physician-scientist, who is primarily concerned with medical research, and the scientific physician, a clinician who fully applies the scientific method in patient care and his understanding of disease (Ibid., p. 111). The scientific physician, Engel claims, is better equipped to combine scientific knowledge and skill in the human domains to effectively help patients.

Figure 5.3 View refuted by Engel.
In his 1987 article, “Physician-scientists and scientific physicians: resolving the humanism-science dichotomy”, Engel addresses the scope of “science”. He contests the pervasive notion that “[s]cience pertains to the study of the body and its diseases” (1987, p. 107), and that the psychological and social aspects of disease are commonly regarded as “outside the bounds of science” (Ibid.). Engel’s view seems to contrast the (contested) view ascribed by Ghaemi to William Osler. Osler was a Canadian physician, one of the founders of the Johns Hopkins Hospital, practicing in the late 19th century. According to Ghaemi (cf. Ghaemi, 2010, pp. 129-144), although other authors downplay the humanism in his view of medical practice (e.g. Bliss, 1999), Osler proposed adding, rather than including medical humanism, to the scientific aspect of medical practice. The figures above demonstrate these two positions graphically. Engel furthermore rejects “relegating to the art of medicine” all that is involved in the physician's human interactions with the patient, be it his so-called “bedside manner” or his consoling of the bereaved” (1987, p. 107, my italics). Yet why then, does he regard calling the skills needed to interact with a patient an “art” as negative and why would this label exclude the human dimension from science? Engel answers this question in the concluding part of the article. He states that most scientific physicians’ knowledge and skills in the human area derive from intuition and thus are based on experience and “innate personal characteristics” (Ibid., p. 111), rather than being based on systematic, testable principles. This renders the approach of these scientific physicians difficult to communicate, test and teach. For Engel, this intuitivism is an obstacle to the advance of our scientific understanding of the human aspects inherent in medicine.

Unfortunately, he does not in any of the articles cited here, offer a system of laws or principles which would contribute to rendering the physician-patient interaction more teachable, nor does he explicitly suggest that general systems theory may offer such a method. Given his own psychotherapeutic training, it is somewhat surprising to note that he refrains from further systematizing the physician-patient interaction as is often done in psychotherapy. A significant part of psychiatric training for residents
specializing in the field consists of developing techniques required to understand (and perhaps also influence) the patients’ experiential world based on their verbal accounts, obtained through physician-patient interaction. It would seem that psychiatry is one of the medical fields in which the centrality of the physician-patient relationship is taken for granted.

In order to return humanness to science, Engel advocates regarding patients’ own verbal accounts of their symptoms, obtained through the reciprocal relationship between patient and physician, as sources of data. Although Engel himself does not explicitly refer to von Bertalanffy’s general systems theory in this context, Engel’s account resonates with von Bertalanffy’s systems epistemology, according to which science is but a “perspective of man” (Engel, 1992, p. 423) and for which knowledge is an “interaction between knower and known, and thus dependent on a multiplicity of factors of a biological, psychological cultural and linguistic nature” (Ibid.). As discussed, Engel acknowledges the limitations of human memory and the fallibility of first person accounts, yet argues that these accounts are nonetheless indispensable and need to be regarded as scientifically valid.

It is worth noting that Engel’s advocacy of the use of data from for instance the psychological and social levels of organisation is not incompatible with biological reductionism. Chapter 2 provided various examples of this, with researchers exploring possible correlations between phenomena at the psychological or social levels and the biological level. The use of terms such as “neurobiological underpinnings” or “underlying” are suggestive of this manner of thinking. For instance, the quote in Chapter 2 from Chue and Lalonde writing about apathy in schizophrenia (2014): “[a]lthough the cellular, molecular, and circuit alterations underlying these symptoms are not clear, studies focusing on motivation, reward, working memory, and goal-oriented behavior may help shed light on the circuitry that is altered in negative symptoms” (p.778). What is different in the biopsychosocial approach is that it does not require reducing this data to biomedical terms.

To summarize, we can conclude that Engel intends the biopsychosocial model as a model aimed at clinical practice. Also, Engel regards clinical practice and the use of first person accounts, which result from an effective patient-physician interaction, as sources of data and advocates scientific study of this data. Thus, on his view, scientific practice need not be limited to or reduced to the biological level of organisation, but may (or must) include data from the psychological and social domains (including psychological and social aspects of the physician-patient encounter) too. Note that the psychological domain involves not only first person accounts, but can also be taken to include
results from psychological tests, the physician’s observations concerning the patient’s psychological functioning etc. Labelling the interactional aspects of clinical practice as open to scientific study, for him entails that it is not some black box procedure or “art” during which a diagnosis and treatment plan are mysteriously drawn from the patient presenting with symptoms. Clinical practice involves skills which on his account can be studied and also, and importantly, taught scientifically.

5.4 The false dichotomy

In the 1987 article, Engel advocates a “more inclusive scientific paradigm” (p. 111), moving beyond the distinction between medical practice and medical science, which he calls a “false dichotomy”. The notion referred to as “science” in both Engel’s 1987 and 1997 articles is limited not just to biomedical research performed in the laboratory - recall that he advocates the addition of the consulting room and bedside to the scientific domain’s laboratories.

I object to Engel’s contention that there is a dichotomy between the scientific and clinical perspectives that is available for being deemed “false” in the first place. At a glance, it may seem that I am projecting on Engel’s account precisely the distinction that he is seeking to excise. However, on closer examination, the dichotomy he is refuting does not resonate with the way that I have described the clinical and scientific perspectives. On my account, borrowed from Glas (1992), the clinical and scientific levels of analysis are not mutually exclusive, they are simply complementary, mutually informative perspectives. Dividing the spectrum of perspectives into hierarchical layers is an interpretative method to ease the discussion of the various aspects at play in the field of medicine. This view is thus perfectly compatible with Engel’s suggestion that what he calls the “humanness” in medicine can be studied scientifically (even though, some three decades after his publications on the topic, we still seem short of widely-accepted methods for its examination as is demonstrated by Malterud’s article). In fact, the hierarchy of levels of analysis account does not exclude any subject or part thereof from scientific study.

In sum, Engel’s coining of a “false dichotomy” appears to be an attempt to simplify his account. This attempt, unfortunately has exactly the opposite effect because it raises many of the above-mentioned questions concerning his intuitions about the status and boundaries of science and clinical practice. His claim, in short, is two-tiered: 1) The biopsychosocial model is a clinical model 2) in order for it to be an effective clinical model, the scope of science needs to be extended to include “human” aspects of medicine and do away with the tension between medicine as an art and medicine as a science.
Chapter 5

This is evident when he states: “to accommodate the human domain, science and being scientific must be redefined” (Engel, 1997, p. 523). Note that when Engel refers to “humanness” or “the human domain”, I take him to intend the psychological and social factors that are relevant to the individual patient’s medical problem in both his everyday existence and also within the context of the physician-patient interaction). Doing away with any dichotomy is thus not necessary for him to make his case, because all he calls for is the inclusion of the social and psychological domains of human functioning in what we accept as “science”. Therefore, I argue that his “false dichotomy” is misattributed and does not follow from his line of reasoning. The dichotomy Engel invokes first suggests to me that he misconstrues the difference between clinical and scientific practice rather than embracing them as complementary perspectives (as I would advocate). This runs parallel to the tension between regarding medicine as an art and medicine as scientific practice, to which Engel responds by trying to render the artistic aspects scientifically rather than regarding them as complementary. And second, it suggests that Engel conflates the use of scientific knowledge in clinical practice and the openness of clinical practice itself with scientific study. To remedy the problem, I suggest embracing Montgomery’s contention that medicine is not primarily scientific practice but is instead primarily a science-using practice (that is art-like because it involves tacit knowledge, intuition and experience) which, I add, is in itself open to scientific study.

Figure 5.5 Perspectives on a patient.

Having established that Engel’s biopsychosocial model seeks to provide an alternative to the reductionism and dualism of the biomedical model and that it is primarily aimed at clinical practice, which Engel regards as a scientifically relevant activity in itself, a new question presents itself. What value does the biopsychosocial model have for science and furthering of our understanding of medicine?
5.5 Intermezzo: Evidence Based Medicine

Although the introduction of Evidence Based Medicine (EBM) is something of an intermezzo in this chapter, it serves to further contextualise Engel’s thinking and prepare the way for a treatment of the question about the use of the biopsychosocial model in scientific practice. This is because EBM is posited as a way of applying scientific knowledge to clinical practice. The term EBM was coined in the 1980s and rose to popularity in 1990s. Many contemporary medical programmes advocate an EBM approach to clinical practice. It is defined as “the process of systematically finding, appraising, and using contemporaneous research findings as the basis for clinical decisions” (Rosenberg & Donald, 1995, p. 1122). These authors contend that “[e]vidence based medicine can be taught to, and practised by, clinicians at all levels of seniority and can be used to close the gulf between good clinical research and clinical practice. In addition it can help to promote self-directed learning and teamwork and produce faster and better doctors” (Ibid., my italics).

For example, a psychiatrist treating a 24-year-old female with depressive symptoms for which she is taking sertraline, whose symptoms nonetheless worsen in the pre-menstrual phase may consult the scientific literature and find that a correlation between lower oestrogen levels (as in the pre-menstrual phase) and depression has been described. He may furthermore find evidence supporting a temporary dose-increase of sertraline, addition of an oral contraceptive to decrease hormonal fluctuations or addition of alprazolam in the pre-menstrual phase to attenuate depressive symptoms (evidence used in this example: Beers, 2016). The clinician can then discuss these options with the patient and select the most appropriate option depending on patient-specific factors (such as desire to conceive, predisposition to benzodiazepine abuse, regularity of menstrual cycle for augmentation of sertraline etc.).

The above example shows how an EBM approach can accommodate an individualised approach. Yet, attempts to base all medical practice on available evidence are doomed to fail because often, evidence which is applicable directly to the individual’s complex clinical question, is lacking. The clinician is then forced to infer from whatever evidence that is available, what to do in a specific case. Take for instance the case of a 48 year-old-female suffering from therapy-resistant schizophrenia (with persistent acoustic hallucinations and negative symptoms which severely affect her quality of life) and cannabis and amphetamine abuse, who experienced a very marginal decrease in her hallucinations one year after commencing clozapine with aripiprazole (a potent combination of anti-psychotic drugs) and remaining abstinent from cannabis and amphetamine. The clinician and patient were confronted with a dilemma because the patient became morbidly obese (+30kg), probably due to clozapine and had to decide whether or not to continue clozapine. The dilemma in short, is between marginal reduction of symptoms.
versus obesity and metabolic health problems. In this case, there is little direct evidence, other than studies which compare effects of different anti-psychotic drugs and weight-loss regimens, to help guide the clinician: the process of weighing the different factors at play for this patient (quality of life with less hallucinations but severe obesity versus more hallucinations but fewer metabolic side-effects) falls to the clinician and patient.

Thus, despite attempts within medical programmes to also introduce biopsychosocial thinking, the EBM approach seems more reticent of the biomedical approach, simply because much of the currently available scientific evidence is biomedical (randomized controlled trials are widely regarded as the gold standard) and does not take into account factors from other life domains or provide guidance how to do so. In this vein, one common criticism for the EBM doctrine is that it alone cannot define good medical practice for it neglects the individuality of the patient. To use Montgomery’s words, it offers a way to make clinical practice “science-using” (Montgomery, 2006, Chapter 3). This fits well with Engel’s pledge for the inclusion of humanness in science.

5.6 How would Engel implement a general systems approach?

Tyreman (2014) laments that in teaching the biopsychosocial model to medical students as part of their standard curriculum, the backbone of the model, general systems theory, is missing, which is why the model flops in practice. Fink (1988) too, mentions the risk of “biopsychosocial” becoming an empty catchword, popular in the medical curriculum, but with little methodological bite, because residents are stimulated to discuss each of the domains separately rather than in a more integrated fashion as Engel intended.

My second line of critique of the biopsychosocial model thus pertains to the place of general systems theory, which was suggested by Engel as the basis for integrating the domains in the biopsychosocial model. More specifically, I argue that Engel does not adequately specify how the systems approach should integrate the biological, psychological and social levels of organisation. This point ties in to the previous line of critique where I pointed out the difficulties of an EBM approach to clinical practice: how are factors from different levels of organisation weighed and compared in practice? How can we determine which factors do and do not bear on the phenomenon under study – i.e. the question of causal relevance? Engel offers no practical guidelines for these issues. Yet, his insistence on including biological, psychological and social data in the approach to a patient raises questions concerning the interdependence of these levels of organisation and how to conceive of that interdependence to make it useful as a tool for clinical (and scientific) practice. Jonathan Bolton (2014) notes that the biopsychosocial model is stratigraphic and that as soon as the levels are distinguished, they are treated as
A critique of the biopsychosocial model

complete and autonomous and hence the problem of reconnecting them immediately emerges (p. 180). Bolton goes on to argue that the incommensurability of the types of knowledge at different levels of organisation is a problem for the biopsychosocial model \((Ibid., \ p. 181)\). For instance, a phenomenon such as consciousness pertains to the person level and cannot be traced back to the level of organs, tissues or cells simply because these sub-systems are themselves not conscious. Trying to identify consciousness at the cellular level would constitute a mereological fallacy. My response to Bolton’s argument therefore, is that Engel called upon general systems theory precisely to avoid such mereological fallacies and integrate knowledge in a way that does justice to the unique nature of level-specific phenomena.

In the face of what he calls the “failure” of the biopsychosocial model, Bolton goes on to argue for a “4p” case formulation. Although Bolton does not explicitly say so, his approach seems a valuable addition to the biopsychosocial model, rather than an alternative. On the “4p” approach, which is very suitable not only for psychiatry but for other medical specialties too, the clinician formulates a case in terms of the preconditions, precipitating factors, perpetuating factors and protective factors (four “p’s”) at play \((Ibid., \ p. 182)\). Outlining these factors seems a useful way to avoid the biomedical model’s biological essentialism and also allows for the integration of knowledge from each of the organisational levels in the biopsychosocial model in the process of understanding disease.

Moving the biopsychosocial model beyond a pseudo-holistic model requires reconnecting its levels. We therefore need a scientifically valid or at least scientifically accepted method for the knowledge-weighing process, which appears to be largely interpretative at present (as demonstrated by the vignette about the 48 year-old female with schizophrenia and Engel’s own study of Mr. Glover \((Engel, 1980)\)). For instance, in the case of Mr. Glover, Engel describes a wide array of aspects from the psychological and social domains which, he suggests, play into the series of events that took place during and after his heart attack. Yet, how can we determine that Mr. Glover’s personality traits affected his decision to delay asking for help and not some other factor (perhaps his mobile phone had run out of battery power or he was too engaged in his work to fully acknowledge the situation)? Answering such questions calls for a notion of causality which can 1) transcend the explanatory levels within systems and 2) is testable in a scientifically valid manner. The word “weighing” used by Engel and myself suggests a numerical approach to this question. But besides a mathematically sound method, a coherent philosophical basis for integration is also called for. Bolton (2014) makes this point by highlighting that general systems theory as described by von Bertalanffy is a meta-theory. It is a theory about systems in general and does not offer any concrete guidance about how to
approach an individual patient.

Interestingly, Perring (2014), in a response to Bolton’s article about the 4p approach, challenges the idea that a “grand theory about the relation of different levels of description” (Perring, 2014, p. 201) is necessary in the first place. Perring’s reason is that Bolton’s primary aim in developing the 4p approach is to focus on and provide a “thick story” about the personal level in clinical practice. My response to Perring is that exploring the relations between different levels of organisation (such as the biological, psychological and social) is not the same as aspiring to a grand unified theory. In fact, the biopsychosocial model allows for agnosticism when it comes to choosing a preferred level of explanation and allows for multiple explanations at different levels to co-exist and complement each other. For, as mentioned in Chapter 4, von Bertalanffy (1972) contends that one system having multiple descriptions is a healthy sign (p. 415).

So in sum, I agree with Engel that a method for uncovering the relations between descriptions at different levels of organisation is called for and that systems theory may provide the foundations for doing so. The main difficulty stems from the fact that systems theory is primarily a very abstract, meta-theory about what a system is, which renders it open to multiple interpretations and operationalisations. In the following chapters I zoom in on the notions of causation and integration and attempt to gather ingredients for a positive account, or in other words, an application of general systems theory, for the mathematical and philosophical integration of the levels of the biopsychosocial model.

5.7 Answer to the first central question of the thesis

The above treatment of two central lines of critique provides the basis for my answer to the first central question of this thesis: what is the utility of the biopsychosocial model at the level of a) clinical practice b) science and c) philosophy? I conclude that the biopsychosocial model, as originally outlined by Engel, despite its limitations, largely succeeds as a model for clinical practice. In the clinical setting, it serves as a useful tool to remind the clinician to take a holistic approach rather than a biomedical approach by paying attention to psychological and social aspects of the phenomenon under study and it also helps to provide a framework for further study of the application of clinical knowledge by the physician in which not only the physician's biomedical knowledge, but also the human aspects of the physician-patient interaction are featured. However, the model faces grave problems as a scientific model. These problems are due to the model’s inability to account for the integration of the data from the biological, psychological and social domains in a scientifically valid manner. Finally, as a philosophical model, the biopsychosocial model is underdeveloped. Although Engel criticises the dualism and
reductionism inherent in the biomedical model which he was seeking to replace and although he points to systems theory (both its mathematical and philosophical tenets), he does not clarify how the philosophical assumptions upon which systems theory is founded translate to the biopsychosocial model.

### 5.8 Definition of disease

A line of critique found in the literature on the biopsychosocial model is that it harbours the risk of an over-inclusive notion of disease (e.g. Ghaemi, 2010, p. 91). In Chapter 4, I noted that one key theme in an early article by Engel (1960) was the lack of a sharp divide between sickness and health in medicine in general (p. 469). Engel (1997) proposes the biomedical model “as a more complete and inclusive conceptual framework to guide clinicians in their everyday work with patients” (p. 522). It is interesting that Engel seems to accept that differentiating more clearly between sickness and health is not one of the most central aspects of practicing medicine, because he does not return to the topic in any of his articles about the biopsychosocial model. Perhaps he concedes that this difficulty is inherent to medicine and insurmountable on any medical model. Perring (2014), in debating the relevance of overarching models for medicine in the first place, points out that there is a difference between a definition and a model. A **definition** is analytic and offers necessary and sufficient conditions for what counts as “disease”. A **model**, in contrast, does not clearly do this (Ibid., p. 199), leaving us without a clear positive account of what “disease” is. Thus, if we adhere to the biopsychosocial approach, according to which being ill is no longer dependent on there being an abnormality at the biological level, how then should we define “disease”? As stated, the biopsychosocial model does not provide guidance on this front. The question then arises whether any problem or abnormality in any life domain can warrant being labelled a “disease”. For instance, how do we deal with situations in which the patient presents with a complaint for which no corresponding abnormality can be found at the biological level?

Differentiating between sickness and health is notoriously difficult in psychiatry. As noted in Chapter 4, Szasz (1961) claimed that, on the biomedical model, mental illness does not meet the criteria for disease and is therefore a “myth”. Ludwig (1975), too exemplifies the problem of differentiating between sick and well, by stating that the field of psychiatry should be limited to those behavioural disorders that ostensibly or presumably arise from brain dysfunction or other “natural causes” (as opposed to “metapsychological, interpersonal or societal causes”) and therefore excluding disorders such as “problems of living, social adjustment reactions, character disorders, maladaptive learning patterns, dependency syndromes, existential depressions” (p.603).
The risk over over-inclusiveness of the biopsychosocial model is closely related to another objection: the dangers of relativism. Because Engel advocates taking into account factors from the biological, psychological and social levels of organisation, it becomes difficult to avoid the downfalls of relativism, where one perspective has the same truth value as the next. Extreme relativism makes analysis and weighing of factors and choosing a perspective on pragmatic grounds impossibly difficult. As for instance, in the case of a young male who is convinced that his former employer has implanted a chip in his brain, which the employer uses to cause a very wide variety of anomalous sensations in his body, thus controlling his behaviour. He is suffering deeply from medical complaints such as vibrations in his foot, constipation, weight gain and muscle aches which he ascribes to the supposed chip in his brain. He has already consulted many different medical specialists and undergone extensive medical testing (to no avail) and has made many life-style changes, including plans to sell his house in an effort to escape his former employer's powers. The psychiatrist believes the patient is suffering from paranoid schizophrenia and wants to offer the patient psychiatric treatment, including antipsychotic medication, to relieve his suffering. The patient does not believe he has a mental disorder and instead insists on getting more medical tests to locate and remove the chip. He furthermore wants legal help to force his former employer to leave him alone. A purely relativistic approach runs the risk of leading to an “anything goes” situation, in which the doctor and patient perspectives are taken at face value and there is no way of determining the best course of action. For obvious reasons, this would severely impede medical practice.

In short, for the field of medicine and psychiatry in particular, I conclude that whereas the biomedical model risks ruling out as disease many afflictions that most psychiatrists would readily accept as “diseases” for a lack of a biomedical abnormality, the biopsychosocial model in contrast offers little guidance as to which of the problems in various life domains count as disease and which do not. This line of critique, however, is part of a much larger discussion on the definition of mental disorder (for an overview of approaches to defining mental disorder, see for instance Philips et al., 2012) and a further treatment of the differences between a biomedical and biopsychosocial approach to this challenge is beyond the scope of this project.

5.9 A missing dimension

The final line of critique of the biopsychosocial model discussed here is that, despite Engel’s lobby for including the psychological and social domains, the model is still lacking one or more domains. Sadler and Hulgus (1992) claim that the biopsychosocial model is of limited utility for the process of clinical decision-making because it “cannot
A critique of the biopsychosocial model

account for clinical problems for which the methods of science do not apply” (p. 1315) and therefore propose complementing the biopsychosocial approach with a three-faced approach involving epistemic, ethical and pragmatic aspects. This would result in the addition of various disciplines to the biopsychosocial model, including anthropology, health law, ethics, philosophy, informatics, communications theory, and business management (Ibid., p. 1322). Other thinkers have proposed to add extra levels of organisation or perspectives to the biopsychosocial model too. Sulmasy (2002), for instance, suggests adding the spiritual domain. Sanneke de Haan, an enactivist philosopher (2013), advocates including an existential layer in the biopsychosocial model (see also Freudenreich, Kontos & Querques, 2010). This existential layer cannot be reduced to either the biological, psychological or social levels or dimensions and therefore warrants separate mention. Including attention to the existential aspects of being ill would further our understanding of disease, by exploring how the patient and people in his or her environment relate to and make sense of disease and the process of being ill. Interestingly, the existential dimension features centrally in the field of psychiatry, where the concept of recovery and the strengths model have gained significant terrain in recent years (Delespaul, 2013). For instance, exploration of the existential dimension may help elucidate why some people who have experienced extremely traumatic events go on to develop post-traumatic stress disorder (PTSD) whereas others do not. There is literature that suggests that how one interprets one’s traumatic experience is one of the strongest predictors for the development of PTSD. Or, why does a proportion of the patients that suffer from anxiety-provoking hallucinations report more intense hallucinations whereas others find a more effective way to live a meaningful life despite their symptoms (e.g. Delespaul, de Vries & Os, 2002)? The existential domain can be drawn upon to answer this question too. Yet another example of the importance of the existential domain is provided by Irvin Yalom, an American emeritus professor of psychiatry who is frequently described as an “existential psychiatrist” and who has written extensively on the topic of existential psychotherapy (e.g. Yalom, 1980). De Haan therefore points out that the existential domain is conspicuously absent from the biopsychosocial model, especially in the field of psychiatry.

In sum, there is a large body of literature containing arguments why certain extra disciplines warrant adding to the biopsychosocial approach and it is beyond the scope of this project to provide an exhaustive account of them. Some thinkers (e.g. Ghaemi, 2010) lament the eclecticism of the biopsychosocial model whereas others (e.g. Sadler & Hulgus, 1992; de Haan, 2013; Sulmasy, 2002; Freudenreich et al., 2010) seek to make it broader in order to do justice to the complexities of clinical practice. Many thinkers are sympathetic to Engel’s systems theory approach but point to the biopsychosocial model’s limited clinical applicability. This line of critique ties in to the previous objection
Chapter 5

cconcerning the risk over over-inclusiveness and relativism, difficulties differentiating between sickness and health and the issues concerning the integration of different levels of organisation and explanation.
A critique of the biopsychosocial model
Chapter 6
Philosophical Models of Explanation
6 Philosophical Models of Explanation

6.1 Introduction

In the previous chapters, I have taken a largely bottom-up approach, departing from current scientific practice, drawing on the literature on apathy in schizophrenia as a case study, and analysing this for the presence of explanatory models such as the biomedical and biopsychosocial model. Having found the unity of scientific insights on apathy to be “mosaic” – to adopt Craver’s term - at best, I moved on to examine the biomedical and biopsychosocial models in more detail, outlining the contexts in which they were developed, zooming in on their advantages and pitfalls in search of a deeper understanding of their philosophical foundations.

It is now time to make a transition to the philosophical literature on scientific explanation to explore whether philosophical models for scientific explanation can be used to fill the gaps in the philosophical infrastructure of the biopsychosocial model. Many philosophers attempting to first float their models of scientific explanation did so in the waters of physics, i.e., they used examples from the field of physics to demonstrate the validity/applicability of their models. However, models that appear to fit scientific practice in physics well, are often a poor fit for other disciplines, such as biology, psychology or medicine. It seems that different disciplines or scientific perspectives call for different models of explanation. My aim here is to examine what is required of a model of explanation for the field of psychiatry in particular.

This chapter is structured as follows: first, I briefly address what it is to explain. Second, five constraints, borrowed from Salmon (1984) and Craver (2007) on explanation are introduced and I explore how these constraints would work in the field of psychiatry. Third, a brief overview of three traditional approaches to explanation in science is provided including a discussion of why they are inadequate for psychiatry. Fourth, I touch briefly on the application (or, more accurately projection) of proposed models onto empirical practice. Within this line of discussion, the question of how to distinguish between good and bad explanations arises. How can philosophical models of explanation provide means to make this distinction?

In a nutshell, some key areas of discussion in the literature on scientific explanation are:

1. What is it to explain?
2. Constraints on (a good) explanation
3. Models of scientific explanation
4. Application of these models to scientific practice
5. Scope for application within different scientific fields

6.2 How does an explanation explain?

There is a vast, multi-facetted literature on the topic of scientific explanation. Terms such as “prediction”, “description”, and “intervention” feature centrally in this debate. Woodward (2003) for instance holds that “we are in a position to explain when we have information that is relevant to manipulating, controlling, or changing nature” (pp. 9-10). To this end, an explanation should show not just what happens, but also how it occurs (e.g. Salmon, 1984, p. 4). Thus, in many fields of science, it is customary to differentiate between description (or classification) on one hand and explanation on the other. Most thinkers hold that description is an aspect of explanation that does not in itself suffice for explanation.

The debate on explanation moves beyond description of phenomena and taps into the thorny philosophical issue of causality. Causal notions are pervasive in everyday language, clinical practice and scientific parlance.

A first question to pose is: do we need an account of causality to develop an account of scientific explanation? Huneman (2010) for instance argues for what he calls “topological explanation”, a non-causal, non-mechanistic model that explains through topological properties. Sober (1983) and Nerlich (1979) too, advocate non-causal forms of explanation. Many other thinkers, however, embrace the idea that explanation must draw on causal relations (e.g. Salmon, 1984; Woodward, 2003; Pearl, 2000).

I agree that reference to causality is helpful in the debate on scientific explanation. However, I stress that I do not seek to provide a definition of causality and furthermore allow that different models of explanation may each have their merits in different situations. This context-dependence resonates well with De Regt’s account of scientific understanding (2017, Chapter 4). He advocates a “contextual theory of understanding”. His approach is pluralistic in that it does not rely on a single model of scientific explanation and holds that different models of understanding can complement one another.

I suggest an agnostic stance regarding the metaphysical aspects of the debate on causality. On this view, it suffices to apply the causality concept to refer to our perception of a type of relation played out in the temporal domain that helps us understand why entities interact in the way we perceive them to and/or helps us understand how a system changes in time. On my interpretation, Salmon’s constraints (borrowed by Craver and...
Chapter 6

introduced in more detail below) can help to axiomatize “causation” without further defining it. I argue for this move in more detail in Chapters 8 and 9. In the next section the five constraints for explanation are outlined.

6.3 Constraints for a good explanation

Craver (2007) draws on Salmon (1984) and articulates an account of “causal mechanical” or “mechanistic explanation” based on five constraints. They form the foundations for his account of mechanistic explanation which he applies to various examples from the neurosciences. Craver suggests that the constraints go beyond merely sorting explanatory models from non-explanatory models. They are designed to sort good explanations from bad ones – thus introducing a normative component.

Later in this chapter, I provide a brief overview of three well-known models of explanation and show how they fail to comply with the constraints on causal relevance as summarized by Craver (2007). In Chapters 7 and 8, I further discuss the constraints in my treatment of mechanistic explanation and Network Analysis (a particular type of statistical approach). First, I outline the constraints, labelled E1-E5, summarized below:

E1: mere temporal sequences are not explanatory (temporal sequences);
E2: causes explain effects and not vice versa (asymmetry);
E3: causally independent effects of common causes do not explain one another (common cause);
E4: causally irrelevant phenomena are not explanatory (relevance); and
E5: causes need not make effects probable to explain them (improbable effects).
(Craver, 2007, p.26)

E1 holds that if one event follows another, the former does not explain the latter, even if the latter (nearly) always follows the former. For instance, many patients diagnosed with schizophrenia present with apathy and attribute this to their medication. Yet, the medication may only be partially explanatorily relevant to the apathy (recall the article on primary and secondary negative symptoms in schizophrenia by Kirkpatrick (2014) outlined in Chapter 2). Kirkpatrick suggests that primary negative symptoms are those that form “an integral part” of schizophrenia for which there are not as yet any effective treatments; secondary negative symptoms are caused by factors such as positive symptoms, medication side effects, depression and substance abuse (Kirkpatrick, 2014, p. 4) and are “usually responsive to treatment of the underlying cause” (Ibid., p. 3). Despite the temporal sequence, it is very difficult for a clinician to determine whether apathy in a patient who has suffered from schizophrenia and taken anti-psychotic medication...
for a long time is due to the disorder or is a side-effect of medication which the patient
has been using for many years. Or in other words, a correlation where one phenomenon
(apathy) follows another (suffering from schizophrenia and taking medication) does
not amount to an explanation unless the explanatory relevance of the preceding event
is demonstrable.

E2 holds that causes explain effects: an explanation must follow the direction of causal
influence. This is a challenging constraint for explanations in psychiatry because
determining temporal sequences with any accuracy and defining the direction of causal
influence, given that causal reciprocity or circular causality appear to be the norm, is in
itself difficult. For instance, it has proven exceedingly difficult to determine the direction
of causal influence at play in the correlation between marihuana abuse and psychosis:
is marihuana best viewed as self-medication by patients suffering from psychotic
symptoms because it provides relief, or do marihuana-users develop psychotic symptoms
because of the marihuana? Or could there perhaps be some form of circular causality
where marihuana makes psychotic people feel better despite also fuelling the psychotic
symptoms? As I discuss further in Chapter 8 on Network Analysis, this constraint in
particular draws into question the philosophical assumption that cause and effect are
distinguishable in the first place. In cases where circular processes may be at play, such
as in the above example, this is not a given.

Another more practical reason that E2 is difficult to satisfy in scientific practice seems
more prominent in psychiatry than in other sciences. Ascribing a temporal sequence to
events at different levels of organisation and isolating phenomena from others is far from
straightforward. This is first because we are dealing with living people, who for obvious
ethical reasons do not readily lend themselves for invasive research at the biological level.
Second, we must therefore depend largely on self-reports, which, given the frequent
lack of illness-awareness in the case of severe mental disorders such as schizophrenia is
problematic in itself. Third, relying on clinical observation to determine precisely which
event followed which, is difficult. The example provided earlier of apathy as a primary
or secondary negative symptom - a distinction which was presumably designed to
amend this problem - further exemplifies this point. If apathy diminishes upon reducing
medication or treating the comorbid depressive episode, can we then conclude that the
apathy was caused by medication or depression? Is being diagnosed with schizophrenia
explanatorily relevant at all in this case, beyond being the pretext for commencing anti-
psychotic medication?

E3 holds that causally independent effects of common causes do not explain one another.
For instance, when a patient commences antipsychotic medication and experiences a
tremor and weight gain (both effects of a common cause: the medication), the tremor does not explain the weight gain. The clause that only causally independent effects do not explain one another is vital here. It raises another practical difficulty for the field of medicine and psychiatry in particular: determining whether effects are in fact completely independent is a challenge. In the example, it may be difficult to rule out the possibility that tremor is related somehow to weight gain - perhaps the patient is so ashamed of the tremor that he has stopped going to the gym or has stopped eating soup because he can barely spoon it into his mouth due to the tremor and has instead taken to eating chocolate. In Chapter 8, I discuss how Network Analysis, a mathematical method, can be used to test whether a relation between two phenomena meets constraint E3.

E4 entails that an explanation explains effects in terms of relevant causes (cf. Hitchcock, 1995). For instance, in Chapter 2 I outlined an article by Limosin, (2014) in which evidence that there is a correlation between childhood abuse and development of schizophrenia was presented. Yet, from the evidence it is not clear if childhood abuse is causally relevant to the development of schizophrenia. A correlation does not amount to a causal relationship, some correlates may be causally unconnected. In the example, the correlation may be due to childhood abuse and schizophrenia being independent effects of another common cause (i.e. they may be indirectly related, as described by constraint E3). As Hitchcock (1995) points out, transmission accounts of causality fail to meet E4. Transmission accounts of causation hold that causal processes transmit some “mark” or “conserved quantity” between states in a system or interacting entities. They do not meet E4, because there are examples of causal interactions in which multiple conserved quantities are exchanged where these conserved quantities or “marks” are not necessarily all causally relevant to the effect.

E5 holds that a cause need not always, in every instance, lead to an effect for the cause to explain the effect. This constraint is very important to keep in mind in a field where relations between causes and effects are often stochastic. Of course, there are many examples where the chance that an effect will not follow from a cause is infinitely small (i.e. there is an almost 100% percent chance that my coffee will spill on the floor if I drop it). In medicine, we need to allow for situations in which this is not the case: a certain type of medication may cause intolerable side effects in only a small percentage of patients, but it is nonetheless the cause of these side effects when they do occur. In many cases in which the relation between cause and effect is stochastic, we have no understanding of whether there is a further mechanism that explains why the effect only follows the cause in some cases. Often, we cannot but take for granted that there is a stochastic relation and therefore need to allow for this option in our models of scientific explanation.
6.4 Three models of scientific explanation

In this section, I discuss Craver (2007, Chapter 2), to whose account of mechanistic explanation I return in detail in the following chapter, and show the ways in which the Representation model, the Covering Law model and the Unification model fail to comply with E1-E5. To illustrate the relevance of these issues for psychiatry in particular, I draw on the dopamine hypothesis of schizophrenia. I choose the dopamine hypothesis of schizophrenia because there is no parallel example of a relatively widely accepted explanation of a phenomenon in the study of apathy and use it as an example of a theory in psychiatry.

The first statement of the dopamine hypothesis of schizophrenia can be traced to 1967, when overstimulation with dopamine was thought to correlate with schizophrenia (Kendler & Schaffner, 2011, p. 42). The dopamine hypothesis of schizophrenia has since been further developed to include sub-hypotheses such as that schizophrenia is due to hypersensitive dopamine receptors, underactive antagonistic neurochemical systems and/or defective feedback loops (Ibid., p. 43).

Models of scientific explanation are designed to resonate with their developers’ definition of “explanation”. One model holds that an explanation involves description in terms of laws of nature applied to starting conditions leading to a prediction of how the phenomenon under study will behave. A second model holds that explanations must classify the target phenomenon under a general representation. This representation may be a mental representation, a diagram, model, written description or equation. A third option entails unifying various phenomena in a given argument schema. The philosophical models for scientific explanation mentioned above and discussed by Craver (2007, Chapter 2) are as follows:

1. Explanation as description in terms of governing laws applied to starting conditions. This model is called the Covering Law model or Deductive-Nomological (D-N) model and is also known as the Popper-Hempel model and the Hempel-Oppenheim model. The Inductive-Statistical model is a sub-type of the D-N model which holds that a specific event was to be expected based on a statistical law.
2. Explanation as representation (Representation model).
3. Explanation as unification (Unification model).

I do not seek to provide an exhaustive account of explanatory models here but instead aim to narrow down my treatment to the most important and relevant accounts and the way in which they can be applied to medicine and psychiatry in particular. Two
further approaches to scientific explanation are discussed in the following chapters: causal mechanical explanation is discussed at length in Chapter 7. Chapter 8 offers a treatment of one specific type of statistical method from which a model can be distilled: Network Analysis.

6.5 Covering Law Model

The Covering Law model holds that an explanation is a description in terms of governing laws, presented by means of law-like sentences, applied to antecedent conditions (cf. Hempel & Oppenheim, 1948). The Covering Law model includes deductive and inductive explanations. In deductive explanations, the phenomenon can be described through applying laws to background and antecedent conditions. In inductive explanations, the phenomenon is taken to follow with high probability given antecedent and background conditions and known statistical relations between these conditions and particular outcomes.

Applying the Covering Law model to an example from psychiatry, the explanandum is the phenomenon we call “schizophrenia”. The explanans is the dopamine hypothesis. For the dopamine hypothesis to warrant the label “explanation” on the deductive-nomological model, the explanandum (i.e. the phenomenon we call “schizophrenia”) must follow, or more specifically, be deduced, from the explanans based on antecedent conditions. Note that explanation and prediction run parallel here: to explain something is to show that a phenomenon is to be expected on the basis of the laws of nature (Hempel, 1965, p. 336).

