Theories of mental disorders remain scientific in spite of both the absence of reductive explanations and the presence of interventional mental autonomy

By
Katrina Faith Winzeler

A dissertation submitted in partial satisfaction of the requirements for the degree of
Doctor of Philosophy
in
Philosophy
in the
Graduate Division
of the
University of California, Berkeley

Committee in Charge:
Professor John Searle, Co-chair
Professor John Campbell, Co-chair
Professor Tania Lombrozo

Summer 2015
Abstract

Theories of mental disorders remain scientific in spite of both the absence of reductive explanations and the presence of interventional mental autonomy

by

Katrina Faith Winzeler

Doctor of Philosophy in Philosophy

University of California, Berkeley

Professor John Searle and Professor John Campbell, Co-Chairs

This dissertation asks the question: what is the nature of mental disorders? I explore 3 lines of inquiry that are opened by this question. The first has to do with whether mental disorders can be explained by the natural sciences. I argue that mental disorders are indeed part of the subject matter of the natural sciences, but not because there are reductions between the kinds of abnormal psychology and the kinds of neuroscience. Rather, mental disorders fall under the purview of the natural sciences because 1) they have identifiable, observer-independent causal structures, and 2) their explanations crucially involve the brain. However, these explanations rely on more than just brain anatomy, which is why we need to go beyond the current scientific trend of claiming that mental disorders are just faulty brain circuits. Instead, we should view mental disorders as brain dysfunctions.

The second line of inquiry concerns the grounding of mental disorders as bad things. Where does the normativity come from? Does it require human evaluation or judgment? Since I defend a theory on which mental disorders are brain dysfunctions, I claim that their badness comes from their being mechanisms that have failed to work as they should. The mechanisms at hand have evolutionary histories and therefore their functions should not be understood with regard to their “current causal roles.” I explore Jerome Wakefield’s “Harmful Dysfunction” theory as an example of a dysfunction-based theory of mental disorders, and I take a critical stance towards the notion of ‘function’ that his theory employs. I then outline a new notion of ‘function,’ one that I call a “genetically-open function,” which I believe can avoid the problems raised against the original notion. This new notion of functions links the dysfunctions that are relevant to mental disorders to losses of flexibility and plasticity in response to changing environmental demands. This notion is meant to be suggestive and exploratory.

The third line of inquiry has to do with the causal structures of mental disorders. How do the various etiological factors and symptoms of a type of mental disorder relate to one another? I claim that mental disorders are neither non-natural kinds (like family resemblances or syndromes) nor essentialist natural kinds. Rather, I argue that mental disorders are homeostatic property cluster kinds, collections of symptoms that make one another’s co-occurrence more likely, and whose collective clustering is held in place via some mechanism. This mechanism is
discoverable *a posteriori*, so these kinds are observer-independent. Because their causal structures exist out in the world, they are natural kinds, even though they fail to have traditional essences. This naturalness bolsters the argument in favor of mental disorders being proper subject matter for the natural sciences.

Once I explore the nature of mental disorders via these three routes, I investigate the interventions taken to treat them. If mental disorders are brain dysfunctions, and hence can be studied by neuroscience and cognitive science, is there any reason not to think that mental forms of treatment (like psychotherapy) will become obsolete with the improvement of science? I argue that yes, psychotherapy is an ineliminable form of treatment. I claim that recovery does not supervene on the internal state of one’s brain, but rather on that state *plus* its causal history. The results of psychotherapy are path-dependent and experiential, and cannot be duplicated by a purely medical/physical treatment process. This claim says nothing against materialism and a scientific account of mental disorders. But it does suggest that even in a fully materialist view of mental disorders, there is some autonomy for the mental, at least with regard to interventions.

My overarching goal in this project is thus to demonstrate the false dichotomy between thinking that mental disorders either must fall outside the scope of science entirely, or else must be reduced to fundamental biological entities. We can argue that mental disorders are natural kinds with observer-independent causal structures, and yet still claim that 1) some of that objective causal structure involves contributions from irreducibly mental or environmental variables, and that 2) there are autonomous mental interventions for recovery.
Acknowledgments

First and foremost, I would like to thank my two dissertation co-chairs, Professor John Searle and Professor John Campbell, for their helpful, challenging, and continual feedback on my work as it took form. This dissertation would not exist without them, and I am very thankful for their minds and their time that they generously offered me. I also would like to thank the following people: Professor Sherrilyn Roush, for her intellectual feedback and encouragement during the years that she served as my adviser before taking up a post at Kings College London; Professor Geoffrey Lee, who was also helpful on an earlier incarnation of my committee; Professor Tania Lombrozo of the Psychology Department, for being my outside reader; Professor Branden Fitelson of Rutgers University, who served on my qualifying exam committee; and Dr. Jennifer Hudin for intellectual and emotional support.

I would like to thank Dave Lynaugh and Janet Groome for above-and-beyond-the-call-of-duty administrative support, more times than I could possibly count.

In addition, I would like to thank my colleagues in the Philosophy Department at UC Berkeley – both undergraduate and graduate – for their contributions to my morale and for their comments on various presentations of mine. The same goes for the audience at 2014’s “Early Career Scholars conference” at the Center for the Philosophy of Science at the University of Pittsburgh.

Finally, I want to thank my family – Abby, Patricia, John, and Josh.
Table of Contents

Introduction: “Ontological Reduction” and “Explanation by the natural sciences” are different methodologies; the latter does not require the former

(i) ‘Reality’ is demystified by the notion of ‘Validity’
(ii) So, what are mental disorders?
(iii) Roadmap

Chapter 1: Medical/Disease Models of mental disorder offer naturalistic, scientific explanations of mental disorders without requiring reduction

1. Introduction: has the brain usurped discussions of mental disorders?
2. Insel: mental disorders involve faulty brain circuits
   2.1 The difficulty of constructing an anatomical definition of ‘brain circuit’
   2.2 Conclusion: an acceptable notion of ‘brain circuits’ requires the notion of functions
   2.3 Dysfunctional brain circuits and the role of neurobiology
   2.4 Classification
   2.5 Examples of empirical hypotheses on which mental disorders are faulty brain circuits
      2.51 Example 1: OCD
      2.52 Example 2: Clinical depression
3. Murphy’s claim: psychiatry is cognitive neuroscience
   3.1 Cognitive science and levels of explanation
   3.2 Functions and dysfunctions in cognitive science
4. Insel and Murphy unified: a moderate Medical Model of mental disorders
   4.1 A Medical Model allows for multi-level explanations
   4.2 A Medical Model does not require fundamental explanations
5. Objections to the Medical Model
   5.1 Objection 1: The Medical Model does not respect the heterogeneity of etiology and symptoms; its explanations are either incomplete or inadequate
   5.2 Objection 2: The Medical Model fails to correctly capture the causal structure of mental disorders
   5.3 Objections 3: The Medical Model doesn’t respect the uniquely mental ontology of mental disorders
   5.4 Objection 4: The Medical Model is too inclusive
6. What are the explanatory alternatives to a Medical Model?

Introduction to Chapters 2-4: There are two interpretation of the question “what is the nature of mental disorders?”

(i) Structure is not evaluative
(ii) Evaluation
(iii) Structure
(iv) A specific answer to the “Disorder” interpretation
Chapter 2 Traditional dimensions in the Disorder Interpretation of the ontological question: “What is the nature of mental disorders?” Making a case for the Harmful Dysfunction Theory

1. Introduction: searching for a general theory of the badness of mental disorders
2. 3 dimensions on which to classify theories that answer the Disorder Interpretation of “What is the nature of mental disorder?”
   2.1 Objectivist/Constructivist
   2.2 Biological-Psychological-Social
      A note on ethical theories of mental disorder
   2.3 Internalism-Externalism
3. Combinations of dimensions
   3.1 OBI
   3.2 OBE
   3.3 CBI
   3.3 CBE
   3.5 OPI
   3.6 OPE
   3.7 CPI
   3.8 CPE
   3.9 OSI
   3.10 OSE
   3.12 CSI
   3.12 CSE

4. Choosing a theory based on combinations of dimensions: A prior commitment to the Medical Model should drive our decision-making
   4.1 Scientific considerations and objectivist intuitions
   4.2 If physical disorders are constructivist, then we cannot expect that constructivism alone could invalidate mental disorders
   4.3 A theory of mental disorders must account for the significant role of harm, disability, and negative consequences; but this imperative does not necessarily mean that harm is constitutive of a disorder
   4.4 Internalism vs. Externalism, and Biology vs. Psychology: why we can postpone decisions on these dimensions at the moment
5. Making a case for the Harmful Dysfunction theory of mental disorder

Chapter 3 Problems with the Harmful Dysfunction theory and the need for a looser conception of ‘function’: introducing Genetically-open Functions

1. Introduction: questioning the concept of ‘function’
2. An objection to the Harmful Dysfunction theory of mental disorder: The Mismatch Objection
3. Functions, fitness, environment
   3.1 Important Concepts
   3.2 Different kinds of fit to environment: wide vs. narrow reference-environments
3.3 Returning to the Mismatch Objection: distinction between evolutionary functions and current functions
3.4 A new notion of function: systemic/current role/a-historical
   3.41 What is an a-historical function?
   3.42 Problems with the notion of a-historical functions
      (a) the problem of liberalism
      (b) liberalism and the choice of system
      (c) a thin conception of ‘validity’
      (d) a-historical functions, summing up
4. A new solution to the Mismatch Objection: Genetically-closed vs. genetically-open functions
   4.1 Evolution of mentality
   4.2 Richters and Hinshaw: natural functions vs. evolutionary functions, and the introduction of openness
   4.3 Genetically-closed vs. genetically-open designs
      4.31 Genetically-open functions
      4.32 How genetically-open functions solve the Mismatch Objection
   4.4 An incomplete theory
   4.5 Objection: are “open functions” really functions?
   4.6 Concluding remarks

Chapter 4 The structural interpretation of the “what is the nature of mental disorder?” question: defending the claim that mental disorders are homeostatic property cluster kinds

1. Introduction: causal structure
2. Natural kinds vs. non-natural kinds
   2.1 Essentialist natural kinds have essences, from which radiate all characteristic properties
   2.2 Non-natural kinds: social constructions, syndromes, and family resemblances do not appeal to categories that exist outside of human classification
3. Natural kinds: alternatives to essentialism
   3.1 Option 1: paranatural kinds are marked by the absence of natural kinds
   3.2 Option 2: HPC kinds are non-essentialist natural kinds composed of a cluster of properties and a mechanism that holds those properties together
   3.3 An example of a homeostatic property cluster kind: biological species
4. Are mental disorders HPC kinds?
   4.1 The property cluster does not distinguish between causes and symptoms
   4.2 The homeostatic mechanism may not be the same in all token cases of a type of disorder
   4.3 The unity of a mental disorder kind is nothing but a shared causal structure
   4.4 As HPC kinds, mental disorders are causal networks, not the entities postulated by latent construct theory
5. Why think that mental disorders are HPC kinds? What benefits do we gain from acceptance of this theory?
   a) HPC kinds respect the autonomy of the mental
   b) The HPC model is more accurate with regard to symptom-symptom causal relations and heterogeneous, complicated etiologies
c) HPC kinds allow for variable members
d) HPC kinds allow for fuzzy boundaries between health and disorder
e) HPC kinds explain co-occurring disorders
f) HPC kinds explain vulnerability and worsening of conditions
g) A theory of mental disorders as HPC kinds is non-reductive yet naturalistic
h) Essentialist natural kinds are a variant of HPC kinds
i) HPC theory allows for symptom-based therapy, consistent with interventionist views of causation

6. An exploratory conclusion: bringing together the Disorder and Structural interpretations of the “what is the nature of mental disorder?” question

Chapter 5 The interventional autonomy of psychotherapy for treatment of clinical depression: resilience, path-dependence, and the importance of veridical experience

1. Introduction: 2 treatments for depression, and the materialism-dualism tension
2. A caveat about the deployment of empirical studies in this work
3. All clinical depression symptoms and interventions can be explained by neuroscience: introducing the higher-lower distinction
4. Empirical hypothesis and explanatory power: ADM and PT function differently because their initial sites of action are distinct; however, some of their ultimate results will be similar due to bottom-up and top-down processes
   4.1 Bottom-up interactions: heightened mood as an enabling condition on cognitive change
   4.2 Top-down interactions: from emotional regulation to mood improvement and resilience
5. Intervenational autonomy of psychotherapy – imperiled by counterfactual medical interventions?
6. Path-dependence and the importance of veridicality
   6.1 Deception
   6.2 Veridicality principle regarding memories
   6.3 Objection: perhaps the distress of non-veridical memories will just become an unpleasant side-effect of brain-tweaking surgery
7. Objection: I unfairly stack the deck by the way that I set up the thought-experiment
   7.1 A mitigated thought-experiment
   7.2 A note on why path-dependence is unique to mental disorders, and what this means for the concept of ‘resilience’
   7.3 What “recovery” means

Conclusion

(i) Concerns of consistency
(ii) Remaining questions and lines of inquiry
(iii) Closing remarks
Introduction

Consider the following two questions:

(a) Are mental disorders real?
(b) Are mental disorders real in the same way that physical disorders are real?