Given the profusion of different versions of the dopamine hypothesis, it is difficult here to provide a single law-like formulation to serve as an explanandum. In its most simple form, on the dopamine hypothesis, patients with schizophrenia must display a functional excess of dopamine in their brains. As Kendler and Schaffner (2011) point out, this is a very abstract formulation (p. 44). Even if we zoom in on one specific sub-hypothesis, for instance the idea that people presenting with symptoms of schizophrenia have increased dopamine receptor function (due to increased numbers of dopamine receptors and/or increased sensitivity), a law-like formulation of the mechanism is lacking. In other words, we have no inkling of any laws that may help predict at what level of dopamine receptor function an individual will have psychotic symptoms. It is here that the Covering Law model already runs into difficulties: phenomena in medicine rarely lend themselves to explanations in terms of clearly defined laws in the way that some phenomena in physics do. We must make do with statistical correlations rather than clean laws that allow us to take into account the roles of relevant factors in order
to explain the explanandum. Take for instance a Covering Law model explanation *why* it took me thirty minutes to get to work this morning. In this case, I draw upon laws pertaining to time and space and via a deductively valid argument I predict that if my work is 50km away and my average speed is 100km per hour it will take me thirty minutes to cover the distance to work.

Furthermore, returning to the dopamine receptor function example, post-mortem studies have found increased D2 receptor density (one of the five types of currently known dopamine receptor) in the brains of patients with schizophrenia compared to controls. Thus, a possible law could be that increased D2 receptor density predicts schizophrenic symptoms. Yet, as Kestler et al. (2001) point out, the increased D2 receptor density could be due to schizophrenia or the exposure to pharmacological agents that bind to the receptors (perhaps secondarily inducing increased density). Although one could design further experiments to test the hypothesis that the increased D2 receptor density may be due to receptor-binding drugs, the Covering Law model does not offer any guidelines for *how* the explanandum follows from the antecedent conditions.

For the everyday context, Churchland (1989) furthermore points out that the Covering Law model is “psychologically unrealistic” because in their everyday activities, people don’t think in terms of antecedent conditions or in terms of general laws which they apply to predict what will happen next. For instance, when playing squash, I do not invoke general laws of nature and apply these to antecedent conditions to predict which way the ball will bounce and then proceed to (try) to hit it back. Even in the domain of medical science, the relevant antecedent conditions are notoriously difficult to demarcate, general laws are not easily found and scientists do not necessarily think primarily in such terms when trying to explain a phenomenon. Or, in Churchland’s words: “the logical acumen ascribed to people on [the Covering Law] account is often substantially in excess of what university students with formal training in logic can display” (1989, p. 199).

Craver (2007) notes that the Covering Law model is lacking as a model for explanation in the neurosciences and I add that his line of argument can be extrapolated to hold for medicine too. He states that the Covering Law model cannot differentiate between laws of nature and accidental correlations and other non-explanatory generalizations (such as mere temporal sequences and mere correlations). It also does not provide an account for what it takes for something to be explanatorily relevant. Thus, for the dopamine hypothesis example, the Covering Law model does not meet the constraints E1-E5 outlined previously.
6.6 Representation Model

Representation models hold that explanations work by recognizing a phenomenon as classifiable in terms of a more general and thus more abstract representation. There is a vast body of literature concerning types of representation and the question of how a representation or model represents the target phenomenon. A model that represents can be either descriptive (linguistic) or non-linguistic, for instance a scale model of a yacht, or anatomical skeleton with elastic bands representing muscles. Often, (partial) isomorphism is invoked to characterise the relation between the target phenomenon and its representation. The term “isomorphism” in turn raises the question of how to quantify degrees of similarity or define isomorphism in the first place. In scale models, the isomorphism can be very literal: the target phenomenon and the model have the same form and proportions between parts and differ only in size. In the example of an anatomical skeleton with elastic bands for muscles, the isomorphism between the elastic bands and muscles is one of analogy which is far less straightforward to define. In some sciences, especially in medicine, models are used when theories are lacking or only preliminary. In medicine, a model may be used to develop and/or test a theory.

With the parallel distributed processing (PDP) account of explanation, Paul Churchland (1989, Chapter 10) offers a representation model which unifies the theory of explanation and theory of perception. The notion of “parallel distributed processing” is borrowed from cognitive neurobiology in which brain function is viewed in terms of neural networks performing tasks in a parallel distributed manner. For Churchland, explanatory understanding entails “prototype activation” that is similar to “perceptual recognition” which, to use Churchland’s example, occurs when one suddenly recognises a person in the distance as one’s close friend (Ibid., p. 189) or when one recognises an outline as that of a horse despite seeing only a small portion of it and understanding that it is the hindquarters of the horse (p.208). Churchland suggests that to explain something is to recognise a pattern and ascribe this to one (or more) of “an organized “library” of internal representations of various prototypical perceptual situations” (Ibid., p. 207). Churchland appears to use “understanding” in an everyday fashion, where understanding implies that one recognises a phenomenon and then behaves accordingly (e.g. moving aside when one is faced with the hindquarters of an unknown horse). Or, in his own words, Churchland describes the “internal representations” as “well-informed”, thus allowing for anticipation of “aspects of the case so far unperceived, and to deploy practical techniques appropriate to the case at hand” (Ibid., p. 210).

One obvious objection, if we try to translate Churchland’s account to the case of scientific explanation, is that recognising a phenomenon as being classifiable under a specific
representation does not amount to explaining it scientifically. This is especially poignant in the case of schizophrenia: recognising a group of symptoms as warranting the label “schizophrenia” and even behaving accordingly (for instance by taking into account the subject’s paranoia and offering medication) still does not amount to explaining why the symptoms are present or explaining what schizophrenia is. To this, Churchland replies that subsuming the phenomenon under a representation is not the same as labelling, for it still aids in “understanding (or perhaps misunderstanding) far more about the explanandum situation than was strictly presented in the explanandum itself” (p. 212). Note that Churchland’s focus is on providing an account of explanation in an everyday sense rather than an account of scientific explanation (see the previous chapter for more on the distinction between knowledge in the everyday, clinical, scientific and philosophical contexts). Unfortunately for his account, Churchland’s reply highlights another problem with the parallel distributed processing model: the account does not offer a way to distinguish between good and bad representations. The PDP model does not have constraints in place to help decide whether the dopamine hypothesis, which was fuelled by the observation that dopaminergic drugs can cause psychotic symptoms much like those of schizophrenia, correctly suggests that the similarity in the symptoms of schizophrenia and drug-abuse is in fact dopamine-based. Empirical testing in which the development of psychotic symptoms upon amphetamine administration was compared in schizophrenic and non-schizophrenic subjects, did in fact show a clinically significant difference in the increase in psychotic symptoms in patients suffering from schizophrenia compared to non-schizophrenic controls. Yet, amphetamine is more potent as a noradrenergic drug than as a dopaminergic drug and further trials with purer dopaminergic agents did not show a greater increase in psychotic symptoms (Dépatie & Lal, 2001). These findings suggest that recognising that psychotic symptoms in people with schizophrenia and psychotic symptoms after amphetamine administration are similar, does not necessarily yield a complete scientific explanation of the phenomenon (cf. Kendler & Schaffner, 2011, p. 48).

So, in sum, Churchland’s (1989) PDP account as an example of a representation model of explanation is suited to a sense of explanation tailored to an everyday (rather than scientific context) in which “to explain” is to behave accordingly after recognizing a phenomenon as being subsumable under an internal representation. Although interesting and perhaps more useful as an account of how we learn, the PDP account fails as an account for explanation in the scientific context. It does not satisfy constraints E1-E5 because it does not offer any means to test whether the representation that is activated by a phenomenon is mistakenly activated in the ways outlined in E1-E5. Churchland himself, in a somewhat brain-centric formulation, concedes that he is not concerned with these matters: “how the brain distinguishes between real causal processes and mere
pseudoprocesses, are secondary matters I shall leave for a future occasion” (Ibid., p. 214).

A second type of Representation model is Cartwright’s Simulacrum account (1983). On the Simulacrum model, “to explain a phenomenon, is to construct a model which fits the phenomenon into a theory” (Ibid., p. 17). On this account, the laws that we apply hold for models of the phenomena we wish to explain, rather than for the phenomena themselves (Ibid., p. 18). Cartwright furthermore makes a case for the notion of a “dappled world” that resembles a patchwork:

“The success of an explanatory model depends on how well the derived laws approximate the phenomenological laws and the specific causal principles which are true of the objects modelled. There are always more phenomenological laws to be had, and they can be approximated in better and in different ways. There is no single explanation which is the right one, even in the limit, or relative to the information at hand.” (Ibid., p. 17)

As I argue in Chapter 9, where I introduce the notion of “perspectival mosaic unity”, I am sympathetic to the notion of “dappled-ness”. However, whereas Cartwright makes a metaphysical claim by situating dappled-ness in the world, I refrain from making such a metaphysical claim and limit myself to a notion of dappled-ness for the epistemic domain only, i.e. for the world as we know it.

In contrast to Churchland’s account, Cartwright’s version of a Representation model is tailored to scientific practice. However, Craver’s critique that the Representation model does not offer means to distinguish between models that aid understanding and those that do not, holds for Cartwright’s Simulacrum account too. The critique that recognising that a phenomenon can be subsumed under a particular representation does not amount to understanding it, can be extrapolated to Cartwright’s account.

All in all, Representation models cannot meet E1-E5 precisely because E1-E5 are intended as constraints in the ontic domain. This is because, on Craver’s (2007) account of constitutive mechanistic explanation, which I discuss in detail in the next chapter, the phenomenon to be explained coincides with the explanation. Craver therefore does not rely on representations for his account of scientific explanation. As outlined above, Cartwright focuses on the epistemic domain, in suggesting that explanation entails applying laws not to the phenomenon itself but to a model or representation of that phenomenon. I embrace this move in arguing for my own account later on in Chapter 9, where I also suggest a modified approach to constraints E1-E5 in order to make a metaphysically agnostic mechanistic approach to scientific explanation in psychiatry.
feasible.

6.7 Unification Model

The Unification model was outlined by Friedman (1974) and further developed by Philip Kitcher (e.g. Kitcher, 1989). Whereas on the Covering Law model explanations are premise-conclusion pairs, on the unification model, “ideal explanations are derivations” (Kitcher, 1989, p. 80). The Unification model departs from the Covering Law model in that it does not “simply list the premises but shows how the premises yield the conclusion” (Ibid., p. 431, my italics). In a nutshell, Kitcher’s account claims that “[s]cience advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number and types of facts we have to accept as ultimate” (1989, p. 432). Thus, unification aids comprehensibility through reduction: the minimal set of argument patterns or explanatory systematisations from which other patterns can be derived/explained is called the “explanatory store” (Ibid., p. 430). Patterns are formed of “schematic sentences” (Ibid., p. 432) in which non-logical terms may be replaced by others according to “filling instructions”. A schematic argument is formed of a sequence of these schematic sentences and contributes, along with the filling instructions and classification of arguments (as premises or conclusions) to a “general argument pattern” (Ibid., p. 432).

Kitcher acknowledges the tension between “minimizing the number of patterns of derivation employed and maximizing the number of conclusions generated” (Ibid.) and advocates choosing the “best trade-off”. At one extreme we find that if a pattern is too stringent, or in other words, the conditions for its instantiation are more difficult to meet, it has little unifying power. In contrast, at the other extreme, if a pattern is relaxed to the degree that it admits any argument, this detracts from the pattern’s explanatory power (cf. Ibid., p. 433).

Furthermore, it is to this trade-off that Kitcher alludes when dealing with the causal relevance constraint (E4). Explanations that include causally irrelevant factors besides causally relevant factors will call for addition of patterns to the explanatory set rather than maximal unification of phenomena under a reduced number of patterns. Kitcher uses the example of “hexed salt” (salt on which a spell has been cast) and explaining why hexed salt dissolves in water, to demonstrate this point. Table salt dissolves in water regardless of whether it has been hexed by a magician and adding a ”hexed-or-not clause” to the explanation in terms of chemical bonds between NaCl and water yields a more complicated explanation that has inferior unifying power because being
NaCl alone is sufficient to explain why the substance dissolves and the “being hexed” requirement commits one “to patterns of explanation that apply only to a restricted class of cases” (ibid., p. 484). Kitcher expands the example to demonstrate that if we assume that any sample of table salt is “hex-able”, being “hex-able” still is not causally relevant to its solution in water. How can the unification model show that being made up of salt is causally relevant to being water-soluble, whereas “being hexed” is not? Kitcher answers that if any substance is hex-able, then hex-ability becomes a redundant clause because not all substances are water-soluble (Ibid., p. 484). Therefore, “being made up of salt” has more unifying power than “being hex-able” when it comes to explaining why a substance dissolves.

Craver (2007) argues that Kitcher’s Unification model nonetheless does not satisfy constraints E1-E5 (pp. 42-49). This is because the Unification model is similar to the Covering Law model and suffers from parallel problems when it comes to satisfying E1-E5. One difficulty is that although Kitcher accepts the constraints for explanation, he refrains from committing to the possibility of there being causal relations in the world, as is demonstrated in the following citation:

“What is distinctive about the unification view is that it proposes to ground causal claims in claims about explanatory dependency rather than vice versa. (...) Thus the picture advanced by the unification approach shows the concept of causal dependence as derivative from that of explanatory dependence, but it does not promote the dubious idea that each of us gains explicit knowledge of causal dependencies through recognition of the structure of the explanatory store”. (Kitcher, 1989, p. 436).

For Craver, Kitcher’s explicit lack of attention to the causal structure of the world is problematic for his ontic approach, because E1-E5 for him, require accepting that there are causal relations in the world that can be scientifically explored. In a footnote, Craver notes that Kitcher, despite denying that there are causal relations in the world (and stating that he accepts E1-E5 as constraints on explanation), still invokes language in his argument schemata that is highly suggestive of causal relations. Words found in Kitcher’s examples, such as “transcription”, “modification”, “translation”, “enables”, and “contributes” suggest a causal relationship between two aspects/items in an argument scheme (Craver, 2007, p. 43). This is an interesting point because Craver thus suggests that Kitcher fails to truly discard causal realism. Nonetheless, regardless of whether one embraces the causal structure of the world as a central feature of explanation, the Unification model, like the Covering Law model, does not satisfy E1-E5.

Craver argues that the Unification model doesn't satisfy constraint E2,
which holds that causes explain effects and not vice versa. In the classic example of the flagpole's shadow, the direction of causal dependency runs from the sun and its position, to the flagpole to the flagpole's shadow. Put very simply, the sun causes the shadow, the shadow does not cause the sun. The Unification model, in order to account for causal asymmetry, must invoke a complicated work-around to show that causes do not explain their effects (cf. Craver, pp. 43-46). Craver raises questions about the degree to which this work-around succeeds, noting that the model offers no means to differentiate between the unifying arguments with and without explanatory power.

The Unification model also fails to meet criterion E5, which holds that causes need not make effects probable to explain them (improbable effects). The model aims to offer the best unification, i.e. the argument under which the most phenomena can be subsumed. Therefore, unifications that explain only a small number of individual phenomena (such as occurrences of a rare disease) automatically have a lower unifying power than those under which more common phenomena, i.e. phenomena with a larger prevalence or phenomena that are similar enough to be explained by a single argument (e.g. occurrences of different strains of the common cold virus), can be subsumed (Ibid., p. 46-47).

Furthermore, even if Kitcher allows for the causal structure of the world, there is a further problem: unification does not always amount to explanation. Take for instance taxonomies: a phenomenon is sorted into a category based on properties which may indirectly contribute to scientific explanation without being explanatory in itself (Craver, 2007, p. 42). Earlier, I made this point using an example from psychiatry: classifying a group of symptoms as warranting the label schizophrenia (given the non-etiological setup of the DSM-IV-TR) does not amount to explaining the symptoms and/or schizophrenia.

The problems for the Unification model are thus largely parallel to those for the Covering Law model when it comes to constraints E1-E5. The problems are further complicated first by Kitcher's refutal of the world harbouring a causal structure (which, for Craver is an important point of departure) and second because unification does not always amount to explanation (taxonomies are an example of non-explanatory scientific achievements).

6.8 Summary

In this chapter, I moved from the medical models of explanation to philosophical models of explanation. The fundamental question of what constitutes a good explanation was addressed based on constraints E1 to E5. I chose to embrace these constraints because
they seem to offer a plausible axiomatization of causal relevance, which I have adopted as important to the project of scientific explanation. Three traditional philosophical models of explanation were introduced and their shortcomings in terms of E1 to E5 were demonstrated. It has proven difficult to devise a model that meets the constraints and that has practical utility despite its high level of abstraction.

In the following chapter, the mechanistic model of explanation is introduced as an example of a model which, as Craver (2007) argues, can meet constraints E1 to E5. Precursors to the mechanistic explanatory model can be found in writings from the late 60s and 70s and beyond. Rather than imposing the rules of logic upon scientific practice, thinkers began to draw upon case studies from science and take a bottom-up approach (as I too have done in this project in the previous chapters) to construct an account of how science progresses. They thus moved away from using the tools of logic and focussing on parts and their law-like interactions, to constructing mechanisms suited to describing the functions of the parts (found at multiple levels) and their contribution to the phenomena under study. I address the question of whether the mechanistic model is suitable for the science and practice of psychiatry. In Chapter 8, I introduce a statistical, dynamic systems-based approach to explanation “Network Analysis”. The question of whether Network Analysis can meet constraints E1-E5 is explored in detail.
Chapter 7
Mechanistic Explanation
Chapter 7

7 Mechanistic Explanation

7.1 Introduction

In the academic setting, the sciences are initially presented as neatly demarcated fields or disciplines. The relations between disciplines and the effects that an important discovery in one field may have for other fields in science are, however, less obvious. The implicit assumption in the second central question of this thesis is that improved integration of different perspectives from these fields can contribute to scientific progress. Although many scientists and clinicians in psychiatry will concede this, there is little evidence of this in current research (as demonstrated in the literature review provided in Chapter 2) or in day-to-day clinical practice. Therefore a notion of integration is required to make the relations between research domains intelligible and to achieve a more coherent understanding of the phenomena we study. On this view, the question of choosing sides between biological and psychological psychiatry, which is still pervasive clinical practice, becomes a non-starter. Both biological and psychotherapeutic approaches to explaining and treating psychiatric disorders are relevant. The aim instead, is to fit them together, to place them on a continuum regarding them as two different, equally valid, perspectives on the same phenomenon. Working towards a model for the integration of perspectives in psychiatry is therefore a central aim of this project. In this chapter, the mechanistic approach to explanation is presented which purports to transcend traditional divisions between fields and perspectives in science. I build on conclusions from Chapter 6 concerning the advantages and disadvantages for psychiatry as a science of various philosophical models of explanation.

This chapter begins with a short introduction on the notion of “levels” because this metaphor is pervasive in scientific literature and further exploration of its consequences seems warranted. The treatment of levels focusses briefly on their use in the biopsychosocial model and includes a review of the use of “levels” in the apathy literature. “Levels of mechanism” (Craver, 2001, 2007, & 2014) are then presented in order to prepare the ground for introduction of the mechanistic approach to explanation. My exploration of the mechanistic approach is based on the accounts of a group of thinkers often referred to as the “New Mechanists”. Having discussed explanatory and methodological aspects of the mechanistic approach, which are presented by the New Mechanists mainly in the context of biology and the neurosciences, I zoom in on the field of psychiatry and examples are again drawn from the apathy research presented in Chapter 2 to investigate the feasibility of the mechanistic approach to explanation for psychiatry. I show that researchers in psychiatry often lack the robustly detectable, physiologically plausible parts of which a sound mechanistic explanation is ideally composed (Craver, 2007).
explore the extent to which this lack of acceptable “parts” can be overcome.

7.2 A Little on Levels

The notion of “levels” is open to multiple interpretations (e.g. Craver & Bechtel 2007; Eronen, 2015). Wimsatt (1976) for instance, holds that “levels” are clusters of rankable features. This ranking can be based on for instance size, aggregation, analysis, causation and organisation. Oppenheim and Putnam (1958) suggest that objects at the same level are governed by the same laws, and that there is thus a correspondence between mereological levels, sciences and theories. On this view, objects at the same level exhibit forces of similar magnitudes and have regular and predictable relations. Some thinkers suggest that theories are situated at levels and everything at a given level may be investigated with the same set of techniques, according to similar disciplinary perspectives. Bechtel and Richardson (2010) point out that viewing natural systems as “hierarchically structured” is a heuristic strategy, which fits well with human cognitive capacities: it is easier to understand the world if we can structure our interpretation of it. The latter thinkers are however agnostic about a whether the natural systems under study are themselves hierarchically structured (Bechtel & Richardson, 2010, pp. 31-32).

For my present purposes, I draw on Craver, who offers an extensive “taxonomy of levels” (2007, p. 171) of which I provide an overview below. The taxonomy is based on three questions to differentiate between types or manners of ranking levels:

1. What are the relata?
2. By virtue of what are two items at different levels?
3. By virtue of what are two items at the same level?
   (Craver, 2007, p. 172)

The first major branch in his taxonomy separates levels of science from levels of nature (levels that are intrinsic to the system itself). Levels of science in turn can be split into levels of products and levels of units. Levels of products pertain to descriptions, explanations and theories yielded by science. Levels of units include fields, disciplines and research projects in the scientific domain. For psychiatry, the products at stake are 1) descriptions, for instance of diagnoses as found in the DSM 5 or descriptions of what apathy is, 2) explanations, for example of how higher doses of anti-psychotic drugs cause side-effects resembling Parkinson’s disease, and 3) theories such as the dopamine hypothesis of psychosis. Craver argues that adhering to levels of science has drawbacks, mainly due to the lack of correspondence between levels of science and levels of nature. Many descriptions, explanations and theories refer to phenomena at multiple levels of
nature. As I show below, in arguing for the multi-level nature of explanations, Engel seems to have assumed that there is a tidy correspondence between levels of science and levels of nature which, as I argue, is a source of confusion in his account.

Levels of nature sprouts several sub-types such as levels of causation, size and composition (as mentioned above). Levels of composition in turn can include subcategories such as levels of mereology, aggregativity, mere material or spatial containment and finally, levels of mechanism. Levels of mechanism is the sub-category of levels of composition on which Craver focuses due to the utility he attributes to levels of mechanism for explanations in neuroscience (2007) and biology (Craver & Darden 2013).

![Figure 7.1 A taxonomy of levels from Craver, 2007, p. 171 (reprinted with permission from publisher).](image)

**7.3 Levels of science versus levels of organisation in Engel**

I pause now, to examine the relation between “levels of organisation” and “levels of science”. I show that in the biopsychosocial model, Engel does not clearly distinguish between them. This lack of a clear distinction appears to be pervasive in clinical practice and psychiatry research. I further demonstrate how loosely researchers employ the levels metaphor in the next section. From a philosophical perspective, confusing matters of ontology and epistemology is problematic because, as I argue in more detail in Chapter 9, it is important to take into account the role of subjectivity. I draw on Craver’s taxonomy and method for differentiating between types of level to clarify the relation between “levels of organisation” and “levels of science” in Engel’s writings.

Engel, in his writings on the biopsychosocial model, refers exclusively to “levels of organisation” (see Chapter 4 & 5). However, his reference to “biological”, ”psychological” and “social” aspects of a phenomenon suggest an appeal to levels of science. To avoid
mistaken projection of philosophical approaches to integrating levels upon Engel’s model, the relation requires some thought. To clarify the difference between levels of science and organisation I draw on the above-mentioned questions Craver used for this purpose.

The first question is: what are the relata? The relata of levels of science are different to those of levels of organisation (Craver, 2007, p. 171). A “level of science”, as Craver calls it, has primarily epistemological connotations: it pertains to knowledge on the part of for instance a scientist or clinician concerning the specific phenomenon under study. The relata, or things sorted into levels, are products of science. “Levels of organisation”, or “levels of nature” in Craver’s terminology, in contrast, have primarily ontological connotations: distinctions between the levels are guided by the phenomenon under study. The relata here are things in nature. Note that I use the word “primarily” to allow for my view that any hypotheses I have about things in the world are necessarily perspectival, or, in other words, mediated by my perception, i.e. by epistemological considerations. As I will argue in more detail in Chapter 9, I hold that epistemology and ontology are mutually informative or co-dependent. My position does not entail that we should drop the ontic-epistemic distinction altogether however.

The co-dependence means that, for science to be a fruitful practice, study of a phenomenon must be taken to yield knowledge with at least some practical utility in the first place. There is no point in talking about a level of explanation if there is no phenomenon for which explanation is required. And, vice versa, there is no point in talking about levels of organisation in the phenomenon under study if these have no bearing upon amassing knowledge about the phenomenon. Furthermore, what it means for something to be a level differs depending on the relata of the levels (Ibid., p. 171). This final point is also relevant for epistemological versus ontological levels.

The second question is: in virtue of what are two items at different levels? This is more challenging to answer for levels of science than for levels of organisation. For levels of organisation, things belong to a specific level in virtue of their organisation. This organisation can be interpreted in terms of size, composition (which in turn can be split into mereology, aggregativity, material/spatial containment and mechanisms) or causation (Ibid., p. 171-2). For instance, based on mereology, things such as planets and social systems are regarded as belonging to a higher level than things such as atoms and individuals. It is here that the correspondence between levels of organisation (or levels of nature as Craver calls them) and science, is best situated. Things in nature warrant an approach from a scientific field that is suited to exploring the phenomenon at hand. A phenomenon may be studied by various scientific disciplines that each interrogate
the phenomenon in a different manner. Thus, explaining *in virtue of what* something belongs to a different level of science requires referring to the level of organisation and the relevant perspective on the target phenomenon under study. A similar response can be offered when answering the third of Craver’s questions: by virtue of what are two items at the same level? Describing in virtue of what two items are at the same level of science (i.e. the biological, psychological or social level) requires referring to the traditional hierarchy of levels of nature. For instance, if the phenomenon under study is best described at the cellular level (e.g. what mitochondria do), then a biologist is likely better equipped for the task than a psychologist.

I return to Engel to see what we can infer about his (implicit) assumptions about the relation between levels of nature and levels of science. In doing so, I refer to Craver’s taxonomy of levels to try and locate the levels to which Engel refers in Craver’s taxonomy. In Chapter 4, I noted that Engel invokes a hierarchy of natural systems, with the biosphere at the top, followed by society-nation, culture-subculture, community, family, two-person, person, nervous system, organs, tissues, cells, organelles, molecules, atoms, and subatomic particles respectively. The hierarchy of “natural systems”, to which Engel alludes is an example of what Craver calls “levels of nature”. The biological, psychological and social levels to which the biopsychosocial model refers in contrast are what Craver would call “levels of science”. In previous chapters it has become clear that Engel is an advocate of multi-layer explanations in clinical and scientific practice. Unfortunately, Engel doesn’t articulate any organisational principles in virtue of which an item belongs at one level (of science or nature) rather than another (hence my reference to “implicit” assumptions on Engel’s part). This appears to hold not only for Engel, but for many scientists – as I show in the next section.

I have shown that Engel does not clearly differentiate between types of levels and in particular, he does not distinguish levels of nature from levels of science. I demonstrated that the different types of levels have different relata and therefore should not simply be lumped together. In the next section, I provide evidence that imprecise employment of the levels metaphor is common in scientific practice. I do so by returning to the apathy literature to explore how the metaphor is used.

### 7.4 Levels in the literature on apathy

Many successful theories in neuroscience span multiple levels of nature (e.g. Schaffner, 1993, pp. 285-286; Craver, 2007 & 2014) and different scientific fields may be dedicated to studying phenomena at the same mereological level (Craver, 2014, p. 27). In Chapter 2 I demonstrated that multi-layeredness also features in the study of apathy
Mechanistic Explanation

in schizophrenia. Although a coherent unified (or “generalized”, to use Schaffner’s preferred term) theory is lacking, many studies reviewed in Chapter 2 examine data from different levels of nature, such as comparing outcomes on an individual’s psychological tests to findings from neuro-imaging such as (f)MRI. As noted in the previous section, Craver’s objection to assuming that levels of nature correspond to levels of science is not a problem for Engel because multi-layered explanations are precisely what Engel advocated in the biopsychosocial model: a multi-layered description or explanation of a phenomenon which itself too is multi-layered (for instance, the phenomenon apathy has aspects from different levels of nature, e.g. at the level of the cell, organ, person and family).

Interestingly, in Chapter 2 on the scientific literature on apathy in schizophrenia, the notion of “levels” in the literature reviewed features mainly in discussions on “levels of consciousness” or “levels of functioning” and not in the context intended here. The notion of “levels” at stake here is not explicit in the literature on apathy. “Levels” as an organisational principle or explanatory aid only feature in my own projections of this way of thinking onto the apathy literature in order to uncover any implicit model for the organisation of knowledge in terms of either a biomedical or biopsychosocial model. The projections or inferences of the use of “levels” here were based on researchers’ use of terms such as “hierarchy”, “underpin” and “underlie”. In short, “levels” themselves do not seem particularly central in researchers’ parlance, and phenomena from different levels of science and organisation, and the perspectives on these phenomena, are frequently lumped together.

Many of the authors of the articles discussed in Chapter 2 refer to a correspondence between “physiological functions” and the “clinical dimension” (of negative symptoms, or more specifically, apathy) (e.g. Chue & Lalonde, 2014) or an effect of various medications on for instance alogia, apathy, social deficits, and anhedonia. Their attempts to discover and explain these correspondences between phenomena typically span multiple levels of organisation. Take for example pharmacological research: the administration of an antipsychotic agent is regarded as an intervention that elicits a change at the biological level of organisation and researchers then proceed to measure its effect at levels including the biological level (e.g. dopamine receptor antagonism), psychological level (e.g. decrease of apathy, anhedonia) and social functioning (e.g. ability to retain employment) (cf. Kirkpatrick, 2014). For another example of research spanning multiple levels of organisation and explanation, see my discussion in Chapter 2, of Limosin (2014) who explores neurodevelopmental and environmental hypotheses of the negative symptoms of schizophrenia. He draws a parallel between brain abnormalities (such as hyper-reactivity of the HPA axis, cerebral atrophy, ventricular enlargement, reversed cerebral asymmetry and hippocampal lesions) found both in traumatised people and people
suffering from schizophrenia in support of the traumagenic neurodevelopmental model of negative symptoms.

I conclude that there is a discrepancy between the importance philosophers attribute to categorizing levels on one hand, and the scarce and imprecise employment of the term in the scientific literature on the other hand. In fact, researchers of a biopsychosocial persuasion appear to take multi-level approaches for granted and the lack of a philosophically sound method for their integration does not appear to feature when they design their experiments. I return to the consequences of these observations in more detail in the Chapter 9 when addressing the potential advantages of paying more attention to the philosophical assumptions underlying the design of experiments and interpretation of scientific findings.

7.5 Levels of mechanism

I return now to “levels of mechanism” before discussing the definition of “mechanism” in the context of mechanistic explanation. In defining levels of mechanism according to the three differentiating questions, Craver states: “the relata are behaving mechanisms at higher levels and their components at lower levels” (2007, p. 189). The relata are neither mere entities nor mere activities, but are instead “acting entities” (more on this activity-entity dualism below). He notes that “the composition relation is not, at base, spatial or material” (Ibid.). The interlevel relationship is characterized by a parts-to-whole relationship between the acting entities. Thus, one acting entity is at a lower level of mechanism if it is a component of a higher level mechanism. Multiple lower level mechanisms may be organized (spatially, temporally and actively) to form a higher level mechanism (Ibid.). A part or sub-mechanism produces a regular or predictable effect, but need not be a materially or spatially demarcated thing; in psychology for instance, interactions between functional parts (rather than spatially or materially demarcated parts) are the rule rather than the exception. In Craver’s words: “[l]evels of mechanisms are also loci of stable generalizations, and consequently can be seen as local maxima of regularity and predictability” (Ibid., p. 190). To count as a “part” of a mechanism, the part must be a component rather than a mere piece that has no active role in the mechanism. Craver states that parts “make identifiable contributions to the behavior of a mechanism” (2007, p. 188). Of course, a component part of the mechanism may not be larger than the mechanism as a whole.

Levels of mechanism are relative. The question “are two acting entities at the same level of mechanism?” is relevant only if the two things are part of the same system. What counts as a system is again dependent on the perspective and scope of the observer.
or “the explanatory context” as Craver calls it (2007, p. 193). So inquiring whether two entities with functional similarities, that are however part of different systems, are at the same level of mechanism, is nonsensical. For example, humans have serotonin receptors in both the gut and the brain, but the level of mechanism at which they are described depends entirely on the greater system that is under study – they are not necessarily at the same level of mechanism just because they are receptors for the same neurotransmitter. To be at the same level of mechanism requires that the components interact to produce the phenomenon under study.

Craver and Bechtel (2007) state that relations within a level of mechanism may be causal, whereas interlevel explanations are constitutive (and not causal). “Mechanistically mediated effects are hybrids of constitutive and causal relations in a mechanism, where the constitutive relations are interlevel, and the causal relations are exclusively intralevel” (Ibid., p. 547).

It is interesting to note that Craver puts much emphasis on the relativity, or “local view” (Ibid., p. 193) of levels of mechanism, as a type of level of nature (see Figure 7.1). The relativity and the lack of an absolute, objective hierarchy of levels of mechanism stem from the idea that what counts as belonging to a specific level of mechanism depends on the explanatory context. Craver (2014) notes: “[l]evels of mechanisms, like part-whole levels generally, are not monolithic divisions in the furniture of the world. Levels of mechanisms are defined only within a given part-whole hierarchy” (p. 18). The importance of perspective seems to contrast Craver’s insistence on a primarily ontic account of mechanism. I discuss the difference between ontic and epistemic accounts and difficulties raised by this distinction in more detail later.

### 7.6 Mechanisms and mechanistic explanations

Mechanistic thinking, which holds that mechanisms rather than laws are the vector of explanation, is hardly new: it was a central feature of Descartes’ thinking. On his view, human minds were about the only thing that could not be rendered as a mechanism. One of the overarching questions of this chapter is: was he right, or can we extrapolate mechanistic approach to work for understanding minds and the domain of psychiatry after all?

Rather than delve further into the history of mechanistic thinking, I jump straight into the more contemporary thinking of the “New Mechanists” and compare their accounts of mechanistic explanation. One of the main differences concerns the focus of the accounts, which can be roughly divided into three types borrowed from Levy (2013),
where mechanistic explanation is an account of (pp. 100-107):

1. Causation (e.g. Glennan)
2. Explanatory relevance (e.g. Craver)
3. Scientific method (e.g. Bechtel, Richardson)

Levy (2013) notes that these views are complementary rather than mutually exclusive and that they have fuzzy boundaries (p. 109). For an example of the first view, take Glennan (1996), who draws upon mechanistic explanation as an account of causation. He offers “a mechanical theory of causation” in which mechanism is the answer to Hume’s search for the “secret connexion” between cause and effect (Glennan, 1996, p. 50). The explanatory relevance perspective focuses on mechanisms underlying a phenomenon and on how to determine which aspects of a mechanism (i.e. parts and their interactions) are explanatorily relevant to the phenomenon (e.g. Craver, 2007). The third perspective is that of mechanistic explanation as a scientific method. Here, the focus is on the use of mechanistic modelling as a scientific strategy to represent and reason about complex systems (Levy, 2013, pp. 104-105).

Besides categorizing approaches based on Levy’s three types (which, as noted, overlap), New Mechanists also differ in terms of whether they prioritize ontic or epistemic constraints on what a good explanation is. Ontic accounts hold that explanation entails accounting for (causal) facts. Epistemic accounts hold that explanation is about accounting for explanatory representations and for how science can be used to represent and reason about relevant facts. I highlight some of the implications of these different perspectives below and attempt to demonstrate their complementarity drawing on thinkers such as Illari (2013) and Tabery (2004). Tabery (2004) for example, shows the merits of synthesizing definitions of “mechanism” by Glennan (2002) and Machamer, Darden and Craver (2002). My discussion of mechanisms and mechanistic explanation in the remainder of this chapter is structured as follows:

• Comparison of different definitions of mechanism
• Entities and activities that make up a mechanism
• Dualist and non-dualist accounts of mechanism
• Ontic and Epistemic perspectives on mechanistic explanation
• Mechanistic approach in current apathy research
• The Mechanistic Model in psychiatry in general
7.7 Comparison of different definitions of mechanism

The definition of “mechanism” by Craver and Darden (2013) is explicitly dualist because it holds that mechanisms “are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions” (p. 15, my italics). The dualism pertains to the idea that entities and activities cannot be reduced one to the other. Craver and Darden’s (2013) definition, which suggests that mechanism is a linear process, contrast those by Bechtel and Abrahamsen (2005) and Craver (2013, p. 139) because the latter explicitly accommodate non-linear processes. Bechtel and Abrahamsen (2005) define a mechanism as “a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena” (Ibid., p. 423). Another definition is that of Glennan (2002): “[a] mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations” (p. 344). Glennan’s approach, in contrast to the two previous accounts, is non-dualist. I return to the implications of dualist and non-dualist accounts in a later section.

The three definitions have in common that they all hold that the mechanism itself is the producer of change and that none of them invoke laws. Glennan (2002) introduces the notion of a mechanical model which has a two-part characterization to distinguish between what a system is doing (i.e. its behaviour) and how the mechanism that accounts for that behaviour or “change” is doing it (2002, p. 347). Note that Bechtel and Abrahamsen refer to “models” of the mechanism in a somewhat different context. Given that they regard explanation as an epistemic activity, they differentiate between mechanisms operative in nature and our use of representations or models for explanatory purposes. Thus, for them, “models” are descriptions of the mechanism operative in the world and they emphasize that mechanisms operative in the world themselves do not explain anything (Bechtel & Abrahamsen, 2005, pp. 425-426). Craver, who adopts a broader use of “explanation” (which he uses almost synonymously with “cause”), would argue that the explanation itself is what produces the phenomenon (2007, p. 27). I return to this point later in the chapter, in the section on “ontic and epistemic implications” but first articulate what constitutes a mechanism in more detail.

7.8 Entities and activities that make up a mechanism

As discussed, on most accounts of mechanistic explanation including the Machamer, Darden and Craver (2000) account, a mechanism consists of entities and activities.
The figure below outlines what a mechanism is and outlines some key terms in the description of a mechanism, such as organisation, structure, location and orientation. The example is based on a generic mechanism. For instance, for the cogwheels in the figure to interact, the wheels must have teeth organised (at regular intervals) and structured (e.g. made of suitable material to withstand pressure) so that they can interlock, the wheels must be placed or located so that they can touch and turn in the correct direction. The wheels’ axles must be in the centres of the wheels for them to articulate effectively at any point in their rotation. Furthermore, the axle must not be locked for the direction of movement (it must be correctly oriented). The continuity in this mechanism can be represented by an arrow between the situation in \( t_1 \) and \( t_2 \) (\( t_1 \rightarrow t_2 \)). \( t_1 \) here, is the idealized starting position, or set-up (Machamer, Darden & Craver, 2000, p. 11). For ease of representation, the stages in a mechanism are often treated as discrete stages. Furthermore, the mechanism must run in a context appropriate to enable the mechanism to work. The mechanism depicted in Figure 7.2, for instance, can run in oil or air, but not in a bucket of sand and there must be available energy to get one wheel moving.

![Figure 7.2 Example of a common mechanism showing how interlocking gears interact to produce the change between states at t1 and t2.](image)

Writing about the mechanistic approach in the neurosciences, Craver (2007) provides the most elaborate account of what a “part” is. He specifies that a part in a mechanism needs to be physiologically plausible, have stable properties, be robustly detectable, and it must be open to intervention (Craver, 2007, p. 131). It must also contribute to the workings of the mechanism and be more than a random or inert piece (Ibid., p. 188). Parts that are merely fictional or that appeal to functions of an agent therefore do not meet this definition. In the verbal description of a mechanism, entities or parts are usually referred to by nouns and activities are mainly described by active verbs (Machamer, Darden & Craver, 2000).

There are two complementary approaches to isolate component functions. One involves identification of components or component functions through first isolating physical
components and then determining their function. This is called the “analytic strategy”. The “synthetic strategy” entails proposing how system behaviour may arise through component operations and then identifying the components responsible for these subtasks (Bechtel & Richardson, 2010, pp. 18-21).

The notion of start to finish conditions (as in Craver & Darden’s definition) is an idealization that works best for linear mechanisms. Yet, a self-organising system’s behaviour is often better described in terms of a non-linear mechanism (Bechtel & Abrahamsen, 2005, p. 435). Furthermore, there are different ways to characterise the relation between the mechanism and the system’s behaviour. For instance, some mechanisms are cyclic – the cyclic mechanism is then said to produce the phenomenon. Others may underlie rather than produce a phenomenon and yet another option is that mechanisms may maintain equilibrium rather than producing change (Craver & Darden, 2013, Chapter 2; Bechtel & Richardson, 2010, p. xxxx). Note that these are very rough categories and that a phenomenon, regarded as a system composed of multiple mechanisms (in turn made up of interacting entities), may be a combination or hybrid of these types. For example, a mechanism underlying a target phenomenon may be geared towards maintaining an equilibrium at some level of organisation through a mechanism that functions in a cyclic fashion. For instance, the mechanism to maintain an equilibrium in the body’s thyroid hormone levels, involves a negative feedback loop in which the hypothalamic hormone that stimulates the thyroid gland to produce thyroxine is itself inhibited by thyroxine to prevent overproduction of thyroid stimulating hormone.

Below, I offer schematic descriptions of each of the above-mentioned subtypes of mechanisms (Figure 7.3). The left figure represents a linear mechanism which has a clear start and finish and various intermediate situations. In the middle figure, there is not a clear start nor a clear finish and if we choose an arbitrary starting point and follow the cycle, we automatically arrive back in the starting situation. The figure on the right shows that a series of entities and their activities can underlie a phenomenon, much like lots of individual bees can produce a swarm. The last figure shows how a mechanism, existing of parts and their activities can serve to maintain an equilibrium.

Figure 7.3 Types of mechanism.
One mechanism can consist of various levels of sub-mechanisms (a nested hierarchy). Thus, a mechanistic explanation can bottom out at the lowest level that is relevant from the perspective of the scientist or discipline (Machamer, Darden & Craver, 2000, p. 13). In this regard, mechanistic explanation differs from traditional reductionism: the description of a lower level mechanism accounts for a different phenomenon than the description of a higher level mechanism (Bechtel & Abrahamsen, 2005, p. 426; Bechtel, 2006, pp. 40-41). For example, the Pavlovian mechanism which explains how the sound of a bell may make a dog’s mouth water, need not bottom out in a mechanism consisting of atoms. The mechanism to explain why a bulb lights up when I flick the light switch in contrast may bottom out at the sub-atomical level. This is not to say that mechanistic explanation does not involve any form of reduction. Decomposing a system into its components and their activities for explanatory purposes is of course a form of reduction (without the –ism). In Bechtel’s words:

Mechanistic reductionism neither denies the importance of context or of higher levels of organization nor appeals exclusively to the components of a mechanism in explaining what the mechanism does. The appeal to components in fact serves a very restricted purpose of explaining how, in a given context, the mechanism is able to generate a particular phenomenon. (Bechtel, 2006, p. 41)

Furthermore, mechanistic explanations may have both etiological and constitutive aspects. Both types are “thoroughly causal” (Salmon, 1984, p. 270). Etiological mechanistic explanations explain a phenomenon by describing how it occurred in terms of its antecedent causes (Salmon, 1984, p. 269; Craver, 2007, p. 107). Constitutive explanations describe the mechanism that produces or generates the phenomenon rather than invoking antecedent causes. Salmon describes the difference as follows:

An etiological explanation is an exhibition of the causal connections between the explanandum and prior occurrences; such an explanation fits the explanandum into an external pattern of causal relationships. A constitutive explanation consists of an exhibition of the internal causal structure of the explanandum; such an explanation exposes the causal mechanisms within the explanandum. (Salmon, 1984, p. 270)

Constitutive explanations are Craver’s main focus. He develops his account within the systems tradition rather than the reductive tradition (where the covering-law model, discussed in Chapter 6 is a common approach to scientific explanation). Constitutive explanation in the systems tradition is “a matter of decomposing systems into their parts and showing how those parts are organized together in such a way as to exhibit the explanandum phenomenon” (Craver, 2007, p.109, italics in original).
Craver’s interest is primarily in the phenomenon and underlying mechanism itself rather than the epistemic constraints on representing mechanisms. Besides constraints E1-E5 (discussed at length in Chapter 6), which should be used to assess whether a mechanistic model fits the causal structure of the world, Craver offers the following further “norms” for the assessment of sound constitutive mechanistic explanations within the systems tradition:

1.) The mechanism must fully account for the target phenomenon (“including precipitating conditions, manifestations, inhibitory conditions, modulating conditions, nonstandard conditions, and byproducts”) (Craver, 2007, p. 161).

2.) The explanation is constitutive: the phenomenon as a whole is described in terms of robustly detectable, physiologically plausible component parts.

3.) Mechanistic explanations differ from aggregate and morphological explanations due to the notion of organization (active, spatial and temporal) (Ibid., p. 162).

4.) Constitutive relevance must be accounted for. A mere functional analysis does not suffice and Craver suggests a mutual manipulability account.

I evaluate the degree to which these norms can be met for mechanistic explanations in psychiatry in a later section, following a further literature study to see whether mechanistic explanations feature in apathy research. First, however, I complete my introduction to mechanistic thinking by elaborating a little on the dualist and non-dualist approaches and also ontic and epistemic perspectives, the role of intelligibility and differentiation between how-possibly, how-plausibly and how-actually models.

7.9 Dualist and non-dualist accounts of mechanism

Mechanisms can be defined in either a dualist or non-dualist manner. On the dualist account, mechanisms are defined as consisting of entities and activities (Machamer, Darden & Craver, 2000; Bechtel & Richardson, 2010). “Activities are the producers of change” (Machamer, Darden & Craver, 2000, p. 3). Entities cannot be reduced to activities nor can activities be reduced to entities, hence the “dualist” label. The traditional reliance on “laws”, as in the Covering Law model is replaced by emphasis on activities that are “constitutive of the transformations that yield new states of affairs or new products” (Ibid., p. 4). Glennan in his later writings replaces reference to laws by “direct, invariant change-relating generalizations” to avoid the downfalls of law-based explanatory models. On the Machamer, Darden and Craver (2000) account, entities are the things that engage in activities (Ibid., p. 3). Entities can interact in specific ways in virtue of their organisation: they “often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate,
and duration” (Ibid.). There must be productive continuity so that there are no gaps in our understanding of how the end situation follows from the starting conditions. Additionally, entities can be individuated based on their properties (although it is strictly not the properties but the entities themselves that lead to change in a mechanism). Activities can furthermore be identified based on the entities/properties they engage with, and the rate and duration of occurrence (Ibid., p. 5). Both entities and activities may have functions or roles in a mechanism. A function is not a property had by an entity, “[f]unctions, rather, should be understood in terms of the activities by virtue of which entities contribute to the workings of a mechanism” (Machamer, Darden & Craver, 2000, p. 6). Also, the authors note that “activities are types of causes”; it is not the entity which causes a change but the activity of that entity (Ibid., p. 6).

In contrast, on Glennan’s non-dualist account, the notion of “activities” (which are explicitly dynamic and a central feature of dualist mechanistic approaches – more on these below) is substituted by “interaction”. “Interaction” is defined as “an occasion on which a property change in one part brings about a property change in another part”, and, this interaction occurs in virtue of a part’s change relating capability (Glennan, 2002, p. 344). Dualism is avoided on Glennan’s account because “interactions” are invoked to reduce “activities” to parts’ properties.

Tabery (2004) suggests regarding the notions of activities and interactions in the different accounts as complementary rather than conflicting. He argues that invoking “activities” does justice to the dynamicity of mechanisms, because describing activities contributes to making a mechanism intelligible (i.e. understanding how change is produced). “Interactions” are required too: “to identify what makes the producer productive” (Tabery, 2004, p. 11, my italics). Zooming in on which properties of one entity bring about a property change in another thus augments the use of “activities” to describe how a change is produced. I follow Tabery (2004) in taking the accounts to be complementary and below zoom in on dualist accounts of New Mechanists such as Machamer, Darden, Craver, Bechtel, Abrahamsen and Richardson.

7.10 Ontic and Epistemic implications

As discussed, there is a rough divide between the New Mechanists who propose that explanation is in the first place an ontic affair (e.g. Craver, 2007, inspired by Salmon) and those that see it as a primarily epistemic activity (e.g. Bechtel, 2006 & 2008). In this section a brief history of the divide is offered. Next the implications of these positions are presented. Finally, I show that trying to argue for the primacy of either is problematic and leads to confusion. Borrowing from Illari (2013), I demonstrate that thinkers on either
side of the debate in fact do not advocate an absolutely ontic or epistemic approach. Drawing on examples from the literature on mechanistic explanation, I argue that the ontic-epistemic duality is best regarded as an heuristic tool that can aid discussion of the differences between accounts rather than as an absolute and clear divide.

Distinguishing ontic and epistemic conceptions of scientific explanation is attributed to Wesley Salmon. On the epistemic conception, an explanation is “an argument to the effect that the event to be explained was to be expected by virtue of the explanatory facts. (…) there is a relation of logical necessity between the laws and initial conditions on the one hand and the explanandum on the other” (1998, p. 53). On the ontic conception “to explain an event is to exhibit it as occupying its (nomologically necessary) place in the intelligible pattern” (Ibid., p. 54). It is interesting that “intelligibility”, which at face value would seem at home in the epistemic conception, features in the ontic conception on Salmon's definition. This relates to his use of and reliance on “laws”: Salmon notes that the term may refer to both “scientific statements describing a regularity in nature” or the “regularity itself” (Ibid., p. 53).

Recall that the accounts of later New Mechanists barely mention laws but focus instead on parts and their interactions which demonstrate “regularities” – a usage that seems compatible with the second definition of “laws” by Salmon. For Salmon, to explain is to “fit the event to be explained into an intelligible pattern” because “we have formulated the law-statements in terms that we understand, or equivalently, that we have seen and identified the lawful regularity described by the law-statement” (Ibid., p. 54). Salmon himself goes on to note that not all laws are causal laws but that some instead reflect lawful regularities (Ibid., p. 55). He also argues that some laws are irreducibly statistical and that statistical explanations are “admissible” and perhaps even “indispensable” for science (Ibid., p. 58). For causal explanation, we therefore need not be determinists.

In his 1984 book, Salmon further develops the epistemic and ontic conceptions. He differentiates between the inferential, information-theoretic and erotetic versions of the epistemic conception. The inferential version holds that explanation aims to “provide nomic expectability for the event-to-be-explained” (1984, p. 84). Salmon dismisses this version because it is unhelpful for explaining low-probability events (hence Craver’s constraint E5). On the information-theoretic version, the aim of explanation is “increasing our information about phenomena of the sort we are trying to explain” (Ibid., p. 97). Although this version works well for low-probability events due to its use of statistical-relevance relations, these relations in themselves have little explanatory force (see Chapter 8 for a detailed discussion of statistical approaches) (Ibid., pp. 97-100). The erotetic version holds that requests for scientific explanation should be formulated as why-
questions. “Whether a given explanation is satisfactory will invariably depend upon the knowledge-gaps a given individual needs or want to bridge” (Ibid., p. 101).

In discussing the ontic conception Salmon states that “some regularities have explanatory power, while others constitute precisely the kinds of natural phenomena that demand explanation” (Ibid., p. 121). He stresses that “[t]he distinction [between regularities that have explanatory power and those that call for explanation] (…) does not coincide with the distinction between lawful and nonlawful regularities” (Ibid.). Rather, ontic accounts require an adequate conception of causality that takes into account both deterministic and probabilistic situations (i.e. situations in which event A is always followed by B and situations in which B only follows A in some of the cases as in spontaneous radioactive decay for instance). He thus invokes causal relevance to differentiate between regularities with explanatory power and regularities that call for explanation. Salmon himself dismisses all but the erotetic epistemic and ontic views, suggesting that the former contribute to explanations that can increase “our manipulative and predictive abilities” whereas the latter increase “our scientific understanding of natural phenomena” (Ibid., p. 134). He furthermore points out that in situations in which there are multiple different explanations for the same phenomenon (different due to different perspectives of the questioners), perhaps the explanatory question asked is not in fact the same for each, effectively ruling out the idea that there may be multiple independent explanations for a single phenomenon.

7.10.1 Mechanism: a thing in the world or a representation?
For the rest of my discussion, I turn back to the New Mechanists. The ontic-epistemic divide corresponds to the distinction introduced by Levy (2013) between mechanistic accounts as an approach to explanatory relevance on one hand and those that focus on scientific method on the other. Illari (2013) points out that the ontic-epistemic divide is not absolute in the case of mechanistic approaches because the debate is not about the metaphysical status of explanations themselves (p. 251). Rather, it is a question of focus (as Levy’s distinction makes clear), and thinkers from both “perspectives” concede that they need both ontic and epistemic constraints. For Craver, to explain is to situate a phenomenon in the causal structure of the world (2007, p. 21). Furthermore, he claims that an explanation refers to the causal structure of the world: the mechanism in the world is the explanation (Ibid., p. 27). An example of Craver’s need for epistemological constraints was presented earlier, where it was noted that what counts as a level of mechanism is perspectival because it is dependent on the explanatory context (Craver, 2007, p. 193). In other words, the kind of explanation that is being sought after by the researcher helps determine what is defined as the system.
Bechtel and Richardson (2010) are more cautious about making realist claims than thinkers such as Salmon and Craver. They hold that to explain is to show by means of representation how the behaviour of the system follows from parts and their interactions (Bechtel and Richardson, 2010, p. 17). Advocates of mechanistic thinking as a scientific methodology, emphasize that explaining is an epistemic activity that is dependent on human cognitive capacities and the state of technology (cf. Bechtel & Abrahamsen, 2005; Bechtel & Richardson, 2010). To avoid confusion about the status of mechanisms, it is important first to differentiate between the mechanism as a thing in the world and models of mechanisms which are representations. On the epistemic account, only the representations (models of mechanisms) can harbour explanatory power and thus serve as an explanation (Bechtel & Abrahamsen, 2005, p. 425). Craver, in contrast, also uses the term “explanation” to refer to the objective mechanism in the world and contrasts this to representations (e.g. texts, models, diagrams) of a mechanism. Thus for Craver, in contrast to Bechtel, an explanation is not a representation but instead a fact and a “full-bodied” thing. The following citation summarizes Craver’s view:

Sometimes explanations are texts—descriptions, models, or representations of any sort that are used to convey information from one person to another. Explanatory texts are the kinds of things that are spoken, written, and drawn. They are the kinds of things that can be more or less complete and more or less accurate. They are representations. Other times, the term explanation refers to an objective portion of the causal structure of the world, to the set of factors that bring about or sustain a phenomenon (call them objective explanations). (…) There are mechanisms (the objective explanations) and there are their descriptions (explanatory texts). Objective explanations are not texts; they are full-bodied things. They are facts, not representations. They are the kinds of things that are discovered and described. There is no question of objective explanations being “right” or “wrong,” or “good” or “bad.” They just are. (2007, p. 27).

7.10.2 The feeling of understanding from ontic and epistemic perspectives

Illari (2013) points out that on Craver’s use, the term “explanation” as used in the citation above, can be substituted by “cause” (p.242). For Craver, the world is full of mechanisms/explanations: “[t]hey just are” (Craver, 2007, p. 27). In defence of the ontic account, Craver emphasizes that the “feeling of understanding” is not enough to distinguish good explanations from bad ones. He draws on the causal constraints (E1-5) discussed earlier to determine which representations of explanations are good. Also, in Craver (2006), he allows that some mechanistic models can explain too (p. 355).
Chapter 7

Bechtel, whose focus is on describing mechanisms, agrees that the feeling of understanding alone is not enough for a sound explanation. The explanation must of course appeal to the causal structure of the world or else the explanation would have no content. "After all, explaining refers to a ratiocinative practice governed by certain norms that cognizers engage in to make the world more intelligible; the non-cognizant world does not itself so engage" (Wright & Bechtel 2007, p. 51). Bechtel offers various reasons for his epistemic approach and his prioritization of representations of mechanisms (versus Craver's prioritization of mechanisms). "The problem with this ontic view is that mechanisms do not explain themselves" (Bechtel, 2008, p. 18). For instance, a phenomenon may remain when the mechanism that produced it is gone and a now-absent cause cannot currently explain in the ontic sense (Bechtel 2006, p. 34). And, mechanisms were around long before we had any scientific explanations (Ibid.).

Furthermore, the notion of intelligibility figures especially in the epistemically oriented accounts of mechanistic explanation (e.g. Bechtel & Richardson, 2010). They propose decomposition and localisation as heuristic strategies “so that the [explanatory] task becomes manageable and the system intelligible” (Ibid., p. 23). The decomposition of a mechanism can proceed both structurally and functionally depending on the focus of the research (Bechtel & Abrahamsen, 2005, pp. 432-433). “Localization” entails linking parts to their operations: “to localize an operation is to assign it to a specific part” (Bechtel, 2008, p.14).

In discussing the epistemic adequacy of the mechanistic approach, Machamer, Darden and Craver conjecture that thinking about mechanisms in terms of their activities is essential for their intelligibility, regardless of whether they are complete or correct. Whether a mechanism is intelligible thus depends primarily on whether there is “an elucidative relation between the explanans (...) and explanandum (Ibid., p. 21) . An intelligible model of a mechanism must show either how possibly, how plausibly or how actually something works (Ibid., p. 21; Craver 2006, pp. 361-362; see Figure 7.4).

Differentiation between these three mechanism descriptions in fact depends on Craver's definition of a part (Craver, 2007, pp. 131-132). If the parts in the description do not meet Craver's criteria (i.e. they are not physiologically plausible, do not have stable properties, are neither robustly detectable, nor open to intervention) then we are dealing with a how-possibly model. There may also be cases in which it is not (yet) clear what type of model is at play: often, further testing is required to determine whether a how-possibly model turns out to be a how-actually model. Furthermore, Kaplan and Craver (2011) concede that “[t]he special sciences would be utterly paralyzed if complete how-actually explanations were the guiding objective” (p. 610). They acknowledge that an
ideally complete (and thus incredibly detailed) how-actually model is a “philosopher’s fiction” (Ibid., p. 609). In scientific practice, idealizing assumptions (similar to filler terms) are employed to enable researchers to “capture broad and robust patterns in the causal structure of a mechanism” (Ibid., p. 610). Thus, the boundaries between how-possibly, how-plausibly and how-actually models are not set in stone (Ibid., p. 612).

<table>
<thead>
<tr>
<th></th>
<th>How-possibly</th>
<th>How-plausibly</th>
<th>How-actually</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How parts and activities <em>might</em> produce the phenomenon</td>
<td>Real parts and their activities are described</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parts and activities need not be proven to exist</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 7.4 Models that show how.

7.11 Mechanistic approach in current apathy research

Many of the examples used by the New Mechanists are taken from the fields of biology and, more recently, the neurosciences. In this section I focus on the use of the term “mechanism” in the apathy literature reviewed in Chapter 2. For each of the articles, I searched for use of the word mechanism and if present, explored the context in which it was used. I show that the term “mechanism” is quite scarce in the literature but next provide examples of mechanisms that are presented yet not labelled as such in the literature.

In Chue and Lalonde (2014) the word “mechanism” appears in the context of pharmacotherapy where the activeness of a mechanism is further highlighted due to the use of the phrase “mechanism of action”. In the article, the phrase is used to describe which receptors are targeted by various pharmacological agents: “Bitopertin, a GRI, is currently in Phase III trials. Its *mechanism of action* is premised on boosting the synaptic availability of glycine, a coagonist of the receptor with glutamate” (Ibid., p. 781, my italics). Again, in the table summarizing “Emerging pharmacological agents for the treatment of negative symptoms in patients with schizophrenia” there is a column titled: “mechanism of action” (Ibid., p. 782). In the section on future research, the term “mechanism” features once more: “The neurobiological mechanisms underpinning
negative symptoms are complex and remain to be clearly elucidated” (Ibid., p. 785). Note that the use of the spatial metaphor (“underpinning”) suggests that a mechanism underlies or constitutes the negative symptoms. This is an implicit yet not inconsequential assumption – for more on this I refer to my treatment of Cramer and Borsboom, 2013 in Chapter 8.

In Kirkpatrick (2014), “mechanism” features once and is used in the phrase “mechanism of action”. It refers specifically to the workings of pharmacological agents: “[h]owever, the need remains for effective treatments of primary negative symptoms, either through psychosocial treatments or through agents with mechanisms of action different than current antipsychotic agents (Ibid., p. 5).

Limosin (2014) too employs the phrase “mechanism of action”: “[t]he mechanism of action of the influenza virus has not been elucidated, but the leading hypothesis suggests that in certain genetically predisposed mothers, antiviral maternal antibodies could cause an autoimmune reaction which is harmful to the fetal brain” (p. 4). Interestingly, in contrast to the previous two articles, the notion of a mechanism here refers not to receptor activity but to the proposition that a virus may cause an immune reaction that impacts foetal brain function. A very vague how-possibly mechanism is thus implied. Furthermore, in the conclusion of the article, the term surfaces again, accompanied by a spatial metaphor that implies that a mechanism may “underpin” the phenomenon under study: “[a]s such, a better understanding of the pathogenic mechanisms that underpin the occurrence of negative symptoms is an essential prerequisite for the identification of therapeutic targets” (Ibid., p. 5).

Foussias et al. (2014) zoom in on phenomenological aspects of apathy and, in contrast to the more neurobiologically oriented literature, the term “mechanism” does not occur in this review article at all. I contend that perhaps researchers dealing with physiological entities may be more inclined to think in mechanistic terms than those working with constructs or functional entities. The same holds for Foussias et al. (2015), in which conceptualization of negative symptoms and options for their clinical assessment and treatment are reviewed: “mechanism” does not appear in this article. In Fervaha, Foussias, Agid and Remington (2013), who review literature on the neurobiology of effort and motivation and measuring tools for effort, the term “mechanism” is not used despite their specific interest in neurobiology. Both Roth et al. (2004) and Liemburg et al. (2015) do not use the term “mechanism” either.

Clarke et al. (2011), who discuss fifteen apathy (sub)scales from a psychometric perspective, mention “psychological and neurological mechanisms”. They state that
“(…) recognition of apathy as a maladaptive consequence of psychiatric, medical, and neurological disorders lowers the risk of failing to recognize the disorders, failing to treat the disorders, or losing an opportunity to understand the psychological and neurological mechanisms that mediate this loss of goal-directedness/interest or motivation” (Ibid., p. 3). This is the first instance in the literature reviewed here, in which a “psychological mechanism” is hinted at. Note the use of the word “mediate” here, which implies an activity between the “loss of goal-directedness” on one hand and underlying “psychological and neurological mechanisms” on the other. The term “mediate” is common in scientific literature and seems to indicate a possible activity between parts themselves or in a part-to-whole relationship which has yet to be uncovered/explored. The term mechanism too seems to fulfil the function of a filler-term here to paste over the gaps in our understanding rather than refer to an interaction between proposed parts of a system (as in the more technical use of the term as discussed in Chapter 7). I add that this “pasting over”, despite its negative connotations may have great pragmatic value that ultimately aids intelligibility. The use of the term “mechanism” in a non-philosophical manner may also serve to avoid further discussion of elements of an explanatory account which are not deemed relevant by the researchers. De Regt (2017), in arguing for his contextual account of scientific understanding, notes that “different people single out different elements [in an explanatory account] as the crucial explanatory factor” (p. 127). I hold that if we were to accept only how-actually models of mechanisms in science, this could seriously obstruct scientific progress.

Selten, Wiersma and van den Bosch (2000), in their article titled “Distress Attributed to Negative Symptoms in Schizophrenia”, use the term “mechanism” twice in their article. As in all instances where the term features in a non-neurobiological context, it is used in a tentative, filler-term fashion and accompanied again by a spatial metaphor: “[p]erhaps the mechanisms underlying anosognosia and anosodiaphoria are the same, in that they prevent the patient from appreciating the full extent of the severity of the impairments. The nature of these mechanisms remains unknown” (Ibid., p. 742, my italics).

In sum, the term “mechanism” is present in only five of the ten articles reviewed. In most instances, it is used in the phrase “mechanism of action” and refers to workings between physiologically plausible, robustly detectable parts at the neurobiological level (of science). In only one article does the term feature, in passing, in the context of a psychological process which is not further specified. It would seem then that the language of mechanisms (in the context of mechanistic explanation) is not pervasive in the literature on apathy and that the idea of a psychological mechanism with (functional) parts is uncommon.
However, the rarity of the term does not necessarily imply that processes that could be labelled as a mechanism are not described in the literature. For, when we project the different types of mechanistic explanations discussed in the previous section onto the apathy literature, the notion of a mechanism that underlies a phenomenon seems to be quite common (even if it is not always labelled as such). Similarly, the term “to mediate”, suggests that there is a part-part or part-whole activity and thus hints at the presence of a (how-possibly) mechanistic explanation. It would seem that the use of these terms in the literature is compatible with both etiological and constitutive accounts of explanation.

Below, I zoom in on an example of an abstract mechanism which is not labelled as a mechanism by the authors and I provide an example from the literature which highlights the focus on parts rather than activities. Fervaha et al. (2013), without using the term “mechanism” offer a depiction of a mechanism in their figure demonstrating “[h]ypothetical components of the motivational process underlying goal-directed approach behaviour” (p. 2652). Figure 7.5 shows a cascade of constructed rather than robust physiological components with “sensory information” as tangible input, and “behaviour” as tangible output, with tiles joined by bidirectional arrows between hypothetical components: reward prediction, valuation of rewards, effort computation, and action plan. The authors refer us to other literature for further reading on the suggested components and neural underpinnings of value guided choice (Ibid., my italics). The bi-directionality of the arrows, which is interesting from the mechanistic perspective, is because “processing during one step may affect any subsequent or previous step, thereby causing continual updating” (Ibid.). The authors appear to offer a mechanism for goal-directed behaviour in the figure and hint at further layers of mechanism in referring to “neural underpinnings”. However, the focus is on discerning component parts rather than on their relations, represented by arrows. Thus the mechanism is incomplete in the sense that the interactions between parts remain in the shadows.

This focus on discerning parts rather than their activities/relations appears to be the case for other examples from the literature too. For instance, Limosin (2014) calls for “integrated vulnerability models” in his review of factors including immigration, urban environment, and disturbances in early parent/child relationships, exposure to cannabis, amphetamines, cocaine and infections in the development of schizophrenia. The literature on which Limosin’s review is based consists largely of statistical correlations between the factors and the symptoms of schizophrenia. The literature he refers to offers a method for interpreting the contributions of separate factors and how these factors (at multiple levels of organisation) interact. His call thus underlines the importance of tackling the question of integration that is central to this project because a mechanism to further explain how the factors (or parts)
mentioned in Limosin's review correlate is not (yet) available. The lack of a method for integration is evident in much of the literature discussed in Chapter 2. In the next section, the feasibility of the mechanistic approach for psychiatry is further explored.

7.12 The Mechanistic Model in psychiatry: testing Craver’s additional norms

In this section, the applicability of Craver’s additional norms for the assessment of sound mechanistic explanations are explored for the field of psychiatry. As discussed, Craver’s focus is not on how to describe mechanisms. Instead he suggests norms for the assessment of mechanistic explanations and to help differentiate them from functional analysis. His aim is to seek an alternative to classical reductionism as an account of constitutive explanation. He does this by specifying a regulative ideal for mechanistic explanation within the systems tradition. In Chapter 3, shortcomings of the traditional reductionist approach to explanation in the field of psychiatry were presented and in Chapter 6, problems with a law-based approach to scientific explanation in biology and medicine were demonstrated. Here, I explore the applicability of this alternative specifically for explanations in psychiatry. Recall that Craver’s additional norms are:

1) The mechanism must fully account for the target phenomenon (“including precipitating conditions, manifestations, inhibitory conditions, modulating
Chapter 7

...conditions, nonstandard conditions, and byproducts” (Craver, 2007, p. 161).

2) The explanation is constitutive: the phenomenon as a whole is described in terms of robustly detectable, physiologically plausible component parts (Ibid.).

3) Mechanistic explanations differ from aggregate and morphological explanations due to the notion of organization (active, spatial and temporal) (Ibid., p. 162).

4) Constitutive relevance must be accounted for. A mere functional analysis does not suffice and Craver suggests a mutual manipulability account (Ibid.).

The first norm, which holds that the mechanism must fully account for the target phenomenon and not just limit itself to input and output conditions (Ibid., p.161), raises a challenge: how *fully* must we take this accounting to be? How can we determine whether a mechanism explains without depending on “filler” terms? When does a mechanism “bottom-out”, or, in other words when are the verbs describing activities and nouns precise enough to no longer be mere filler terms? Craver (2006), when commenting on “ideally complete descriptions of a mechanism” concedes that “[f]ew if any mechanistic models provide ideally complete description of a mechanism. In fact, such descriptions would include so many potential factors that they would be unwieldy for the purposes of prediction and control and utterly unilluminating to human beings” (Ibid., p. 360).

For psychiatry, consensus on what a mental disorder even is or where to situate it (e.g. Philips et al., 2007; Kendler et al., 2011) is lacking. Consequently, defining the system within which to study mental disorder (i.e. the input condition) is difficult. Biologically oriented thinkers tend to regard the brain as the seat of mental disorders, others invoke the mind, psychodynamically oriented practitioners and enactivist thinkers take a broader approach where a psychiatric disorder pertains to a human within his/her context. If the boundaries of the system within which we seek the conditions that are involved with the development or manifestation of mental disorder are unclear, then how can this norm for a good explanation be met? And even when empirical research has demonstrated a statistically significant correlation between precipitating conditions, manifestations, inhibitory conditions, modulating conditions, nonstandard conditions, and by-products and the target phenomenon, how can we check that these *fully* account for the target phenomenon in the case of psychiatry. Many of the target phenomena and the conditions for their manifestation are in themselves poorly understood or vaguely demarcated, or, in the words of Kendler et al. (2011), they are multilevel and “fuzzy” (p. 1143). These problems are not easily solved. Therefore, to move towards a mechanistic explanation, the notion of “fully” accounting for the target phenomenon may need to be interpreted somewhat more loosely.
The second norm shall receive the most elaborate treatment here serves as a reply to third and fourth norms too. It holds that for a constitutive explanation, the phenomenon must be described in terms of robustly detectable, physiologically plausible component parts. As I will show, these two clauses are intertwined. Also, the norm runs parallel to the difference between how-possibly, how-plausibly and how-actually models. If we accept Craver’s second norm, to count as an how-actually model, the components must meet the criteria. They may not be “mere how-possibly fictions” \((Ibid., p. 161)\). As pointed out in the previous section, the additional norms present a challenge for psychiatry, a) because we struggle to define the system within which to study psychiatric phenomena in the first place, let alone decompose said system; and, b) because we do not know if the reductionist enterprise of finding a biological realisation at the cellular level of organisation is only a practical impossibility (due to lack of knowledge/technology) or a theoretical one.

In my survey of the literature, we saw that the phenomenon “apathy” is difficult to decompose into localisable parts. To ease the discussion, let us assume that apathy is a phenomenon that pertains to the human as a whole. Thus, the physical system within which we would seek a mechanistic explanation encompasses the individual and his direct environment. Most of the proposed components of apathy are functional, for instance amotivation, anhedonia and alogia. Engel would likely have regarded them as examples of psychological or social phenomena. For a mechanistic explanation, according to the additional norms (the second one in particular), we would need to specify robustly detectable, physiologically plausible parts that have the above functions or capacities within the system in which we have situated our target phenomenon.

First, I explore if the second norm can include Engel’s psychological and social features as part of a mechanistic explanation. Given that functional parts are often regarded as “mere constructs” or “fictions”, we need to address the question of what the status of a “construct” is in mechanistic explanation. In the apathy literature, examples of functional parts that do not (as far as we know) correspond to a definable “physiologically plausible” part within the target system (i.e. at a sub-human level of organisation) abound. Yet, reliance on “functional parts” and “constructs” is something Craver seeks to avoid because he argues that functional analysis does not amount to mechanistic explanation \((Ibid., p. 128-129)\): a phenomenon is not explained by capacities or functions, it is the capacities of the parts, which in turn interact in virtue of these capacities, that explain \((Ibid., p. 129)\). Craver suggests that explanations based only on capacities instead of on parts and their interactions cannot uphold constraints E1-E5. For instance, mere temporal sequences cannot be ruled out, real and pseudo-capacities cannot be adequately distinguished and functional
analysis lacks a method for weeding out causally irrelevant capacities \( (Ibid., 134-140) \). Craver (2006) adds that there are not as yet any accounts which can differentiate how-possibly from how-actually functional analysis (p. 362). How, then are we to deal with such functional parts for which there is no known biological realisation in the field of psychiatry?

My first line of argument addresses the assumption that functional parts are “more constructed” than physiological parts. I claim that physiological parts too are constructs, demarcated in virtue of their capacities. All components are selected through the heuristic tools of localisation and decomposition, which is ultimately fuelled by the explanatory task at hand. Therefore, the difference between physical and functional parts lies in their degree of constructed-ness. Unlike physical parts however, psychological and social parts are defined solely in terms of function for lack of a known “biological realisation”. I add that our choice of physiological and functional parts is of course informed, constrained even, by the causal and organisational structure of the world as we perceive it, or else the parts and explanations in which they feature would not be able to do any explanatory work in the first place.

I suggest that the parts of a system are always at their core functional and they can be placed on a spectrum (akin to the how-actually to how-possibly range) with functions with a virtually undisputed biological realisation at one pole and those for which we have no inkling of a possible biological realisation at the other pole. Thus, any part is always a construct that is more or less loosely, depending on our ability to decompose and localise, based on the world as we experience it and suppose it to exist. How we carve up nature, depends on the explanatory task at hand. This position raises a question about the ontological status of “parts” at the latter pole.

My second line of argument pertains to the requirement that parts should be “physiologically plausible”. This is confusing for a number of reasons. First, it is not clear what renders a part “physiologically plausible”. Craver takes “plausibility” to entail that “the parts’ existence should be demonstrable under the conditions relevant to the given request for explanation of the phenomenon” (2007, p. 132). So it would seem that what is plausible depends on a) demonstrability (which seems very similar to “detectability”, hence my statement that the two clauses are deeply intertwined) and b) the explanatory task at hand. The plausibility requirement thus introduces a clear perspectival or epistemic constraint. It also appears circular: what is plausible depends on the explanatory task at hand, but the explanatory task at hand is dependent on having plausible parts as a starting point. Second, and related to the former point which shows that what counts as plausible is subjective, it is unclear how “plausibility” and the soundness of a mechanistic explanation
Mechanistic Explanation

are related. There are countless examples of mechanistic explanations that start out as how-possibly models and are promoted to how-actually models with parts that at first seem implausible, far-fetched but that can eventually be detected and prove plausible after all. Craver concedes that “[t]here is no clear evidential threshold for saying when one is describing real components as opposed to fictional posits, or for detecting when one is pushing one’s hypothesis a bit far” (Ibid., p. 131). So far, physiological plausibility seems a weak constraint that does not sit well with Craver’s ontic focus and does little to help distinguish between how-possibly and how-actually models of mechanism.