What are these questions actually asking, and are they asking the same thing? While I can now state that no, they are not asking the same thing, accounting for what these questions are actually asking is a more difficult task. This difficulty arises partly because of the ambiguity of the word ‘real.’ First, I will be concerned to distinguish these questions from concerns about reductionism. Following that discussion, I will then attempt to translate these questions into more understandable forms, using the language of validity and observer-independence.

But first, what are these “mental disorders” of which I speak? Many of us are familiar with “quintessential” mental disorders, those that have survived multiple reformulations of the Diagnostic and Statistic Manual (DSM), the tool that clinicians use to diagnose disorders: Clinical Depression, Generalized Anxiety Disorder, Obsessive Compulsive Disorder, Schizophrenia, Bipolar Disorder, Addiction, Phobia, Autism. There are also the personality disorders, like Narcissistic, Histrionic, Schizoid, and so on. Then, there are supposed “mental disorders” whose categorization as such might puzzle us: Dyslexia, Behavioral Addictions (like gambling, looking at pornography, sex), Oppositional Defiant Disorder, Body Integrity Identity Disorder. What about just having a weak memory or being absent-minded? We might ask: are these mental disorders (rather than neurological ones) or, are these disorders at all? Perhaps some of them are mental disorders, perhaps some are disorders but not of a mental kind, and perhaps some are not disorders. The pertinent question is: what decides? And how can we best guide our inquiry?

(i) “Ontological Reduction” and “Explanation by the natural sciences” are different methodologies; the latter does not require the former

Recall (b) from the opening: Are mental disorders real in the same way that physical disorders are real? We might re-state this question in ontologically reductive terms: are mental disorders just physical disorders of the brain? That is, are there type-identities between the kinds of psychology and/or psychiatry and neurobiology and/or genetics (like clinical depression” = serotonin deficiency”)? These are empirical questions, and their answers depend on the results of empirical studies, many of which are in their earliest stages. At the present moment, we have no certain identities to which we can point. Perhaps this lack of results gives us reason to be pessimistic about the existence of such identities, or perhaps scientists do not yet have the tools to discover the identities. So, must we wait until the science is in to answer the questions that I have posed?

The answer depends on what our feelings are about reductionism. If psychological-biological reduction fails in the case of mental disorders, then does that failure preclude mental disorders from being real? No. Those who would answer in the affirmative are likely confusing the related goals of reduction and explanation. We might in fact restate (b) in an alternate way: (b1): Is psychology/psychiatry a branch of the natural sciences (as are the sciences that study physical disorders)? Or, is there something about the ‘mental’ that bars us from having a science of mental disorders? I argue that psychiatry falls under the purview of natural sciences, but not because it reduces to a lower-level science. Rather, psychiatry falls under the purview of the
natural sciences because the kinds of explanations that we can offer for mental disorders are relevantly similar to those that we can offer for physical disorders.

One does not have to secure type-identities between psychiatry and neuroscience (or genetics) to achieve explanatory clarity for mental disorders by lower-level sciences. For instance, think about token reduction, on which every mental event is token identical to some brain event. The collection of tokens that fall under a type may have no systematic unity at the reducing level. This lack of unity might be due to multiple realizability. Multiple realizability is the theory that a psychological kind is ontologically constituted by heterogeneous lower-level kinds. Unity is found only at the higher level, establishing a psychological kind as a kind that is relatively independent from its physical details. The evidence that we have thus far on the neural realization of mental disorders does in fact point to high levels of genetic and pathophysiological heterogeneity. For example, there are many brain dysfunctions exhibited by those with clinical depression, but not all clinical depression sufferers show the same patterns of dysfunctions. As another example, schizophrenia is linked to hundreds of small genes – none of which look to be necessary or sufficient for the disorder. These outcomes could indicate that mental disorders are multiply realized, but they could also indicate that perhaps there are identities, just at far wider grains of description than scientists have claimed up until now.

Regardless of what we do or do not decide about the truth of this multiple realizability thesis, I do not feel that a failure of reduction entails a failure to understand mental disorders by appeal to the physical, biological, and genetic levels. If each token case of mental disorder can still be explained by its individual biological and genetic details, then mental disorders are indeed explained by what is happening at the level of the brain.

No doubt, the lower-level sciences have helped us to understand mental disorders, but the more interesting question to ask is whether they are all that we need for explanation. Do the physical/biological levels leave something out? Appeal to the supposed failure of type-reduction is not the only consideration that supports a view on which an autonomous mental level is necessary to an understanding of mental disorders. Three other purported arguments in favor of such mental autonomy are as follows: 1) It may be that pertinent variables – ie: trauma, humiliation – cannot be given a biological account. Mental disorders are characterized by the causal powers of specifically mental causes, symptoms, and sustaining factors. 2) Mental disorders are bad things to have, and therefore, values must make their way into an account of mental disorders. The introduction of values into this inquiry makes the study of mental disorders no longer objective, and hence non-scientific. And finally, 3) it is impossible to understand mental disorders without considering their environmental context, meaning that the disorder does not just supervene on the individual. Perhaps in some way, external facts partially constitute the disorder.

It is my goal to explore these strategies and to argue that even if they are successful, these strategies do not threaten psychology’s status as part of the natural sciences. For instance, though “traumatic experience” may not be identical with a neurobiological kind, we can talk about how trauma – in tandem with other kinds of experiences, pathophysiology, and genetic factors – alters brain functions. I claim that “explanation” involves being able to provide an understanding of what is going wrong with the physical substrate (in this case, the brain) along with being able to provide an account of the causal structure of the disorder. I argue that psychiatry is as successful for mental disorders with regard to these explanatory goals as the realm of physical medicine is for physical disorders. This is true even if the causal structures cited by explanations of mental disorders involve unreduced psychological and environmental factors.
Thus, I claim that the kinds of “mental autonomy” that previous writers have brought up do not imperil the scientific status of psychiatry. While I believe that the trend towards representing mental disorders as *just* brain disorders gets it wrong (again, I believe that other explanatory levels have a place in the natural science of psychiatry), I do not advocate going to the other side of the duality. That is, I do not advocate a view on which mental disorders are most fundamentally disturbances in one’s phenomenology, rendering the role of the brain irrelevant. What I would like to do is to push physicalist commitments as far as possible and *then* see what remains of “mental autonomy.” My eventual goal is to present a novel approach to defending mental autonomy, which is via *interventions* used against mental disorders.

(ii) ‘Reality’ is demystified by the notion of ‘Validity’

That said, I now want to return to question (a) Are mental disorders real? To begin, we can note that the reality of the *symptoms* of mental disorders are typically not called into question; undoubtedly, people suffer from depressed and manic moods, hallucinations, delusions, and so on. What is unclear is whether or not sets of these symptoms are manifestations of objectively-existing conditions (of either the brain or of the psyche) or not. In other words, is the existence of mental disorders observer-independent?

One might interpret the question through the lens of traditional Philosophy of Science, as one concerning realist vs. antirealist stances on the existence of theoretical posits in psychiatry. Through this lens, an anti-realist view would push the claim that a term like “schizophrenia” means nothing *but* a collection of symptoms that clinicians dubbed useful to group together. The grouping, based on utility or pragmatic choices, would correspond to nothing above and beyond those symptoms. It would involve no ontological commitment to some entity that is schizophrenia.

Alternatively, one might view the question of “observer-independent” mental disorders through a more normative lens. A less extreme anti-realism than the kind that I described in the previous paragraph could hold that mental symptoms are systematically related to more fundamental entities (say, brain conditions), but that these entities do not qualify as *disorders* until some further evaluation is made. Thus, perhaps mental disorders are observer-independent taxa, but their status as disorders is determined by which conditions we take to be bad (because we think that they are immoral, or involve too much suffering, or are disabilities, or…).

These questions about the objectivity or reality of mental disorders are, I believe, best construed as questions of *validity*. The validity of scientific theories or diagnostic instruments (like the DSM) should not be confused with their reliability. Reliability concerns the consistency of a theory. The DSM diagnoses mental disorders on the basis of whether an individual matches a set of descriptive criteria. The DSM is a reliable tool if independent clinicians produce the same diagnoses for the same patients. However, reliability is not itself indicative of *accuracy*, which is what validity concerns. When we ask if a particular psychiatric classification system is valid, we are typically asking about whether its categories track naturally-existing categories. In other words, does the classification system get it right? Do “diagnostic constructs” serve as shorthand for symptoms, or do they pertain to conditions that are causally responsible for these symptoms? And if the latter is right, can we identify the diagnostic constructs with concrete entities or will they always just pertain to *latent* variables?

I should mention that how we should understand “validity” in this context is itself a contested issue. Some philosophers of psychiatry claim that realist conceptions of validity are
to be too strict and that we should embrace conceptions of validity based at least partially on pragmatics or clinical utility.¹ Others claim that there are *many* relevant conceptions of validity and that we have the resources to adopt a pluralist approach to the issue.² Because these are novel positions, I want to at least begin by looking at the traditional realist conception of validity: “a diagnostic category is valid only if it represents a real entity.”³ And I will do so with the same goal as I emphasized in the last section: this traditional conception of validity does *not* commit us to reductionism or to the privileging of certain levels of explanation. I defend a view on which a “real” entity has identifiable causal structure; this structure need not lie at just the biological level and it need not be homogenous across cases.

For instance, take prognosis of a mental disorder. We might think that similar outcomes and courses of illness validate the taxonomic category under which those individual cases fall. But these outcomes and courses need not be measured in biological terms. We hear things like: “The majority of people with phobias respond well to CBT,” or “A structured family life serves as a protective factor in the onset of schizophrenia,” or “Episodes of humiliation can cause depressive relapses.” CBT, family life, and humiliation are ostensibly *not* biologically-reducible concepts. And yet, these concepts figure into legitimate generalizations and predictions about mental disorders, thus potentially validating the existence of those disorders. So to restate what I have been saying all along, we can offer a *scientific* theory of mental disorders – one that meets traditional constraints of validity – without being reductionists.

(iii) So, what are mental disorders?

Now that we understand the distinctions between, and the meanings of, questions (a) and (b), we can now ask a third question: (c) What *are* mental disorders? This question has two interpretations. One is concerned with the origin of the evaluation that mental disorders are bad things to have. The other is concerned with understanding the causal structures of mental disorders, how their symptoms and etiological factors relate. This project will deal with both of these questions, in addition to scrutinizing (a) and (b) in more detail.

What are the guidelines for how one should proceed through such an inquiry? Well, we need to be somewhat constrained by intuition, at least with regard to prototypical mental disorders. It would be problematic if a theory of mental disorders either failed to include paradigmatic cases – like Depression and Generalized Anxiety Disorder – or else included conditions that just did not seem like mental disorders, even on liberal understandings of the term. For this reason, the DSM should serve as a resource to which a theory of mental disorders is somewhat answerable. Intuition also tells us that the borders between having certain mental disorders and not having them can be fuzzy, as can the boundaries between different conditions. A good theory should respect this intuition as well. However, the DSM is itself in a state of revision. Along with this instability, it is also a *descriptive* manual, meaning that it does not posit entities beyond collections of symptoms, and it has no fully worked-out answer to the evaluative question that I cited earlier. Thus, it would be a mistake to rely too heavily on either intuition or the DSM in our exploration of the nature of mental disorders.