Another issue with Craver’s requirement that a part is “physiologically plausible”, pertains to the term “physiological”. The notion that a part must be physiologically plausible suggests that Craver excludes psychological and social parts (for which a biochemical account of their biological realisation within the individual human is lacking) from mechanistic explanations. The latter are not mentioned at all in his additional norms and as discussed above, he eschews mere functional explanations. Can we then still conceive of mechanistic explanations which are multi-layered in Engel’s sense, i.e. in a sense that moves beyond the physical, in which psychological or functional parts interact with physiological parts? Or must we reduce everything to the physical first to develop an adequate mechanistic explanation on Craver’s definition? The idea that parts must be physiologically plausible is reminiscent of precisely the reductionism that Engel was seeking to excise.

An example from Engel (1980) discussed in Chapter 3 highlights the point (for clinical practice in this case and not scientific practice, but the point may be extrapolated): Engel describes the case of Mr. Glover, who has been hospitalised following a myocardial infarction (a phenomenon at the organ level) which interacts with phenomena at the person level (e.g. anxiety, memories of his previous infarction). He invokes causally relevant factors from the biological, psychological and social domains to explain what produced the phenomenon of myocardial infarction in Mr Glover. Craver’s second constraint would entail that a constitutive mechanistic explanation of Mr Glover’s infarction would not be possible because components such as memories of his earlier infarction and his experience of patienthood cannot be rendered in terms of physiologically plausible parts at the sub-human level of organisation. I would add that although they may not be physiologically plausible, the above-mentioned psychological components of myocardial infarction are nonetheless robustly detectable: one could interview Mr Glover or observe his behaviour to detect them.

Wright and Bechtel (2007), however expose difficulties in extending mechanistic explanation to psychological phenomena and draw on “motivation” – which they call “a
state” and later also “a construct” - as an example of a psychological mechanism (2007, p. 65). Their example however proves convenient because the discovery of brain reward circuitry provides the sought-after robustly detectable, physiologically plausible parts:

Confining motivation research to better-delimited characterizations of the target phenomenon with more tractable constructs was one way in which psychologists were able to get a handle on motivation. This coincided with the move toward models of mechanistic activity in lieu of nebulous instinct-need-desire taxonomies and descriptions of lawful stimulus-response patterns — a move fostered by the serendipitous discovery of brain reward circuitry in the mid-1950s. (Ibid., p. 67)

Wright and Bechtel (2007) continue to describe the discovery of component physiological parts and their operations in the brain. Their account confirms the conclusions drawn above: that for an adequate mechanistic explanation of a psychological phenomenon, the physiologically plausible parts that correspond to the psychological components must first be detected. Wright and Bechtel nonetheless, somewhat over-optimistically conclude that: “[i]n sum, explaining motivation mechanistically requires illuminating the organizational collusion and interaction of these various composite systems that engage their environment at increasingly higher levels” (Ibid., p. 72).

Craver’s third norm that mechanistic explanations differ from aggregate and morphological explanations due to the notion of active, spatial and temporal organization (2007, p. 162), highlights that a mechanistic model seeks to demonstrate the interactions between parts rather than merely describing their form or function. Given the above-mentioned difficulties in discerning parts and their interactions in the field of psychiatry, this is an attractive, yet difficult to achieve target. As shown in Chapter 2, in the study of apathy, despite attempts to examine different aspects of the phenomenon at different levels of organisation, none of the articles express a mechanism that would meet this standard. At best, in the literature sub-phenomena are proposed, a method for their measurement is suggested and finally the correlations between the values are calculated. Explanations or even hypothetical descriptions of how these proposed correlated parts interact in a how-possibly mechanistic model, do not feature in the literature.

So in sum, the requirements that a mechanism must 1) fully explain a phenomenon, 2) consist of robustly detectable, physiologically plausible parts, 3) draw upon principles of active spatial, and temporal organisation, and, 4) that constitutive relevance must be accounted for, are exceedingly difficult to meet for explanations in psychiatry. In this section I focussed on the second norm in particular and dissected it. I pointed out that Craver’s dismissal of functional components is problematic for psychiatry. I suggested
that physiological parts are constructs too: decomposition and localisation depend on demarcating their functional contribution the workings of the mechanism as a whole. I also suggested that the plausibility requirement is weak (given Craver’s ontic focus) because it is an epistemic rather than an ontic constraint driven by the explanatory task at hand and is therefore unhelpful in differentiating between how-possibly and how-actually explanations. Finally, I expressed concern about the risk of reductionism to the biological implied by requiring that parts are “physiologically plausible”. The difficulties that arise due to these additional norms leave us with two options: either, mechanistic explanations are per definition unsuitable for explanation in psychiatry, or, the definition requires adaptation to allow for inclusion of non-physiological, functional parts so that the mechanistic approach is feasible for a broader range of phenomena in psychiatry.

For now I cannot but agree with Murphy (2012), who observes that the explanations for which the mechanistic approach holds promise, are those that pertain to cognitive theories of attention, perception, motor behaviour and working memory, for example. In contrast, research with a stronger normative component, such as theories of planning, judgment, emotion, interpersonal relationships and problem-solving in contrast do not lend themselves well to a mechanistic approach, perhaps due to the poor decomposability and localizability of precisely their normative “parts” (Murphy, 2000, p. 198). Note that this conclusion fits well with the observations of the way “mechanism” is currently used in the apathy literature: the term arises almost exclusively in the context of biochemical mechanisms. Given this fairly disappointing conclusion, I now move on investigate other New Mechanists’ treatment of functional parts to adapt and augment the mechanistic approach to make it more suited to psychiatric research and practice.

### 7.13 Mental Mechanisms

To explore whether there is any hope left for mechanistic explanation of mental phenomena in the face of the problems in the previous section, I turn to Bechtel’s book *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience* (2008). I introduce some of Bechtel’s key terms such as his definition of information, distinction between vehicle and content, the notion of “representation” and his “heuristic identity theory” in order to provide a brief sketch of his account and compare his account with Craver (2006, 2007, & 2013). Bechtel’s starting point is the difference between mechanisms in biology, and mechanisms in psychology and the cognitive sciences. I show how Bechtel, much like Craver, ultimately requires a physical mechanism and regards the functional mechanism as a means, or heuristic tool to aid the discover of the physical mechanism to which it is related in a way which Bechtel does not specify by
relying on his “heuristic identity theory”.

Bechtel (2008) provides detailed examples from memory and visual processing to demonstrate how mental mechanisms are discovered. The process runs parallel to the discovery process for physical mechanisms outlined earlier in this chapter. He emphasizes that phenomenal decomposition (i.e., decomposing a phenomenon into subtypes of phenomenon) is but a “prolegomenon” (p. 61) to mechanistic decomposition (with functional and structural decomposition as subtypes) (Ibid., p. 50). Note that this is in agreement with Craver (2006), cited above. The former does not amount to explanation but serves only to define the explanandum. The exploratory process begins by first delineating the phenomenon for which a mechanism is responsible. As shown in Chapter 2, where the history and definition of the concept of apathy was discussed, this can be a huge challenge in itself due to both normative aspects and the inherent “fuzziness” of the parts and their interactions. This may provide clues on how to proceed to the next step: decomposition of the mechanism. To understand how a mechanism works, one may intervene on starting conditions or intervene on (part of) the mechanism itself and then detect how the behaviour of a component part or the mechanism as a whole is affected. Usually, various modes of intervention must be applied and their results combined in a process which again presents challenges - to gain an understanding of the workings of the mechanism (Bechtel, 2008, p. 39).

Bechtel characterizes mental mechanisms as mechanisms that process information or convey content and suggests that the physical stance and the information processing stance are complementary in psychology and the cognitive sciences (Ibid., pp. 22-23). Craver (2013) too, adopts the idea that there are two “stances” (pp. 134-135), noting that the two co-exist in an intermingled fashion in scientific practice. For Bechtel, the complementarity of the stances entails that “a mental activity also involves material changes” (Bechtel, 2008, p. 23) or in other words, that any mental operation corresponds to a physical embodiment.

For his definition of “information” Bechtel draws on Dretske (1981) adopting a causal, regularity-based notion of “information”. Bechtel states that it is the “regular effects of a cause that make it possible to infer features of the cause from features of the effect” (2008, p. 24). Bechtel’s appeal to regularity here appears to depart from Craver’s constraints E1-E5 (Craver, 2007, p.26). In particular, it does not evidently meet constraints E1 (mere temporal sequences are not explanatory) and E5 (causes need not make effects probable to explain them). How then, should we regard causality from the information processing stance given that the previous discussion of the constraints focussed on their applicability in the physical stance? Interestingly, many of the interventions used to
clarify the mental mechanism are at the physical level, for instance brain lesions and neural stimulation, and their effects are then examined from the information processing stance. Examples of interventions at the information processing level include measuring subjects’ ability to memorise items in a variety of situations. This situation is exemplary for psychiatry research: Chapter 2 offered examples from the apathy literature in which interventions from different domains (e.g. psychological tests and fMRI imaging) are combined seemingly seamlessly in practice in the search for an explanation of the phenomenon apathy.

Next, Bechtel again draws on Dretske (1981) and links information processing to the notion of representation: “when the interest is on the information an effect carries, the effect is of interest not for its own intrinsic properties but for the ability to identify or respond to the cause. The effect stands in or re-presents its cause, and hence is often called a representation” (Bechtel, 2008, p. 24). He introduces the distinction between “content”, which pertains to how an organism’s behaviour responds to environmental cues and “vehicle”, which pertains to the physical structure or realisation of the mental operation. “What is critical to understanding an information processing mechanism is how content is changed as a result of causal operations that change vehicles, for it is in virtue of content that the mechanism is linked to the states outside the mechanism” (Ibid., p. 25). Thus, Bechtel’s distinction between content (viewed from the information processing stance) and vehicle (best viewed from the physical stance) offers a way to avoid the reduction to the biological level because it makes it clear that both stances are equally crucial. Although Craver’s additional norms for a sound mechanistic explanation seem to exclude this manner of avoiding reductionism, Craver’s commitment to multilevel explanation may mean that he would not dismiss this method.

The distinction between content and vehicle is a helpful heuristic to highlight the need to allow for mental mechanisms besides physical mechanisms. In a similar vein, Bechtel’s “heuristic identity theory” offers a method to assume an identity relation between the mental and the physical for pragmatic reasons without drowning in debates about mind-body dualism. This is because the assumption that there is an identity relation between the mental operations and brain structures is problematic. Bechtel therefore holds that such an identity claim is an heuristic tool (2008, p.69), coining his position “heuristic identity theory”. The position appears a good way to avoid discussions of an ontic nature on the relation between the mental and the physical:

One of the virtues of viewing identity as a heuristic claim is that it can guide not only the elaboration of the two perspectives which are linked by the identity claim, but it can use each to revise the other. Thus, mechanistic research does not require
that investigators identify the component operations correctly before they attempt to localize them in the brain. As long as the initial hypothesis as to the operation performed is even roughly in the right ballpark, an identity claim can play a fruitful role in generating evidence that leads to revisions and refinements of the initial claim. (Ibid., p. 71)

The advantage of relegating important assumptions to the “heuristic tool” category is at the same time a disadvantage because it leaves open many questions as to the relation between proposed mental and physical explanations of psychiatric phenomena. Differentiating between “vehicle” and “content” is intuitively pleasing, but also problematic because it is unclear whether such a distinction, which works well in the hardware-software case also holds for psychiatric phenomena. The distinction between vehicle and content, if interpreted as an absolute divide, would obstruct a useful interpretation of the constitutive relationship between psychological phenomena and physiological parts. I suggest that, in the case of psychiatry, we need to leave open the possibilities that the two are highly mutually dependent, or conversely, that there may be no one-to-one mapping (i.e. phenomenal demarcation may not follow spatial demarcation in the brain) of the content to a specific brain region due to the plasticity of our brains and the influence of environmental causal factors. I contend that Bechtel’s inclusion of mental mechanisms in his account of mechanistic explanation, despite all the questions it raises about the nature of information and representation, should be adopted in order to render the mechanistic approach to explanation feasible for psychiatry.

7.14 How Now? Where we stand

Time for a résumé. To provide a case study, I reviewed recent literature on apathy (Chapter 2) in schizophrenia and examined the explanatory models at play in current research in psychiatry (Chapter 3, 4, & 5). An extensive exploration of the biopsychosocial model revealed serious flaws which make it of limited use to serve as a model for explanation and prompted a move to more abstract, philosophical models of explanation. In Chapter 6, three traditional philosophical models of explanation and their limitations were discussed: the covering law model, unification model and representation model. I adopted five constraints on what a good explanation should provide. These constraints, E1-E5, zoom in on how to detect causality. Armed with these constraints, I continued the search for a sound explanatory model suitable for psychiatry that allows for the integration/reconciliation of scientific perspectives or levels and arrived at the mechanistic approach.

The mechanistic approach has emerged as one of the most promising contenders as
an explanatory model for psychiatry because of the way it accommodates a multi-
level approach. However, although beautiful examples of mechanistic explanations in 
neuroscience have been provided by prominent New Mechanists (e.g. Craver, 2007; 
Craver & Darden, 2013; Bechtel, 2008), I have been unable to uncover any parallel 
examples in the literature on apathy. When used in the literature on apathy, the term 
“mechanism” seems to be reserved for predominantly “biological” aspects of the field 
and in some instances it is no more than a filler term used to indicate areas in which 
in our understanding is incomplete. The lack of even the most non-committal type of 
mechanism in the apathy literature, the how-possibly mechanism, may be due to the 
many gaps in our understanding, or because the phenomena under study in psychiatry 
simply are more “fuzzy” - i.e. further understanding will not clear the fuzziness altogether 
because the fuzziness is inherent to the phenomena. The applicability of the mechanistic 
approach to explanation in psychiatry seemed riddled with problems at this point and, 
due to its reliance on physiologically plausible parts (by Craver in particular), hearkened 
back to exactly the reductionism and essentialism we were seeking to avoid. This lead 
me to look further into the literature on mental mechanisms, specifically Bechtel’s 
account. Bechtel (2008) provides examples of mechanisms in which the mechanism 
shows how information is processed rather than describing activities of physiologically 
plausible, robust parts. By regarding the information processing stance and the physical 
stance as complementary positions, he made a case for the inclusion of mechanisms 
involving “information processing”. His account seems to sit more comfortably with 
actual psychiatric research, in which the different perspectives and stances are combined 
all the time.

Given the problems with a mechanistic approach in psychiatry, especially concerning 
the problems of detecting and defining the boundaries of the system under study in 
the first place and the parts and their interactions that produce the phenomenon, my 
search for an explanatory model that aids integration of scientific perspectives continues. 
The following chapter delves into the possibilities offered by Network Analysis as a 
mathematical framework for improving our understanding of complex systems. 
Network Analysis is an eccentricity in this context because, as I will argue in more detail 
in the next chapter, it is a mathematical methodology rather than a philosophical model.
Chapter 8

Network Analysis
8 Network Analysis

8.1 Introduction

This chapter outlines some of the key ideas of Network Analysis and their relevance and applicability in the field of psychiatry. Network Analysis is steadily gaining ground as a methodology which does justice to the complexity of phenomena in various sciences including psychiatry. It is somewhat of an outlier in the context of the previous chapters, which focused on models of explanation because network analysis in itself is a mathematical methodology and not a philosophical model for explanation. However, as I discuss later on in this chapter, the mathematic methodology of network analysis, when combined with philosophical assumptions regarding how the mathematical analysis is applied to data, may indeed harbour explanatory or proto-explanatory powers. The study of the explanatory potential of network analysis is therefore relevant to this thesis and especially for the second central question of this project regarding the possibilities for integration of knowledge from different levels of organisation.

The previous chapter outlined the mechanistic approach to explanation, according to which a phenomenon is explained in terms of its parts and their activities. Difficulties for this approach include the observation that psychiatric phenomena often do not lend themselves easily to decomposing the phenomena into robustly detectable and physiologically plausible parts (a requirement for Carl Craver (2007)). This is because many of the proposed “parts” in psychiatry are difficult to demarcate and localise: they are (perhaps inherently) “fuzzy”. Furthermore, even when New Mechanists allow for inclusion of functional parts (e.g. Bechtel, 2008; Craver, 2016; Wright & Bechtel, 2007), they advocate a model where these functional parts are linked to a biological substrate, thus again leading us back to the difficulties of determining such a substrate.

What mechanistic explanation and network analysis have in common, is that they both take inspiration from a complex systems theory approach to science and both signify a move away from traditional reductionism. As argued previously, the ideas of complex systems theory and avoidance of traditional reductionism are basic tenets of the approach to scientific explanation in psychiatry advocated in this project. A central question in this chapter is: can network analysis meet the demands of constraints E1 – E5 (Craver, 2007, p. 26)? To explore this, I first introduce network analysis, addressing advantages and disadvantages in the process before outlining the three most common types of network used to model phenomena in psychiatry. With this background information the degree to which these three networks can meet E1 – E5 is tested, followed by a discussion of the findings. This discussion leads me to call
the status of “causality” into question. I conclude by emphasizing that network analysis in itself is a mathematical method and not an explanatory model. I argue that network theory, i.e. the application of these mathematical methods to phenomena in psychiatry, offers proto-explanatory powers when we are sensitive to the philosophical assumptions at play concerning: 1) choosing how to apply the mathematical methodologies, 2) choosing which empirical methodologies to use and 3) interpreting the findings. I explore whether network analysis offers a feasible method to test whether constraints E1 - E5 can be met in the model of a phenomenon. I then demonstrate some disadvantages of embracing constraints E1-E5 (pertaining to causal relevance). To avoid these difficulties, I suggest regarding causality as a heuristic tool and as a term that may be axiomatized to avoid the problems that arise when trying to define it in absolute terms.

8.2 What is Network Analysis?

“Network analysis” is the mathematical analysis of data represented in terms of a graph, where the analysis is based on graph-theoretical knowledge about the properties of graphs. This knowledge is often referred to as “graph theory”. Network analysis can be applied in myriad fields to all sorts of data. The focus in this chapter is on network analysis as a methodology for the representation and interpretation of empirical findings in psychiatry in particular.

“Network theory” in turn, is a part of graph theory. It is deployed to model complex networks of objects or phenomena. Network theory entails application of graph theory to empirical/hypothetical data modelled in terms of a graph in order to describe or predict the model’s behaviour. Borsboom (2017) uses the term “network theory” to refer to a theory about the essence of psychopathology. According to his view, mental disorders are best conceptualized as strongly connected symptom networks. He offers convincing arguments to support the idea that such a network theoretical conceptualisation of mental disorder offers explanatory powers (of description, prediction and intervention) that can thus help us discover new interventions.

In this chapter I outline key terms in network analysis and offer some examples from the literature. Network approaches in psychiatry thus entail using mathematical analyses to identify or explore relations between phenomena, for instance symptoms of psychopathology. Networks, or “graphs” can be visualised by dots representing nodes and lines representing edges. In the psychiatry literature that draws on network analysis, this manner of modelling is common and helps demonstrate the structural properties of a network. Graphic representations make optimal use of humans’ ability to process highly dimensional information visually (Epskamp et al., 2012, p. 2).
To perform a network analysis for psychiatric phenomena, one must first identify the nodes, for instance as symptoms as described in the DSM. Second, one must define what kinds of relationship the connecting “lines” or “edges” represent. Examples are nodes that connect if symptoms are part of the same diagnosis in the DSM, or if clinicians or patients rate the symptoms as causally connected, or the edges may represent the frequencies of their co-occurrence (Borsboom & Cramer, 2013, pp. 98-100). Borsboom and Cramer note that one can zoom in or out on a network of nodes and edges. Zooming in, we see that separate symptoms (nodes) themselves may in turn be dissectible into networks of symptoms. Zooming out, a larger context comes into perspective, in which for instance social effects of the symptoms of the original network come into focus. The networks are thus part of a nesting hierarchy, organised in terms of “layers of complexity” (Ibid., p. 102).

In short, visualisations of networks may include:

1. Nodes that can represent either variables or entities. Entities may be symptoms of a disorder or aspects of a phenomenon. Two nodes are called “adjacent” or “neighbouring” if they are connected by an edge. Often, individual nodes are referred to by letters or numbers. Size and colour labelling for instance can be used to represent further information on a node.

2. How centrally nodes are placed in the diagram: strongly correlated nodes may be programmed to cluster in the middle, nodes that are more weakly connected are peripheral. This is because, in some visual representations, stronger correlations are represented by shorter edges, which brings strongly correlated nodes closer together (Epskamp et al., 2012, pp. 5-6). When nodes represent symptoms, the development of a symptom represented by a central node suggests a higher risk of developing other symptoms due to the stronger correlations with other symptoms).

3. The edges (or lines) represent the relationships between nodes. Relationships may be based on for instance which symptoms are clustered in the DSM (see Figure 8.1), reports by clinicians or self-reports from patients about which symptoms they observe/experience together or in sequence, frequency of correlations of symptoms in time.

4. Edges may be directed when relationships between the nodes are asymmetrical, or undirected, when it is unclear whether the edges are directed or when one wants to leave open the possibility of reciprocity (cf. for instance Costatini et al., 2015; Epskamp et al., 2012).

5. Edges may be unweighted (whether or not there is an edge depends on whether or not there is a connection) or weighted (the strength of the connection is indicated by the thickness of the line. Colour may be used to indicate whether a relationship is
positive or negative depending on whether the phenomena represented by the nodes induce or inhibit one another.

Figure 8.1, reprinted from Borsboom and Cramer (2013), provides an example of a network representation of the DSM-IV symptom space. Here, symptoms in the DSM-IV are represented as nodes (coloured circles). They are connected by non-weighted, undirected edges (grey lines) whenever they feature in the same DSM-IV disorder. “The color of nodes represents the DSM-IV chapter in which they occur most often” (Ibid., p. 100). The size of a node corresponds to its degree; i.e. the number of other nodes to which it is connected. In Figure 8.1, substance-related disorders, mood and anxiety disorders are highly connected (and are thus represented by larger nodes and can be seen to form clusters within the network).

Network analysis can be used to study how a network of connected nodes changes in time and to uncover trajectories (a wave of change) through nodes after initial activation of one or more nodes (cf. Schmittmann et al., 2013). In the oft-cited example of depression, the nodes represent symptoms of depression. A major life event, such as the loss of a loved one, may trigger anxiety and loss of appetite. These in turn may be connected (perhaps causally) to other symptoms such as insomnia, fatigue and weight loss, thus triggering a wave of symptoms (turning from not-present to present). If enough symptoms are triggered or activated within a pre-specified time-frame, it may be called a state-transition: the person moves from not being depressed to meeting the...
diagnostic criteria for clinical depression in the DSM-IV-TR. The “depression network” may differ from person to person (Ibid., p. 50).

All in all, it must be noted that the network analysis approach is extremely versatile: what counts as a node, and how edges are modelled are open to such a huge array of options, that it is very difficult to do justice to all of the possibilities here without ending up with an overly theoretical and abstract account (cf. Boccaletti et al., 2006). As I discuss later, network analysis’ neutrality with respect to the content it represents, harbours both advantages and disadvantages. So far, I have therefore given a mere indication of the possibilities for visualising a network in the case of a psychiatric phenomenon.

In the next section, I zoom in on the field of psychometrics, in which network analysis can be used to detect patterns in empirical data from psychology and psychiatry. I discuss three types of mistake that are commonly made in psychometrics so that we may explicitly avoid such errors.

8.3 Psychometrics

“Psychometrics” refers to the science of measuring cognitive capacities and mental processes. Network analysis provides an incredibly versatile mathematical method to deal with the data yielded by attempts to measure cognitive capacities. Borsboom (2009) draws attention to the question: “what does it mean for a psychological test to measure a psychological attribute?” (p. 3, italics in original). There is far more to psychometrics than mathematical analysis. Given the ubiquity of psychometrics in the field under study and the link to network analysis as a mathematical method for representing empirical data, “psychometrics” warrants brief treatment here. This is primarily because psychometrics is not purely a “measuring” enterprise. Instead, it is a rather more complex enterprise that harbours philosophical assumptions. There are myriad philosophical assumptions that are made before the analysis can even begin. For instance in formulating the research hypothesis, in choosing an empirical method for collecting data and in choosing the most suitable type of mathematical analysis.

For instance, in Chapter 2, in which the scientific literature on apathy in schizophrenia was sampled, we saw that questionnaires (e.g. those “measuring” negative symptoms) are an oft-employed tool for data collection. These questionnaires are already wrought with assumptions about the phenomenon they purport to measure (for instance that assumption that the items on the scale such as blunted affect, emotional withdrawal and social withdrawal can usefully be scored numerically and independently). These assumptions are very rarely articulated, much less argued for. In fact, most researchers are
oblivious to the philosophical assumptions on which they are operating (cf. Borsboom, 2009; Lamiell, 2013). This is to the detriment of the results they find and progress in the field more in general. In this section, the following three common psychometric mistakes, which are in turn based on implicit philosophical assumptions, are discussed: “statisticism”, “operationalism” and appeal to a “latent variable”.

### 8.3.1 Statisticism

Statisticism is the view that “knowledge of the statistical properties of trait constructs can enhance the power and scope of scientific psychology’s capacity to account for (predict, explain, understand) individual behavior” (Lamiell, 2013, p. 66). Lamiell (2013) warns us against statisticism’s inherent dangers in personality research but adds that his line of argument may be broadened to hold for other traits or phenomena as well. Given the ubiquity of trait measurement in the apathy literature, I thus take his line of argument to be relevant to apathy research too.

Lamiell first emphasizes that statisticism poses a conceptual problem rather than that it indicates a problem with statistical methods themselves. He posits the central tenet of the argument against statisticism: “population statistics cannot serve as a basis for claims to empirical knowledge about individuals” (Ibid., my italics). Borsboom (personal communication) instead suggests that population statistics can serve as a basis for knowledge about individuals, but that this knowledge is indirect and that we should indeed proceed carefully when it comes to interpreting statistical relations at the individual level. For instance, a clinician who knows that psychoactive agents can cause psychotic symptoms, cannot, when faced with an individual patient who frequently smokes cannabis and is now suffering from psychotic symptoms, determine that this individual’s psychotic symptoms are caused by cannabis. The statistical data can however, lead the clinician to advise the patient to abstain from smoking cannabis because it is correlated to psychotic symptoms.

A danger that can result from statisticism is the reification of the traits that are measured statistically. Traits such as “cognitive and behavioural flexibility”, “diminished expression”, and “anhedonia” are at risk of being regarded as real entities rather than useful fictions. I return to this point below, when discussing appeal to “latent variables” but first address a second pitfall of psychometrics.

### 8.3.2 The Measurement Problem

The “measurement problem” refers to the problem of differentiating between a phenomenon and the numerical test-score used to measure it (e.g. depression and an individual’s numerical score on a depression measurement scale). Borsboom (2006)
regards confusing the phenomenon and its numerical test-score as a psychometrical faux pas, based on “operationalism”. He rejects “operationalism”, stressing how important it is not to substitute such a score for the phenomenon because a theoretical attribute is not synonymous with the way it is measured (2006, p. 428). Instead it is essential to investigate how the score follows from the phenomenon. In turn, an understanding of this relation is necessary for accurate interpretation of the data. For instance, if sum scores can be linearly ordered, this does not mean that the attribute being measured lends itself to such ordering. Classical test theory, which is still pervasive in psychology, holds that the theoretical attribute and its test-score are axiomatically fixed. Modern test theory in contrast, allows that the relation between the two may take on many forms (Ibid., pp. 428-429). This realisation is helpful for a network approach to psychiatry. For instance, we need to recognise when defining a node whether it represents an empirically known entity, a theoretical construct, or a value on (part of) a theoretical construct’s measurement instrument. We must furthermore recognise that such values often are oversimplifications - they may reduce a complex, multi-factorial phenomenon such as insomnia or a hypothetical construct such as general intelligence to a binary value on a single node. In some cases, further deconstruction (into separate nodes perhaps) of such oversimplified phenomena may be necessary to develop a more accurate model of the over-arching phenomenon under study.

In an experimental set-up, the expected presence or absence of an underlying latent variable needs to be accounted for explicitly. Factor analysis, a sub-type of network analysis, was used extensively in DSM development. Factor analysis is applied in situations where the presence of a latent variable is assumed. I return to the pitfalls of such an assumption below, but first address what appears to be an inconsistency in Borsboom and Cramer’s (2013) account which pertains to the measurement problem.

A useful spin-off of the network approach, is that it does away with psychiatric disorders’ identity crisis: discussions about whether depression is a construct or a thing in the world dissolve (Borsboom & Cramer, 2013; Borsboom, 2017). They write: “[t]he idea that mental disorders are network structures provides a new answer to the old question of whether mental disorders are real, and if so, in what sense” (p. 115, my italics). My critique is that the citation suggests an identity relation between network structures and mental disorders. And, assuming such an identity is to fall prey to a version of the measurement problem, thus confusing the phenomenon and a method used to represent the phenomenon. I quickly note, that elsewhere in the same article, the authors are more careful, stating for instance that “[t]o the extent that these causally active symptoms have been charted adequately, we thus already have a quite good idea of what constitutes mental disorders” (Ibid.). This citation, due to the use of the verb “charted”, more explicitly differentiates the phenomenon from its representation. For my own purposes,
I follow Borsboom and Cramer (2013) in embracing network analysis, but am careful to add that a network is a model or representation of the phenomenon under study. I am agnostic with regard to any matters metaphysical, i.e. pertaining to the organisational structure of the target phenomenon itself.

Having dealt with the above intermezzo, I discuss the issue of the latent variable in the next section.

### 8.3.3 The Latent Variable

Latent variables are variables that are mathematically inferred through measurement of other observable variables. Latent variables need not correspond to actual things or entities and instead often correspond to constructs. The disorders described in the DSM are latent variables – they lend their existence to observable variables – symptoms such as “hallucinations” or “disorganised speech” are taken to reflect an “underlying” psychotic disorder.

As Borsboom and Cramer (2013) argue, the network approach, which eschews commitment to latent variables, has “the potential to provide a way about thinking about disorders that does justice to their complex organisation” (p. 93). In the case of understanding psychiatric disorder, they depart from the pervasive notion that symptoms result from some “underlying” root cause or disorder (cf. Frewen, 2013; Borsboom, 2006) as is the case in analyses involving factor analysis. Borsboom and Cramer (2013) offer two propositions:

(a) Given the current evidence, we should forestall the conclusion that symptoms of the same disorder are uniformly caused by a single psychological or biological condition (or a single constellation of such pathological conditions), and (b) psychopathology symptoms causally influence one another. Hypothesis \(a\) is merely a matter of scientific prudence, given the absence of credible mono-causal explanations for how psychopathology symptoms arise. Hypothesis \(b\) cannot reasonably be denied by anyone familiar with the symptoms that are typically listed in diagnostic systems in psychology. (Ibid., p. 97)

Mental disorders are not something over and above their symptoms - perhaps “under and below” is a more appropriate expression given the common yet misleading idea that the disorder underlies the symptoms. “As an important corollary, this means that disorders cannot be causes of these symptoms” (Ibid., p. 95, my italics). Traditionally, on what is called the reflective model, a disorder is assumed to be a latent variable which produces symptoms and which needs to be uncovered to further our understanding.
Chapter 8

of the disorder. This idea is depicted in Figure 8.2 (from Cramer et al., 2010, p. 139), where “disorder A” is connected to a series of symptoms $X_1$ to $X_5$ by directed edges suggesting that the disorder leads to the symptoms. The symptoms thus reflect a latent variable and a change in this latent variable would be taken to effectuate a change in the symptoms. DSM disorders are modelled reflectively: the symptoms stem from the latent variable – the psychiatric disorder. Schmittmann et al. (2013) contrast this with the formative model in which the (latent) variable is modelled based on the symptoms (pp. 43-44). In formative models, the direction of causality between latent variable and the observable factors is reversed: changes in these factors cause a change in the underlying latent variable. An example of such a latent variable is socio-economic status (SES) where changes in education, income and occupation (the measurable factors) lead to a change in the SES.

For psychiatry, the notion of a disorder as “underlying” symptoms has not proven fruitful for understanding what a psychiatric disorder is. So far, few robustly detectable, physiologically plausible entities have been found to take the place of proposed latent variables (e.g. a genetic defect, or an abnormal neurotransmitter metabolism underlying symptoms at the behavioural level). Another poignant example is the case of mental disorders. The diagnosis of a mental disorder is based solely on the observable symptoms and not on the detection of the latent variable (the mental disorder) by some other means. Nonetheless, an overwhelming majority of researchers and clinicians (not to mention patients themselves) reify latent variables and attribute causal powers to them. For instance, clinicians frequently write that a patient’s hallucinations are due to their schizophrenia, or that depression caused someone to attempt to take their own life. In fact, many of the constructs that emerge from the latent variable approach have a deeply ambiguous ontological and epistemological status rendering them of limited use for research purposes.

Schmittmann et al. (2013) point out that both the reflective and formative latent variable models harbour three problems: “the role of time, the inability to articulate causal relations between construct and observables in terms of processes, and the subordinate treatment of relations between observables” (Ibid., pp. 45-46). Philosophically minded thinkers (e.g. Borsboom et al., 2011; Costantini et al., 2015; Cramer et al., 2010; Markus & Borsboom, 2013) argue for abandonment of the still pervasive appeal to latent variables for other facets of psychiatry too, for instance in personality and comorbidity research. Latent variables, which are traditionally taken to underlie behaviours measured in personality research, are unnecessary if one accepts that the behaviours themselves are directly related variables. For instance, instead of positing “extraversion” as a latent variable which leads to “enjoying parties” and “having lots of friends”, “enjoying
parties” itself may increase the number of people one meets which in turn leads to more invitations to parties without any need for invoking an underlying construct such as “extraversion” (Costantini et al., 2015, p. 14). See also Figure 8.3 (from Cramer et al., 2010, p. 140) for a representation of a network analysis in which symptoms of a disorder themselves interact.

Figure 8.2 A model of comorbidity between disorders A and B, under the standard assumptions of latent variable modelling. The circles represent the disorders (i.e., latent variables) and the rectangles represent the observable core symptoms of those disorders (i.e., X₁ to X₅ for disorder A, and Y₁ to Y₅ for disorder B). In this model, comorbidity is viewed as a correlation between the latent variables, visualized by the thick bidirectional edge between disorders A and B. (Cramer et al., 2010, p. 139; reprinted with permission from publisher).

Figures 8.2 and 8.3 show the difference between latent variable modelling and a network analysis approach. In the latent variable model, the disorders (recall that these cannot be taken to be anything more than constructs) are assumed to interact causally in the case of comorbidity. In contrast, in Figure 8.3, the “disorder” encompasses the collection of symptoms (represented by the circles around the three clusters of symptoms). The symptoms themselves, some of which are a feature of more than one psychiatric disorder (as was depicted in Figure 8.1), causally interact and are thus joined by edges in the schematic representation. These symptom interactions, instead of interactions at the level of “disorders” are taken to play a role in development of comorbid disorders.

In short, the theoretical and practical obstacles to furthering our understanding of psychiatric disorders raise questions about the feasibility of the idea that there is an underlying latent variable feeding the jungle of symptoms in the first place. The network analysis approach allows that there may be a causal interplay between aspects of a phenomenon (e.g. symptoms) themselves without appealing to the idea of an underlying construct (such as a DSM disorder, or such as apathy). The goal of network analysis in...
answering questions about the nature of psychiatric disorder, then, is to extract clinically and scientifically relevant information from the interplay between symptoms, regarded as a “network of autonomous causal entities” (Schmittmann et al., 2013, p. 47). It is this network that best represents the psychiatric disorder. These insights, which build largely on examples such as DSM diagnoses and personality research may be translated to apathy research too. An important take-home message here is that we must carefully interrogate the philosophical assumptions (for instance that there is some disorder causing psychiatric symptoms in the manner that a brain tumour can cause a headache) that are inherent in any choice of statistical representation in order to more fully understand the blind spots in our research methodologies. The methods themselves are often not the problem. Instead, our application of methods, without acknowledging their limitations in the context of the research question at hand may be obstructing scientific progress. Thus, use of network analysis in psychiatry raises challenges at the empirical level (collection of data), the philosophical level (choosing a method to represent and interrogate this data and being aware of the implications of such a choice) and finally the mathematical level (application of methodologically sound statistical analysis).

In discussing statisticism, the measurement problem and the latent variable, three downfalls of the employment of statistical methods in psychiatry have been briefly addressed. The account is far from exhaustive but nonetheless draws attention to some of the problems that are most frequently discussed by network analysis advocates. In the next section, I return to network analysis and zoom in on three ways in which a network may be modelled in psychiatry research. Advantages and disadvantages of the network approach are articulated and I explore how each of the three types of network analysis
holds up to constraints E1-E5 for the detection of causal relevance.

8.4 Types of networks

Very roughly speaking, in the literature on network analysis in psychiatry, three main types of network appear most commonly. They may be distinguished based on how the edges that represent a relationship between nodes are configured. These types are 1) undirected graphs, 2) directed graphs and 3) directed acyclic graphs (DAGs), a type of directed graph.

I emphasize that the examples of network analysis from the psychology and psychiatry literature to which I now limit my discussion merely scratch the surface of the possibilities afforded by network analysis. The above-mentioned types of graphs are the most common, but I highlight that the data that are fed into such an analysis may prove informative in yet new ways when they are represented in different manners and interpreted using other statistical methods. Although this is very exciting, a more detailed exploration of the possibilities is well beyond my expertise and beyond the scope of this thesis. I therefore base my tentative conclusions in the remainder of this chapter on the ways in which network analysis is already employed in psychiatry rather than attempting to do justice to the full array of possibilities that further use of network analysis in psychiatry may yield.

8.4.1 Undirected Graphs

Undirected graphs have edges which have no orientation; the edges are lines without arrows (see Figure 8.3 for an example). Adjacent nodes are connected by edges which may be weighted (i.e. numbers are assigned to the edges). An edge represents a relationship: for instance it may show that the phenomena represented by adjacent nodes co-occur in the context under study. A possible direction of causality, which is represented by an arrow in directed graphs, does not feature in this type of network. Undirected graphs therefore do not show causal relationships because the nodes are connected by edges that imply statistical correlation and not a causal relation. Such graphs also do not display temporality in the examples that I have come across in the psychiatry literature: they offer no information on the order in which nodes are activated. Undirected graphs lend themselves well to displaying cross-sectional data. For instance, to answer the question: which symptoms were present at a particular time in a particular individual, or the question: which symptoms co-occur in a particular study population? Figure 8.4 provides an example of such a graph based on time series data collected over two weeks from a clinical patient (reprinted from Epskamp et al., 2018).
Figure 8.4 Network representation of time-series data analysis from a clinical patient collected over two weeks. Blue links indicate positive relationships, red links indicate negative relationships, and the width and saturation of a link indicates the strength (absolute value) of the relationship (Reprinted from Epskamp et al., 2018, p. 3, no permission from publisher required.

8.4.2 Directed Graphs

In a directed graph the edges harbour an asymmetry: the nodes are ordered so that the connection from A to B is not the same as that from B to A (Boccaletti et al., 2006, p. 180). The edges between the nodes (A, B, C and D) are usually represented by arrows, for example, $A \rightarrow B \rightarrow C \rightarrow D$. Here again, although it is tempting to interpret the arrows as causal relations, it must be noted that the arrows represent statistical relations and that causal information can at best be inferred indirectly (more on this later).

A graph in which two nodes are connected by more than one link (i.e. one from A to B and one from B to A, e.g. Figure 8.6), or in which there are loops (a link from a node to itself) is called a multigraph. A directed graph can also contain a cycle: it has a set of nodes connected by directed edges that allows a closed walk (i.e. starting and ending at the same node) of three or more nodes without repeating any edges during one instantiation of the cycle. However, the cycle may be repeated as a process follows the cycle in time. Node A may thus be activated at $t_1$ when the situation it represents occurs, and again at $t_5$ following activation of node D at $t_4$ (see Figure 8.5).

Cycles may be arranged in three different ways: as negative-feedback cycles, positive-feedback cycles and contingent-feedback cycles (Clarke, Leuridan, & Williamson, 2014). Negative-feedback cycles often contribute to maintaining an equilibrium or
homeostasis. Positive-feedback cycles tend to lead the system away from equilibrium. Contingent-feedback cycles, in contrast, do neither of the above and the behaviour of the parts are influenced largely by external, environmental factors rather than internal positive or negative feedback.

Cycles abound in the sciences, especially in biology and medicine. Examples include metabolic cycles such as the Krebs cycle; life cycles of organisms and cycles controlling homeostasis such as the feedback loop controlling thyroid hormone secretion. Pathological processes too are often modelled as cycles. Clarke, Leuridan, and Williamson (2014) use the example of brain swelling after trauma. Trauma -> swelling in the brain -> increased pressure -> decreased blood flow -> hypoxia -> death of brain cells -> further swelling (Ibid., p. 1656).

A graph may also contain a loop, in which a single node is found to be connected to itself. Such loops may suggest an area that warrants further investigation. For a simple example, take the case of sneezing. One sneeze is often followed by another. If “sneeze” is modelled as a node, then multiple sneezes may be represented as an edge linking the “sneeze” node to itself. Yet, further investigation of the phenomenon of sneezing may yield a more complex network, in which persisting nasal irritation caused by pollen or mucosal inflammation due to a virus may in fact provide a more accurate or complete representation of relevant factors.

Figure 8.5 Directed (cyclic) graph. A: anhedonia (no enjoyment from formerly enjoyable activities); B: avolition (no will to do anything); C: inactivity; D: apathy.

Figure 8.6 Nodes adjoined by bidirectional edge. I: lack of appetite; J: anhedonia.

8.4.3 Directed Acyclic Graphs
The third type of graph that is employed frequently in the literature, is the directed acyclic graph (DAG). The DAG is a special subtype of directed graph: in contrast to directed graphs, which may contain cycles and loops, the DAG has no cycles. DAGs
have a linear configuration (see Figure 8.7 for an example). Given the interpretational difficulties that arise with directed graphs, especially those that contain non-linear node configurations such as cycles, DAGs offer advantages when it comes to detecting possible temporal and causal connections (e.g. Pearl, 2009, p. 14). Although Pearl (2009) acknowledges that “statistical relations do not logically imply causation” (p. 42), he concedes that causal relations may be inferred from the data through various different (mathematical) approaches. I return to the question of causality in network analysis in more detail below.

Directed acyclic graphs can be used to identify relations between the phenomena represented by nodes. See for instance Figure 8.7, a DAG in which the nodes represent caffeine intake, rumination, insomnia and daytime sleepiness. Let us assume that each node is a dichotomous variable: it can be either “on” or “off”. In the example, we may start out by drawing in the directed edges based on our causal intuitions. Next, empirical data (often cross-sectional population data) can be fed into an analysis to see which nodes are “activated” in which order and to see whether the “insomnia” node can be activated in absence of caffeine intake or rumination. Pearl (2009) employs Bayesian inference, through which the probability that an event, such as insomnia, was indeed responsible for the observed outcome, i.e. daytime sleepiness in the example, is calculated. The value of the variable H (daytime sleepiness) is dependent only on the value(s) of its parent node(s). This is called the Markov Condition. In the example, the value of variable H is conditional on the value of G, its parent node. The value of G can in turn be calculated from the values of E and F.

Directed acyclic graphs are often interpreted in terms of interventions. This interpretation fits well with Woodward’s (2003) interventionist account of causation which holds that that causality requires counterfactual regularity: i.e. if event A had not occurred, event B would not have happened. Pearl (2009) provides a detailed guide to statistical analysis of DAGs and ways to infer causal relations from the data. An intervention may entail fixing the value of a node. For instance, in the example in Figure 8.7, “insomnia” may be intervened upon through interrupting the sleep of subjects so that the insomnia node is switched “on”. The effects of this intervention on the value of node H (daytime sleepiness) can then be measured. The “causal effect” on the occurrence/value of node H that results from intervening on node G can be measured by calculating the difference in outcome on H for different values of G. Pearl (2009, Chapter 3) refers to such interventions as fixing a “do-operator” and then calculating the effects of such an intervention statistically from the data.
So far, the network approach seems promising as a method for exploring and contributing to explanation of phenomena in the field of psychiatry. It is in keeping with dynamical systems theory because, especially in the case of directed cyclic graphs, the mathematical framework can represent non-linear processes such as (causal) reciprocity which are taken as hallmarks of complexity.

Advantages of a network analytic approach to psychiatry thus include that:

- Disorders are conceptualised as systems of causally connected, or clustered symptoms rather than as effects of a latent disorder (Borsboom and Cramer, 2013, p. 93 & p. 97; Borsboom, 2017).
- Network representations can do justice to the complex organisation of mental disorders.
- Network theory of psychiatric symptoms holds that disorders may exist as symptoms rather than as entities: a single disorder may not necessarily have a single causal background and disorders are not merely labels for an arbitrary set of symptoms either (cf. Goekoop & Goekoop, 2014; Borsboom, 2017).
- The network approach, due to its neutrality with regard to the content it is used to represent, allows for the intrinsic “fuzziness” of systems of symptoms without putting this down to methodological flaws. Our inability to find true boundaries between disorders is because there are none. (Borsboom & Cramer, 2013, p. 97).

However, there are also some difficulties for network analysis applied to psychiatry data as an explanatory aid in psychiatry. First, defining the parameters to be able to discern patterns reliably, is tough because it is a highly complex and somewhat arbitrary process. In the example of depression, labelling the activation of nodes as a state transition when five or more DSM symptoms of depression are present, is relatively straightforward because it simply follows the definition of the DSM. Here the problem of arbitrariness involved in deciding on a cut-off is parked at the DSM’s door rather than at the network.
analyst’s. However, there are countless examples where defining what counts as a state transition is a challenge in itself. The “joints” of the psyche do not lend themselves to clean-cut carving. In this vein, Schmittmann et al. (2013) observe that “[t]herefore, in order to study networks of variables, simplifying assumptions will often be made, and simulation methods can be helpful in investigating the plausibility of network properties” (Ibid., p. 48). Also, network analysts need to be aware of the implications of their decisions regarding what the edges in a graph represent. The difficulties discussed here are entirely empirical and philosophical, and are not due to flaws in the mathematical methods themselves.

A second difficulty is that, after defining the nodes, one must rely largely on patient reports, and/or clinical observations for lack of other markers to decide whether a symptom is sufficiently present to “turn on” the node. For some nodes, the symptom may be more easily quantified (e.g. weight-loss could be specified as a weight reduction of 5% or more in three months), than others (e.g. fatigue). Note again, that the problems entailed by reliance on patient reports of clinical observations and the choice of a threshold for node-activation are independent of network analysis as a mathematical method.

A third source of problems is the question of determining whether two nodes that are statistically connected are causally connected too (e.g. Spirtes, Glymour & Scheines, 2001; Shipley, 2002; Pearl, 2009). Schmittmann et al. (2013) note that inference of causal relationships is possible if the relationship between the nodes is assumed to be linear, as opposed to cyclical, and normally distributed (p. 48). To detect causal relations, for three correlated variables A, B and C, this requires testing whether one can occur without the other and defining the direction of causality. I return to this point at more length below.

A fourth concern, related to the previous issue, is at a more abstract level: various thinkers have questioned the explanatory power of network models. Rathkopf (2015) argues that networks have explanatory power when it comes to explaining whole-system behaviour in “highly abstract and idealized” systems (Ibid., p. 10). He argues that more abstract and idealized systems also yield deeper explanations precisely because they are not system-specific. He draws upon the example of disease transmission as a target phenomenon, where, as he claims, it is more interesting to discover general patterns than it is to model the spread of a particular disease at a particular time. It is hard to see how this would translate to the situation of psychiatry at present, where researchers are attempting to use a network approach to model specific instances of psychiatric phenomena to first get an idea of their basic (causal) structure. In other words, obtaining a more solid
basic understanding of the structure of (psychiatric) phenomena seems useful before moving on to exploring more abstract whole-system data. Without this step, we may risk making misguided generalizations, that are insufficiently empirically grounded. Nonetheless, Borsboom’s hypothesis that psychiatric disorders are best modelled as networks of symptoms which are taken to be causally connected, may be regarded as an abstract example in the sense intended by Rathkopf (2015). On the “Network Theory” of psychiatry account, network modelling can yield predictions about the behaviour of a system of causally connected symptoms.

Use of network analysis in psychiatry has focussed on the behaviour of particular (clusters of) nodes within the system as well as whole-system behaviour. So far, the “whole systems” I have encountered in the psychiatry literature represent very specific situations. By “specific” I mean that a graph may be used to represent the occurrence of depression symptoms in one particular subject, or the symptoms that occur in the DSM for instance. Individual nodes are usually variables that correspond to heterogenous phenomena (e.g. fatigue and anhedonia may be represented as nodes). This contrasts the use of network analysis in some other fields, such as genetics and cognitive neuroscience. The items represented by nodes there, tend to be more homogenous (e.g. nucleic acids, genes, neurons) real entities and are thus easier to compare qualitatively. Also, networks employed in fields that have more robustly detectable, physiologically plausible (and homogenous) entities at their disposal, often have far more nodes because more data is available (whole genomes or brain connectomes can thus be represented). In the latter case, the idea of making empirically useful inferences from statistical data (in the ways intended by Huneman (2010) and Rathkopf (2015)) takes on a different form than in psychiatry research.

In sum, this section addressed some advantages and disadvantages of network analysis specific to the field of psychiatry. Importantly, network analysis’ neutrality with regard to what is represented by a node is both a virtue and a vice. It is a virtue in that the neutrality entails that entities that are difficult to compare because their organisational structure and/or constitutive mechanism is unknown or because they are “mere” constructs, can still be represented side by side as nodes. It is meanwhile a vice because it means that the scale of the networks remains limited because nodes are so heterogenous. Mere statistical interpretation of patterns in the data is in practice fruitless unless the properties of the individual nodes are included in the interpretation of the data. Sensitivity both to the constraints given by the nature of the field of psychiatry, and to our current state of knowledge are essential to determine which types of networks are most suitable for the representation of psychiatric phenomena. Although network analysis is at root a mathematical enterprise, rendering its employment productive thus
requires consideration of philosophical aspects too. In the next section, I move on to the explanatory power of network analysis.

### 8.6 Network analysis and Craver’s causal constraints

In this section, the idea that *correlation does not imply causation* is explored for network analysis. To test the merits of the network analysis approach in the context of explanatory models in psychiatry, I explore whether, in the form in which it is currently employed in psychiatry, it is equipped to meet constraints E1-E5 (introduced at length in Chapter 6). Recall that these constraints hold that:

- **E1:** mere temporal sequences are not explanatory (temporal sequences);
- **E2:** causes explain effects and not vice versa (asymmetry);
- **E3:** causally independent effects of common causes do not explain one another (common cause);
- **E4:** causally irrelevant phenomena are not explanatory (relevance);
- **E5:** causes need not make effects probable to explain them (improbable effects).
  
  (Craver, 2007, p.26)

For the purposes of this project, I have embraced constraints E1-E5 and noted that to earn the label “explanatory model” a model must organise the content in such a way that it is possible to meet them. The questions of whether network analysis can meet E1-E5 and whether network analysis counts as a stand-alone explanatory model are thus related. Note that thinkers who do not accept E1-E5 as a “litmus test” for an explanatory model would not regard these questions as related. In other words, thinkers who do not adopt causality as an essential ingredient for explanation reject this assumption. For instance, Huneman (2010), as is discussed below.

I demonstrate below that network analysis as currently employed in the psychiatry literature and as a stand-alone approach does not meet E1-E5 and therefore conclude that in the present form, it does not count as an explanatory model on the definition I have adopted. Other thinkers however take a different approach, arguing that correlations can be shown to imply causation (e.g. Spirtes, Glymour & Scheines, 2001; Shipley, 2002; Pearl, 2009). At the end of this section, I compare Shipley’s axioms of causation with E1-E5 and demonstrate that E5 in particular is not explicitly accounted for.

Another approach to tackling the question of whether network analysis can count as an explanatory model, is that of Huneman (2010). He claims that network analysis provides a non-causal, non-mechanistic explanatory model which he calls “topological explanation”. This is:
An explanation in which a feature, a trait, a property or an outcome X of a system S is explained by the fact that it possesses specific topological properties T. What “possess topological properties” means is the following: S has elements, parts, features or capacities, moments of its regular behavior, or of the set of its possible behaviors, which are likely to be represented in a graph, a network or a variety S' in a space E" (Ibid., p. 216).

Although Huneman (2010) concedes that “topological explanation” is non-causal in a strict mechanistic sense, he does hold that topological explanations can show how an explanandum is caused by topological properties (p. 222). He claims that topological and causal explanations are more accurately regarded as poles on a continuum (Ibid., p. 225). I am sympathetic to this idea because it fits my aim of exploring the area between the poles where topological properties in a network representation of a target phenomenon and a mechanistic representation mutually inform each other and are complementary.

Next, I explore if the three major types of network most commonly applied in psychiatry can, in their current form, comply with E1-E5. I show that testing the constraints presents different difficulties for each.

8.6.1 Undirected Graphs and E1-E5
I begin with undirected graphs, in which edges represent correlations between the nodes. To determine the nature of correlations in any type of graph, further empirical data and mathematical analysis are required. This raises the question whether, in the face of potentially non-causal correlations, we can test constraints E1-E5.

To explore whether an undirected graph can have explanatory value in psychiatry based on the applicability of E1-E5, the very first step is uncovering potential causal pathways in the data by ruling out spurious correlations. In other words, we need to know whether a specific correlation holds when other symptoms are controlled for. The network which remains after ruling out spurious connections and/or non-causal correlations is called a “partial correlation network” (e.g. Costantini et al., 2015, pp. 15-16). I borrow an example from Borsboom and Cramer (2013): a partial correlation network can show that nicotine-stained fingers and a nasty cough are not strongly related to each other, but that each is more strongly related to smoking (pp. 105-106; see Figure 8.8). Here, the strength of correlations can be represented by the thickness of the edges (i.e. in a weighted graph). Note that there is often an empirical side to distilling a partial correlation network from an undirected (weighted) network: it requires additional empirical data. In the smoking example, a sufficiently robust data set is needed to show
Chapter 8

that the frequently co-occurring phenomena, coughing and nicotine-stained fingers, do occur separately and that both are individually more strongly related to smoking than to each other (the latter is in fact a spurious relation, because having stained fingers does not cause a cough nor does a cough lead to nicotine staining).

Figure 8.8 Example of undirected graph after ruling out spurious relations. S: smoking; C: cough; F: nicotine-stained fingers.

Can such a partial correlation network, which is still undirected and at most hints at potential causal relations between nodes, meet constraints E1-E5? The short answer is: no. E1 cannot be met because an undirected network of this sort does not offer temporal information (let alone an opportunity to differentiate between a mere temporal sequence and a causal relation). In the smoking example, the graphic representation does not entail information about the temporal order in which phenomena occur. As mentioned in a previous footnote, interval graphs are an example of graphs that are undirected and that can hold temporal information. I have not encountered any in the psychiatry literature but suggest that they may indeed be usefully employed to represent temporal information. For instance, an interval graph may be used to show that smoking precedes and co-occurs with a cough and nicotine-stained fingers (see Figure 8.9).

Figure 8.9 Interval graph representation of occurrence of phenomena where lines represent duration and location in time.

E2, the asymmetry requirement, cannot be met either due to the lack of directed edges which mean that cause and effect cannot be differentiated. Epskamp et al. (2018) in their discussion of partial correlation graphs point out that any undirected graph may be compatible with multiple directed graphs5 and that “assessing the direction of effect

5 The undirected graph A-B-C may be compatible with the following directed graphs: A→B→C, A←B→C, and A←B←C. Without additional interventional data, and relying solely on observational
Network Analysis

often requires an assumption that the causal structure is acyclic” (Ibid., p. 5). E3 can be met: if two effects of a cause are indeed independent, one would expect not to find these nodes joined by an edge after producing a partial correlation network to rule out spurious connections. In the smoking example represented in Figure 8.8, there is indeed no edge directly connecting the “cough” and “nicotine-stained fingers” nodes. Causal relevance between phenomena (E4) can at most be suggested but not demonstrated due to the lack of directed edges (as in the smoking example above). To test for E4, further (interventional) data is required. The graph in Figure 8.8 suggests which data might be interesting to collect to test which phenomena are causally relevant. The graph for instance suggests we collect data that can test if nicotine-stained fingers occur in non-smokers or whether a cough usually occurs before or after commencing smoking. E5 can in theory be met in a partial correlation network of an undirected graph. This is because the strength of a correlation is not necessarily dependent on prevalence. For instance, there may be a strong correlation between smoking and nicotine stained fingers even if not all smokers develop nicotine-stained fingers because an overwhelming percentage of people with nicotine-stained fingers does indeed smoke. For E5 to be met however, the data set must include an instance of the rare (or infrequently occurring) phenomenon in order for the strong correlation between that phenomenon and its cause to become apparent. If, for a moment, we assume that nicotine-stained fingers are rare, then, for a given data set that has no instances of “nicotine stained fingers”, it is impossible to determine whether constraint E5 can be met from the data represented in that particular graph. In sum, E5 can be met for correlations which occur in the given data set. However, we cannot draw any conclusions about the absence or presence of a causal relation between two variables from a data that does not contain an instance of a rare variable.

8.6.2 Directed (Acyclic) Graphs and E1-E5

I turn now to the two types of directed graphs to examine whether they fare any better in the face of E1-E5. Again, I note that I draw on directed graphs of the types I have encountered in the psychiatry literature and gladly invite the possibility that there are ways to adapt graphic representations and collect extra data so that E1-E5 can indeed be met.

Recall that the arrows in directed graphs may represent causal direction or mere temporal direction. Although I zoom in on the various facets of causality as specified by E1-E5 in this project, the literature sometimes simply specifies that an arrow represents a causal relation (without articulating a definition of “causal”). The case for acyclic directed graphs is more straightforward than that of directed graphs containing cycles, precisely

---

data, it is difficult to determine which of these options is most accurate (cf. Epskamp et al., 2018, p. 5)
because interpretation of cycles raises questions about how to understand the cycle and navigate the treacherous waters of causal reciprocity which are barely intelligible.6

Directedness of edges between nodes does not necessarily imply a cause-effect relationship, but in the literature on directed acyclic graphs, causality often is implied by the arrows. Although the arrows may represent temporal order, E1 holds that temporal order alone does not imply causality. Graphs with directed edges in which the arrows represent temporal order thus do not in themselves offer a means to differentiate between mere temporal sequence and causal relationships (and vice versa, interval graphs for instance, have undirected edges but harbour temporal information). Are there any options to rule out non-causal, mere temporal sequences (if event A always precedes event B, this need not necessarily imply that A causes B)? As a first option, it would seem that the use of partial correlation networks can contribute to this, but that they still only offer an approximation. E1 can only be met if one simply assumes that the temporal sequences that remain after spurious correlations have been corrected for imply causality.

A second option for meeting E1 requires collecting data at various times and looking for patterns in node activation. If enough data is collected, one may be able to differentiate between causal relations and mere temporal sequences through testing whether node B is causally dependent on or causally relevant for node A’s occurrence; i.e. can B also occur in absence of, or before A? This second remedy however, entails collecting additional data and representing it in a different subset of graph. As pointed out earlier, network analysis in itself does not rule out inclusion of temporal data, but it is the way in which we use network analysis to represent and interrogate empirical findings that is at stake here.

A third option, is collecting empirical data from an intervention that is designed specifically to rule out mere temporal relations. This is common in the literature on directed acyclic graphs which can model the probability that B will occur given occurrence of A (as discussed earlier in this chapter). Borsboom (personal communication) points out that DAGs are mostly used to represent population data and are particularly well-suited to represent the effects of interventions in this data (cf. Pearl, 2009). Although use of intervention data is superior in theory, the design of (ethically acceptable) interventional experiments is a considerable challenge in practice. This is due to difficulties of finding and applying interventions at a level so local that they affect only one variable and ruling out direct effects of an intervention on other variables.

---

6 For my purposes, which pertain to the field of psychiatry, I do not take into account the notion of “backward causation” or “retro-causation” and assume that cause always precedes effect.
As an aside, I add that analysis of DAGs may also help determine whether the phenomena represented by nodes are in a part-part relationship or a part-whole relationship (if for instance they occur at exactly the same time in a time-sensitive data-set). In his discussion of the “do-operator”—which, as discussed earlier is a label for fixing a variable so that the effects of this fixation can be determined—Pearl (2009) takes “disjoint variables” as his point of departure (p. 70). In other words, for his interpretation of the do-operator, he requires that variables are not, for instance, in a part-whole relation. However, I suggest that there may be suitable statistical methods to detect whether variables are indeed disjoint in the first place. This is because, in the case of psychiatric phenomena, this often is not evident at the outset.

The situation for directed graphs containing cycles is somewhat more complicated than for directed acyclic graphs precisely due to the cycles. At a very local level, for instance when zooming in on the directed edge between two nodes, we can say that phenomenon A is followed by the occurrence of phenomenon B. When we zoom out and regard the entire cycle, it becomes more difficult to say what follows what. This is because, as we run through the cycle, we inevitably return to the node from which we departed. Interpreting the role of temporality in graphs containing cycles thus calls for zooming in (and treating the graph as acyclic) or using a different representation entirely.

So in sum, although directed graphs fare better than their undirected counterparts in meeting E1, precisely because the directedness of edges offers additional information on the relation between nodes, further analysis (for instance using partial correlation networks and interventional data when this is empirically possible) is required. I conclude that E1 can be met for directed acyclic graphs after this additional analysis, but that E1 cannot be met in cyclic graphs because the cycles need to be zoomed in on and unravelled, or represented altogether differently (e.g. in a layered graph where each new layer represents the status of the network at a new time-slot) for the temporal aspects to make sense.

Directed acyclic graphs can meet E2 for obvious reasons: their directedness immediately portrays asymmetry between cause and effect. For graphs containing cycles, E2 is met only at the local level for the same reasons as E1 cannot be met. If we regard the entire cycle, E2 cannot be met because the cycle suggests that any node is either directly or indirectly both a cause and an effect of itself and the other nodes in the cycle.

So, although directed graphs fare better than their undirected counterparts due to their inherent asymmetry, these arrows are still not enough to rule out mere temporal sequences or differentiate causes from effects despite the asymmetry afforded by the
arrows. Use of partial correlation networks may serve to help meet constraint E1 in directed acyclic graphs.

Directed graphs with and without cycles differ in their ability to meet E3 because, as I discuss in more detail below, cause and effect are difficult to differentiate in a cycle - except at the local level, for instance between two nodes connected by an arrow. Directed acyclic graphs therefore can meet E3, whereas directed graphs containing cycles cannot. In the example from Figure 8.8, we can conclude that yellow fingers and coughing are not causally connected to each other, but they co-occur because they are causally connected to a common cause, i.e. smoking.

E4 can be met only in cases where we (rather boldly) assume that the arrows between nodes do in fact show causal relationships and not mere temporal sequences (i.e. E1 is taken for granted). In such cases, the arrows in directed graphs of the acyclic subtype can be taken to suggest which factors are causally relevant. For instance, in Figure 8.8, smoking can be taken to be causally relevant for coughing and nicotine-stained fingers (and not vice-versa, i.e. E2 is also met in this example). Any nodes that are not directly connected to each other can be deemed causally irrelevant to each other, for instance, if we added another node to Figure 8.8 representing “wearing blue socks” we would not find any connections to the other nodes because sock-colour has no bearing on whether one smokes, coughs or has nicotine-stained fingers.

E5 (an effect need not be probable for a cause to explain it) cannot be met for directed (acyclic) graphs for the same reasons that it cannot be met for undirected graphs. E5 seems the most difficult constraint to meet for the network analysis approach because it relies on having enough data to find causal connections that occur only rarely. The difficulties here are primarily practical. Network analysis as a methodology offers plenty of ways to represent weak connections, yet in practice determining whether these already statistically weak connections meet the requirements for a causal connection is very difficult. In network visualisations where the probability of a connection is demonstrated by the length of an edge (e.g. a longer edge means the connection occurs less frequently) or thickness or colour of an edge, such improbable connections may be shown, but, I reiterate that it does not follow that they are necessarily causal too.

8.6.3 Differentiating between cause and effect based on network analysis

As noted, if we take the liberty to assume that a directed edge within a cycle represents a causal relation (for instance in Figure 8.5), a further practical issue arises: in a cycle,
Network Analysis

it is more difficult to differentiate between cause and effect (cf. Eberhardt & Scheines, 2007). This is why constraints E3 and E4 (if we boldly assume that a temporal sequence implies a causal relation) could be met in an acyclic graph but not in a cycle. For instance in schizophrenia, anhedonia may turn out to be both an effect and a cause of apathy (if one does not enjoy doing things (anhedonia), one will not want to undertake activities (avolition) and become inactive, which in turn reinforces apathy). Each node, through activation of other nodes, can thus lead to a new instantiation of itself. Ultimately each node is (indirectly), on a causal interpretation of the directed edges, both a cause and an effect of itself. Nonetheless, the idea of a multigraph as a model of psychiatric symptoms is in line with the idea (e.g. in Cramer et al., 2010, shown in Figure 8.3), that symptoms of a psychiatric disorder may influence each other. Cycles are perhaps best taken to indicate areas that call for further empirical research.

For the reasons outlined above, much of the scientific literature on network analysis focuses on directed acyclic graphs (e.g. Pearl, 2009, p. 12-13). Borsboom and Cramer state: “[c]ausal analysis is easiest when the pattern of causal relations among variables creates a directed acyclic graph” (2013, p. 107). Although various statistical methods for transforming a partial correlation network into a directed acyclic graph have been proposed, their utility appears limited for the field of psychiatry. As Opgen-Rhein and Strimmer (2007) point out, these methods are tailored to graphs representing a large sample size with small numbers of variables. Such large data sets unfortunately are not the norm in psychiatry research. And, as stated, restricting network analysis to directed acyclic graphs, would neglect the idea that in many mental disorders, causal reciprocity between phenomena and cyclic patterns are the rule rather than the exception (Borsboom and Cramer, 2013, p. 104).

This point is discussed in more detail below.
Before concluding this section, I wish to point out that my attempts to see if E1-E5 are applicable in graphs of the sorts found in the current psychiatry literature are based on a limited understanding of the full scope of what network analysis has to offer. Another reason that the comparison is doomed to be rough at best, is that it is akin to comparing apples with oranges. The language of cause and effect and other terms in constraints E1-E5 would require axiomatization in mathematical terms to facilitate a more accurate comparison. An example of a mathematical axiomatization of causality (and comparison of this to E1-E5) is outlined in Section 8.5.

8.7 Discussion: causality and network analysis in the science of psychiatry

All in all, directed acyclic graphs appear the best equipped to meet constraints E1-E5 (see also Pearl, 2009). We have seen that directed graphs containing cycles unfortunately do not conform to E1 - E5 (only E2 is met). This raises a tension: directed graphs containing cycles appear to be more suitable to represent possible causal reciprocity (e.g. Borsboom & Cramer, 2013, p. 104), whereas DAGs are better placed to meet E1-E5. A disadvantage of restricting network analysis to directed acyclic graphs, is that in many mental disorders, causal reciprocity between phenomena and cyclic patterns is the rule rather than the exception (Borsboom and Cramer, 2013, p. 104; Clarke, Leuridan, & Williamson 2014). For example: poor sleep leads to poor daytime concentration and lower productivity, resulting in anxiety which then leads to rumination before bed and poor sleep.

Borsboom (personal communication) offers a counter-example: he notes that directed acyclic graphs representing time-series data can demonstrate causal reciprocity in a non-cyclic fashion. The cross-lagged panel model is a specific instance of this counter example. In this model, two or more variables are measured on two or more occasions in order to detect relations between these variables over time. These relations may turn out to be correlations or even causal relations. Thus, all in all, I conclude that DAGs offer the best opportunities for detection of causal relations between variables/phenomena, even though meeting constraints E1-E5 is not straightforward.

This leaves us with various options, such as challenging E1-E5 in the first place, or finding means to represent cyclic processes in DAGs (as in the cross-lagged panel model) to benefit from the methods available for detecting possible causal relations. Regarding the former, the question becomes: should we thus question the feasibility of constraints E1-E5 in the context of network analysis and reject the notion of causality that they imply? One problem with constraints E1-E5 is that they imply that the difference
between causes and effects is unambiguous. I reject this possible implication and do so whilst still retaining E1-E5. To do this, I suggest that causality is best regarded as a heuristic tool rather than as something that separates causes from effects in any absolute manner. Thus, cause and effect can be disambiguated locally only, and on a case-by-case basis, for instance using intervention data where possible. In this manner, we can still describe a causal relationship between two nodes in a larger cycle. Thus, if we grant that what counts as a cause and what counts as an effect is relative rather than absolute and unambiguous, we can defend the use of directed graphs containing cycles and retain the notion of causality embraced by E1-E5.

Another option to deal with the difficulties of detecting causal relations in undirected graphs containing cycles is outlined by Clarke, Leuridan, and Williamson (2014). They suggest methods to transform cycles into DAGs so that Bayesian inference can be applied. One of these entails unwinding the cycles. Where necessary, extra directed edges between nodes are then added to meet the Markov condition.

Like myself, Shipley (2002) and Spirtes, Glymour and Scheines (2001) prefer to avoid the metaphysical discussion about causality. Rather than seeking to define causality, they instead offer axioms and then incorporate these into the mathematical analysis of correlations. A correlation is thus causal if:

1. It is transitive: “if A causes B and B causes C, then it must also be true that A causes C” (Shipley, 2002, p. 5).
2. It is local (or “Markovian”): “events are caused only by their proximate causes. (…) Thus, if event A causes event C only through its effect of an intermediate event B (A→B→C), then the causal influence of A on C is blocked if event B is prevented from responding to A” (Ibid., pp. 5-6).
3. It is irreflexive: “an event cannot cause itself” (Ibid., p. 6).
4. It is asymmetric: “if A is a cause of B, then B cannot simultaneously be a cause of A” (Ibid.).

Although there is not a 1:1 correspondence, there is an overlap between these four axioms and constraints E1-E5. An interesting contrast is that E1-E5 offer a “diagnosis per exclusionem” of causally relevant factors because they are formulated negatively, whereas the axioms are formulated positively. Roughly, the first and second axioms

8 “Locally” here can be taken to refer to the relation between an node ad its parent(s), thus satisfying the Markov condition that was introduced earlier in this chapter.
9 Giving an outline of the methods they suggest falls beyond the scope of this project.
taken together imply E1, E3 and E4 combined. The third and fourth axioms imply constraint E2. In other words, I suggest that for any example in which axioms 1 and 2 are met, constraints E1, E3 and E4 are also met. And, any example which satisfies axioms 3 and 4, also satisfies E2. E5 is absent from the axioms; they neither include nor rule out E5 and would need to be supplemented explicitly to account for improbable effects\(^{10}\). Again, note that such a comparison is akin to comparing apples to oranges because constraints E1-E5 are formulated in terms (such as *cause*, *effect* and *temporality*) that do not lend themselves immediately to translation into mathematical terms.

All in all, the idea of axiomatizing rather than defining causality for the purposes of detecting possible causal relations between nodes in a graph fits is one to which I am sympathetic. It fits with my agnosticism regarding metaphysical matters and is compatible with the idea of regarding the notion of causality as a heuristic tool that works only in situations in which the context is well specified. However, it still leaves us with the thorny issue of meeting constraint E5, a constraint that is of importance in the field of psychiatry and which I therefore am unwilling to drop. This is why I ultimately argue that network analysis in the way it is currently employed in the literature is not in itself an explanatory model but that it instead can be used to augment and complement the mechanistic approach outlined in the previous chapter. The way in which network analysis and mechanistic explanation can be used together is the theme of the next section, after which I discuss the idea of Mechanistic Property Clusters proposed by Kendler, Zachar and Craver (2011).

### 8.8 Do Network Analysis and Mechanistic Explanation converge?

There is a clear convergence between the descriptions of the network analysis approach and the mechanistic model of explanation: both embrace a notion of organisation. Both allow for nested hierarchies of sub-systems within systems. In both, the part-whole relationship and more specifically the interactions between parts are central. There are, however, limits to this convergence: a mechanistic explanation seeks to describe *how* parts interact. The network approach as a mathematical methodology, in contrast, does not incorporate any data on this and is in fact agnostic on this front. As has been discussed at length earlier in this chapter, network analysis offers a method for mathematical manipulation of data and is neutral or agnostic with regard to the content being represented. Understanding *how* parts of a mechanism interact is however crucial

\(^{10}\) This is based on rough observation and it is beyond the scope of my thesis to argue for this in more detail because this would require axiomatizing other terms in constraints E1-E5 for a more accurate comparison.
for the integration that is the topic of the second central question of this project. It is only when assumptions pertaining to how the mathematical methods are applied to data from the field of psychiatry are articulated, that we can draw conclusions about whether (and perhaps how when interventional data is available) parts interact based on a network approach. The field of “Mechanistic Property Clusters”, the topic of the next section, refers to precisely this melting together of data and methodology.

Furthermore, although a mechanistic model can be represented as a network, the opposite, that a graph can be represented as a mechanistic model does not hold. This is because a mechanistic representation of a graph requires additional information about how the entities interact. The network approach however, is more versatile in other ways: any type of relation between nodes can be incorporated, whereas in the mechanistic model, the relations between interacting entities are causal. Network analysis can serve well as a means to identify relations and uncover their nature up to a point, as discussed at length in the previous section, where limitations (regarding meeting constraint E5 in particular) were noted. As such, they can lead the way to further research aimed at explaining the entities and their interactions that produce the correlations detected through network analysis. It is this complementarity that I advocate and to which I return in more detail in Chapter 9. Before concluding this chapter, I discuss the idea of Mechanistic Property Clusters and question whether it is an example of a hybrid explanatory approach (i.e. combining the ideas of mechanistic explanation and network analysis) for the purposes of studying phenomena in psychiatry.

8.9 Mechanistic Property Clusters: a hybrid approach?

In this section I introduce “Mechanistic Property Clusters” (MPCs) and explore their scope by also delving into the related notion of homeostatic property clusters. I argue that the MPC concept is posited primarily as an answer to the question of how to categorize natural kinds and that it therefore does not count as an explanatory model despite combining ideas from both mechanistic thinking and network approaches.

“Mechanistic Property Cluster” refers to a collection of symptoms of mental disorder from the biological, psychological and social levels of organisation for example, that are related mechanistically (Strijbos, 2016). As noted, the MPC notion can be regarded as combining the ideas of mechanistic explanation and constraints to determine causal relevance on one hand, with the latent variable-eschewing methodologies of network analysis and a network representation of data from psychiatry on the other hand. The network aspect is that properties (regardless of whether they pertain to the biological, psychological or social level of science) are represented in a matrix. The idea is that
causally connected properties will form clusters on analysis. These clusters suggest that there is a mechanism at play and that the properties that are clustering are at the same “level of mechanism”. Mechanistic Property Clusters were introduced as a variety of natural kind. The MPC view fits well with Borsboom’s (2017) network theory of mental disorder, which holds that rather than being a latent variable that causes symptoms, mental disorder is best described or represented as a network of symptoms that (reciprocally) affect each other.