Besides keeping the DSM in mind, any theory of mental disorders also must take into account the most up-to-date empirical studies on the neural correlates of mental disorders and their interventions. In this work, I am not interested with conceptual analysis or linguistic usage. I am interested in constructing and defending theories that pertain to actual disorders as they occur in the world. I have alluded to theories that argue that there are actually no such things as
mental disorders, or that they are social constructions, or tools for those in power to subjugate others, or that they are in a state of flux depending on changing medical judgments. Nonetheless, I take as my starting point the possibility that psychology and psychiatry have the potential to be branches of natural science. That is, they have the potential to be grounded at least partially in some more fundamental science, in this case biology and/or neuroscience. I take this as my starting point based on a judgment that the empirical evidence for the neural underpinnings of mental disorders is too compelling to ignore.

(iv) I offer the following roadmap through the dissertation:

CHAPTER 1 deals with the following questions:
Are mental disorders just physical disorders of the brain?
What would it mean for psychiatry and psychology to be branches of natural science, and to be fully integrated with the rest of medicine?
Which features characterize a Medical Model, or Disease Model, of mental disorders?
What does the empirical data on neural correlates of mental disorders lead us to conclude about the possibility of a reduction between psychiatry and neuroscience?

I make the following claims:
1a) Medical Models of mental disorder fall on a spectrum from more to less extreme. More extreme theories focus on type-reduction and fundamental explanations. Less extreme theories involve grounding in a hard science along with imposition of causal explanatory constraints.
1b) Abnormal psychiatry fits a weaker version of the Medical Model of mental disorder.
1c) Multiple realizability does not impugn the possibility that mental disorders are physical disorders. It does not even impugn the possibility of type reductions because theorists may err by mismatching the grains at which the reducing and reduced kinds are described.
1d) If there are lower-level kinds that are identical to mental disorders, then these kinds are at a wide grain of description.
1e) Empirical evidence from neuroscience is able to provide many contenders for these identities beyond the simplistic neurotransmitter-deficiency theories. Many of these contenders are currently referred to as “brain circuits,” though these circuits must be understood in functional, rather than physiological, terms.
1f) These considerations about neural correlates provide support for psychiatry’s being a branch of natural science and/or medicine (at least with regard to the constraint about grounding in a harder science).

CHAPTER 2 deals with the question: What underlies the evaluative component of mental disorders? What makes it the case that they are bad things to have? Is this badness dependent on facts, judgments, or both?

I create a framework for theories that purport to answer this question. The framework has 3 dimensions (Objectivist/Constructivist, Biological/Psychological/Social, Internalist/Externalist), which can combine into 12 combinations. I claim that:
2a) For the purposes described in this introduction, we need an Objectivist theory.
2b) The Harmful Dysfunction theory of mental disorder is the strongest contender for answering this question, since it is objectivist and grounds mental disorders in biological dysfunctions. It also takes into account our intuition that the consequences of mental disorders matter to making them what they are, when it adds the need for a judgment of ‘harmful.’
In CHAPTER 3, I raise objections against the Harmful Dysfunction theory in order to revise it into a novel form that does not fall prey to these objections. I claim that:

3a) HD is vulnerable to the Mismatch Objection, whereon there is the possibility that a mental disorder is in fact a well-functioning mechanism that is operating properly in an improper environment. Therefore, there needs to be a better account of how functioning is linked to environment(s).

3b) We cannot respond to this objection just by claiming that the environment is to blame; this answer misses the point about mental disorders being internal dysfunctions of those who are afflicted by them.

3c) We cannot dismiss the objection entirely, or else we lose the benefits of a “dysfunction-based” theory of mental disorders.

3d) We therefore must construct a new notion of ‘function’ that is able to sidestep the Mismatch Objection. However, the attempt to do this via deference to ‘current role/systemic’ notions of functions fails.

3e) The best way to deal with the Mismatch Objection is via a looser construction of ‘function,’ that I call “genetically-open.” Genetically-open functions are programmed to undergo ontogenetic development in the fulfillment of their functions, and their end states are not genetically-reducible. Their functions involves a built-in flexibility of response to novel environments. I suggest that some psychological functions are of this nature.

In CHAPTER 4, I switch to asking the following questions:
What is the structure of mental disorders? What kinds of things are they, and what is their causal structure? Are they natural kinds?
How are mental disorders individuated? What does a proper nosology (classification) look like? How does diagnosis work?

I claim that:

5a) Mental disorders are not essentialist natural kinds, nor are they mere family resemblances.

5b) Mental disorders are homeostatic property cluster kinds (HPC kinds). HPC kinds are natural kinds. HPC kinds have two features: a set of co-occurring symptoms that make one another more likely to occur, and a mechanism holding these symptoms in place.

5c) The homeostatic mechanism may be separate from the property cluster, or it may be composed of the causal linkages within the cluster itself.

5d) A theory of HPC kinds is objectivist and explains fuzzy boundaries, instability of diagnostic systems, variability of token members of a type of disorder, co-occurring disorders, and symptoms that are in direct causal relations with one another – including casually-efficacious mental properties.

5e) Although a theory of HPC kinds is merely structural and descriptive, some features of HPC kinds (widening reach, positive feedback loops) can explain badness, risk, and vulnerability.

In CHAPTER 5, I look at the question: Is psychotherapy merely pragmatically useful, in comparison to medical interventions, like psychotropic medications? I claim that:

6a) Psychotherapy is a necessary intervention in many cases of clinical depression. Its results cannot be duplicated by any physical/medical intervention.

6b) Psychotherapy (PT) and antidepressant medications (ADM) act primarily via different levels of the brain (what I call “higher” vs. “lower”). Interventions at each of these levels also exhibit
top-down and bottom-up causal processes, which explains why the ultimate results of the two types of interventions may be similar. All of these mechanisms are neurobiologically explicable.

6c) Empirical studies demonstrate that PT offers more protection against relapse than ADM. I argue that this fact can be explained by an empirical hypothesis on which PT changes brain areas corresponding to resilience.

6d) Because of the neurobiological explicability of PT, we could counterfactually produce the total neural end states of PT via some kind of “brain tweaking surgery.” Thus, one might argue that PT itself is not necessary to recovery from depression.

6e) I defend the path-dependency of PT by appeal to a particular results-based success-criterion of recovery that involves the ability to successfully navigate the social world, an ability that is imperiled with the false memories that counterfactual brain surgeries would involve.

6f) Therefore, at the very least, the mental has a degree of autonomy when it comes to interventions on mental disorders.
Chapter 1: Medical/Disease Models of mental disorder offer naturalistic, scientific explanations of mental disorders without requiring reductions

1. Introduction: has the brain usurped discussions of mental disorders?

Thomas Insel, the Director of the National Institute for Mental Health, writes that the claim that “mental disorders are brain disorders” is a “disruptive insight.” I take it that this insight is “disruptive” because it turns the earlier paradigm – that mental disorders are irreducibly psychological phenomena – on its head. In this chapter, I want to delve further into the statement that “mental disorders are brain disorders,” in order to discover 1) exactly what the claim means, 2) whether we are justified in holding it to be true, and 3) if it is true, what implications it has for the kinds of explanations that we can give for mental disorders. When we say that “mental disorders are brain disorders,” are we saying that mental disorders are nothing but disorders of the brain, or are we merely saying that mental disorders have rough neural correlates? If the former is true, then it looks like explanations for both physical disorders and mental disorders will have the same structure; explanations for symptoms will be given in terms of some physical “disease type” entity. Theories that endorse this stance are referred to as “Disease Models” of mental disorder. They are also called “Medical Models” of mental disorder, as mental disorders now become fit subjects to fall under the purview of those natural sciences that form the backbone of physical medicine (physiology, biology, neurobiology).

Just a note about terminology: this chapter is based around the fact that the last decade or two have seen a flourishing of claims like “mental disorders are just like physical disorders” or “psychiatry should be a part of medicine.” However, the latter statement is confusing, as the word ‘medicine’ is often used to refer to a practice, one with a long anthropological history. The practice is employed in order to treat those who suffer. Such an instrumentalist, treatment-based orientation towards the field of medicine can make the claim seem trivial – of course psychiatry is a part of medicine, if medicine is just what aims to alleviate suffering.

I believe that the triviality of the claim evaporates when we view ‘medicine’ differently, with the help of the term “Disease Model.” It is true that a doctor, or a psychiatrist, need not understand the mechanism behind a person’s suffering in order to treat that person. But with physical medicine, especially with the advent of germ-theories of disease and the increase in knowledge of genetics, medicine has become not just a practice, but also a collection of theories with ontological commitments about entities and causal mechanisms. When one claims that psychiatry is a part of medicine, then, this person claims that psychiatry can offer scientific theories for mental disorders on par with the theories of physical disorders. Then, the question becomes one of exploring what is meant by ‘scientific.’ This term could indicate something as broad as a project’s cohering to the scientific method or something so strict that is requires materialist type-reductions. With the hope of striking a balance between these two methodologies, I will explore what it means for a field to be part of the natural sciences by making my own specific claims about what scientific explanations require.

When we question the meaning and truth of the claim “mental disorders are brain disorders,” we are asking a philosophical question about the ontology of the world. But the investigation is also empirical – the answers that we can offer about the neural basis of specific disorders depend on the results of empirical research that is still in its early stages. Brain scans have become more sophisticated and genetic studies can now look at thousands of biomarkers for disorders. Progress on the potential physical correlates of mental disorders has come quite a way since theorists worked backwards from the effects of antidepressants to postulate
neurotransmitter theories of depression, for example. I will present particular theories of the neural correlates of mental disorders, but because the research is still at its beginnings, I will not be concerned to defend the truth of any of these theories. Rather, my goal is to explore the question of what kinds of explanatory structures theories of mental disorder require in order to be considered part of the natural sciences. For example, would such explanations be allowed to cite psychological factors, or would the entire explanation have to be couched in neurobiological terms?

Another way of stating my aim is to say that I am interested in the constraints that a Medical Model or Disease Model of mental disorders imposes on explanations. For example, do such explanations need to be reductive? The argumentative structure given by those who take a hard-lined Disease Model perspective of mental disorders is as follows:

P1) Mental disorders are brain disorders.
P2) Brain disorders are physical disorders.
P3) We can provide scientific explanations for physical disorders.

C) Therefore, we can provide scientific explanations for mental disorders.

I am curious about whether these “scientific” explanations, cited in the conclusion, are complete and/or adequate. And, if they are not, what would the alternative explanatory frameworks would look like, and what kinds of components would fill in the holes left by a purely neurobiological approach? Furthermore, if the neurobiological explanations are incomplete, does this incompleteness force psychiatry outside of the realm of natural science?

I have appropriated the term “Medical Model” from the work of Dominic Murphy. Murphy claims that these models contain “a set of commitments about the nature of mental illness that treat it as a disease,” and that diseases “are causally explained by their underlying pathophysiology.” Furthermore, such “abnormalities in underlying neurobiological systems…are responsible for the observed patterns of signs and symptoms.” These quotes are just more complicated ways of stating the premises given in the argument above. On such a picture, mental disorders are no different from other physical disorders, except for the facts that 1) they are realized in the brain (instead of other organs) and 2) they have characteristic mental symptoms. Still, the most important point is that these disorders bear a systematic relationships with kinds at a scientific (in this case, the neurobiological) level. These systematic relationships are what make mental disorders amenable to scientific explanations. In this chapter, I will assert that “scientific explanations” must exhibit two features: 1) they defer to a lower-level (physical or biological) of explanation, and 2) they present a causal structure for the disorder. It is to these explanatory features that I now turn.

I will start by presenting two purported Medical Models for the explanation of mental disorders. These will be Insel’s “Brain Circuits” theory and Murphy’s “Psychiatry as cognitive neuroscience” theory. My hope is that by presenting both of these models, we will get a feel for what binds all Medical Models together, as well as where they might permissibly differ.