8.9.1 MPCs and the status of mental disorders
Kendler, Zachar and Craver (2011) first proposed the idea of “Mechanistic Property Clusters” in response to the problems with regarding psychiatric disorders as essentialist, constructionist or practical kinds. On the essentialist view, symptoms of a disorder result from an underlying essence present in all individuals with the disorder. Constructionists hold that the nature of a disorder is dependent on the social and cultural network within which it presents. Those that embrace practical kinds, emphasize that utility (for defining treatment, or researching etiology) is the guiding factor when categorizing mental disorders: there is no one true categorization. They point to drawbacks for each position, stating that psychiatric disorders are embedded in the causal structure of the world. They nonetheless refute the essentialist position calling it too idealistic for the purposes of psychiatry because it mistakenly assumes that there is a single causal agent for each mental disorder (Ibid., p. 1145).

Kendler et al. (2011) thus describe the MPCs as non-essentialist: the clusters of properties do not have deterministic essences (p. 1146). They nonetheless suggest that there are “robust explanatory structures to be discovered underlying most psychiatric disorders” (Ibid.). They offer two examples of how a MPC can identify a mental disorder. In one, what characterizes the disorder, is the “more or less stable patterns of complex interaction between behavior, environment, and physiology that have arisen through development, evolution and interaction with an environment” (Ibid., p. 1147). In this case, there is no underlying essence that causes the symptoms to cluster, the clinical features cause each other. In the other, a series of causes interact, producing an “underlying state” which in turn causes clinical features that may interact causally.

The authors caution us to expect that despite adhering to the MPC view of natural kinds, the explanatory structures will be messy and that it will take hard work and some degree of idealization and abstraction to bring them into focus (p. 1146). Yet, although phenomena may have fuzzy boundaries (they may be intrinsically fuzzy), this does not detract from their stability across time and cultures (cf. Schmittmann et al. 2013, p. 53).
Whereas Schmittmann et al. (2013) explicitly eschew the idea of psychiatric disorder as a latent variable and accept that the cluster of symptoms is the psychiatric disorder, the MPC account by Kendler, Zachar and Craver leaves open the possibility of finding a latent variable underlying the symptoms (or other aspects) of a psychiatric phenomenon (Ibid., p.1147). Recall that I claimed that an important merit of network analysis (of the types described above) compared to factor analysis (a narrower, sub-type of network analysis), is that it is better suited to investigating complex phenomena in which there may be causal relations between parts and in the part-whole relationship (e.g. Borsboom, 2006; Borsboom & Cramer, 2013; Cramer et al., 2010; Frewen et al., 2013). This is because network analysis avoids assuming that there is a latent variable at play, which is an important contrast with the factor analytic approach that takes the existence of a latent variable as a given. Thus, one important issue is that in cases where Kendler, Zachar and Craver allow for the presence of a latent variable, it is unclear how the MPC approach differs from the “regular” factor analytic approach. This is of importance because the question of whether or not one expects to find a latent variable which causally connects the features of a phenomenon guides the way empirical research is set up. I reiterate the point made by Schmittmann et al. (2013) that faulty assumptions about the presence of latent variables in psychiatric disorders may be why empirical progress has been disappointingly slow.

To remedy the problem, I suggest adopting a version of the MPC account which avoids inclusion of latent variables. We can thus retain the idea of properties that co-occur based on causal connections between them (here too, I advocate a pragmatic, context-dependent interpretation of “causality”) and use the methods of network analysis to examine which properties are related.

In the next sub-section, I delve into the history of the MPC line of thought by tracing the idea back to Boyd, who developed “homeostatic property clusters” and offers a more in depth definition of the concept. The goal of this further exploration is to determine whether the idea of MPCs can perhaps extend beyond being an answer to the philosophical question of how to categorize natural kinds.

**8.9.2 Homeostatic Property Clusters**

The idea of MPCs is based on the philosopher Richard Boyd’s “homeostatic property cluster” (HPC) (e.g. Boyd, 1989 & 1991). HPC-kinds are kinds defined in terms of clusters of properties that co-occur because they “correspond to inductively and explanatorily relevant causal structures” (Boyd, 1989, p. 16). Their co-occurrence is brought about by a “(homeostatic) mechanism”. The term *homeostatic* is used somewhat loosely and refers to the metaphorical and sometimes literal *homeostasis* which results
from co-occurrence of the properties: “[e]ither the presence of some of the properties (…) tends to favor the presence of the others, or there are underlying mechanisms or processes which tend to maintain the presence of the properties (…), or both” (Boyd, 1989, p. 16).

The idea of HPCs contrasts the idea of clusters of properties that are unified by convention or conceptually, by invoking necessary and sufficient conditions (Boyd, 1991, p. 141). What counts as an HPC-kind is thus partly dictated by the causal structure of the world rather than being dependent solely on human theories or convention. The upshot of Boyd’s position is that the HPC-kind, in contrast to kinds based solely on convention, is useful for purposes of prediction, description and control. Boyd concedes that the causal structure of the world is “messy and complex” (Ibid., p. 143). There are cases in which some properties in the cluster are absent or in which some of the mechanisms are inoperative (Boyd, 1991, p. 142). Boyd embraces the indeterminacy that the HPC account introduces:

(…) it will sometimes happen that neither theoretical nor methodological considerations assign the object being classified determinately to the kind or to its complement, with the result that the homeostatic property-cluster definition fails to specify necessary and sufficient conditions for kind membership. Both the property cluster form of such definitions and the associated indeterminacy are dictated by the fundamental epistemic task of employing categories which correspond to inductively and explanatorily relevant causal structures. In particular, the indeterminacy in extension of these natural definitions could not be remedied without rendering the definitions unnatural in the sense of being scientifically misleading. (Ibid., p. 142)

Boyd’s account is primarily ontic in focus because what counts as an HPC depends on the underlying causal mechanism. However, he also allows perspectival aspects: the property cluster, besides being maintained by a common underlying mechanism, must, on his definition also have some epistemological role: “there are (theoretically or practically) important effects which are produced by a conjoint occurrence of (many of) the properties (…) together with (some or all of) the underlying mechanisms in question” (Boyd, 1989, p. 16).

I embrace Boyd’s tolerance for the messiness and complexity of what we take to be the causal structure of the world. I suggest that for psychiatry, we must tolerate the indeterminacy to which Boyd refers. On my view, we need to accept that what we regard as parts in mechanisms and also what we take to be the boundaries of the system of parts and their interactions, are determined by a combination of the causal structure of
the world (as we perceive it) and the epistemic task at hand. A degree of arbitrariness or indeterminacy is simply inescapable.

In sum, Boyd’s account of HPCs, like the MPC account, is an account of natural kinds. The questions to which HPCs help provide an answer are therefore still primarily philosophical. They pertain to how we can employ categories that are informed by the causal structure of the world and that are explanatorily relevant, rather than being based on convention alone. The above exploration of the HPC account again underlines the importance of accepting that there is indeterminacy (in our choice of parts and the boundaries of the system under study) and that we must tolerate rather than seek to obliterate the indeterminacy in our attempts to explain psychiatric phenomena. In the next section, I briefly explore Craver’s (2009) response to the HPC view of natural kinds zooming in on the notion of mechanism at play.

8.9.3 Back to property clusters in psychiatry

The idea of classifying natural kinds based on the mechanisms that produce property clusters, caught other thinkers’ attention. Craver (2009) zooms in on the definition of mechanism that is implied in the HPC account. More specifically he asks 1) when are two mechanisms of different types? 2) and, when does one mechanism end and the next begin? He shows that for both questions, an objective answer is impossible and advocates a perspectival approach where we accept that the answers to the above questions depend on the explanatory task at hand. He concludes his call for a pluralistic HPC view: “[i]t will not tell us in any definitive way how many kinds of memory there are, whether the emotions and concepts are natural kinds, or how to refine the DSM. Instead, there will be multiple incompatible answers to these questions depending on which mechanism one attends to, on how one describes the phenomenon, and on where one draws the boundaries of the mechanism” (Ibid., p. 591). Craver then questions what work the mechanism is doing in the definition of HPC-kinds if what counts as a kind is primarily dependent on the explanatory task at hand and the required “grain of abstraction” rather than on the underlying mechanism (Ibid., p. 589).

Interestingly, this is a more epistemically oriented response than Craver’s (2007) defence of a primarily ontic account (as discussed at length in the previous chapter). In the closing paragraph of the article, he remarks that the position that he defends also entails

11 This is in line with the Quine-Duhem thesis, according to which scientific hypotheses cannot be tested in isolation. They are inescapably dependent on auxiliary hypotheses, for instance concerning what we take to count as parts and how we define the boundaries of the phenomenon under study (Quine, 1951; Duhem, 1954).
that “our interests and objectives contribute partly, but ineliminably, to the kinds of mechanisms we find in our world” (Craver, 2009, p. 592).

All in all, I conclude that the idea of MPCs was formulated primarily as a response to the philosophical question regarding the ontological and epistemological status of mental disorder rather than being a fully-fledged practical model for interpreting data from empirical psychiatry. For my purposes, the main merit of the MPC view is that it emphasizes that the indeterminacy that we encounter in trying to categorize phenomena is inherent to the field of psychiatry. Although I am sympathetic to the idea of combining network analytic and mechanistic approaches, the MPC account does not offer any guidance on how to combine them. Hence my conclusion that the MPC view does not count as a model to aid scientific explanation.

8.10 Conclusion

In this chapter, a brief introduction to network analysis was provided, including an outline of the three most commonly used types of network in psychiatry. The explanatory power of those three common types of network was explored drawing on constraints E1-E5. I claimed that network analysis as currently employed in the psychiatry literature in itself does not meet the requirements I have adopted for an explanatory model. Network analysis, although perfectly equipped to represent statistically improbable relations between nodes, is limited in its ability to differentiate between spurious relations on one hand and improbable effects on the other. The beauty of network analysis is that it is neutral with respect to the content that is represented. Relations between very diverse entities can be modelled as nodes and then the statistic relations between them can be explored. This makes it ideal to contribute to the integration of phenomena from different levels. The challenge of interrogating a philosophical topic such as causation mathematically was a central theme of the discussion. I advocated viewing causation as a heuristic tool that aids our understanding of the phenomena and suggested that the translation of this philosophical concept to the domain of mathematics can be achieved by axiomatizing “causation” into mathematically testable axioms. This move, of reclassifying philosophical concepts (or dichotomies) as heuristic tools will be addressed in further detail in the Chapter 9, in which I sketch my own, network- and mechanism inspired, account of explanation in psychiatry. I concluded the chapter with an exploration of the Mechanistic Property Cluster view. Although it initially appeared to hold promise as a hybrid account that combines network analytic and mechanistic approaches, its scope turned out to be limited to answering a question about natural kinds. In the next chapter, I therefore work on bringing together the advantages of mechanistic thinking and network analysis for the purposes of offering an account of perspectival integration to aid explanation of psychiatric phenomena.
Chapter 9

Perspectival Mosaic Unity
9 Perspectival Mosaic Unity

9.1 Introduction

In this chapter, I aim to bring together the findings from previous chapters to formulate answers to the central questions that fuel this project. These questions are: 1) what is the utility of the biopsychosocial model at the level of a) philosophy, b) science and c) clinical practice? 2) How can we understand the use of the biological, psychological and social levels of explanation in science and account for the coherence and possibilities for integration of these different levels of explanation? 3) What are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy? Recall that the first of these questions was answered in Chapter 5.

Before discussing the other two questions, I first offer a rough outline of the view of explanation that I have developed during this project. Although I refer to it as “my account”, I emphasize that it is far from being a fully developed account. It is best viewed as an attempt to adapt the ideas of mechanistic explanation (augmented with network theoretical approaches) in order to make the mechanistic model of scientific explanation more suitable for the science of psychiatry. Further articulating my account is beyond the scope of this thesis because my primary focus here is on answering the above-mentioned central questions. Although the aim is thus to offer a philosophically sound guideline for scientific practice in psychiatry, the model may also, as a spin-off, inform clinical practice, in which explanation of phenomena in the individual patient is called for.

In the next section, I introduce the notion of “perspectival mosaic unity”. I coin this phrase to underline the importance of a relativistic, pragmatic and dynamic view of scientific explanation in psychiatry. I introduce six basic assumptions on which my views and thus also my answers to the central questions are based. I then return to the ontic-epistemic divide, a theme which surfaced in Chapter 7, and which warrants further discussion in light of my fundamental assumptions and my own account in this chapter. I note that on my view, mechanistic explanation implies a notion of teleology and I investigate three contexts in which the notion of teleology has different implications. I suggest how an adapted version of mechanistic explanation and network theoretical approaches may complement each other. I provide brief intermezzos, first delving into the problem of interlevel causation and second discussing an example from the apathy literature of how the mechanistic and network approaches may be combined. After this stage-setting preliminary work, I consider the podium ready to reformulate the second
central question of this thesis. During the process of writing this thesis, the original version of this question became increasingly obsolete because of my move towards perspectival integration rather than integration of levels. Having reformulated the second central question, I finally proceed to answer it and then move on to answering the third question.

9.2 Mosaic Unity

During the project, the need to reconsider the phrasing of the second central question resurfaced on numerous occasions. Its formulation entails assumptions that need to be clearly articulated because they tap into fundamental philosophical debates. To illustrate the first of such debates, I use an expression from the subtitle of Craver’s 2007 book: “mosaic unity”. Craver contrasts “mosaic unity” to the “reductive unity” of neuroscience. “The central idea is that neuroscience is unified not by the reduction of all phenomena to a fundamental level, but rather by using results from different fields to constrain a multilevel mechanistic explanation” (2007, p. 231). Thus, on Craver’s account, it is the constraints on explanations which differ per field, that have a unifying effect without detracting from the contributing fields’ autonomy (Ibid.). He emphasizes that taking issue with the goal of reductive unity does not imply rejecting the idea of the unity of science (Ibid., p. 230). He sums up four advantages of non-reductive, mosaic unity:

First, it provides a straightforward way to interpret “levels” and, accordingly, the idea of interlevel integration. Second, whereas reduction models involve global relationships between theories at different levels, the mosaic model accommodates the fact that interlevel relations are often formulated piecemeal, within local mechanisms, by adding constraints on interlevel relations. Third, the mechanistic account accommodates both upward- and downward-looking interlevel integration. Finally, the mosaic view details the varied forces driving the co-evolution of work at different levels. (Craver, 2007, p. 256)

Furthermore, integration of findings from different disciplines can occur both inter- and intralevel. Intralevel constraints include componency, spatial, temporal and active constraints; interlevel constraints include accommodative (top-down and bottom-up), spatial, temporal and mutual manipulability constraints (Ibid., pp. 249-256). Interlevel integration can take place upward and downward. In the latter case, integration is achieved by showing how lower level components fit into a higher level mechanism and describing how a lower level mechanism produces a higher level phenomenon respectively (Craver, 2007, p. 257). Rather than relations of identity between levels, the relation is based on a part-whole relationship (recall the treatment of different types of
level in Chapter 7 where Craver’s “levels of mechanism” were introduced). If the evidence for a mechanistic explanation can be upheld from different scientific perspectives, this indicates a robust explanation. In Craver’s words, “it is a sign of epistemic success” ([Ibid.], p. 233).

The mechanistic approach to constitutive explanation purports to provide a philosophically sound foundation for this mosaic unity. However, the terms “mosaic” and “unity” reveal a tension: “mosaic” suggests small, separate parts or elements whereas “unity” implies a whole. Let us briefly explore how far we can take this metaphor to shed more light on the tension. Craver states that the elements in the mosaic represent the empirical findings from different scientific fields within the neurosciences ([Ibid.], p. 228). The unity results from using findings from different, autonomous fields to constrain a multilevel explanation. I address three aspects of the metaphor. The first is about how to fit parts of the mosaic together. The second concerns the point at which the wholeness of a mosaic becomes apparent. Third, perceiving the wholeness requires finding the right vantage point.

First, during the manufacturing period, the pieces are separate and may be grouped according to a wide array of parameters such as size, shape, colour, and material. How then, does one determine how to fit them together? Craver might reply by drawing on the notion of “constraint”. The characteristics of each element in a mosaic play a role in determining the best fit for others, i.e. the pieces automatically constrain each other. Pieces that fit together more tightly are aesthetically more pleasing because they require less cement. There is not one single “right” way in which the pieces fit together, but clearly some configurations are more effective than others for producing a recognisable picture in a mosaic form. For instance, when producing a mosaic in which a yellow circle is depicted, the artist will limit his choice of the next element to his pile of yellow tiles even if there is a blue tile that fits the shape of the adjoining pieces more tightly. Thus, the idea of “constraint” avoids a situation in which “everything goes” and tiles can be randomly combined to produce an intelligible image and yet doesn’t prescribe one absolute way in which to combine the tiles. In discussing E1-E5, Craver uses the idea of constraints in a primarily ontic sense: the system itself determines how parts restrict each other based on its organisation. The notion of “constraint” in the mosaic metaphor also allows for a more epistemic interpretation: how the tiles (representing evidence from different disciplines) are pieced together is determined by both the nature of that tile and the intelligibility of the chosen configuration of tiles so that the mosaic becomes

---

12 Autonomy of the contributing fields is required to ensure that the constraints are independent (Craver, 2007, p. 231).
“coherent” for the beholder.

Second, the wholeness or unity of a mosaic is usually not yet apparent when the mosaic is being made. The producer of a mosaic has an idea in mind of the end result and this guides how he fits the parts together. At the very local level, shape, size, and colour determine which parts go where. Parts that appear to fit well together based on their properties may be placed together. However, in the mosaic metaphor, the whole (the picture the artist wants to produce) also determines the process of fitting the parts together. The process of fitting the parts together thus depends not only on the properties of the individual parts, but also on the project of using the individual tiles to make a coherent picture. This assumption, that making a coherent picture requires having in mind an idea of what the end result should look like, cannot simply be taken to hold for science. For science, we are aiming to expand our understanding and increase our knowledge. In contrast to the artist making a mosaic, scientists do not yet have in mind an idea of how the results will look. Rather, the central aim of scientific practice is to yield a yet-unknown, but more complete picture/explanation of the phenomenon under study as a whole.

Third, even once the mosaic is completed, the wholeness (formed of pieces) may be appreciated only when viewed from an appropriate distance, under suitable lighting circumstances and so forth. The idea of “wholeness” taps into a thorny philosophical difficulty. On my account (for which I argue in more detail below), “wholeness” requires a beholder. There is no view from nowhere. The mosaic itself, on my interpretation, does not hold an intrinsic wholeness. Its wholeness is instead subjectively mediated. And furthermore, besides requiring a beholder, the wholeness of the mosaic becomes apparent only when the beholder is suitably placed to perceive it. The observer can be fairly sure that he has found the appropriate vantage point because this is the point from which the picture as a whole has recognisable content. When we carry this idea over to the domain of science, it suggests that unity is dependent on intelligibility and a sense of wholeness. Yet, science is not complete: it is a patchwork in the making. So how can we determine the optimal vantage point from which to appreciate the unity of perspectives so that they may form a coherent whole? This in turn raises the question whether appearing coherent to the beholder is indicative of a correct explanation of the phenomenon under study. In other words, just because a model appears coherent, this does not prove that it offers the best possible portrayal of the phenomenon it was designed to represent.

Note that although the term “constraint” may at first glance seem to imply a limitation, an exclusion of options, this is not necessarily implied. Etymologically, the term
“constrain” derives from “binding together” and this binding may afford new strengths to the material that is bound together. In the context of constraints that contribute to good explanations, the presence of such constraints may thus serve to weed out poor aspects of a proposed explanation (such as factors that are not explanatorily relevant to the phenomenon under study) and offer a clearer view of explanatorily relevant options.

As with all metaphors, there is a limit to how far the metaphor can usefully be carried to support the idea that is being proposed. The issues discussed above underline the importance of a pragmatic and relativistic approach: what counts as a whole is context-dependent and subjectively mediated. Ascribing boundaries to a phenomenon or to a collection of scientific perspectives on a phenomenon and taking these to count as a whole is an arbitrary enterprise, guided by the explanatory task at hand. Thus, on my interpretation, “mosaic unity” highlights beautifully the importance of the part-to-whole relationship in mechanistic accounts of constitutive explanation and also demonstrates that part-whole relationships are not absolute but are instead open to multiple interpretations. Also, the use of a visual, art-inspired metaphor conveys the importance of the observer, despite Craver’s initial focus on a primarily ontic account of explanation. The mosaic’s meaning, its function as a representation of a phenomenon is dependent on there being an observer. It opens up the discussion on the ontic-epistemic divide which featured in Chapter 7. Also, given that I take explanation to be an important goal of scientific practice, the quest for perspectival mosaic unity in turn calls for an account of teleology that is in line with the mechanistic approach to explanation.

In the next section I present some of the basic assumptions of my account and then zoom in on key debates and my own position and finally I show how they affect the phrasing of the second central research question of this project.

9.3 Six fundamental assumptions

To structure my discussion, I address six assumptions upon which my account is founded and briefly develop each of them so that I can later show how they are relevant to the phrasing and answering of the second central question of this project.

I. There is no such thing as a view from nowhere
   The “view from nowhere” was a phrase coined by the philosopher Thomas Nagel in his 1986 book. My contention that “a view from nowhere” is a theoretical posit that is unattainable for humans, is deeply related to many of the assumptions below. Nature, on my account is always “nature as we know it”. Thompson (2010) iterates this point: nature “must be seen in relation to the subjectivities and research practices
of the scientific investigators to whom it is disclosed” (p. 19). Thus, the closest we can get to a view from nowhere is to try and get a view from as many places as possible. By this I mean that it is common practice in the (medical) sciences to rely not on the observations of one single person or machine, but to replicate results and test models in as wide an array of circumstances and account for inter-observer differences. For mere mortals, a view from nowhere is a contradiction in terms: for something to be a view in the first place, requires subjective mediation and a vantage point for that view. I therefore position myself as an agnostic regarding questions of “pure ontology”. My eschewing of “objectivism” is related to my reclassification of the ontic-epistemic divide into the “heuristic tools” category, as discussed in Chapter 7 and in further detail in this chapter.

2. Nature as we perceive it has infinite sets of “joints” for carving and these sets may be incompatible with each other

The idea of carving nature at its joints suggests that nature has a set of joints and that carving it is as easy as carving up your average Sunday roast. However, the account I defend holds that nature can be “carved-up” in many different ways (cf. Quine, 1951). This is in line with the idea of underdetermination of scientific theory: the evidence we have available may be insufficient to determine which theory is most accurate. Returning to my Sunday roast: the manner of carving it up that works best depends on the explanatory question at hand and the tools available for carving (which relates to the following assumptions). If the task at hand is to serve your guests, the roast may be carved in such a way as to separate meat from bones. If, on the other hand, the task at hand is to determine nutritional value or the effects of heat on animal protein for instance, then different “joints” must be used for splitting the roast into its component parts. The different methods of analysis may be mutually incompatible: after carving up the roast to make it suitable for protein analysis, it can no longer be served as a meal. Or, in other words, what count as the relevant parts contributing to the mechanism in a system carved up from one particular explanatory perspective, need not also count as parts in a different mechanism in the same system.

Salmon (1984) draws on an example from Hanson (1958) to illustrate that what is relevant to an explanation depends on the standpoint from which an explanation is called for.

There are as many causes of $x$ as there are explanations of $x$. Consider how the cause of death might have been set out by a physician as “multiple haemorrhage,” by the barrister as “negligence on the part of the driver,” by a carriage-builder as
“a defect in the brakeblock construction”, by a civic planner as “the presence of
tall shrubbery at that turning”. (Hanson 1958, p. 54; quoted by Salmon, 1984,
p. 127)

Each of the causal factors in the carriage accident is relevant to an explanation
of what happened. Salmon demonstrates that when one zooms in on the precise
why-question fuelling the need for explanation, there are differences in the
questions that are relevant for each of the above-mentioned experts. Different
perspectives on a single phenomenon, such as the cause of death in a carriage
accident can yield different answers to the same question (cf. van Fraassen,
1980). “The particular why-question determines which explanation is being
sought; when that is determined, the explanation must take all objectively
relevant factors into account” (Salmon, 1984, p. 129). At a higher level
of abstraction, each of the factors may be equally relevant, whereas, for the
particular experts answering questions on the cause of death, some factors may
be more relevant to the explanatory task at hand from their standpoint.

On the view of mechanistic explanation that I adopt, the focus of the example
above on causal relevance runs parallel to the carving of nature at its joints.
This is because what count as parts (bones) and their interactions (articulating
joints) in a mechanistic explanation of a phenomenon depends entirely on
the explanatory task at hand. This relativistic view sits more comfortably with
epistemically oriented New Mechanists (e.g. Bechtel and Richardson, 2010)
than with accounts of thinkers of an ontic persuasion (e.g. Craver, 2007).
Besides the system under study itself, the explanatory task in this manner
constrains what counts as explanatorily relevant.

3. Any research is inherently theory-laden
Werner Heisenberg, writing in 1958, made this point succinctly: “what we observe
is not nature itself but nature exposed to our method of questioning” (p. 58).
To investigate a phenomenon with a view to explaining (some aspect of) it, one
must first define the target phenomenon and then specify hypotheses about that
phenomenon. In medicine, a pattern of symptoms may be the point of departure.
Often, the theoretical background of the researcher or the types of investigative
methods and technologies available play a role in the choice of target phenomenon
and hypotheses. Research is thus both theory and technology-driven. A geneticist
is more likely to focus on finding a genetic anomaly whereas a neuroscientist with
access to an fMRI scanner may choose a hypothesis that could be confirmed or
disproven through an imaging study. Even in the case of qualitative research, where
the data is collected with a view to letting the data guide hypothesis formulation, choices are still made by humans about which data to collect and how to collect and interpret them. I contend that any research is inescapably affected by the implicit and preferably explicit assumptions on the part of the researchers. It is inherently theory-laden (cf. Hanson, 1958, Chapter 1; Popper, 2014, reprint of original from 1959; Kuhn, 1970). My goal is not to escape theory-ladenness. This is, as I argue, impossible because it would require a neutrality that rules out any form of measurement and interpretation. I instead suggest that a heightened sensitivity or awareness of implicit assumptions can improve the effectivity (and perspectival integration) in empirical research.

4. Any form of measurement and even “mere observation” inescapably affects the target phenomenon

In the early 20th century, quantum physicists made a compelling case for this claim for processes at the sub-atomic level. Perhaps the most famous example is Schrödinger’s thought experiment featuring a cat. The idea fuelling the thought experiment (deemed a paradox by many) is that sub-atomic particles can exist in multiple states at once, in a state known as “quantum superposition”. This superposition collapses into a single, definite state on observation. In the thought experiment, a cat is hooked up to a device that releases a lethal dose of poison the moment a Geiger counter registers the radioactive decay of a single radioactive atom. The subatomic particles of the radioactive atom are taken to be in a quantum superposition, which would lead to the conclusion that the cat, whose life depends on the position of the atom’s parts, must be dead and alive simultaneously, until an observer assesses the situation.

The above example illustrates the notion that observation can affect the phenomenon itself at the level of very small particles. I do not mean to suggest that the (usually much larger) systems we study are in a quantum superposition that collapses as soon as we take a peep. I do however adopt the idea that any form of observation (including measurement with instruments) affects the target phenomenon. The significance of the effect may differ depending on the context and is in itself challenging to detect and quantify.

A more classic example of the above is measuring the temperature of your tea with a thermometer. The act of measuring affects the temperature of the tea. This is because: first, the temperature of the material of which the thermometer is made does not exactly match that of the tea; second, the thermometer itself absorbs heat; and, third the perturbation of the liquid caused by dipping in the thermometer may
lead to cooling down of the tea. One could of course gauge the effect of thermometer insertion by placing a very sensitive thermometer in the tea and then using it to measure the effect of the insertion of a second thermometer. Yet this again provides only an approximation of the effect of measuring.

Examples more relevant to the field of psychiatry abound. For instance, when sliding someone into an fMRI scanner to measure the effect of some intervention, we must take into account that placement into the scanner itself too is an intervention which may cause some measurable effect besides the effect of the intervention that serves to test the null hypothesis. The introduction of (double) blind trials in which a control-group is included serves to help differentiate between non-specific effects resulting from participation in a study and effects of the intervention itself. Of course, these non-specific effects are far broader than only the effects of measurement.

In sum, with the assumption that measurement affects the phenomenon under study, I highlight that observation and/or measurement thus also affects the research results pertaining to that phenomenon.

5. What is intelligible to humans is not absolute but is instead dynamic
For Craver (2007) “explanations exist even if we cannot represent them cognitively” (p. 34). For my account, and based on the previous assumptions, Craver’s position takes the primacy of the ontic a step too far. Interestingly, and as discussed in Chapter 7, Salmon (1998) weds “intelligibility”, which at face value holds primarily epistemic connotations, to his ontic account. This is because, on his account, to explain is to “fit the event to be explained into an intelligible pattern”.

Instead, I take a more Bechtelian approach, adopting the notion that for something to count as an explanation, it must be intelligible to us. I allow that we often need computational tools in order to render complex data sets and systems consisting of more than a few parts intelligible (cf. de Regt, 2017). Furthermore, I hold that what counts as intelligible is not a static feature but can evolve with time and technological developments. Intelligibility is thus a dynamic notion. I furthermore follow Illari (2013) in claiming that “[i]f empirical engagement continually forms what we find intelligible, simple and unified, then epistemic constraints on explanation (…) are deeply entangled with ontic ones” (p. 254). I return to the ontic-epistemic divide in more detail in a later section.

6. There is a fundamental reciprocity between mind and world
For my construal of an account of subjectivity, I embrace the ideas of embodied embedded cognition and enactivism. Embodied embedded cognition is a
theoretical position, in line with dynamic systems theory, which holds that there is a fundamental reciprocity between mind and world. Cognition is “embodied”; it is not merely brain-based but is instead dependent on the physical body’s internal milieu. For instance, fatigue due to sleep deprivation has been shown to slow reaction-time (e.g. Williamson & Feyer, 2000). “Embeddedness” refers to the idea that the cognitive agent interacts with its environment. The body-world interaction, which is a reciprocal relation of mutual constraint, influences cognitive processes. Alboni et al., (2017) for example, showed that the effects of administration of specific serotonin re-uptake inhibitors (SSRIs), a class of anti-depressant drugs, to mice were influenced by levels of environmental stress. SSRIs worsened the condition of mice exposed to stressful environments whereas SSRI-treated mice in enriched environments showed marked improvements in their depression-like symptoms. This example shows that it is not just the internal milieu, or \textit{embodiment}, that affects cognition (in this case cognitive symptoms associated with depression). The environment affects, or constrains how a drug influences cognitive processes.

“Enactivism” is an umbrella term that takes the ideas of embodied embedded cognition a step further (e.g. Kiverstein & Clark, 2009). Enactivists hold that cognition itself is dynamically co-constituted through the interaction between an organism and its environment (e.g. Thompson, 2010). Noë (2006) for instance, holds that perceiving is not passive but is instead a skilful activity in which we investigate the world looking for information. Noë’s interpretation sits well with the idea of perspectival integration for the achievement of mosaic unity for which I argue later in this chapter. This is because it highlights the importance of scientific experience as a function of how we relate to the world of which we also form a part\textsuperscript{13}.

I advocate taking into account the ideas of embodied embedded cognition in particular both in the development of a model for explanation in psychiatry

\textsuperscript{13} Although it falls beyond the scope of this project, it would be useful to further extrapolate the idea of scientific practice as an enactivist enterprise. This could entail the observation that any mechanistic representation of a phenomenon under study must also contain the scientist and the way in which (s) he resonates with the phenomenon under study. Any scientific findings concerning the phenomenon would then be the product of dynamic co-constitution of the phenomenon in its context \textit{and} the scientist collecting information. The system represented as a mechanism as a whole would thus encompass scientist and phenomenon in a much broader context than is described by the New Mechanists in the current literature. An obvious disadvantage of taking the ideas of co-constitution this far, is that it in practice makes for impossibly (?) large and practically unintelligible complex systems riddled with feedback loops.
and in the scientific practice of psychiatry itself. These two contexts correspond to the last two of Glas’s (1991) four levels of analysis in medicine (i.e. everyday language, clinical practice, scientific practice and philosophy) outlined in Chapter 5. Furthermore, the assumptions outlined above, and especially 1, 3, 4, and 5 are in line with the embodied embedded tradition.

In this chapter, I draw on the ideas of embodied embedded cognition (EEC) and enactivism for the two levels of analysis mentioned above: philosophy and scientific practice. I use these for the development of a philosophical model of mechanistic explanation and for specific instances of scientific explanation (Glas, 1991). EEC proponents emphasize the importance of our “scientific experience”. In science, which was traditionally purported to be “objective”, our self and the reciprocity between self and the world we are studying scientifically, constitute scientific experience (Merleau-Ponty, 1962; Thompson 2010). Merleau-Ponty highlights the importance of taking into account the experiential domain: “[t]he whole universe of science is built upon the world as directly experienced, and if we want to subject science itself to rigorous scrutiny and arrive at a precise assessment of its meaning and scope, we must begin by reawakening the basic experience of the world of which science is the second-order expression” (1962, p. viii). I too hold that the experiential aspect of our scientific pursuits cannot be neglected: we are embedded in the world and our cognition is co-constituted through this embeddedness in the phenomena we seek to study. Any scientific reflections are thus inescapably imbued or entangled with our own subjectivity. To speak in the language of complex systems theory (see Chapter 8): the self-world system is minimally decomposable due to nonlinear processes and circular causation and therefore, subject and object are deeply intertwined and cannot be meaningfully disentangled. The self-world reciprocity thus entails that a strict subject-object divide becomes obsolete: our embeddedness in the world is too complex to decompose in such terms. This rejection of an absolute subject-object divide will feature in my rejection of the ontic-epistemic divide in mechanistic accounts later in this chapter.

All in all, I claim that when “doing science” scientists must 1) take into account that we are doing science - there is no view from nowhere. 2) Nature itself does not present a single optimal set of joints at which it is best carved but instead there are multiple, often incompatible chopping choices to be made. 3) The way in which we proceed our research is fuelled by expectations, perhaps in the form of theories or just mere inklings - it is theory-laden. And 4) when we do get started with the empirical dirty-work, we must accept that our observations and measurements in themselves affect the target phenomenon. 5) Intelligibility is a dynamic notion, dependent on the current state of
science and technology. In a nutshell, the assumptions illustrate that every step in the scientific process entails an interaction between world and scientist. Moreover, in both the development of a model for scientific explanation for psychiatry and in the scientific practice of psychiatry itself I advocate adhering to the philosophical ideas of embodied embedded cognition.

9.4 The ontic-epistemic divide

In Chapter 7, reference to the ontic-epistemic divide was made in three related yet distinct contexts: 1) mechanistic explanation, 2) the levels metaphor, and, 3) constraints on a good explanation. In this section, I first discuss how the New Mechanist’s reference to the divide in the context of mechanistic explanation has evolved from Salmon’s original definition and then delve into the three contexts in a little more detail for the purposes of clarifying my own, enactivism-inspired, position in the debate.

I noted that the divide is attributed to Salmon (1984; 1998) and cited his definitions of the ontic and epistemic conceptions of scientific explanation. Illari (2013) observes that “[w]hile this distinction between ontic and epistemic explanation originates with Salmon, the ideas have changed in the modern debate on mechanistic explanation, where the frame of the debate is changing” (p. 237) and advocates “seeing the relationship as one of integration” (Ibid.).

I examined the “frame-shift” and showed that it coincides with a move away from a law-based approach to a mechanism-based approach to scientific explanation (Salmon, 1998, p. 54; Bechtel, 2008, p. 10). Thus, Salmon’s and Bechtel’s definitions of the epistemic conception differ. As discussed in Chapter 7, Bechtel (2008) eschews the ontic view and states that “mechanisms do not explain themselves. (…) Explanation is an activity of scientists who must contribute mental labor in advancing explanations” (Bechtel, 2008, p. 18). For Bechtel, explanation consists of providing a representation (for instance linguistically or diagrammatically) of the contribution made by parts of a mechanism to its operation so that scientists may “construct and assess hypotheses about mechanisms and design experiments upon them” (Ibid., p. 21). Contrast this to Salmon, for whom the epistemic conception entails that an explanation is “an argument to the effect that the event to be explained was to be expected by virtue of the explanatory facts. (…) there is a relation of logical necessity between the laws and initial conditions on the one hand and the explanandum on the other” (Salmon, 1998, p. 53).

Based on the three contexts discussed below and drawing on an enactive approach where the project of scientific explanations requires accounting for our scientific experience,
I will argue for regarding the ontic-epistemic divide as a heuristic tool that emphasizes the difference between things in the world and our perception of those things. A benefit of regarding the ontic-epistemic divide as a heuristic tool, is that we can avoid the metaphysical debate on how to regard the divide. This fits my deeply pragmatic, metaphysically agnostic approach. Invoking the notion of a heuristic tool entails that we are free to employ that tool, without necessarily understanding everything there is to understand about the materials to which we are applying that tool. In the case of a tool, it is enough to have an inkling that a particular tool will be effective for the task at hand (i.e. using a saw rather than a screwdriver to shorten a plank).

9.4.1 Context I: mechanistic explanation in general

The first context in which the ontic-epistemic divide features here, is the broadest: mechanistic explanation. As discussed at length in Chapter 7 and reiterated briefly above, there are important differences between ontic and epistemic accounts of mechanistic explanation. On an ontic approach (e.g. Craver, 2007), it is the mechanism that constitutes the phenomenon under study itself, i.e. that does the explaining. Note that Craver, in later writings takes a looser approach to this very strictly ontic and Salmon-inspired approach and more explicitly allows for epistemic aspects (e.g. Craver, 2013, 2014, & 2016). On an epistemic approach, it is the representation of the mechanism (a text, or diagram for instance) which does the explaining. Despite the previously discussed problems, the two positions are clear enough. For my own account, given my agnosticism about the ontic domain, my focus is on an epistemically oriented account. I therefore move on to the second context.