2. Insel: mental disorders involve faulty brain circuits

As compared to Murphy’s Medical Model of mental disorders, Insel’s is relatively simpler, so I will begin with it. I will be basing the following exposition on Insel’s writing, supplemented by more technical work on the subject of brain circuits. Insel writes that “the malfunctioning of entire [brain] circuits may underlie many mental disorders.” Note the relative
weakness of this claim: we should not necessarily interpret Insel to be asserting that mental disorders are malfunctioning brain circuits. While “underlie” could indicate identity or constitution, it could also just mean that the brain circuit is either a diachronic cause or a correlative factor of the disorder, or even just one symptom of many.

The empirical work on neural correlates of mental disorders makes much reference to brain circuits, though very few sources attempt to state explicitly what a brain circuit actually is. What are the identity conditions on a brain circuit, and how do we individuate them from one another? In much of my own reading on the topic, I have been tempted to conclude that the scientific community tends to take ‘brain circuit’ to be a primitive concept not subject to further analysis. However, it is philosophically irresponsible to dismiss further investigation of the meaning and extension of the term. I aim to show that different concepts of ‘brain circuits’ fall out of different versions of the Medical Model of mental disorder. With regard to Insel and Murphy, I believe that for all that Insel has said thus far, he has in mind an anatomical/physiological/neurobiological criteria for understanding and individuating brain circuits. I think that Murphy, with his cognitively-inclined theory, likely is thinking more along the lines of brain circuits being akin to functionally-defined mechanisms. Of course, because Insel claims that circuits “malfunction,” there is at least an implicit emphasis on functions/functionality for him as well. But, one of my central claims will be that any plausible notion of ‘brain circuit’ must be understood functionally rather than merely anatomically.

So, Insel believes that faulty brain circuits “underlie” mental disorders. Insel writes that “brain regions that function together to carry out normal mental operations can be thought of as analogous to electrical circuits…the details of each disorder’s ‘circuit diagram’ or map are still emerging.” This quote is difficult to unpack due to its use of the term ‘analogous.’ Is Insel just suggesting a helpful way to think about the brain? Likely, he means the analogy to be more than just a tool, since brains do in fact conduct electrical (and chemical) signals between neurons. Thus, when Insel speaks of brain circuits, I interpret him to be talking about the literal wiring patterns of the brain – the routes via which energy/information flows. We could be focused on local wiring – small loops of activity – or conductance between individual pathways, or coordinated conductance across spatially-segregated pathways via synchronous activity. The upshot of this discussion of brain wiring is that when it comes to what is systematically going wrong in mental disorders, the answer is typically not given just in terms of insults to gross anatomy. The brain of a mentally ill individual is likely to be physically intact. Insel makes an analogy to the difference between myocardial infarction (destruction of heart tissue) and cardiac arrhythmias to elucidate his claim that mental disorders, unlike neurological diseases, are not equivalent to brain lesions. Mental disorders are like brain arrhythmias.

2.1 The difficulty of constructing an anatomical definition of ‘brain circuit’
As I stated, it is difficult to find a conceptual analysis of ‘brain circuit’ in either the psychological or neurobiological literature. Instead, the term seems to be taken as a primitive concept in these fields. Some relatively uninformative purported definitions say things like brain circuits are “major routes” of electrical activity in the brain, or that brain circuits are paths “between the brain and the target tissue through which chemical and electrical signals are relayed.” I call these definitions “uninformative,” because they say nothing beyond the basics of what neuroscience has already told us: the brain conducts electrical impulses that allow for communication between different brain areas. Presumably though, the brain is more than just a
mass of holistic wiring – presumably, we can break it down into components. The question is, what could guide us in this componential analysis of wiring patterns?

Anatomy will not do the job. Anatomy might help us in distinguishing pathways that are visible to the naked eye, i.e.: paths of neurons that stretch from the spinal cord up to very specific cortical regions, or pathways that have relatively few branching connections (some loops of neurons may be like this). Another possible indicator that a collection of neurons are circumscribed in nature, and not just via our descriptions, is that the collection is modulated as a unified system. According to this consideration, we could discover circuits by looking for groups of neurons that are targeted as a unit by some extra-circuit source for neurotransmitter-based or hormonal control/regulation. Still, the fact remains that the brain is highly interconnected and every neuron connects to so many other neurons that it is impossible to individuate most brain circuits from one another on the basis of anatomy alone. I claim that we simply require the notion of ‘function’ in order to understand what brain circuits are.

Although it is beyond my scope to construct a positive theory of ‘brain circuits,’ I do want to emphasize this point about the inadequacy of anatomy to define brain circuits by showing how difficult the construction of an adequate theory would be. Thus, one might start by saying that (Definition) Brain circuits:

(i) are a collection of neurons
(ii) that are connected.

The details of each of these conditions require much more working-out. For instance, note that in (i), I have not said that a brain circuit is a particular collection of neurons that constitute only one circuit. In other words, it is not as if brain circuits require that their constituents take part in only one circuit. This might be the case for certain circuits, for instance, those that are composed of dedicated neurons that are specialized for just one task. However, in many more cases, different brain circuits might overlap in their constitution; hence, a particular collection of neurons might partially compose multiple brain circuits. As a result, the circuits to which these neurons belonged would be context-dependent, depending on which circuits were active at any given time. Condition (i) is important insofar as it just points to the basic constituents of brain circuits.

Condition (ii) is most difficult of all to work out, and on my view, is the undoing of an anatomical theory of brain circuits. Presumably, if we describe ‘connectivity’ loosely enough, all of the neurons in the brain will count as being connected to one another. So we need a further restriction. Imagine that we said that a brain circuit is a collection of neurons that are all directly connected to one another. By “direct” connections, we would mean that the axon of one neuron was in contact, electrically or chemically, with the dendrites of another neuron. Of course, they are not technically “directly” connected because of the synapse between them, but they are directly connected insofar as there are no mediating neurons between them. This move would be a way of getting into touch with Insel’s emphasis on “wiring.” Presumably, this restriction would be too strong, as the propagation of electrical and chemical signaling progresses along chains of neurons. So if we consider 4 neurons, they might connect like this:

A \rightarrow B \rightarrow C \rightarrow D. A is directly connected to B (and other neurons not included in the chain), which is directly connected to C, which is directly connected to D. But A is directly connected to neither C nor D and B is not directly connected to D. So the neurons are not all directly connected with one another. Nonetheless, this kind of chain is a piece of the simplest kind of neural collection that we would intuitively take to be a brain circuit.
So perhaps it would be better to claim that a brain circuit is a

(i) collection of neurons

(ii) that are bound by a chain of direct connections.

The problem that would then confront us is that it is likely that you could pick any two neurons in the brain and “trace” out a path of direct connections between them. When does such a path become too “big” to count as a single circuit any longer? It seems to me that any attempt to choose an arbitrary number of direct links from the beginning of the chain as a cut-off point for counting as being in the same circuit would be unjustified.

Alternatively, we might try to come up with a definition of “connected” that involves appeal to a relatively high degree of activation or inhibition that one neuron has upon another. Such a standard might allow us to carve out significant connections between neurons vs. the potentially inevitable ability to trace a path between any two neurons at random. We could think of the significant connections as being like “roads” that are deeply carved by continued traversal. But again, what counts as relatively high activation or inhibition? Once again, it looks like we face a problem of arbitrariness.

The best evidence in support of an anatomically-based account of brain circuits comes from the study of Frontal-Subcortical circuits (FSCs). These circuits are neuronal pathways that project from the frontal cortex to the basal ganglia and back to the frontal cortex again (with an indirect foray to the thalamus). The circuits are therefore loop-like, which makes them intuitive candidates for examples of brain circuits. What I mean by ‘loop’ is a topographically circular structure where neurons are directly connected in relatively unbranching chains. The neurons are substantially “closed” in that very few neurons are themselves directly linked to non-loop elements. We might think that because of this closed-offness, that the neurons are “encapsulated.”

Thus, we envisage a circular path with nodes, where these nodes are neurons. Although the loop is integrated with the rest of the brain, such integration is not available at every point in the loop. The circuits contain “open-loop” elements, defined as “afferent or efferent projections.”14 These projections are responsible for communication between – along with coordination of - different circuits. This global modulation of the interactions of the circuits is controlled by neurotransmitters, particularly dopamine, acetylcholine, serotonin, and norepinephrine. Within the circuits, the most prevalent neurotransmitters are the inhibitory GABA and the excitatory glutamate. Depending on whose findings you read, there are between 5-7 FSCs that are currently defined on the basis of the cortical region with which they are associated: motor, oculomotor, DLPF, AC, and OF. Some studies claim that each of these 5-7 circuits is composed of “multiple parallel segregated circuits,”15 which are differentiated by more fine-grained distinctions on the basis of cortical location.

In trying to determine what a brain circuit is via this example, note two things: 1) FSCs are referred to as “closed loops,” which indicated that distinct circuits are physiologically “cut off” from one another, and 2) anatomy and physiology are doing the work of individuating the brain circuits – if you count the physical loops, then you have counted the brain circuits.

However, the problem here is that FSC’s represent a low percentage of the total neuronal volume in the brain. It may be that we can thus give an anatomical account of some brain circuits, but certainly not all.

2.2 Conclusion: an acceptable notion of ‘brain circuits’ requires the notion of functions
All of these considerations point to why we need a third condition in our definition of brain circuits, one that would give us a non-arbitrary way of circumscribing collections of chains of directly-connected neurons, one that would allow us to call things other than visible neuronal loops “brain circuits.” I suggest a third condition like: (iii) the directly connected neurons are unified by a common function. Note that the unit of functionality here is the brain circuit, *not* the individual neurons (individual neurons A and B might be part of brain circuit X, but neither A nor B has function F; X does). So it is not that the individual neurons share a function; rather, together, they all realize a common function.

Consider how the following purported definition of a brain circuit makes explicit reference to functions: neural circuits are the “primary supracellular mediators of the brain’s diverse functional capacities. A circuit typically refers to a set of interconnected components that together subserve a specific function.” Or, to repeat Insel’s quote: brain circuits are “brain regions that function together to carry out normal mental operations.” So functions are important to individuating brain functions. What the relevant functions are, and how they are individuated, is of course a related and important question. For instance, one might claim that the function of brain circuits in the hippocampus is to aid in spatial navigation. But these very same pathways light up when organisms rehearse spatial pathways in memory, so there seems to be a memory function here as well. Nothing that I have said so far restricts us to saying that a brain circuit can only have one function. It could be that brain circuits – like collections of neurons – have functions that are context-dependent. Presumably, larger neural pathways are built up from units of brain circuits, and it is possible that these pathways have their own functions.

Furthermore, even with the example of FSCs, which I suggested might exemplify an anatomical conception of ‘brain circuit,’ neuroscientists *still* appeal to the notion of functions to elucidate the unity of the loops/circuits. For instance, when discussing the “open loop” elements of FSCs, Lichter and Cummings write that “brain regions linked by these afferent or efferent projections are functionally related…circuits integrate information from anatomically disparate but functionally related brain regions.” Hence, even in cases where topographical features aid in brain-circuit-individuation, condition (iii) is nonetheless the most significant.

**2.3 Dysfunctional** brain circuits and the role of neurobiology

As I stated at the end of section 2.1, I am not suggesting that *all* brain circuits are so easily individuated via an inference from physiological descriptions. But the fact that at least *some* might be suggests that there could be a theory of dysfunctional brain circuits that does not defer to a science outside of neurobiology for an account of “dysfunctional.” This account would likely define dysfunctionality on the basis of statistically significant deviation from normal (again, defined statistically) circuit activity. One might worry that such an account would be circular. We cannot determine when a circuit is malfunctioning, absent some prior theory of what is considered normal. One might also worry that the account would be overly inclusive: there are all kinds of ways of departing from statistically “normal” neuronal behavior that do not thereby qualify the possessors of those brains to be the bearers of mental disorders. Presumably, the way that the scientists would determine “normal” circuit activity would be to compare scans and data from healthy and unhealthy people. Thus, they would already need to have a set of people diagnosed with the supposed disorder/dysfunction – this move would privilege a non-biological basis of diagnosis in driving the biological research. This worry might be what drives people like Murphy to construct more substantive hypotheses about dysfunctional neural mechanisms.
What counts as dysfunctional could be over/underactivation of the circuit compared with normals, or problems with open-loop elements of modulation, or disturbances of coordination with other brain regions, etc. In fact, because it might be relative activation in one brain circuit as compared to others that is the marker of functionality, it could be that any absolute measurement of activity will not be a proper diagnostic marker. In any event, my point is, to reiterate, that we could push for a purely neurobiological understanding of dysfunctional brain circuits. Progress in both individuation and diagnosis would come from looking at brain scans of the activity within and across circuits to come up with a model of normal activity.