9.4.2 Context II: the levels metaphor and levels of mechanism in particular

The ontic-epistemic divide features centrally in Craver’s taxonomy of levels (see Chapter 7). For the purposes of this project, I extrapolate to psychiatry Craver’s contention that levels of nature, science and theory do not coincide in the neurosciences (2007, p. 231). Keeping in mind that different types of levels do not necessarily coincide will help ward off much of the confusion caused by endorsing a more absolute account of levels and assuming a one-to-one mapping of levels of nature to levels of science. However, there are two aspects of Craver’s reliance on the levels metaphor that I discuss in this section which are both related to his assumption that the ontic and epistemic can be separated. I first take issue with the focus on levels of mechanism. And second, with a tension within Craver’s account of levels of mechanism. Regarding the former point, note that Craver posits “an ontological hierarchy working behind the scenes (...) that guides the development of theories, informs the criteria for evaluating explanations, and underlies the roughly hewn idea that theories are organized into levels” (Craver, 2014, p. 8).
For my account, this claim is too limited, it covers only half the story by neglecting considerations of scientific experience.

The second and related problem is within the levels of mechanisms account. It occurs because on one hand, mechanistic explanation entails situating phenomena within the causal structure of the world (Craver, 2007, p. 21; Craver, 2017, p.163). Hence, levels of mechanism, the notion of levels Craver deems most informative for mechanistic explanation, belong to the “levels of nature” category (2007, p. 171). On the other hand, he advocates a perspectival approach which better takes into account the importance of our scientific experience in developing scientific explanations. The following quotes from various writings demonstrate this. For instance: “the description of mechanisms is ineliminably perspectival” (Ibid., p. 259). And, “parts appear as parts only relative to a decomposition framed by reference to some highest-level property or activity” (Craver, 2014, p. 15). Here, the idea of framing suggests an agent doing the framing (through his theoretical background and based on available methodology and technology) and making the reference. Another example: “[e]xactly how many levels there are and how they are to be individuated are empirical questions that are often answered differently for different phenomena” (Craver, 2001, p. 63). Moreover, stating that the mosaic unity of science serves “the epistemic function of making mechanistic explanations robust, that is, of constructing explanations that are able to withstand scrutiny from multiple independent lines of evidence and multiple independent disciplinary perspectives” again demonstrates the importance of agent-mediated constraints on explanation (Craver, 2007, p. 232). Thus, appeal to levels of mechanism in the context of an ontic account (as intended in Craver (2007)) is problematic because “levels of mechanism” have both ontic and epistemic aspects – to speak in the language of the New Mechanists. The problems here dissolve when we take an EEC stance, according to which we drop the strong ontic-epistemic divide and opt for a perspectivalism that is deeper than intended here by Craver: the perspectivalism of the EEC proponents is inherently saturated with the (causal) reciprocity between self and world.

Moreover, the problems within the levels of mechanism account, pertaining to their situation in the ontic-epistemic debate raise the question what work the “levels of mechanism” are doing in the mechanistic approach to explanation in the first place. Or in other words, do we need to adhere to a specific account of levels in order to define the mechanistic approach to constitutive explanation? Eronen (2015) states that providing an account of levels serves four main purposes. First it provides a framework

14 This is in keeping with my assumption that nature does not offer a single set of joints for carving but that carving methods are constrained by the explanatory task.
for understanding reductive or mechanistic explanation. Second, it captures significant features of the organisation of the world. Third, it helps us understand talk of levels in the sciences. Fourth, it provides a framework for understanding downward causation (Eronen, 2015, p. 41).

Despite these interpretative functions, I contend that we do not need them to defend a mechanistic approach to scientific explanation. Craver emphasizes that the “levels metaphor” is a descriptive tool, “central to the explanatory practices of the special sciences and defensible as a metaphysical picture of how phenomena studied in the special sciences are constituted” (2014, p. 1). Eronen (2015) in contrast, rejects the ubiquity of levels of mechanism and mechanistic explanation presupposed by Craver and Bechtel and instead suggests that it is more accurate to say that scientists ubiquitously appeal to relations of composition and scale and that these notions suffice for scientific practice (pp. 46 & 54). Recall that my review of the apathy literature supports this: reference to levels is relatively rare and I did not encounter any articles that employed the term in the way intended by the New Mechanists.

Bechtel (1994) suggests that “[t]he view of mechanistic explanations puts a different perspective on the way we conceive of levels and relations between levels in science” (p. 15). He goes on to add: “[e]xplaining a phenomenon is in part a matter of finding the correct level for understanding particular interactions. Moreover, the resulting model is inherently interlevel” (Ibid., p. 16). I find Bechtel’s idea that scientists are already actively occupied with “finding the correct level” overly optimistic. My literature study shows that scientists themselves rarely appeal to levels in this way and are thus not concerned with “finding the correct level”. The process of situating research and (hypothetical) explanations of a phenomenon into the explanatory landscape of science more often takes place post-hoc, if at all. It also frequently occurs that researchers contribute mosaic tiles (of explanations of a very specific target phenomenon) to a large pile rather than placing their own tile beside other tiles to form a picture (the so-called mosaic unity). Cochrane reviews\(^\text{15}\) in the medical literature are an example of how scientists fit research findings together to achieve some form of unity. Occasionally researchers describe the gap in the mosaic which they aim to fill with their research a priori - often this occurs in the context of funding applications. A final point on the matter is that the

\(^{15}\) Cochrane Reviews are “systematic reviews of primary research in human health care and health policy, and are internationally recognized as the highest standard in evidence-based health care resources. They investigate the effects of interventions for prevention, treatment, and rehabilitation. They also assess the accuracy of a diagnostic test for a given condition in a specific patient group and setting” (definition from http://www.cochrane.org).
representation of a phenomenon in terms of a mechanism does not necessitate analysis in terms of levels of mechanism but only in terms of parts, wholes and their interactions (e.g. Eronen, 2013 & 2015).

9.4.3 Context III: normativity and teleology in the ontic domain
The tension mounts when we move beyond defining what an explanation is and the ontic-epistemic divide in the levels metaphor, towards the third, normative context: constraints to define a good explanation (Illari, 2013, pp. 240-241). The frameshift here is due to Craver’s (2007) focus on causal relevance as a requirement for a good explanation (as is reflected in constraints E1-E5). Here, ontic and normative aspects come together and how the inclusion of normative aspects call for an account of teleology too. The line of reasoning that brings me to discuss teleology can be summarized as follows:

• For Craver, a philosophical account of mechanistic explanation is normative: it offers a means to distinguish good explanations from bad ones.
• Good explanations are explanations that contribute to the goal of explanation: they offer a means to describe, predict and manipulate/control/intervene in the (workings of the) explanatory target.
• Due to this goal-directedness, I contend that explanations on the mechanistic account are teleological.

The above is uncomplicated so far. However, when combined with Craver’s (2007) contention that his account is ontic (i.e. that mechanisms are operative in the world and that these mechanisms, rather than their representations, themselves explain), the normative and ontic facets of his account seem conflicting. Recall that the narrowest formulation of the ontic approach is exemplified by the following citation, which demonstrates the contention that the target of explanation and its explanation itself coincide:

Other times, the term explanation refers to an objective portion of the causal structure of the world, to the set of factors that bring about or sustain a phenomenon (call them objective explanations). (…) There are mechanisms (the objective explanations) and there are their descriptions (explanatory texts). Objective explanations are not texts; they are full-bodied things. They are facts, not representations. They are the kinds of things that are discovered and described. There is no question of objective explanations being “right” or “wrong,” or “good” or “bad.” They just are.” (Craver, 2007, p. 27)

As previously discussed, the citation offers an account of explanation is that is too narrow for my own purposes. First, because Craver’s reference to “an objective portion of the
causal structure of the world” is at odds with my contention that scientific practice is about the world as we experience it. As discussed above, I have stated that I am agnostic about properties of an “objective portion” or about the “causal structure of the world” outside of the explanatory, (subjectively mediated,) realm. Second because, as I claimed in Chapter 7 in agreement with Illari (2013, p. 242), the term “explanation” is best substituted by “cause” in Craver’s citation. Third, in the citation, Craver himself rules out that this particular type of explanation can be evaluated normatively because it just is. Thus, in this citation, in which the explanation and the target of explanation coincide, Craver’s rejection of normativity in the phenomenon under study is at odds with his desire to distinguish good explanations from bad ones. For an explanation to be “good” or “bad”, the explanation must be something over and above the target of that explanation, that can offer the powers of description, prediction and intervention required of an adequate explanation.

In this section, I have thus identified a source of tension between ontic and normative principles in Craver’s account as being due to his proposed relation of identity between the explanation and its target phenomenon. First, I rejected this identity claim, nonetheless welcoming inclusion of normative aspects. I instead embraced a deeply pragmatic approach to defining a good explanation: an explanation is good because it contributes to intervention, prediction and description. Or put more simply: an explanation is good because it works. Second, I have shown that a strict interpretation of the ontic-epistemic divide fails in each of the three contexts outlined in the previous section. A weaker interpretation of the ontic-epistemic debate is more suitable for the purposes of developing an account of mechanistic explanation. On this weaker interpretation, the divide is regarded as an interpretative tool rather than a framework for the introduction of absolute categories between which mechanistic thinkers must choose. This move takes Illari’s (2013) contention that mechanistic accounts must take into account ontic and epistemic considerations one step further, by questioning the status of such a divide in the first place. The move is based on the embodied embedded approach to cognition (see my sixth basic assumption above). Rejecting ontic-epistemic dualism as an absolute divide and regarding it as a heuristic tool that indicates two poles on the self-world continuum and instead adopting the terminology of a self-world reciprocity at all of Glas’ (1991) levels of analysis (i.e. philosophy, science, clinical practice and everyday being) rids us of many of the difficulties discussed above. In the next section, I zoom in on the issue of teleology in mechanistic explanation, which in this discussion was closely linked to normativity (because goodness is linked to purposiveness). To structure the discussion, I examine the roles of teleology in three contexts.
9.5 Teleology in mechanistic explanation

The question about teleology in mechanistic explanation (ME) can be posed at three levels of analysis. The first pertains to the philosophical mission of developing a model of mechanistic explanation, the second pertains to an instance of a mechanistic explanation and the third pertains to the phenomenon under study.\(^\text{16}\)

<table>
<thead>
<tr>
<th>Context</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development of model of ME (philosophy)</td>
<td>Providing a model for good explanations that contributes to description, prediction and intervention of the target phenomenon.</td>
</tr>
<tr>
<td>Specific instance of ME</td>
<td>Explaining a phenomenon in terms of parts, and their interactions. Parts are delineated in terms of their function in the whole. Calling upon purposiveness aids explanation.</td>
</tr>
<tr>
<td>Target phenomenon</td>
<td>I refrain from answering the question whether the purposiveness which we perceive in living organisms reflects intrinsic purposiveness that exists independently of our (scientific) experience.</td>
</tr>
</tbody>
</table>

Figure 9.1 Teleology at three different levels of analysis.

9.5.1 Context I: teleology in the development of a model of mechanistic explanation

The first question about the purpose of a mechanistic explanation poses few problems and has been discussed above. New Mechanists largely agree that mechanistic explanation serves to describe, predict and intervene in the phenomenon under study. I have embraced these aims of mechanistic explanation and agreed that good explanations are those that best achieve these goals. I have also adopted Craver’s constraints E1-E5 as essential tools for detecting causal relevance and thus contributing to the achievement of the goals of explanation. However, a consequence of this pragmatic stance is that it allows conflicting explanations for a specific phenomenon (let us assume that they are “equally good”\(^\text{17}\)) to co-exist. The account does not provide means to decide between the two; the best explanation is simply the one that best stands up to the explanatory

\(^{16}\) Another way of framing the conflict in Craver’s 2007 account is in terms of the above questions. The second and third questions, which clearly pertain to different domains (i.e. that of the explanation and that of the target phenomenon, respectively) become conflated on a strict reading of the citation from 2007 according to which explanation and target mechanism coincide.

\(^{17}\) I use the phrase “equally good” very loosely. I explicitly refrain from treating questions on measuring how good an explanation is and using this information to compare two competing and possibly conflicting explanations for a target phenomenon.
task at hand. An extreme consequence, somewhat reminiscent of the current situation in the scientific literature of psychiatry (!), is that we risk ending up entangled in a jumble of complementary and conflicting explanations. A more positive way to label this consequence is to refer back to the mosaic metaphor. In this case, the tiles of the mosaic represent explanatory perspectives. Sorting through the jumble of “mosaic tiles” to achieve unity can be regarded as an essential task for scientists. Besides forming their own explanations for a target phenomenon, scientists are then also charged with the (perhaps even more challenging) task of fitting their explanations into the explanatory landscape in science.

Note that the phrase “fitting explanations into an explanatory structure in science” can be regarded as a substitute for Craver’s project where to explain is to situate a phenomenon in the causal structure of the world. The transition is in line with the fundamental assumptions of my position explored in an earlier section, of which agnosticism with regard to “the causal structure of the world” as something independent of subjectivity is a key tenet. “Levels of mechanism” form a tool to help represent the way in which explanations of aspects of a target phenomenon (which in turn is composed of parts and their interactions) fit together - a little like families of nesting dolls. My use of “levels of mechanism” departs from Craver’s. Mechanisms here pertain to explanation and science rather than to nature.

Thus, the first question about accounting for teleology in the development of a philosophical model for mechanistic explanation is not in itself problematic. Above, I showed that the question, to which my response is pragmatic, nonetheless brings to light further consequences of this pragmatic approach: what counts as a good explanation is not absolute and the task of melding together explanations of related target phenomena is presented as a key task for the scientists offering explanations. Fitting explanatory perspectives together can yield a form of unity. The unity pertains to the whole of that particular mosaic patchwork, but does not rule out that there are many ways in which the perspectives can be put together, each yielding its own mosaic unity.

9.5.2 Context II: teleology in specific instances of mechanistic explanation

I move on to the second question: what is the role of teleology in specific instances of mechanistic explanation? Researchers seem drawn to (or maybe they simply cannot escape) using the language of function in their writings – perhaps for purposes of sense-making. They describe things in terms of their role, function or purpose within a broader system, thus introducing teleology. Although both parts and their interactions are frequently described in terms of their function within the system, the discussion
here will focus on parts. This is because the conclusions I draw for teleology of parts also hold for the interactions between those parts. In addressing the role of teleology in specific instances of mechanistic explanation, I argue that mechanistic explanation and teleology are in fact entangled rather than incompatible. I show how Craver’s implicit suggestion that mechanistic and functional descriptions differ is misleading. In the discussion, I include an outline of an enactivist approach to the problem of bringing together teleology and mechanistic explanation. I suggest making a distinction in terms of the degree to which we know of, or suspect, a physiologically plausible part instead. An overarching goal of this section is to render mechanistic explanation a more feasible model of explanation for psychiatry research and in the process avoid reverting back to a biomedical model in which explanations must bottom out in the robustly detectable and physiologically plausible parts that are conspicuously absent in psychiatry.

Early mechanists, such as Descartes, found it particularly difficult to bring together the notions of mechanism and teleology in accounting for the behaviour of living organisms. Thompson in contrast, weds the two in his book Mind in Life (2010). He draws extensively on systems theory and mechanistic thinking in arguing for a deep continuity of mind and life. He invokes circular causation, nonlinear processes, and self-organisation to unite mechanistic thinking and teleology (2010, pp. 138-139). His key claim is that “life and mind share a set of basic organizational properties, and the organizational properties distinctive of mind are an enriched version of those fundamental to life” (Ibid., p. 128). Thompson eschews “objectivism” because it “takes the organism for granted as a ready-made object out there in the world. No concern is shown for how the category “organism” is constituted for us in scientific experience” (Ibid., p. 164). Instead, he appeals to a more reflexive self-world dialectic. Teleology occurs in this self-world dialectic: a bacterium moves towards sucrose because sucrose means “food” and ultimately “survival” for that bacterium (Ibid.).

Craver (2013) concedes that mechanistic descriptions are inherently teleological in the neurosciences (p. 140). Writing about the example of neuronal firing he states: “[t]he mechanism works from beginning to end, where the end is not what the mechanism invariably does but what we think it is supposed to do” (Ibid.). Bechtel and Abrahamsen support this. They define a mechanism as “a structure performing a function in virtue of its component parts, component operations, and their organization” (Bechtel &

---

18 See chapter eight for a treatment of the importance of circular causation and non-linear processes for explanation of complex systems.

19 I take this contention one step further, suggesting that explanations are inherently teleological regardless of the scientific discipline.
Abrahamsen, 2005, p. 423, my italics). Wright and Bechtel (2007) explicitly argue for inclusion of functional parts, but they emphasize that they serve mainly to aid further research aimed at detecting robust physiological parts. Craver introduces a functional stance: “a stance that there is a behavior that the mechanism as a whole exhibits (that it is the mechanism of a behavior) and that the components of the mechanism are organized and interact such that they exhibit its overall behavior” (Ibid., p. 156). He posits three “projects” to which it can contribute: the first is to “understand how functional and mechanistic descriptions are related to one another in physiological sciences” (2013, p. 135). The second shows “how functional descriptions can be explanatory even when there is no etiological story to tell about how the functional item came to be”. And he adds that “functional description can serve as a form of causal-mechanical explanation; it is a means of situating an item in the causal structure of the world” (Ibid.). The third “is to make explicit the evidential criteria by which functional and mechanistic descriptions are evaluated” (Ibid.).

Craver’s reference to a separate “functional” stance seems at odds with his mission of showing that mechanistic explanation works well for the neurosciences (in which teleology is “inherent”). Especially because his formulation of the functional stance’s projects implies that the functional stance is something separate from the mechanistic stance20. I argue that this is misleading, and reject the idea of a separate “functional stance” precisely because the language of function is already inherent to the language of mechanisms. As the citations and detailed discussion in Chapter 7 show, the decision to call something a “part” depends on our ability to decompose the system under study in terms of parts with separate functions and localise parts (Bechtel & Richardson, 2010).

If we embraced the notion of a functional stance as something different from the mechanistic stance, this would mean that practically all explanations posited in psychiatry are proto-mechanistic and that they would satisfy the criteria for a sound mechanistic explanation only if we find robustly detectable and physiologically plausible substrates for the merely functional parts that they currently feature. This does not seem conducive to furthering our understanding of psychiatry in which accounting for more than just biological aspects is essential.

Instead, I suggest that entities and their activities that constitute a mechanism that can be situated in the how-actually to how-possibly range. This range is a measure

---

20 Craver for instance refers to a “happy coexistence of functional and mechanistic descriptions in our contemporary physiological sciences” (2013, p. 134) and an “intermingling of functional and mechanistic descriptions” (Ibid.).
of “constructedness” rather than a measure of the degree to which a (sub-)system is functional. Parts with a virtually undisputed biological realisation reside near one pole, and those for which we have not yet found a biological realisation reside at the other\textsuperscript{21}. Even parts with a clear biological realisation can be functional too. Again, this is because the parts that feature in mechanistic explanations are chosen based on their function\textsuperscript{22} in the mechanism (Craver, 2007; Bechtel & Abrahamsen, 2005; Bechtel & Richardson, 2010). We can concede that ideally, mechanistic explanations draw on physical parts that reside near the how-actually end of the spectrum. These mechanisms are, due to their more predominantly physical nature, closer to the mechanism archetype (articulating gears) than mechanisms composed of functional parts with no (known) physiological substrate. Yet, the difference between these more and less physically delineable parts has little to do with function because all parts are parts that have a role within the system anyway. But what does this mean for the status of mechanisms composed of “how-possibly”, constructed parts? For instance the mechanism described by the Oedipus complex? Or mechanisms described in terms of other psychological faculties?

Given this line of argument and my rejection of a separate “functional stance”, we need perhaps not accept that explanations in psychiatry are unlikely to meet the criteria for a sound mechanistic explanation. In a scientific climate in which having robustly detectable, physiologically plausible parts is the ideal, mechanisms comprised of psychological parts are sub-ideal. However, on my more pragmatic approach, such mechanisms can nonetheless count as “good” if they contribute to description, intervention, and prediction of the target system’s behaviour.

A drawback of the inclusion of constructed, functional parts, within mechanistic explanations is that it departs almost entirely from the generic mechanism that comes to mind (with articulating cogwheels, that physically move in a particular direction and interact in an easily intelligible manner). Does the reference to “mechanism”, which is, at heart a metaphor, make sense in the context of non-physiological, constructed, functional parts. In stretching the “mechanistic” aspect of mechanistic explanations, do we not risk throwing out the baby with the bathwater?

I argue that this baby can be saved. The idea of mechanism, of parts and their interactions and of reciprocity between parts and whole need to be viewed in a more abstract manner. We can still retain the ideas of parts or entities, which interact to account for

\textsuperscript{21} See chapter seven for discussion of these terms

\textsuperscript{22} In Craver’s words: parts “make identifiable contributions to the behavior of a mechanism” (2007, p. 188)
the behaviour of a system. In fact, I suggest that this is where network analysis comes in. Network analysis, although not in itself an explanatory model (as discussed at length in Chapter 8), can help model more abstract parts and their articulations (or causal relations) using mathematics. I return to the role that network analysis can have in the quest for a model of scientific explanation suitable for and conducive to furthering our understanding of psychiatric phenomena. I return to this point in Section 9.6, but first further divulge my own position regarding another aspect of teleology.

Now that I have accepted teleology as integral to mechanistic explanation, it is time to reflect further on its status within the account. To clarify my own position, I first reintroduce the distinction between constitutive and regulative concepts in a manner that resonates with how these terms were employed in Chapter 7. Constitutive concepts concern what makes up our knowledge (or an explanation, as in the context of constitutive explanation discussed in Chapter 7). Regulative concepts are those that guide our inquiry (e.g. constraints E1-E5) (cf. Thompson, 2010, p. 137). I follow Maturana and Varela (1980), who state that “purpose” or “aims” are notions that belong to the domain of our discourse about our actions (pp. 85-86), and I thus regard teleology as a regulative principle. “Purpose” and “aims” are “descriptive notions used to characterize a system in relation to some context of use defined by an observer” (Maturana & Varela, 1980, pp. 85-86). In contrast to Thompson (2010), I do not answer the question of whether teleology is also a constitutive concept that exists outside of scientific experience23.

The purposiveness we see in living organisms is thoroughly context-dependent. The following example highlights the context dependency of teleological aspects (and also shows how normative and teleological aspects are intertwined). For a single tree, dying can be deemed “bad” because the tree ceases to be, whereas surviving (or maintenance of homeostasis) is regarded as the intrinsic purpose of a tree. In contrast, in the context of a forest, the death of one tree may make more light and water available to surrounding trees allowing them to flourish at the cost of one single tree. So in this larger system, the death of that specific tree is “good” and serves the survival of multiple other trees. For a single tree, normative notions such as “good” and “bad” and “functioning” versus “dysfunctioning” have different meanings than for a collection of trees. The example highlights that there is a difference between the intrinsic and extrinsic purposiveness of a specific system. Intrinsic purposiveness is a spin-off of the tree’s self-organizing properties; the purposiveness is generated by the tree itself and pertains to itself. Extrinsic purposiveness refers to the idea that the tree serves a purpose outside of itself (e.g. being felled to ensure the survival of other trees or for its wood to be used as a building

23 I return to this point in the next section on teleology in nature.
material) (cf. Thompson, 2010, p. 145). Given that what counts as a single system is context dependent (and as I argue, is not given by nature in any absolute sense), what counts as intrinsic and extrinsic too cannot be absolute. I therefore wish only to claim that teleological notions (pertaining to the function of a part in a mechanism) make sense only when the context of the system under study is taken into account. This holds even for intrinsic purposiveness because, the context is still required to demarcate the boundaries of the system to which intrinsic purposiveness is attributed. The boundaries are given by the observer and the way in which he is interrogating the phenomenon under study.

Due to my enactivism-inspired approach, due to which teleology is positioned as inherent to the self-world relationship and integral to our scientific experience, the discussion about the status of teleology in mechanistic explanation led me to also comment on my position regarding the idea of teleology in nature. I continue this discussion in addressing the third and final context for teleology.

9.5.3 Context III: teleology in nature
The third question on teleology is: (how) can we account for function or purpose in the phenomenon under study? Craver follows the proto-mechanists such as Descartes and states that “the causal structure of the world is disenchanted and purposeless” (2013, p. 134). Thompson, in line with his enactivist position in which the ontic-epistemic divide dissolves, in contrast argues that “purposiveness is a constitutive property the whole possesses because of the way the system is organized” (2010, p. 146). As noted in the previous section, I take yet another approach. I adopt an agnostic stance about whether the notion of teleology makes sense in the context of the causal structure of the world. I am unsure of how to conceive of a notion of teleology that is independent of (scientific) experience. In other words, I have no means of knowing whether teleology is 1) strictly human (e.g. Craver, 2013; Maturana & Varela, 1980), 2) whether it also belongs to nature independently of a self-world relation or 3) whether it is intrinsic to the self-world dialectic and our scientific experience which is superimposed upon this dialectic (e.g. Thompson, 2010). Unsurprisingly, this view is extrapolated from my broader agnosticism about all matters metaphysical. My discussion of teleology remains limited to the idea of purposiveness as a feature of lived (scientific) experience rather than the question of whether it is a feature of nature independently of our subjectivity.24

In sum, in this section, three different contexts in which questions about teleology arise

24 I concede that human subjectivity is brought forth by nature and is therefore belongs to the natural domain. When I refer to “nature” however, I intend the world as it is, independently of lived experience (the “ontic” pole of the ontic-epistemic spectrum).
Chapter 9

were analysed. The first context is the philosophical task of developing a model for mechanistic explanation. Given that there is little disagreement amongst New Mechanists about the goal of explanation, teleology is not problematic in this context. For the second context, concerning teleology within instances of mechanistic explanation, I first showed that teleology is inherent to mechanistic thinking and rejected the idea of a separate functional stance. Next I drew on enactivist thinking and reflected on the status of teleology. I posited that teleology stems from the self-world relationship, where it is a regulative principle that thus guides our inquiry. For the third context, teleology in nature, I took an agnostic stance, limiting my discussion of teleology to lived (scientific) experience.

In the following section, I address the role of Network Analysis in light of my call for a more abstract interpretation of the “mechanism” metaphor which comes with the explicit inclusion of purely functional, constructed parts in the model for mechanistic explanation.

9.6 Network Analysis and mechanistic explanation as complementary methods

The suggestion, ventured earlier, that we must accept a more abstract idea of “mechanism” for mechanistic explanation so that mechanisms consisting of “merely” constructed, non-physiological parts can still be approached mechanistically, invites employment of network analysis. It paves the way to address the third central question of this project pertaining to the implications of an adapted account of scientific explanation and clinical understanding of apathy.

As an aside, I note that although this primarily philosophical study does not of course itself yield new knowledge about apathy, it does offer a potentially useful methodology for investigating phenomena. Despite the fact that my own clinical background has led me to incorporate clinical practice into the central questions of this project, my main focus has been on the scientific domain. I have argued earlier that the scientific and clinical levels of analysis differ and that they constitute different perspectives on a target

---

25 Combining a mathematic methodology with a more primarily philosophical explanatory framework was suggested by Salmon (1984). He suggested combining his proposed statistical relevance account with his proto-mechanistic account of explanation in a manner that is similar to the suggestion I am making here. An important difference is that further development of both models have made them more suitable for the task of suggesting a model of explanation in psychiatry that accommodates the need to integrate aspects from different levels of explanation.
phenomenon (usually a patient suffering from psychiatric symptoms). I have chosen not to remove the clinical aspect from the central questions because I am convinced that furthering our scientific understanding of psychiatric phenomena by anchoring our research in firm philosophical grounds, will also positively influence clinical practice. By asking the right questions in the scientific domain, we can hopefully learn about psychiatric disorders and more effectively help those suffering from them.

Recall that network analysis was introduced in Chapter 8. It is a complex systems theory-inspired mathematical method to model and measure the relations between entities. The potential of network analysis for explanations in psychiatry lies in its neutrality: any entity/part can be modelled as a node in a network and relations between nodes can then be interrogated mathematically. The nodes need not represent physiologically plausible entities but can instead represent any conceivable thing (even if we are not sure of its status as a relevant part). Combinations of different types of entities (for instance biological, psychological and social) can all be represented in a single graph, with their (hypothetical or empirically tested) correlations represented by their connecting lines. Whether network analysis harbours explanatory power in such situations (without offering a complete explanatory model) depends on how we are using it to interrogate the explanandum (Craver, 2016). Network analysis thus offers a means to explore the explanatory structure of the phenomenon at hand. It can help determine:

1) Whether parts represented by nodes may be causally related (see Chapter 8 for more details on detection of causal relations): i.e. a change in a variable pertaining to one part corresponds to a change in variable in another part.

2) Whether entities modelled as nodes are best regarded as in a part-part relationship or in a part-whole relationship. A network that takes into account the temporal dimension can help to determine whether a relation is causal in the sense that a change in one node is followed in time by a change in another (suggesting a possible part-part relationship between the entities represented by those nodes), or whether a change in one node corresponds directly to a change in another, suggesting that the entities represented by those nodes are in a part-whole relationship.

3) Given the two points above, network analysis can serve to determine whether the network accurately displays a part-whole relationship in which nodes represent relevant constituent parts and the network as a whole represents the system under study as a whole. Thus, if two entities represented by nodes turn out to be better represented in a part-whole relationship rather than a part-part relationship, the network representation must be adapted accordingly. As noted, such an adaptation would ideally show all the
relevant parts of a chosen “whole” phenomenon, so that the whole is represented by the graph in its totality.

In Chapter 8, I claimed that Network Analysis is not in itself an explanatory model in the sense intended in this project because it does not offer means to test all of the constraints E1-E5. Nonetheless, I adopted Huneman’s (2010) suggestion that network analysis and mechanistic explanation are complementary. Having discussed Mechanistic Explanation and Network Analysis in the previous chapters, it is now time to explore how they can complement each other to suit research in psychiatry. Before moving on, I next offer a short intermezzo on the thorny issue of interlevel causation in relation to mechanistic thinking and network analysis. I draw on Romero (2015) to dissolve the tension.

9.6.1 Interlevel-causation: a final hurdle?

In discussing the complementarity of network analysis and mechanistic methods, it is important to draw attention to a possible source of tension in the accounts: interlevel causation. Above, in discussing how network analysis can help clarify the causal structure of the phenomenon under study, the question of differentiating between parts and wholes arose. Although this may seem trivial, in the field of psychiatry, this is often far from clear. In the previous section, I made the assumption that temporality may help in this differentiation. I make a case for this point below, by first outlining the source of the tension, drawing on Romero (2015).

Romero (2015) illustrates that there is a tension between three commonly endorsed and individually plausible views. The three positions, which I discuss in more detail in below, are:

1. The interventionist view of causation (Woodward, 2003): ideal interventions are those in which a change in one variable produces a change in another, such that, when other influences are controlled for, the change in the former is a sufficient and necessary condition for the change in the latter. If these conditions are met, the former can be said to be a direct cause of the latter.

2. The “mutual manipulability account of constitutive relevance” (Craver, 2007). On this view (see also Chapter 7), a change in a component of a mechanism, for instance through an intervention, produces a change in the whole mechanism and vice versa. This account suggests that there is reciprocal interlevel causation, i.e. between part and whole.

26 Huneman (2010) argues that network analysis is a form of non-causal, non-mechanistic explanation, which he calls “topological explanation”.

216
3. The idea of interlevel causation, i.e. that entities at different levels in the part-whole hierarchy are causally related, can produce “problematic redundant and cyclic causal structures” (Romero, 2015, p. 3732).

In a nutshell, the views don’t immediately sit well together. One issue (as demonstrated in Chapter 7) is that despite Craver’s (2007) endorsement of interventionism for mechanistic explanation, “ideal interventions” in the sense intended by Woodward are vanishingly rare, especially in psychiatry. In practice, it is often impossible to isolate and intervene on a single part without there being any other side effects. A second issue is that, given the “fuzziness”\(^{27}\) of the parts and a lack of understanding of other intervening factors, it is difficult to achieve “ideal” settings in which other factors are indeed controlled for. For mechanistic explanations, we must thus make do with less-than-ideal manipulations (or naturally occurring changes) in variables to then test if they produce changes in others. Third, is the so-called “redundancy problem” (Romero, 2015, pp. 3739-3740), which arises in the case of top-down interlevel causation. Given that the whole cannot simply be reduced to its parts and that we accept that the properties of each part are determined by their own sufficient physical causes, we are left with a problem. How can the whole causally influence its constituent parts without causal over-determination?

A fourth issue is what Romero calls the “cyclicity problem” (Ibid., p. 3740): the occurrence of (redundant) cyclic causal structures across levels. To this, and following Romero, I offer two responses which both draw on the role of temporality. One response to the problem is to hold that the cycles are not redundant but that they portray feedback relations in the mechanism (see Chapter 7.8 for an example). In the case of feedback, a change in the whole is followed in time by a change in one or more of its parts (to respond to the former change). However, it does not seem plausible to subsume all cases of top-down causation under the heading of feedback loops. Therefore, a second response is that we can take changes in the whole to be *synchronic* with changes in its parts. Thus in this situation, a change in the whole immediately corresponds to, but does not “cause” in the sense of E1-E5 discussed in previous chapters, a change in one or more parts. This response in fact does away with interlevel causation leaving us with feedback loops as the only cases of true interlevel causation.

The way in which I dissolve the issues of interlevel causation thus follows from conclusions I have previously argued for:

\(^{27}\) That is, they are not robustly detectable and physiologically plausible.
First, for an account of mechanistic explanation, we are not dependent on levels of mechanism (except as a useful heuristic tool). Second, given the possibility of feedback loops and cyclic causation, we need to adopt a very relativistic and pragmatic interpretation of what causes and effects are. What counts as a cause from one perspective, may be an effect from another. Third, this confronts us with the boundaries of intelligibility: our capacity to think non-linearly is limited and we often depend on mathematical methods, such as those used in network analysis, to predict the behaviour of systems represented in a non-linear fashion. This should not however be an argument to eschew all processes non-linear, because it may be the case that accepting non-linear representations can significantly further our understanding of psychiatry. Fourth and finally, empirical practice is far messier than most philosophers allow. Practical difficulties more often than not prevent the application of the elegant theories developed by philosophers (e.g. the idea that systems’ behaviour is governed by laws; or the idea of ideal interventions in which interventions affect one part only and in which all other conditions can be controlled for; or the idea that causes and effects are clearly demarcated). I thus arrive at a position regarding interlevel causation that is formulated neatly by Romero (2015): “a mechanism’s activity and its components arranged and working in the right way are the same physical event. As such, an intervention on a component changes both the component and the whole. And an intervention on the whole changes both the whole and at least one of its components. Nonetheless, even though the intervention counts as a single physical event, we can measure its effects either on the whole mechanism or the components using different techniques (…) and register them in changes in variable at different levels” (pp. 3746-3747, italics in original).

An illustrative but highly schematic example (see Figure 9.2): when designing a graphic representation of factors that play a role in apathy, one may include nodes such as: “altered dopamine transmission”, “meeting diagnostic criteria for schizophrenia”, “use of antipsychotic medication”, and “apathy”. One may then attempt to add in the edges based on empirical findings. For instance, one may note that receiving the schizophrenia diagnosis leads to (i.e. the nodes are joined by a directed edge) administration of antipsychotic medication. And anti-psychotic medication has been shown to alter dopamine transmission (i.e. these nodes too, are joined by an arrow), which in turn is linked to an increase in apathy (an undirected edge because they co-occur but we don’t know if one causes the other or whether they are correlated due to being in a constitutive relation). However, the occurrence of apathy and receiving a diagnosis of schizophrenia also co-occur (undirected edge due to synchronous occurrence) independently of anti-psychotic medication use. Differentiation between sequentially and synchronously occurring phenomena that can be placed at different levels of mechanism can help determine whether we are dealing with a feedback loop or a part-whole relation in...
which a change in the whole corresponds to a change in at least one part and a change in a part corresponds to a change in the whole.

Figure 9.2 Example of graphic representation of apathy.

9.6.2 An apathy-based example of combining network analysis and mechanistic explanation

In the remainder of this section I draw on an article from the apathy literature which is already philosophically informed and explore how network analysis and mechanistic explanation could be combined. To investigate how network analysis and my adapted version of mechanistic thinking may work together, I draw on an article to provide an example: “Weakness of Will, Akrasia, and the Neuropsychiatry of Decision Making: An Interdisciplinary Perspective” (Kalis et al., 2008).28

This article was chosen because it explicitly offers “an interdisciplinary perspective”. The authors adopt “a common theoretical framework” for the process of decision-making, according to which it is divided into three stages: option generation, option selection, and action initiation. Problems may occur at each of these stages and are labelled “akrasia as accidie”, “decisional akrasia” and “last-ditch akrasia” respectively. “Akrasia as accidie” refers to the hypogeneration of options for action due to listlessness or indifference - the subject suffering from it is unable to come up with options for action in the first place. “Decisional akrasia” refers to dysfunctional option selection – despite being able to conjure up enough ideas, the subject chooses which action to take poorly or not at all. “Last-ditch akrasia” refers to not acting on one’s decision at the last moment, ditching it for some other course of action (or no action at all) against one’s better judgement.

The authors acknowledge that their model harbours assumptions about the nature of the phenomena under study that are not (yet) supported by empirical data, and, that there are alternative, sometimes conflicting ways of defining concepts and selecting theoretical frameworks. To organise the aspects of decision-making, the authors provide the figure

28 “Akrasia” means “lack of self-control”.

(9.3) shown below.