The most devastating objection to this kind of theory is the charge that it fails to be properly explanatory. On such a theory, we can know that something is going wrong with someone’s brain, but we cannot give a proper characterization of what is going wrong, except to say that brain circuit activity deviates from normal. But why is this kind of activity difference a problem? Is there some more robust notion of ‘dysfunction’ that can be deployed when we speak of faulty brain circuits? We can make progress on this point by asking what brain circuits are for, which is why I am so insistent on making ‘brain circuit’ a functional notion. And in fact, I think that Insel would likely be behind such a project. I do not believe that Insel’s version of what brain circuits are is as constrained as the one given by neuroscientists who focus on FSC circuits. When Insel outlines the “circuits” involved in depression, OCD, and PTSD, he does seem to be more focused on couching the neural mechanisms in terms of tasks that they typically perform, rather than in terms of their physiology alone. We can see this point if we return to his specific wording: “brain regions that function together to carry out normal mental operations can be thought of as analogous to electrical circuits.” If we take this wording literally, it differs from the anatomical theory in 2 ways: 1) it unifies a brain circuit via reference to a shared function, and 2) “normal mental operations” might not be understood with regard to the biological level of description. Perhaps a characterization of “normal” will require reference to a computational or intentional level.

Hence, I do not necessarily think that Insel himself is claiming that neurobiology on its own will be explanatory of the brain circuit dysfunctions underlying mental disorders, without some concomitant functional framework. But the reason that I explore this possibility is because I believe that it is an isolable kind of Medical Model (one that claims that we can know if brain circuits are “going wrong” just by studying the neurons and their connections, without reference to what those neurons are supposed to be doing), and that its shortcomings will help to elucidate Murphy’s upcoming framework.

2.4 Classification

The claim that mental disorders are - or involve, or are caused by, or are correlated with - faulty brain circuits (understood functionally) has implications for what a nosology or taxonomy of mental disorders will end up looking like, as well as how diagnosis will take place. If brain circuits are the realizers of mental disorders, there are multiple ways in which the relation between the levels of neurobiology and abnormal psychiatry could be understood. The simplest nosology would be to distinguish mental disorders on the basis of which circuit was faulty. According to this nosology, there would be a 1:1 relation between circuits and disorders (and perhaps subsets of disorders based on sub-circuits of brain circuits).

But the theory also allows for more complex inter-level relations. Individual disorders might be identical to the interplay of one or more dysfunctional circuits, which might explain the high level of patients with co-occurring disorders (For instance, we can imagine a classification
like: Depression is circuit 1 and 2 dysfunction, anxiety is circuit 2 and 3 dysfunction; this overlap in circuit 2 explains why anxiety and depression have many symptoms in common, and why many people with depression also have anxiety and vice versa). Or, in order to get the identities right, we might have to zoom out even further, to collections of FSCs plus their open-loop elements plus perhaps their modulation by neurotransmitters. All of these taxonomies would postulate the existence of “mental disorder-neural kind” correlations (which could end up being constitutive identities or diachronic causes) based on brain circuits, but they would just differ in the scope of the neural kinds at play. The proper neural scope might be extremely wide – so wide that some might claim that there is no way that such a wide realizer could play a role in a type-reduction. In any event, instead of diagnosing on the basis of symptoms as we currently do with the DSM, we would have a properly scientific diagnostic system based on biological markers.

2.5 Examples of empirical hypotheses on which mental disorders are faulty brain circuits

2.5.1 Example 1: Obsessive-Compulsive Disorder (OCD)
Insel offers empirical evidence in support of his claim that mental disorders crucially involve faulty brain circuits. He does this for a handful of disorders, two of which I will present now. These examples are not intended to make us conclude one way or another about whether brain circuits should be understood as playing a causal role vs. a constitutive role in mental disorders; we will get to a discussion of reductionism later on.

Obsessive-compulsive disorder is marked by the combination of obsessive thoughts with compulsions to perform certain actions, like repeatedly checking that appliances are turned off, washing one’s hands, enacting elaborate rituals, etc. Studies have shown that the FSC correlated with OCD is a circuit that links the Orbital-Frontal Cortex (OF) with the Caudate Nucleus of the basal ganglia, and that those with OCD have increased activity in this circuit. The story is actually more complex than this; it involves both direct and indirect pathways from this circuit to the thalamus. The direct pathway to the thalamus is disinhibiting, while the indirect pathway is inhibiting. Those with OCD show more relative tone in the direct pathway, making the thalamus more active than normal, which may be where the compulsive drive to perform certain behaviors is coming from.

However, the most interesting empirical data that I have come across with regard to OCD has to do with the brain-localization of variants of the disorder. Baxter et al. propose an “OCD-Tourette’s-Tics Spectrum,” to cover all cases where there are pathologically uncontrollable thoughts, actions, or verbal or motor tics.\textsuperscript{19} This commonality is explained by a shared FSC (the OF-CN circuit that I just described, related functionally to increased disinhibition), but the particular regions of cortex and striatal areas involved determine the content of the disorder (ie: “sexual-ingestive urges” are associated with a link between the lower limbic cortex and the olfactory tubercle, while “complex social/territorial concerns” are associated with a link between the OF cortex and the ventromedial caudate nucleus).\textsuperscript{20} These findings support an eventual nosology for OCD wherein subtypes of the disorder are mirrored by malfunctioning subcircuits of the larger FSC. It also provides an interesting kind of “isomorphism” between brain locations (and their corresponding functions) and different manifestations of one kind of disorder.

2.5.2 Example 2: clinical depression
When I was considering the possibilities for a nosology of mental disorders based on dysfunctional brain circuits, I considered possibilities other than the “1 disorder: 1 brain circuit” relation that we saw with OCD. I claimed that with some mental disorders, we might have to
“zoom out” to a larger brain correlate, one that is nonetheless still based in brain circuits. I want to use clinical depression as an example of this process of zooming-out. “Depression” is a heterogeneous psychological category, meaning that it is marked by a wide variety of symptoms: emotional, cognitive, vegetative, motivational, etc. As such, cases of so-called depression differ markedly from one another, both in terms of symptoms and intensity. Complicating the picture even more, some cases of depression are secondary to other disorders (like Huntington’s Disease or post-surgery cardiac patients).

Because of this heterogeneity, it might turn out that the diagnostic category of “depression” actually pertains to a variety of distinct disorders. For example, if empirical studies demonstrate that these various kinds of depression correlate with different circuits in the brain, we might rethink our nosology. On the other hand, it is also possible that the various kinds of depression will turn out to correlate with distinct sub-circuits of one larger circuit, like we saw with OCD. Another possibility is that all cases of depression could involve malfunctions in one large, holistic circuit that can be influenced at many different places.

I now want to discuss how some empirical evidence from the study of FSCs point to the latter conclusion. This evidence also offers an explanation as to why serotonin has been the focal point of treating depression for so many years. It was not too long ago that psychiatrists claimed that “depression = low serotonin concentrations,” given that SSRIs (medications that prevent serotonin from being degraded, hence allowing it to act longer in the synapses) were so effective. But this new theory can explain the importance of serotonin without endorsing a “chemical imbalance” theory.

In his paper, Insel points to a part of the prefrontal cortex referred to as “area 25.” He makes the bold claim that “neuroscientists now think of depression as a circuitry disorder involving abnormal activity in area 25 that disrupts its vast connected network.”21 Area 25 is the supposed “governor” of a vastly connected network containing neural pathways in the hypothalamus, brain stem, amygdala, insula, hippocampus, and many parts of the frontal cortex. Many sufferers from depression show high activity in area 25. Depending on whether or not this abnormal activity spreads to other parts of the network – and if so, where – the various symptoms of depression will manifest: forgetfulness (if the hippocampus is affected), problems relating to fear/anxiety/self-esteem (if the amygdala and insula are affected), biased information-processing and distorted assessments (frontal cortex), and more brute physical symptoms (due to abnormal activity in the brainstem and hypothalamus).

Work by Tao et al. discusses one of the possible circuits involved in the cognitive and emotional aspects of depression.22 These scientists point to the fact that many depressives have an “uncoupled” hate circuit, meaning that the part of the cortex that is cognitively in control of labeling certain stimuli as hateful becomes unconnected from the limbic system’s production of feelings of hate and loathing.21 The particular part of the limbic system that is involved in this uncoupling looks to be the insula, which is normally responsible both for feelings of disgust and for interoceptive awareness (awareness of oneself). Hence, what ensues from such a decoupling is a free-floating feeling of negativity in the absence of any appropriate stimuli. This emotion is then attributed to the self, leading to self-loathing and lowered self-esteem.

Thus, the “hate circuit” might be part of the cognitive architecture of the larger “PFC area 25 system” of which Insel speaks. The importance of the entirety of this network is what leads Mayberg to claim that depression is a “systems level disorder,” where there is not only dysfunction in individual circuits but also a “failure of the coordinated interactions of a distributed network of cortical-subcortical pathways.”24 Thus, the brain circuits at issue might
not even themselves be faulty; instead, there may be a loss of synchronicity amongst the various circuits that compose a larger network of pathways. Interestingly, after antidepressant therapy, some people show decreased brain activity across all regions of a particular network, even when the disorder was marked by some areas with heightened activity and others with lowered activity relative to normal. To pick up a point that I mentioned earlier, these results might be evidence that what matters for mental health is not the absolute level of activity in any particular brain circuit, but rather, the relative similarity in activity levels across multiple circuits.25

Returning to Insel’s model based on area 25, though there is a connected network of downstream elements from area 25, nothing ensures that all pieces of the network will be activated together. There might be many branching pathways, some independent of one another. Thus, the neurobiology can provide us with reasons for claiming both 1) that there are different species of depression (some downstream elements are not affected, for whatever reason), and 2) that depressions all have something biological in common, if we consider a wide-grained, holistic neurobiological picture. Furthermore, on this view, we can explain why low serotonin was the best candidate for a depression correlate before more complex studies have shown us the superior importance of brain circuits: PFC area 25 is rich in serotonin transporters, providing evidence for the claim that modulation of the entire system takes place in this location via this neurotransmitter. However, the modulation at this “high” point in the system via serotonin does not mean that things cannot go wrong independently in downstream parts of the circuit and with other neurotransmitters. In these cases, treatment with SSRI drugs would be unlikely to be effective.

A note about this theory of depression: the total explanation just given of depression is not solely neurobiological, even with its emphasis on brain circuits and physiology. For instance, I quoted material that uses the term “hate circuit.” Clearly, this circuit is so-named not for an anatomical or physiological feature, but rather, for a purported function given in psychological terms. This recognition supports my contention that brain circuits cannot typically be defined in purely anatomical or neurobiological terms, that they require a unifying function. However, this function itself might not be able to be given in neurobiological terms either; the capacities for multiple realizability of the circuit are just so high that we require a more abstracted characterization of the circuit. This point will become even more important later on, when I make the claim that “scientific” explanation of mental disorders need not be reductive in order to count as on par with the scientific explanations for purely physical disorders. Thus, we have reached a point where it is justified to conclude that most people who write about “brain circuits” have something more robust in mind than physiology as the grounds of individuating brain circuits. This is a good time to have a look at Murphy’s Medical Model.

3. Murphy’s claim: psychiatry is cognitive neuroscience

3.1 Cognitive science allows for multiple levels of explanation

Recall that this paper is investigating the plausibility of, and structure of, scientific explanations of mental disorders. I have just finished presenting Insel’s version of a Medical Model of mental disorders, a version that claims that mental disorders are (or involve in some crucial way) faulty brain circuits. When I discussed the taxonomy that might come out of such a theory, I explored the option of a 1:many relation between mental disorders and brain circuits, that is, a classification that went beyond individuation on the basis of malfunctioning individual brain circuits. I also discussed how Insel does not yet offer a theory of what makes a brain circuit
“faulty” or “dysfunctional.” For now, the theory can be grounded only in the statistical norms of neurobiology.

I now introduce Murphy’s Medical Model of mental disorder. I do this not necessarily to juxtapose it to Insel’s in order to pit one against the other. Rather, I want to show how Murphy’s theory might be able to answer some of the questions that were raised about Insel’s theory, and hence to productively build upon it. I also want to use Murphy’s theory to demonstrate the fact that mental disorders can be explained by the natural sciences without having to reduce to the kinds given by a natural science.

My interpretation of Murphy’s position is that instead of claiming that mental disorders involve faulty brain circuits, he claims that mental disorders are brain dysfunctions. This difference might seem like a semantic quibble, but it is not, because not everything that can go wrong in the brain has to do with circuitry, although circuitry clearly plays some kind of role. As pointed out before, we may indeed have to zoom out from circuitry in order to find a stable unit of neurobiology that correlates to most instances of a disorder. The unity of this correlate would not have to do with its being a circuit, but would rather have to do with its playing a function. For instance, cognitive neuroscience combines cognitive science, computational modeling, and psychology. It attempts to individuate/decompose cognitive functions on the basis of information-processing tasks and then localize these functions to certain regions of the brain. This process (“decomposition and localization”) is not reductionist in the sense of prioritizing molecular biology, but it is reductionist in offering a mechanistic explanation for a disorder.

As regards this claim about priority, Murphy privileges neither molecular genetics nor biochemistry. For Murphy, cognitive neuroscience provides us with the relevant functions of the brain. One benefit to deferring to cognitive neuroscience is that this field deploys intentional concepts in its explanations. Thus, cognitive neuroscience allows us to use mental language while still being committed to the underlying neuroscientific ontology as being what realizes those intentional states. For instance, if we were to say: “Schizophrenia is a failure in one’s ability to self-monitor,” we would be committed to there being a mechanism in the brain that realized the function of self-monitoring, and that something was going wrong with this mechanism in the patient with schizophrenia. As for exactly what this mechanism is like at either the computational or neural level, this is an as-yet unanswered empirical question. It may be that schizophrenia is a set of malfunctioning mechanisms, each realizing a narrow function, and each responsible for certain symptoms; or it may be that the malfunction is more global, a widespread flaw in a larger mechanism.

Murphy has been influenced by Marr’s “tri-level” approach to psychological explanation. Marr (using vision as an example) claims that a mental process can be described at 3 levels — the functional/task level, the algorithmic level, and the implementational level. It is important to recognize that these are 3 levels of description that all describe one and the same system. The neurobiological level is the implementational level for mental disorders. And the folk-psychological level is the intentional level. But, cognitive scientists point out that there is an intermediate level, a computational level, which involves subpersonal processing of representations. Now, it may be that neurobiology on its own fails to be fully explanatory of disorder for the reasons that I spoke of earlier — there are not enough systematic correlations between the neurobiological level and the psychological level when we consider the brains of all sufferers of a particular mental disorder. But the additional explanatory power of the functional and algorithmic levels can account for why heterogeneous neurobiological substrates can be unified into a scientifically-relevant kind: they realize a particular role or function for the
organism. And when it comes to what is “going wrong” in mental disorders, the best characterization of the problem might not lie with the neurobiology per se, but rather, with the computational processes that the neurobiology realizes.

As I stated, Murphy doesn’t privilege a particular explanatory level over the others; in fact, he believes that an explanation of mental disorder will likely involve components from all levels. But what all of these causal factors will have in common is that they all contribute to the same function/mechanism. And the function can be studied scientifically, regardless of whether it (and its dysfunctions) are multiply-realized or not.

But what are these functions/mechanisms according to Murphy, how are they to be individuated, and are they the same things as brain circuits?

3.2 Functions and dysfunctions in cognitive science

I opened this section by stating Murphy’s view in the following way: mental disorders are brain dysfunctions. Functions are undoubtedly realized in the brain and dysfunctions are undoubtedly at least associated with abnormalities in brain circuits. But although these functions may be given by neurobiology in some cases, in other cases they may require characterization by cognitive neuroscience. Of course, some mental disorders can be explained by gross insults to anatomy, or obvious electrical difficulties. On the other hand, details concerning the physical implementation of other disorders might fail to explain why those conditions are disorders unless we ask what the physical realizers are for. Thus, in order to explain why a mechanism is malfunctioning, we need to know both (1) what the mechanism is and (2) what it means for that mechanism to function.

With regard to (1), we need some justification for carving up the relevant functions/mechanisms in abnormal psychiatry in a certain way. One ad hoc maneuver for carving up functions would be to claim that for every mental disorder now in the DSM, there is a corresponding function, of which the disorder is the dysfunction. This would be a bit like arguing a priori for a set of psychological functions based in folk-psychology, or like arguing from either a cognitive science or evolutionary psychology perspective without taking into account neural architecture at all. These strategies carry no scientific clout, or at least I maintain that it is unclear how we would validate these schemes. Compare these kinds of theories with the earlier neurobiological view, on which we individuate functions on the basis of neural architecture (“there are X number of circuits, and we shall designate a function for each of them…”). This move may seem as ad hoc as the purely psychological individuation.

Murphy argues for a methodology of reflective equilibrium. Recall that he uses the term ‘decomposition’ to refer to dividing one’s psychology into a series of functions. He writes that: decomposition cannot be the autonomous analysis of the abstract tasks that minds carry out…we should reject top-down approaches to decomposition, as assumed by classical computational theories…The final decomposition of our mental life into components should be guided by the interrelation of cognitive hypotheses and physical facts…compared to the Marrian picture, this is a reductionist view that puts much tighter constraints on the relations between levels- our understanding of realization feeds back into and constrains our understanding of the abstract demands of cognition.30

In other words, although a classification of mental functions given by a purely cognitive level might help guide scientific inquiry, these classifications need to be answerable to—and modifiable by – the ensuing scientific results. There are a myriad of examples from the history of
science that demonstrate how the results of empirical research can show us that there are really
two functions where we once thought that there was just one, or vice versa. For instance,
although one might think that phobias and anxiety are variants of one another, it turns out that
fear and anxiety are governed by different functional areas of the brain. On the other hand, the
results of empirical studies can also help us link up functions that were previously believed to be
unrelated but actually have neural pathways in common. For example, OCD and motor disorders
are realized by overlapping parts of the brain, as are the odd thought processes of schizophrenia
and normal motor feedback mechanisms. When Murphy talks about his view being “more
reductionist” than Marr’s, he means that we must never cognize the intentional level as
autonomous from the lower-levels. He is not necessarily saying that we should privilege the
implementational level, just that theories of mental disorder must be accountable to this level.
Thus, if theories at the intentional level conflict with data from the algorithmic and
implementational levels, then this conflict is reason to discard them.

So now we can ask once more: what do all tokens of a type of mental disorder have in
common? The answer that I have provided in this section, via Murphy’s work, is that presumably
they all are dysfunctions of a particular functional mechanism. However, because a mechanism
might be quite neurologically complicated, these dysfunctions will likely display a high level of
physical heterogeneity. This heterogeneity explains why we need the computational and
intentional levels of explanation. In the following 2 chapters, I will explore these claims about
the relevant functions in much more detail. For now, though, I hope to have provided a reason
for thinking that the reflective equilibrium between neural circuitry and the functions that those
circuits realize shows why the natural sciences play a large explanatory role in theories of mental
disorders.

4. Insel and Murphy unified: a moderate Medical Model of mental disorders

4.1 A Medical Model allows for multi-level explanations

I interpret the main point of Murphy’s project to be that a Medical Model fits with a
plethora of explanatory patterns: the pathology could be classed in neurobiological,
neurochemical, molecular, or cognitive scientific terms, or it could make appeal to more than one
of these sciences. The dysfunctional mechanism could be locally or globally distributed across
the brain. For different disorders, subtypes might be explained by appeal to different sets of local
malfunctions or else via appeal to distinct places of insult in a more global mechanism.

Explanations can be multi-level (ie: implementational, computational, intentional). They
can also involve variables from the biological, psychological, social, and environmental levels,
so long as the role of those variables in the pathology of the function are made clear. A
“biopsychosocial” view of mental disorders was never intended as an alternative to a
“biological” view of mental disorders. In most cases, brain pathology cannot be explained just by
appeal to something suddenly going wrong just in the brain; oftentimes we have to look beyond
the skull for the relevant causes and effects.

The crucial conclusion that I push is that the need for variables from other levels does not
take away from the scientific status of the explanation. There is nothing unscientific about, for
example, the explanation that stress (a psychological variable, oftentimes the product of social
and environmental variables as well) leads to problems in the brain because it raises cortisol
levels. Some studies have also shown that stress has an epigenetic effect (whereby an
environmental variable alters gene expression) on the cortisol system. When the brains of suicide
victims were studied, for example, it was noted that “those who had been abused in childhood
had unique patterns of epigenetic tags in their brains." My point is that there is nothing unscientific about such studies, even though they are heavily focused on psychological variables; they show how the psychological variables impact the neurobiological substrate.

Murphy claims that "disease models are conceived of as templates to fill in at a number of levels of explanation, with the causal relations among levels established wherever possible." Note that the important point being made is that Medical Model explanations for mental disorders are causal. Explanations for physical disorders are also causal. Murphy is claiming that explanation of a mental disorder occurs when we can both 1) show how a multitude of combinations of etiological factors can causally converge on the relevant pathological realizer and 2) show how the pathological realizer, as the proximate cause of the symptoms, actually causes those symptoms in the typical course of the disorder. Lest these conditions sound too reductive, we need to keep in mind that the pathological realizer is some functionally-defined mechanism going wrong; it need not always be via the same neurobiology.

4.2 A Medical Model does not require fundamental explanations

Note, then, that the Medical Model involves two constraints on mental disorders being proper explananda for the natural sciences: 1) the central role of the brain, and 2) the availability of causal explanations. Although I have emphasized this point many times before, I want to be clear that a Medical Model is not a reductive model, not with regard to either of the 2 constraints listed above. I noted that a causal explanation for a mental disorder might cite multiple sources and multiple levels. Thus, it does not privilege "lower" levels, the lowest of which has traditionally (for mental disorders) been taken to be either molecular biology or neurobiology.

Murphy asserts that many people are under the false assumption that scientific explanations of phenomena, including mental disorders, are inadequate unless they are fundamental explanations. Fundamental explanations are an appeal to simplicity, an attempt to find that one etiological factor of a disorder that can account for all of its characteristics. Murphy quotes Woodward when he talks about fundamental explanations being robust. A robust explanation "is one that continues to apply when one moves from an idealization to the real world, and is thus largely unaffected by variation in context." Robustness means that regardless of what other environmental contingencies might occur, the explanation will still hold. In this case, we are trying to explain the presence of a mental disorder. If there is a fundamental explanatory factor, X, then if X is present, the probability of the mental disorder coming about is extremely high no matter what else is going on.

Examples of disorders with robust explanatory factors are Mendelian genetic disorders, one of which is Huntington’s Disease. If a person has the gene for Huntington’s Disease, that gene is going to become penetrant regardless of what happens in the person’s life (barring early death from some other cause). The same is true of many infectious disease pathogens. There is an inevitability to the causal powers of robust explanatory variables. But in addition to being robust, fundamental explanations are also multidimensional, which means that they can explain all of the effects of a disorder. Thus, the one explanatory factor would explain the presence of every symptom. Again, we can take the example of Huntington’s Disease – when we look at the molecular actions of the deficient protein that is the result of the faulty gene, we can explain all of the symptoms of the disease in terms of that one protein’s actions.