![Diagram showing three levels of explanation in the model: phenomena, processes/stages, and brain regions.](image)

Figure 9.3 Three levels of explanation are distinguished in our model: (1) phenomena, in philosophical or psychopathological terms; (2) processes or stages of decision making; and (3) associated brain regions. SMA: supplementary motor area; PFC: prefrontal cortex. Reprinted from Kalis et al., (2008) with permission from publisher.

The authors furthermore state that they are interested in whether and how these different aspects can be integrated. Meanwhile, they note that:

Despite the interesting overlap between philosophical and psychiatric points of view, the aims and starting points in philosophy are different from those in psychiatry. This implies that neuropsychiatric insights will not be relevant for all aspects of the philosophical debate on akrasia. In part, the philosophical problem of akrasia is a problem that requires a philosophical (i.e., conceptual) solution. (Ibid., p. 412).29

Kalis et al. review philosophical literature and scientific research on their topic to tentatively fill in the items on their matrix. They review the scientific literature, drawing on research using a large range of methods spanning the three levels of explanation in their matrix. For example, they include research using psychological or cognitive tests (e.g. for capacities such as creative thinking, abstract problem solving, measurement of the role of time on option selection), brain imaging techniques, lesion studies, factor analytic studies of the Apathy Evaluation Scale.

---

29 Earlier, the authors state that the philosophical debate on akrasia focuses on the questions whether it is “possible to act against what you think is best” and how actions that go against one’s values should be explained (Kalis et al., 2008, p. 403).
Interestingly, the term “mechanism” is used in a variety of contexts in their article. For instance, it is used to express the idea of a process underlying (e.g. Ibid., p. 408)30, causing (Ibid.), or contributing to a clinical phenomenon (e.g. Ibid., p. 410). However, in line with my suggestion to allow for a more abstract interpretation of “mechanism”, the authors’ use of “mechanism” does not necessarily entail that the mechanism reflects a physical or brain-based process. For example, the term is employed in the contexts of “generating options” to act on and “risk assessment” – i.e. at the phenomenological/behavioural level. Whilst mechanisms are mentioned, the authors are careful to note the many gaps in our understanding. No mechanisms (in the sense of a mechanistic explanation intended in Chapter 7) that bring forth the phenomenon akrasia or “decision making” are posited in detail, simply because none are known. Rather, the authors sketch the philosophical (conceptual) landscape and suggest how empirical findings may be pieced together and situated within this landscape, with a view to uncovering more detailed mechanistic explanations in the future. This exemplifies the interpretation of aiming for “mosaic unity” that I argued for earlier in this chapter.

The matrix in which motivation is decomposed into three parts and where phenomenological and neurobiological correlates are suggested could be a good starting point for a network analytic approach. By representing the findings, such as activation of the supplementary motor area, pre-frontal motor cortex and psychopathological phenomena such as impulsivity, loss of goal directedness, and impulsivity as nodes, a map can be made and (temporal) correlations expressed as lines between the nodes. Studies could then be designed to see whether activation of brain areas occurs in the same temporal order as the model suggests, and whether correlations between nodes are suggestive of a causal relationship. Network analysis can also be used to uncover which phenomena most frequently co-occur (cf. Schmittmann et al., 2013; Borsboom & Cramer, 2013).

30 Kalis et al. (2008) write: “we use the term apathy only as a description of psychopathological phenomena, not as a designation of a specific underlying mechanism” (p. 408, italics in original).
This suggested use of network analysis to uncover relations between phenomena fits a more abstract interpretation of the mechanism metaphor precisely because network analysis is neutral with regard to which entities are chosen as parts (or nodes). This makes it possible to combine entities and test for correlations in a way that transcends traditional levels hierarchies. The method does justice to the idea that parts and wholes can reciprocally affect each other. Although network analysis helps detect relations between entities, it cannot provide an account of the organisation of these nodes. This is where exploration from a mechanistic perspective comes in.

For instance, two nodes representing two entities that were thought to be separate entities, may turn out to be the same thing (i.e. they have an identity relation), the one may cause (activation of) the other, or, one may turn out to be a part of the other (i.e. they have a compositional relation). Network analysis cannot directly answer the question which of these three options is most accurate in a particular situation. Yet it can point us towards such situations and help direct our further research.

In sum, the adaptations of the mechanistic account that I suggested in Chapter 7 and the present chapter are aimed to make the mechanistic account more suitable for a systems theory approach and also for the field of psychiatry. Using an example, I showed that combining mechanistic explanation with the mathematic methods of network analysis can help us make intelligible a more abstract notion of mechanism where psychological entities can reciprocally influence each other and where parts can affect the whole and vice versa. Network analysis can help test whether proposed (how-possibly) entities belong within a mechanism (i.e. their inclusion in the mechanism contributes to the mechanistic explanation's ability to describe, predict and intervene in the target system). This section thus provided suggestions in answer to the third central question of this project: What are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy? In the following section, I return to the second central question that fuelled this project and address how the second central question in particular requires reformulation in order to make sense in light of the account of scientific explanation in psychiatry that I have proposed.

9.7 Reformulating the project's second central question

The second central question, which largely fuelled this project is: how can we understand the reference to the biological, psychological and social levels in the scientific literature on apathy and account for their coherence and possibilities for their integration? In this section, I first briefly sketch how I have gone about addressing the question in this thesis. This summary will serve to show how the question itself no longer makes sense
in light of my findings. Due to the manner in which my account has been developed, I reformulate and pose the central question in a post-hoc fashion. The reformulation of the question thus appears at the very end of this section – I have sought out a suitable reformulation of the question based on the account of explanation in psychiatry that I have developed.

In attempting to answer the question about integration of levels, one option was to choose a single preferred level, and then to explain all empirical findings in terms of that level. Given the arguments presented against the biomedical model in Chapter 3, it is clear that I reject this option (see also Bechtel, 1994). A second option was based on the biopsychosocial model, which, at a glance, seemed a valid alternative because it proposes taking into account biological, psychological and social aspects. However, the biopsychosocial model does not clearly articulate what kind of levels are at stake. Engel merely pledges for “consideration of psychological, social and cultural factors” (1977, p. 132). He continues: “establishing a relationship between particular biochemical processes and the clinical data of illness requires a scientifically rational approach to behavioural and psychosocial data” (Ibid.). Engel furthermore frequently states that on his model, we need to “take into account” the biological, psychological and social factors. He refers to the “psychobiological unity of man” (Ibid., p. 133) and “the struggle to reconcile the psychosocial and the biological in medicine” (Ibid., p. 134). On a very narrow interpretation of the second question, fuelled by Engel’s own words, the question may suggest that we are trying to take “biological findings” and integrate them with “psychological/social findings”. Another considerably more charitable interpretation is that it is not the findings themselves which need to be reconciled, but the perspectives and domains31 from which they stem. This is in keeping with Engel’s “psychobiological unity of man” and the need to reconcile the things that a scientific approach has taken apart. The troubles in differentiating between these different interpretations can be put down to Engel’s own lack of differentiation between levels of ontology and levels of epistemology (or “levels of nature” and “levels of science” as Engel calls them) in his presentation of the biopsychosocial thinking. Even more problematic is that the biopsychosocial model provided no guidance as to how we should integrate (the findings from) these different domains. The central question thus remained unanswered.

I went back to the drawing board and proceeded by reviewing the notion of “levels” itself. Craver’s (2007) taxonomy of levels served as a starting point, because it shows

---

31 Note that this formulation may misleadingly propagate a reification of the notion of “domain”, as an entity from which “findings can stem”. This suggestion goes against the grain of my account, in which such domains are mere heuristic tools.
that there are many possible interpretations of the levels metaphor. The ontic-epistemic divide was a central feature in Craver’s taxonomy of levels (and in the New Mechanist literature in general). I found it to be a source of confusion and took issue with adherence to an absolute ontic-epistemic divide. An enactive approach, according to which scientific experience arises from the inescapable self-world reciprocity, was shown to be a promising alternative. This is because in practice, the process of defining what belongs to a particular level of mechanism is informed by both nature itself and the explanatory task that the observer is seeking to answer. Both the levels metaphor and reference to an ontic-epistemic divide were accordingly reclassified as heuristic tools that aid our discussion of topics without necessarily corresponding to absolute categories in the world.

The mechanistic approach to explanation emerged as a model that may provide a means to tackle the second central question of this project. The decision to regard levels and the ontic-epistemic divide as heuristic aids, affects the phrasing of the second central question, which in its current state, pertains to interlevel integration. The account has evolved so that integration of levels is no longer the primary focus. Instead, on an enactive view, the integration of perspectives has become the crux.

Besides clarifying what it is that we are trying to integrate in the second central question, it is also important to explore the notion of “integration” itself. To integrate is to bring together, to combine and make into a whole. This is open to multiple interpretations, some of which are at odds with my intended goal. Two easily dismissed possible interpretations of the project of integration are that the items that are to be integrated must be shown to either be the same or, instead, be reducible to one another. I wish to emphasize that I do not argue for an identity of perspectives. I concede that there is a single target phenomenon, that may look different from different perspectives. Thus the yields of (scientific) investigations of a specific phenomenon from different angles are different and I have no intention of integrating them in the sense of showing that the perspectival yields themselves are one and the same thing. Regarding the second interpretation, it should be clear, given the discussion of problems of traditional reductionism that this is not the interpretation of “integration” intended here either.

The sense of integration intended instead involves a bringing together of perspectives so that we can understand what happens at the seams between perspectives to form a coherent account of the whole. As discussed, this is similar to Craver’s suggestion of mosaic unity, only on Craver’s use of the metaphor, the tiles are pieces of empirical data, whereas on my account, the tiles would be “perspectives” on aspects of the phenomenon under study instead. I prefer the focus on “perspectives” because it more accurately
Perspectival Mosaic Unity

reflects what I take to be the challenge at hand. This challenge involves understanding how different perspectives can usefully be strung together and fit into the scientific landscape, rather than regarding the challenge as fitting together empirical data.

The focus on the perspectival includes paying attention to why a phenomenon may look different from different perspectives. A form of unity may be achieved by fitting together different perspectives on either a single phenomenon or, instead, by fitting together perspectives on various closely related phenomena. Also, each perspective is connected to an agent who has his/her own way of observing and (scientifically) interrogating the phenomenon under study. The meaning of the phenomenon is inherent to the observer’s perspective which is in turn influenced by his/her relation to the target phenomenon and his/her role in examining it. As an example of the former, a geneticist, an fMRI expert, a clinical psychiatrist, a psychologist and a family member will all have very different perspectives on an individual with negative symptoms. The idea of perspectival integration is that these accounts can exist side by side and that they can be placed together to achieve the best fit (i.e. forming a mosaic unity that has more explanatory power than the separate perspectives). An example of the latter form of perspectival integration is that the negative symptom cluster can be dissected into separate, related but roughly delineated “parts” of the “target system”. Then, perspectives on the separate symptoms, which are likely to partially overlap, are again arranged together to form a larger “picture” of the target.

In the examples of perspectival integration\(^{32}\) outlined above understanding “why” a phenomenon may look different from different perspectives can guide how best to reconcile\(^ {33}\) or integrate these perspectives. This is important, because different perspectives can sometimes yield (seemingly) conflicting data that in turn lead to the formulation of conflicting models and theories for the explanation of a specific phenomenon. Another axis, along which different perspectives (or, in this case, roles of the observer) lead to different findings, follows from Glas’ (1991) differentiation between the philosophical, scientific, clinical and everyday levels of analysis.

\(^{32}\) On my interpretation, “integration” takes on a weaker, non-reductionist, form, where integration occurs through creation of perspectival mosaic unity.

\(^{33}\) The term “reconcile”, is interesting in this context because “re-” suggests a “going back”, a “doing again”; “conciliare” stems from the Latin “to bring together”. The use of this word in the context of scientific explanation highlights the notion that the world in our experience is unified (i.e. it has no preconceived single set of joints to speak in the terms of the second assumption discussed above), and that in scientific experience, we “pull apart” a phenomenon which, for a more complete understanding must then be put back together again.
In light of the turn in focus in this project to an enactivism-inspired approach to mechanistic explanation, the original phrasing of the second central question has been rendered inadequate. I instead reformulate the question having provided an answer in the process of arguing for its reformulation: How can we account for the integration of different perspectives on a target phenomenon?

9.8 Answering the central questions that fuelled this project

In Chapter 5, Section 8, I discussed my answer to the first central question. Here, I argued that the biopsychosocial model is not sufficiently developed to work as a philosophical or scientific model because, despite making a compelling case against reductionism, the model offers no guidance on how to integrate the biological, psychological and social levels. As a clinical model, it works primarily as a reminder to clinicians to take into account various life domains of a patient rather than limiting themselves to biological findings.

The answer to the second question in its reformulated form can be summarized as follows: any philosophical model for scientific explanation must take into account the self-world reciprocity and our scientific experience. The practice of science is therefore inescapably perspectival. In order to make sense of all these different perspectives on the phenomena we study, appeal to network analysis and mechanistic explanation have been suggested. Network analysis is invaluable because, as a mathematical method, it is neutral with respect to the content it is used to study. It can therefore help point us in the direction of interesting relations between represented items based on statistical analysis. Relations that appear relevant can then be further explored through the lens of a mechanistic approach. The mechanistic approach, in which parts and their interactions, and the part-whole relationship are central, does justice to the complexity of the phenomena under study in psychiatry. My response to the second question is that we can draw on an adapted version of mechanistic explanation supplemented with the methods of network analysis to yield new perspectives and also to help fit together different perspectives. Together, these perspectives form a varied landscape, or a “patchwork”. There may not be a single way to best fit together the perspectives. The best “patchwork” of perspectives is simply the patchwork that is most fruitful to the scientific tasks of prediction, description and intervention pertaining to the phenomenon under study and the explanatory task at hand. Due to my agnosticism regarding realist claims, I thus anchor my normative constraints epistemically and pragmatically rather than, as Craver (2007) does, in the ontic domain. Besides being pragmatic, the rough sketch of an account of explanation in the science of psychiatry that follows from the second central question is also contextually sensitive and relativistic.
The third central question of this thesis is: what are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy? I note that it is difficult to offer concrete implications for clinical practice of the use of the account of scientific explanation outlined in this chapter. My tentative response follows from my answer to the second question plus my expansion of the example based on Kalis et al. (2008) outlined above. In a nutshell, I have called for avoidance of traditional reductionism in both clinical and scientific practice. I have instead suggested methods that are more philosophically sound and that are better equipped to deal with the complexity inherent to psychiatric phenomena. These methods include taking the mechanistic approach one enactivism-inspired step further: by applying the attention to the importance of part-to-whole relations to the self-world relation too. I adopted an agnostic stance with regard to matters metaphysical and emphasized that the levels metaphor and appeal to an ontic-epistemic spectrum are heuristic tools. These adaptations called for a perspectival interpretation of the need for mosaic unity. This is a significant departure from Craver’s (2007) use of the phrase “mosaic unity”. I suggested that fitting different perspectives on a single target phenomenon or on related phenomena together in the scientific landscape (with attention to building on firm philosophical foundations) is an important task for scientists. This process yields the perspectival mosaic unity that can help further our understanding and improve clinical practice for those suffering from psychiatric symptoms.

As noted throughout this thesis, there are important differences between the tasks of scientists and clinicians. Scientific practice of psychiatry focuses on tasks such as defining and explaining phenomena, investigating which phenomena may be subsumed under a particular explanation, and on using statistical methods applied to (preferably) large data-sets to detect correlations between relevant factors. Clinicians can be regarded as translators: they attempt to usefully apply the fruits of scientific research to the individual patient. To do so, they must integrate clinical findings from the individual with their knowledge of larger scale patterns and explanations from scientific research with a view to classifying the phenomenon, determining prognosis and suggesting and implementing the most effective possible interventions.

These differences mean that the implications of the answer to the second central question are different for clinical practice and scientific practice. For scientific practice, I have suggested using an adapted version of mechanistic thinking and employing the methods of network analysis with a view to finding explanations that are “good” based on their pragmatic value. I have stated that scientific practice may benefit from becoming more sensitive to the philosophical assumptions (for instance regarding reductionism, mind-body dualism and part-whole relations) on which study design and interpretation are
often based. The findings from philosophically sensitive research can then in turn be used by clinicians in their approach to the individual patient. I would be very excited to see whether such an approach can indeed help speed up scientific progress and ultimately help reduce suffering due to psychiatric symptoms.

Another, more direct clinical implication of my advocacy of network analysis as a method in psychiatry is that it can be used to provide personalised treatments for individual patients. For a single individual, a network representation of symptoms and their relations can be used to determine how to intervene. A fictional example: Tom suffers from psychotic symptoms, insomnia, and is currently having difficulties paying his rent. We can intervene on each of these symptoms by offering antipsychotic medication, improving sleep hygiene and helping him sort out his financial problems respectively. Using a network representation can help us decide to first tackle his financial problems because the network representation in his case shows that the financial problems increase his insomnia and that insomnia in turn leads to more psychotic symptoms. Tackling his financial problems rather than prescribing antipsychotic medication (with its grave side-effects) may prove more effective in reducing his suffering.

9.9 Conclusion

In this thesis, I set out to explore the current use and applicability of various models for scientific explanation in the field of psychiatry in order to find answers to three central questions concerning 1) the utility of the biopsychosocial model which is taught to medical students around the world 2) the possibilities for perspectival integration in the scientific practice of psychiatry and 3) the implications of improved perspectival integration for the scientific and clinical domains of psychiatry.

Chapter 2 commenced in a bottom-up fashion, with a cross-section of the scientific literature on apathy in schizophrenia as a case study of the current state of affairs within the field. I began by zooming in on the two most common models of explanation in medicine, the biomedical and biopsychosocial model that are both a standard part of the medical curriculum. In Chapters 3 and 4, more in-depth discussions of these models were offered. In particular, I took issue with the reductionism to the biological level that is inherent in the biomedical model. I showed that, although the biopsychosocial model purports to avoid the reductionism of the biomedical model and also emphasizes the importance of a holistic approach, it is incomplete in various ways. In Chapter 5, I responded to the first of the three central questions fuelling this project and discussed the ways in which the biopsychosocial model is incomplete. Recall that the question was “what is the utility of the biopsychosocial model at the level of a) philosophy, b) science
and c) clinical practice?"

Regarding the first sub-question, I proposed that the biopsychosocial model is underdeveloped as a philosophical model because although Engel makes an excellent case against traditional reductionism, he does not sufficiently specify how the biological, psychological and social aspects are to be brought together through a systems theory approach. Second, the lack of a solid philosophical infrastructure may be linked to the problems with regarding the biopsychosocial model as a scientific model. It does not offer a scientific method that specifies how researchers are to integrate or reconcile information about the biological, psychological and social domains. Finally, Engel’s description of the model offers examples of how a doctor may pay attention to aspects from each of the domains when examining and talking to his patient. I therefore concluded that the value of the biopsychosocial model pertains mainly to the clinical domain, where it serves as a powerful reminder of the importance of a holistic approach, warning clinicians not to focus on the biological level only.

Chapter 6 offered an overview of more abstract, philosophical models for scientific explanation, given the gaps in the philosophical infrastructure of the biopsychosocial model. I directed my attention to the Covering Law model, Unification model and Representation model, demonstrating that they are of little help for the reconciliation of the domains implied in the biopsychosocial model. In the chapter, I adopted Craver’s (2007) constraints on explanation and turned to an approach that is more in line with systems theory: mechanistic explanation. Chapter 7 delved into the ins and outs of this approach, where both its promise and disadvantages for the field of psychiatry were discussed. Next, in Chapter 8, I moved on to network analysis, which is somewhat of an outlier because it is not an explanatory model but rather a mathematical methodology and provided a sketch of the possibilities it offers for identifying relationships between observations or phenomena.

In this chapter, I developed the findings from the previous chapters. I outlined further adaptations (building upon my critique of mechanistic explanation in Chapter 7) to the account of mechanistic explanation that I suggest render the model more suitable for scientific practice in psychiatry. I then put forward my own rough model, that builds on perspectival integration rather than integration of data of levels, to tentatively answer the reformulated second and the third central questions of this thesis.
E
Epilogue
Epilogue

Now we have come full-circle, it is time for a brief reflection on what can best be described metaphorically as a journey. My travels across lands scientific and philosophical took me past unexpected intermezzos. It was challenging to stay on route given all the enticing side-turns that invited further exploration. Although the reasons for asking the questions that fuelled this project remain the same, the questions themselves and my terminology have evolved during my research and writing. In these concluding remarks, I touch upon the question of how this journey has changed my approach to clinical practice.

I embarked on writing this book during my first years as a newly fledged psychiatrist. Three days of the working week were spent setting-up a so-called “F-ACT” (flexible assertive community treatment) team. My team’s goal is to offer top-notch mental health care to out-patients with severe mental illnesses. The challenges were myriad and included navigating the politics of a large mental health institute, implementing FACT principles in our team of seasoned professionals, and of course: achieving constructive working relationships with all of the patients in our care. The other two days were spent recovering from the above and immersing myself in piles of articles and mountains of books in an attempt to answer the questions driving this thesis.

Although this thesis is primarily a philosophical enterprise, it was also shaped by questions stemming from clinical practice. My choice of the vague and multi-facetted topic “apathy in schizophrenia” for the literature study is the most obvious example of this. The difficulties began when I tried to find a clear definition of apathy from which to depart. This proved impossible: few of the articles even offered a definition, and when they did, the definitions from different articles were incompatible. At first, I lamented my choice of topic, thinking that I should have chosen a “cleaner” more clearly circumscribed and better understood phenomenon (much as Craver (2007) had done with his example of neuronal firing). However, switching to a psychiatric phenomenon that is well-understood and clearly circumscribed was not really an option for lack of such a phenomenon in our field. This led me to realise that in fact the situation for apathy is quite representative and that this was a good reason to persist with it.

I commenced with my analysis of the literature and found remarkably few traces of the models that medical students are taught about. I was both disturbed and reassured by the almost complete dearth of explicit epistemological considerations in the literature. I was disturbed because I cannot imagine how we can do good research unless we explicitly take into account the underlying assumptions regarding hypothesis formulation, methodology and data interpretation. Yet I was also reassured, because the
literature proves that one can secure research grants, be a renowned medical scientist and contribute to the field nonetheless.

Looking back, I think writing this thesis has subtly yet deeply affected how I am as a clinician. I have of course learnt much about apathy. I have found it especially useful to expand my toolkit for interviewing people suffering from apathy in order to explore its various facets such as anhedonia, amotivation, and existential consequences of dealing with apathy, especially in those patients who suffer from it directly. I have noticed that there are huge differences in the way people experience negative symptoms: some patients feel incredibly weighed down because they want to do things and simply can’t bring themselves to do so or because they no longer enjoy things to the degree they did previously. They sometimes feel drawn to excessive use of caffeine or drugs. Others however, are indifferent to their own negative symptoms. The fact that they spend most of their time in bed or on the sofa and no longer do the things they did before falling ill, does not bother them. They don’t feel bored, or frustrated, but often suffer indirectly, because they fail to meet social norms (showering regularly, keeping their house clean, staying in touch with family) or don’t do the things that society requires of them (paying bills, filling in tax forms, caring for their children). I have found it useful to differentiate between these different experiences of negative symptoms in order to tailor the treatment I offer to their individual needs.

On another level, this project has attuned me to how much the metaphors and analogies we use shape the way we understand phenomena. Often, they are filler-terms that serve to gloss over the gaps in our understanding. Often, they are used somewhat carelessly and risk carrying implications that are not accurate or are unintended. An example that is common but also carries important implications is that of “underlying”. A process or entity is often said to “underlie” another. What this means is rarely articulated and seems to depend very much on the context. The term is common in the literature, but also in everyday parlance. Scientists, clinicians and patients alike use it. Another particularly confusing example is when causal powers are attributed to a mental disorder. For instance: “I can’t concentrate because of my ADHD”. Such a statement carries far-reaching and disputable assumptions about what ADHD is, and is circular if having concentration problems is one of the criteria for receiving the diagnosis ADHD in the first place.

My response to this heightened sensitivity to the nuances of language is that I have become more aware of my own use of filler-terms and have also been more able to question the assumptions (or gaps in my understanding) leading to my employment of them. I feel better equipped to pay attention to the context in which others use filler-
Epilogue

terms and also explore these gaps. It has sparked interesting discussions in my team and ultimately led to the use of different interventions. In the case of a patient who says “I can't concentrate because of my ADHD”, we can then explore what he means and how his concentration issues affect him. We can invite patients to take on a more active stance towards their symptoms, empowering them. This is more constructive than accepting that they “blame” their problems on some abstract disease entity.

In the introduction, I noted that dabbling in philosophy helped me develop a tolerance for the aporia that the field of psychiatry induces. My research led me to accept that many of the phenomena we see are poorly defined, multi-factorial, and unfathomably dynamic. For clinical practice, I now find myself trying to develop something akin to a (very rough) explanatory model for each individual patient. In it, I attempt to fit symptoms and protective properties, relevant contextual factors (both positive and negative) and effects of interventions. The model is constantly adjusted to new findings and the patient is invited to participate in this adaptation. My philosophical research has armed me with arguments in favour of taking analytic and holistic approaches in the development of these individual explanatory models. The analytic perspective serves to detect parts (in an abstract interpretation of the term), their internal workings and their interactions. The holistic perspective takes the whole within its context into account. The importance of a descriptive diagnosis rather than a mere summary of DSM classifications is of course glaringly obvious. I venture that mechanistic thinking and perspectival integration contribute to the rather more challenging project of formulating such a descriptive diagnosis. The descriptive diagnosis in turn opens the way to determining more accurately which interventions I expect to have the best effects. For instance, for my young female patient with anger-management issues, understanding that her anger seems to spring from a combination of very low self-esteem, inability to express her feelings and severe anxiety, helps me to direct our attention to intervening on her low self-esteem, expression of emotions and anxiety rather than prescribing an anger-management course.

In the previous chapter, I called for “perspectival mosaic unity” to emphasize the combination of different yet related perspectives in order to create a dynamic yet coherent scientific landscape. Based on the way my own approach to the individual patient has changed due to writing this thesis, I add that the aim of achieving perspectival mosaic unity holds for clinical practice too.
Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Hughlings Jackson, J. (1883). On some implications of dissolution of the nervous system, 36, 64–66.


Bibliography


Bibliography

69–83.


Stuss, D. T., Van Reekum, R., & Murphy, K. J. (2000). Differentiation of states and causes of apathy. In
Bibliography


Bibliography
Summary
SUMMARY

The eccentric manners of explanatory models: towards an account of perspectival mosaic unity in psychiatry

In this thesis, I advocate a pragmatic, relativistic, post-dichotomous and enactive version of mechanistic explanation in psychiatry. I also advocate augmenting mechanistic explanations with the methods from network analysis. Use of network analysis combined with a mechanistic approach can aid us in detection of the parts and help uncover the interactions that together comprise psychiatric phenomena. The account, which is outlined rather roughly, follows in response to the central question of how we can account for the integration of different perspectives on a target phenomenon in psychiatry.

Below, I outline the central questions which I set out to address and the contents of the chapters that follow. As I discuss in the Chapter 9, my findings during the project lead me to reformulate the second of the central questions. The central questions with which I began, and to which I refer in the chapters to come, are:

I.) What is the utility of the biopsychosocial model at the level of a) philosophy, b) science and c) clinical practice?

II.) How can we understand the use of the biological, psychological and social levels of explanation in science and account for the coherence and possibilities for integration of these different levels of explanation?

III.) What are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy?

Chapter 2: in order to stay as close to psychiatric practice as possible, I survey a literature sample. Because a large proportion of the patients I encounter meet the diagnostic criteria for schizophrenia and because the negative symptoms are a huge source of misery, the literature sample zooms in on apathy in schizophrenia. I offer short introductions of schizophrenia, the negative symptom cluster and apathy in particular before outlining my search strategy. Ten articles that meet my search criteria are then briefly summarized with a view to detecting possible philosophical assumptions at play and evidence of biomedical or biopsychosocial thinking. I show how evidence of specific medical explanatory models is indirect and meagre at best.

In Chapter 3, the biomedical model is introduced. It is a model with origins in the 19th century and with even earlier roots that date back to Cartesian thinking. The
biomedical model’s main limitation is its adherence to reductionist methods: the model seeks to explain diseases solely in terms of pathophysiological mechanisms and thus neglects other facets such as interpersonal and psychological aspects. The problems of reductionism are the theme of the remainder of the chapter. Different contexts in which the term features are examined and various types of reductionism are distinguished. I advocate continuing the search for a model that avoids reductionism and thus does not seek to explain by reducing all phenomena to occurrences at a particular level of organisation such as the biological level.

Chapter 4 introduces the biopsychosocial model. The model was developed by Engel and first presented in his 1977 article. Engel expands the model in many of his later publications and this literature is examined. I discuss how the biopsychosocial model aims to avoid the problems of biomedical thinking. The model is inspired by the doctrine of dynamical systems thinking and Engel himself refers to the writings of von Bertalanffy. I introduce this doctrine and some different variants of systems theory. The degree to which systems theory is incorporated in the biopsychosocial model to help integrate the different domains turns out to be somewhat limited. The chapter concludes with a short comparison of the biomedical and biopsychosocial models.

The fifth chapter delves into the shortcomings of the biopsychosocial model, which, as I note is the model that is commonly taught to medical students. The first line of critique is that Engel does not differentiate between clinical and scientific practice – in fact, he deems this a false dichotomy. I instead embrace the notion of different roles and perspectives and show that differentiating between them is a heuristic tool that can help avoid confusion. On Engel’s account, it is unclear whether the model is intended as a clinical or scientific model. A second objection is that Engel alludes to, but does not further develop, the ideas of systems theory in the biopsychosocial model. He offers little guidance on how to bring together the biological, psychological and social domains. Third, the model fails to help provide a definition of what disease is in the first place and risks inviting an over-inclusive interpretation of “disease”. Finally, I address the question of whether the biopsychosocial model should not include other dimensions, such as an existential dimension. In the concluding section, I address the first central question of the thesis, proposing that the biopsychosocial model is best regarded as a guide for clinical practice rather than a scientific model.

Having discussed the two most common medical models of explanation, Chapter 6 turns to more abstract, philosophical models of explanation. Key facets of explanation, such as description, prediction and intervention are discussed. I borrow heavily from the philosophers Wesley Salmon and Carl Craver by adopting their five constraints
Summary

on a good explanation. Put simply, the five constraints are aimed at detecting causal relevance. Armed with these constraints, I discuss three philosophical models of scientific explanation: the Covering Law model, the Representation model and the Unification model. I offer examples from the field of psychiatry and show how these models fail to meet the five constraints. There are of course many more models but I focus on these three because a more exhaustive treatment is beyond the scope of my project. In the following two chapters, I discuss two further models in greater detail as they hold more promise for the field of psychiatry: mechanistic explanation and a statistical method called network analysis.

Chapter 7 offers an in depth exploration of mechanistic explanation. The chapter zooms in on the notion of levels, their appearance in the psychiatric literature and also on Craver’s idea of “levels of mechanism”. Furthermore, I investigate what a “mechanism” is, showing that it consists of “parts” or “entities” and their interactions. The accounts of different mechanistic thinkers (e.g. Glennan, Machamer, Darden, Bechtel, Abrahamsen, and Richardson) are compared and the consequences of the differences are discussed for the field of psychiatry specifically. I revisit the apathy literature sample, searching for evidence of mechanistic thinking, and provide examples of proto-mechanisms. I take issue with some inconsistencies within Craver’s account and also with the aspects of the accounts of mechanistic explanation that hinder the model’s applicability in the field of psychiatry. The most important problem for mechanistic explanation is the lack of robustly detectable, physiologically plausible parts in mechanisms for psychiatric phenomena. I suggest that there is an inherent fuzziness in psychiatry.

Network analysis, the topic of Chapter 8, is a statistical approach that entails using mathematical analyses to identify relations between phenomena. However, before any data can be represented in terms of a network, it must be collected. Psychometrics is the field that studies how mental phenomena may be measured. I argue for the importance of sound psychometrics, highlighting downfalls such as statisticism, the measurement problem and the idea of the latent variable. Next, three types of network that are most commonly used to model psychiatric phenomena are distinguished. In order to determine whether network analysis can be regarded as a model of scientific explanation, I investigate if the three types of network can meet the constraints for a sound explanation introduced in Chapter 6. I conclude that none of them can fully meet the requirements imposed by Craver and adopted by me. I include a discussion on Mechanistic Property Clusters, a hybrid between mechanistic thinking and statistical approaches. I then proceed to further develop the idea of a hybrid model for explanation that draws upon both Mechanistic Explanation and Network Analysis to explain a psychiatric phenomenon.
Chapter 9, titled “Perspectival Mosaic Unity” serves to bring together the findings from the previous chapters and attempts to answer the remaining questions fuelling this project. I begin by offering an enactivism-inspired set of basic assumptions upon which I base my own attempt to offer a hybrid account of explanation that is suitable for psychiatry. Many dichotomies that are pervasive in traditional philosophy and also in scientific practice, are afforded the status of heuristic tools rather than absolute categories (for instance the ontic-epistemic divide and the subject-object divide). I argue that the perspectival approach, based on the inescapable reciprocity between self and world at which I thus arrive, runs deeper than the dichotomy-dependent notion embraced by Craver. I finally return to the central questions of the project. Given my findings, I adapt the phrasing of the second central question: How can we account for the integration of different perspectives on a target phenomenon? In answer to the question, I explore the idea of perspectival integration as a means to develop a more unified scientific landscape.

I then move on to the third central question: what are the implications of the answers to the above questions for the scientific explanation and clinical understanding of the phenomenon of apathy? I draw on an example from the scientific literature to illustrate the implications of my refutal of biomedical and biopsychosocial thinking. I instead offer a post-dichotomous account of mechanistic explanation and supplement it with network analysis which is more suitable for the thorny issues presented by the field of psychiatry. In this account, I suggest that the use of statistical methods can help determine which parts are relevant to a mechanism in the first place by using statistical methods to identify possible causal relations.

Finally, I conclude the last chapter and my thesis with reflections on how my findings have affected my clinical practice and suggest that an adapted account of mechanistic thinking can be of help in formulating a descriptive diagnosis and making individualised treatment plans with patients.
Acknowledgements
Acknowledgments

I would like to take this opportunity to thank the many people that supported me during my research and writing endeavours.

First I thank Gerrit Glas. He has played a great role in moving me both to become a psychiatrist and to undertake and complete my dissertation. Over the years, we have talked extensively about philosophy and psychiatry (and of course about other topics such as cars, houses and the human condition). Our meetings usually took place at Gerrit’s home, but sometimes impromptu in the car whilst racing along to get to some other event on time. I look back on these meetings with great fondness. I am also very grateful to Henk de Regt. He meticulously worked his way through various versions of the book (at truly impressive speed!) and offered fantastic feedback during our meetings at the VU.

Although we only met recently, I wish to thank Lena Kästner for our (skype) conversations about philosophy and psychiatry. The analogy of “breeding transparent fish” brings a smile to my face every time. I thank Denny Borsboom and Carl Craver for their helpful feedback. I thank Derek Strijbos for his encouragement and guidance in navigating the process of writing a dissertation.

I thank Bram Bovendeerd for our Monday morning mentoring sessions. His insightful observations and reflections about an unfathomably rich array of topics (i.e. not just pertaining to the process of writing a dissertation or of improving care for people suffering from autism, but also including the virtues of teaching martial arts to children, the politics of working in a large institution, and how to deal with annoying adolescent cyclists), have been heartening and important to me.

I am grateful to have had the opportunity to take part in the Science Beyond Scientism project at the VU. Collaborating with a such a diverse team of talented philosophers for a year and getting a taste of life in academia was an enriching experience.

I am of course deeply indebted to my partner Guus, especially for his emotional support. He has helped me endure my frequent episodes of Imposter Syndrome and has put much effort into reassuring me that the defence ceremony is in fact not the end of the world. Our talks about network analysis, artificial intelligence and causation have been very inspiring.

I much appreciate my parents’ unconditional backing. They maintained the ability to
inquire how the writing process was progressing despite my often unforthcoming or utterly boring answers. And of course I am very happy that my mum designed the cover for the book.

Special mention of my colleagues from the FACT teams of Zwolle is warranted. They managed to keep things at work afloat despite my patchy availability (due to spending two days a week on my PhD). Without these wonderful teams and their endless humour and problem-solving skills, I would not have been able to write this book. I also thank the people that were entrusted into my care: they have, often unknowingly, inspired me every way imaginable.

I also want to thank my friends Marianne, Krista, Hilde for providing moral support, for being interested and for facilitating many desperately necessary (and often horse-related) procrastination sessions. Ditta, with her penchant for high-paced, thematically dense and ultra-critical conversation, preferably against the backdrop of a local fine-dining establishment, is incredibly dear to me. Frederique, I find your incredible energy and awe-inspiring openness to life, your understanding of matters spiritual and your knack for humorous descriptions of mundane situations exhilarating.

Finally, I would like to acknowledge that this project would have been impossible without the financial aid from Dimence.
Curriculum Vitae
Curriculum Vitae

Josephine Lenssen is a Netherlands-based psychiatrist and philosopher. She was born in 1985 in Aberdeen, Scotland. In 2003 she graduated from the European School of Bergen, Netherlands. She holds MA degrees in philosophy (2009, cum laude) and medicine (2010) from Leiden University. She completed her psychiatric training in 2015 (with Dimence Mental Health Institute) and now leads a multi-disciplinary team that specializes in the treatment of people suffering from severe mental illnesses according to principles such as shared-decision-making and rehabilitation, remotivation and resocialization. Her research has always been at the intersection between philosophy and psychiatry. In her research, and also in this dissertation, she draws on her experience as a psychiatrist and focuses on adaptation of mechanistic explanation and network analytic approaches to make them more suitable for use in the clinical and scientific practices of psychiatry and defends a notion of perspectival integration.