Murphy notes, however, that fundamental explanations are the exception rather than the norm, even in physical medicine. When it comes to the possibility of a disorder’s reduction to molecular genetics, there are far fewer cases than not where there is the gene for X. Rather, what
is typically seen is that there are a multitude of individually non-necessary genes with varying degrees of causal impact linked to the presence of the disorder. In addition, these genetic factors are usually not robust; most genes merely lend vulnerability and susceptibility to a certain condition, and then additional factors – other genes or environmental occurrences – are needed in order for the gene to have its effect. If the Medical Model is attempting to align mental disorders with physical disorders, then the explanations of mental disorders should not be held to standards to which even the explanations of physical disorders are not held. If it is the case that explanations for physical disorders are accepted as being scientific without being robust, the same should be true for mental disorders. Therefore, the lack of fundamental explanations for mental disorders is not a point against the Medical Model. Indeed, given the plethora of social, psychological, biological, and genetic factors that impact mental disorders, it would be surprising if there were fundamental explanations.

In addition to the causal explanatory constraint for Medical Models of mental disorders, I also noted the ontological constraint, the condition that the brain plays an important role in explaining the disorder. I do admit that this constraint may appear flimsy; after all, what qualifies an explanatory factor as being “important” or “crucial” rather than just subsidiary or only somewhat relevant? On the view at hand, the brain plays a crucial role in these explanations because the brain realizes the functions that malfunction when one is mentally disordered.

It is my hope, then, that this section has defended the two constraints put on a Medical Model of mental disorder, and furthermore, explained why couching the framework in terms of reduction – the intuitive strategy for medicine as a whole – is tangential to the topic at hand.

5. Objections to the Medical Model

5.1 Objection 1: The Medical Model does not respect the heterogeneity of etiology and symptoms pertaining to mental disorders; its explanations are either incomplete or inadequate

Murphy foresees a nosology of mental disorders built on causal explanation. I interpret him here as using “causal” in a constitutive way, since he does say that a mental disorder is a disease and that a disease “is a brain process, not caused by a brain process.” In other words, we do not have a diachronic process from cause to effect, where the cause is the brain process and the effect is the mental disorder. Rather, “the neuropathology is just what the disease amounts to.” So like Insel, Murphy believes that with more research, future diagnoses will occur by medical tests, like brain scans, that look for the presence of such pathology.

I believe that highlighting this constitutive use of the word ‘causal’ in this context – with regard to what constitutes the mental pathology – can get us around a possible objection to Murphy, which I alluded to in section 4.1. That objection goes like this: Murphy, you claim that mental disorders are caused by pathological brain processes. But isn’t this move reductionistic of you? Mental disorders are caused by all sorts of things, some of which are not biological – experiences of humiliation or trauma, for example – and some of which are not even internal to the individual – environmental factors, etc. So how can you possible assert that the brain causes a disorder?

In fact, Murphy’s picture look like this:

Causal factors (social, environmental, psychological, biological, genetic) \(\rightarrow\) Brain pathology/dysfunction (= mental disorder) \(\rightarrow\) Symptoms (social, environmental, psychological, biological, genetic). In other words, both the etiologies and symptom-sets themselves are allowed to be vastly heterogeneous and to differ from case-to-case. We are to imagine a
funneling-effect where a myriad of collections of causes from all kinds of levels converge onto a common brain dysfunction (which itself may have vastly heterogeneous neural realizations), which then “re-opens” to various collections of symptoms, depending on contextual factors and perhaps the unique etiology of this particular case. The fact that the pathology itself is taken to be generally the same across individuals says nothing about the causal processes and ensuing effects being exactly identical across cases. When Murphy claims that a Medical Model is scientific in virtue of its being able to offer a ‘causal’ explanation, he is using ‘causal’ in a diachronic way. But he is not committed to causal simplicity on either side of the pathology. And when he claims that mental disorders are caused by brain dysfunctions, he is using ‘cause’ in a constitutive way, thereby not committing himself to a claim that only lower-levels can have an etiological impact.

A distinct way of capturing the objection that the Medical Model offers inadequate or incomplete explanations of mental disorders is to claim that it leaves out the power of social judgment to define mental disorders, and that it just assumes that a science of mental disorders is forthcoming. I will investigate this theme – of objectivism vs. constructivism – more in the following chapter. But suffice it to say, if this is a problem for mental disorders, then it is a problem for physical disorders, too. In other words, even if constructivists have a point, this point need not break the mental disorder-physical disorder analogy. If anything, it would mean that we would have to rebuild our conception of medicine in general, starting with physical disorders.

5.2 Objection 2: The Medical Model fails to correctly capture the causal structure of mental disorders

In the last section, I said that Murphy offers the following causal structure for mental disorders: Causal factors (social, environmental, psychological, biological, genetic) \(\rightarrow\) Brain pathology (= mental disorder) \(\rightarrow\) Symptoms (social, environmental, psychological, biological, genetic). However, we might question whether this kind of linearity is an accurate representation of the structure of mental disorders. For instance, someone might claim that Murphy actually is privileging the biological level by placing the brain pathology in the middle of this causal pathway. Why not think that the brain pathology is either a diachronic cause or effect of something else? On this kind of view, we might conceive of mental symptoms as constitutive of the disorder. Or we might claim that there is no necessary condition that constitutes the disorder, that instead, all causal factors, regardless of what level they are at, are in a causal network with one another. Can it not be the case that symptoms might “reach back” and have a causal impact on the earlier etiological factors, or upon the brain itself?

In Chapter 4, I will explore this objection in more detail via defense of the claim that mental disorders are “homeostatic property cluster kinds.” In that chapter, I aim to show that the causal structure of mental disorders is better explained by a multi-directional multi-level net-like structure, rather than the linear structure given above. However, Murphy could make his theory compatible with the claim that mental disorders are best represented by a net-like causal structure in a couple of different ways. He might claim that the brain is still at the center of the pathology by arguing that a faulty neural mechanism is always responsible for the homeostatic clustering of symptoms. Alternatively, he might just say that so long as some neural variable has a central causal role in the network, then the network can be studied with the methods of the natural sciences. The burden on him then would be to delimit the standards on a variable’s being “central” in one of these nets, and offering some argument as to why the brain would meet these standards in all cases of mental disorder.
5.3 Objection 3: The Medical Model doesn’t respect the uniquely mental ontology of mental disorders

Someone might claim that what makes a disorder properly mental is that it has a unique mental etiology – that intentional states play key causal and maintenance roles in the disorder and that the ‘mental’ in ‘mental disorder’ does far more than just make reference to a kind of symptom. Such a person might for this reason think that natural science is not the correct arena from which to theorize about mental disorder. Mental disorders could never be just brain disorders. An example of such a person is Nomy Arpaly, who notes that the causal logics of mental states and non-mental states are distinct from one another, as the former involve “warrant,” “desirability,” “content efficaciousness,” and “reason responsiveness.”38 Typical scientific explanations offer causes, not reasons, so mental reasons (1) seem not to have a place in science, while (2) they look to play an integral role in explanations of mental disorders. Taken together, these suppositions mean that scientific explanations of mental disorders are inadequate or incomplete.

However, I do not necessarily think that Murphy would be threatened by this objection, as on his theory, the intentional level can make contributions to causal explanations. In fact, the emphasis on brain dysfunctions vs. brain abnormalities (where the norm is merely biological-statistical) – that is, the emphasis on cognitive neuroscience vs. biology as the more relevant branch of science – might just be able to offer a rejoinder to this objection. It could be that we can define the proper functioning of some system in intentional terms. A disorder might be detectable when the rational relationships between mental states describable at this intentional level break down. But again, the mere existence of this level of description does not negate the fact that there is one process happening, describable also in computational and implementational terms. In particular disorders, a given level might be more explanatorily appropriate if more of the etiology and symptoms cluster around that level. But just because a disorder favors the intentional level in this way, this does mean that the Medical Model cannot account for it.

5.4 Objection 4: The Medical Model is too inclusive

In response to the way in which I responded to the Arpaly-style objection just mentioned, one might wonder: does the Medical Model really say anything at all? Or does it purport to be able to account for every kind of explanatory factor whatsoever, thereby negating its own strength with a lack of constraints? One might after all recall the direct quote from Murphy about “filling in templates” at multiple levels of description. If we are allowed to add in any relevant causal factors – those from outside the individual, those that are intentional/reason-responsive, etc. – what makes this a scientific theory rather than a pluralistic one? I believe that Murphy would respond to this objection by once again citing the importance of the brain. Regardless of the heterogeneity of etiology and symptomology, nonetheless, there is something in common with regard to all tokens of a type of mental disorders, and all types of mental disorders, which is that there is breakdown of some functional system realized by the brain. The necessity of being able to cite this breakdown anchors the explanation. Whatever else might be said, the brain must play a crucial role in the explanation.

Of course, any materialist view can claim that mental states (including mental disorders) are token identical to brain states, without thereby allowing those brain states to be explanatory of the disorder. Presumably, Freud, who cited relevant causal factors for mental disorders in irreducibly psychological terms (‘defense mechanisms,’ ‘unconscious drives,’ etc.), thought that the mind and brain were importantly linked. An analogous situation occurs with genetics.
Because all personality traits are in some way genetically-influenced, all personality traits are demonstrated to have a degree of heritability. Even traits that are clearly not genetically-determined are heritable, i.e.: religiosity. But genetic influence is not the same as genetic determination, and neural realization is not the same as the brain’s being explanatorily relevant.

What we need to combat these kinds of objections is an appeal to the *systematicity* of the neural correlates to which we make reference. The categories at the higher level of abnormal psychiatry must line up in *systematic* ways with kinds at the lower level (whether these are brain circuits/networks or brain functions). Earlier, I stated that categories at both levels are in a state of flux, but are nonetheless in reflective equilibrium with one another. Changes at each level inform and are constrained by changes at the other. The empirical data do tentatively support the existence of such correlates. These correlations indicate that appeal to kinds at the neural level (particularly at the level of brain *functions*) can perform the traditional tasks of scientific kinds: they can support inductive generalizations, help make predictions, etc. For Murphy, it is not enough that a causal explanation provide the etiology of a disorder; it also must provide the typical course of the disorder. This is why the Medical Model is committed to more than just a minimal materialism, and why it requires more of a relationship between the mental and physical levels than just token realizations in every case.

One of the benefits of the Medical Model that I am pushing for in this paper is that it need not say the same thing about all mental disorders. This is a good thing, because ‘mental disorder’ refers to such a vast collection of disorders that it would be surprising if there was just one account of them (an appeal to some fundamental explanatory factor, for example). As Kirsten Weir writes, “a one-size-fits-all approach does not apply. Some diseases may be more purely physiological in nature.” Weir then quotes Richard McNally when he says that “certain disorders such as schizophrenia, bipolar disorder and autism fit the biological model in a very clear-cut sense.” This is because in these diseases, “structural and functional abnormalities are evident in imaging scans or during postmortem dissection.” In other words, for these more “biological” mental disorders, we can ascertain concrete things that are going wrong with the brain, just like we can do with the physiological substrates that malfunction in physical disorders. On the other hand, Weir writes that for disorders like depression and anxiety, the “biological foundation is more nebulous.” I take it that this nebulosity has to do with the heterogeneity of causal factors that we have been discussing, and also the lack of available fundamental explanatory variables (though I am skeptical that we have indeed located any such variables in schizophrenia, bipolar disorder, and autism).

In any event though, for different mental disorders, the biological causal factors (whether one or many) can as a unit be more or less fundamental in the explanation of the disorder. This spectrum is accounted for by the Medical Model, which aims to uncover causal structure at whichever levels exist. Some mental disorders will be more like physical disorders, and some disorders will be relatively less like physical disorders. This is not a problem for the Medical Model at hand. As I have stated, it is actually a virtue, since it does not purport to succeed in the impossible task of finding one explanatory framework for all mental disorders. Though some mental disorders will have one fundamental genetic or biological cause, most will not.

6. What are the explanatory alternatives to a Medical Model?

In supporting a particular kind of theory about mental disorders, it is also important to cite the alternatives to that theory. After all, if someone really thought that the Medical Model was an “anything goes” framework, they might be tempted to say that it includes all of its
alternatives. One alternative to the Medical Model that I have been supporting is the more “extreme” type of Medical Model that focuses on reduction and fundamental explanations. I have shown why I think that such a theory is not empirically tenable (but for perhaps a few disorders).

But when I think of true alternatives to the Medical Model, what comes to mind most are theories that avoid the brain altogether: for instance, traditional theories in abnormal psychology that focus wholly on the intentional level with no regard for any other level (psychoanalysis, some cognitive theories of depression), theories that cite “sickness” in something other than the individual (social psychology, family psychology), and “anti-psychiatry” theories that assert that mental disorders are constructions of social judgments, targeted towards what is taken to be deviant behavior. The following chapter will look at the explanatory possibilities for mental disorders given by those with these concerns. In essence, the chapter will serve as a further defense for the claims of the current chapter, both by exploring the downfalls of alternative conceptions in more detail and by defending a dysfunction-based theory of mental disorders from an alternate perspective.
Introduction to Chapters 2-4: There are two interpretations of the question “what is the nature of mental disorder?”

I would like to take a brief interlude to introduce the theme of the upcoming 3 chapters. Though I will give an indication of my claims in this interlude, full exposition and defense of those claims will come in the chapters themselves. In the previous chapter, I focused on questions that arose in the introduction to this work: are mental disorders just physical disorders, and do mental disorders fall under the purview of the natural sciences? In that chapter, I argued that we can indeed provide scientific explanations for mental disorders, so long as we recognize two points: 1) scientific explanation does not require reduction to molecular biology or neurobiology, and 2) scientific explanations can involve variables from “higher-level” fields, so long as the causal impact of those variables on the brain is explored.

With all of that said, I now want to look at the question: what are mental disorders? Or, what is the nature of mental disorder? These questions are ambiguous and suggest to us a couple of different tasks. One such task is conceptual: what is the meaning of the term ‘mental disorder,’ as it is currently used? This linguistic task is not one with which I will be concerned. Instead, I am interested in exploring the ontological territory of mental disorders; as we saw in Chapter 1, I have already made and supported one ontological claim: that mental disorders are among the kinds of things that can be studied by natural science and medical fields.

However, even when we read the above question as ontological, it still lends itself to two different interpretations, one structural and the other seemingly evaluative. When I say ‘structure,’ what I have in mind is causal structure, that is, the form of the causal relations holding between the various symptoms and etiological factors of a mental disorder. As an example, let’s consider clinical depression: etiological factors might include low serotonin and an under-functioning frontal cortex. Symptoms might include lowered mood, sleep disturbances, and cognitive distortions. When we ask about the causal structure of clinical depression, we wonder, for example, whether all symptoms radiate out from a common cause, or if sets of symptoms radiate from a few distinct causes. We wonder whether a particular cause is always more proximate than all others, and if the symptoms ever can engage in direct causal relations with one another. For instance, it could be that serotonin deficiency causes low mood and also causes cognitive distortions, which cause rumination, which then causes insomnia.

(i) Structure is not evaluative

With the previous example in mind, it looks as if an answer to the structural question need only be descriptive or formal and that it need not make reference to the content of the properties (the specific symptoms or etiological factors at hand). What I mean is that the causal relations can be represented schematically, by assigning variables to the symptoms and etiological factors, and then connecting them according to their causal linkages. For instance, the causal structure of traditional natural kinds, like elements of the periodic table, can be given in the following diagram, where E is an essence and P_n is a characteristic property.
This is a simple causal structure that pertains perhaps only to certain chemical kinds. But when we consider the many etiological factors and symptoms involved in mental disorders – as well as the likely interactions between symptoms – we can imagine far more complicated diagrams, with an infinity of different possible structures. I will leave the question of whether or not all types of mental disorders have the same general causal structure – and what this structure might look like – until my 4th chapter. However, my point at present is to emphasize the fact that whichever kind(s) of structure(s) that we land on for mental disorders, the structure on its own does not necessarily imply anything evaluative. Of the structure given above, for example, it makes no sense to ask whether it is good or bad. The kinds of things that exhibit this structure might themselves be good or bad according to some perspective or theory, but the structures themselves cannot be good or bad.

(ii) Evaluation

Conversely, the evaluative question of asking what makes a mental disorder a *disorder* requires some notion of wrongness or badness. My goal, then, is to delineate these two interpretations of the “what is the nature of mental disorder?” question (Structural interpretation, Disorder interpretation), offer and defend theories that answer them, and see if there might be some way to connect the two interpretations.

The literature contains comparatively more work that is directed towards what I am calling the “Disorder Interpretation” of the question “What is the nature of mental disorders?” Those who write about this interpretation try to explain what constitutes the badness of mental disorders. Traditionally, these theories have tried to provide necessary and/or sufficient conditions on something’s being a mental disorder. Thus, the theories aim to provide *definitions* of mental disorders. We can see this definitional goal by considering some examples. We have already seen people who make claims like: “Mental disorders are faulty brain circuits.” Or, “Mental disorders are brain dysfunctions.” Other purported definitions that arise in the literature are “Mental disorders are conditions of the mind that lead to high levels of disability and distress” and “Mental disorders are conditions that have received negative societal evaluation.”

One problem with trying to stake a claim in this debate about disorder is that it is not at all obvious that “what determines that a mental condition is a disorder/bad/negative?” is itself just asking one thing. Instead, it looks as if there are a variety of different dimensions at play. For instance, some theorists ask this question because they wonder about whether or not mental disorders are objectively-existing or constructed entities (are they “real”?). Other theorists are interested in the proper level of explanation of the badness of disorders – for instance, are biological, psychological, or social norms involved in some way or another? Yet others focus on the question of internalism vs. externalism, how deep of a role context is allowed to play in the constitution of disorder. These dimensions need to be separated, their ambiguous terms must be clarified, and individual theories must be held accountable to knowing where they fall on these various dimensions. This will be my aim in Chapter 2.
(iii) Structure

One might think that I have made a mistake in separating the question “what is the nature of mental disorder” into 2 interpretations (the structural vs. the disorder interpretation). In fact, I was led to thinking about the Structural Interpretation of the “What is the nature of mental disorders?” question when I had my attention drawn to the fact that most theories answering the Disorder Interpretation push an essentialist line. In putting forth purported definitions, such theories attempt to isolate the defining feature of all mental disorders. Paradigms that attempt to make mental disorders exactly analogous to physical disorders (like the strict Medical Models that claim that mental disorders are identical to brain disorders) are particularly prone to this kind of thinking. But mental disorders are typically not so simple. For example, distinct cases of depression tend to look much more disparate than do distinct cases of tuberculosis. Two people with clinical depression might not have any symptoms in common, and their courses of illness might be entirely distinct. Furthermore, the causes of their depressions could be as different as brain injury vs. divorce. With tuberculosis, there is an expected symptom profile that at least most patients express, and it is always caused by a particular strain of mycobacteria.

What I take from the above discussion is that if the etiology of a type of mental disorder is complicatedly multifactorial and its symptomology heterogeneous, a traditional essentialist picture will be too simple to characterize the structure of mental disorders. What we then need to explore is whether or not mental disorders can still be natural kinds even if they fail to have essences.

I will talk a great deal in my fourth chapter about homeostatic property cluster kinds (HPCs). A homeostatic property cluster kind is a non-essentialist natural kind, of which biological species are the prime example. Such kinds involve a characteristic set/cluster of properties that tend to show up together and make each other’s presence more likely. A homeostatic mechanism maintains the causal structure, holding the properties together. Although a member of the kind must have some of the properties in the cluster, no one property serves as a necessary condition for inclusion in the kind. This lack of a distinct necessary condition is what might make one think that exploring possible answers to the Structural Interpretation is an entirely new way of answering the Disorder Interpretation of the question “What is the nature of mental disorders?” This is because most answers to the Disorder Interpretation attempt to offer necessary conditions on something’s being a mental disorder. I made this point before, when I said that such answers tend to be purported definitions. But what if we could give an answer that was not constrained by having to offer a necessary condition?

It is true that such a move would take us away from the limitations inherent in the traditional way of answering the Disorder Interpretation. However, there is a principled reason that I am separating the Structural and Disorder Interpretations, which has to do with the level of application. Most answers to the Disorder Interpretation intend to be general, that is, to apply universally to the kinds of things that fall under the general category ‘mental disorder.’ We ask – what do all mental disorders have in common? And then we formulate purported definitions intended to apply to all cases.

However, we need to be careful; the Structural Interpretation is not a distinct way of answering the Disorder question, because the interpretations are pointed towards different levels of categorization. The Disorder Interpretation focuses on the general kind ‘mental disorder,’ while the Structural Interpretation focuses on types (categories) of mental disorders. For example, it could be that types of mental disorders (ie: clinical depression, schizophrenia, bipolar disorder, PTSD, etc.) are homeostatic property cluster kinds, and hence cannot be given
definitions in terms of necessary and sufficient conditions. Nonetheless, on this picture, it might be that what makes all types of mental disorders (taken as a unit) disorders can in fact be so captured. The unity of the wider kind (‘mental disorder’) might be based on different principles than the unity of the narrower kinds (e.g. ‘depression’). Mental disorders, generally, might be defined as brain dysfunctions, while types of mental disorders would have structural descriptions involving the specific details of their homeostatic mechanisms and characteristic property clusters.

There is a further reason to defend the claim that the Structural Interpretation is not just a novel kind of answer to the Disorder Interpretation. We have just seen that the two interpretations cast their questions at different levels of description. But furthermore, if the designation of ‘disorder’ has anything to do with human judgment, then talking about the structure of causal relations alone will not yield insight into what makes a condition a disorder. While some of the symptoms of a disorder might look evaluative (i.e.: the disordered thinking of schizophrenia must be based on deviation from some prior notion of “ordered thinking”), the structure in which those symptoms are embedded is just descriptive/formal. Thus, the structure is one thing, and then an evaluation (either social or based on naturalistic norms) about some of the properties of that structure determines whether it is, or counts as, a mental disorder.

(iv) A specific answer to the Disorder Interpretation

So we must explore these two different interpretations of “What is the nature of mental disorder?” separately. Only then does it make sense to ask how our conclusions about them might come together. In the ensuing chapters, I would like to do just that. I will start this discussion by attempting to map the territory of existing theories of the Disorder Interpretation, along the three different dimensions that I introduced earlier (Objectivist-Constructivist, Biological-Psychological-Social, Internal-External). Then, I will talk about which considerations speak in favor of adopting theories that combine certain positions on these dimensions. In this endeavor, I will be driven by the methodology of the Medical Model, defended in Chapter 1. I will claim that this methodology leads us to tentatively accept Wakefield’s Harmful Dysfunction (HD) theory of mental disorder as the strongest contender for a theory that answers why certain conditions are mental disorders. The HD theory claims that a mental condition is a disorder when it fulfills two necessary conditions: 1) the presence of a dysfunction, and 2) the judgment that the dysfunction is bad for a person.

However, the HD theory is not without its problems, and I will explore its various problems in my third chapter. These issues mostly revolve around the theory’s understanding of the term ‘function.’ I ultimately argue that HD is untenable based on its shaky conception of the functions that purportedly malfunction in mental disorders. More specifically, we must recognize that functionality is always understood with reference to a particular environment. A function-bearer always enacts its function in some environment or context. But how closely tied to any given environment is a purported function? The primary objection that I explore (the “Mismatch Objection”) is that HD ties functions too closely to the environment in which they originally evolved, a move that has the consequence of making mental disorders mere mismatches between functioning mechanisms and unanticipated environments. Instead of rejecting the HD theory outright, I use this objection to suggest a novel reworking of this “dysfunction-based” theory of mental disorder.
My strategy is to suggest that at least some of the functions that go wrong in mental disorders should be conceptualized as “genetically-open functions.” Genetically-open functions have two features: they are programmed to undergo ontogenetic development in the fulfillment of their functions, and their end states are not genetically-reducible. Thus, the proper way to regard the environment by which we should judge the functionality of the relevant mechanisms is by considering a set of counterfactual environments. Therefore, whether someone’s psychology is functional or not is determined neither solely by how he actually is nor by how he would be with regard to the early evolutionary environment. Instead, functionality is determined with how one would be in a variety of possible situations. I will explain how this conception of functions allows us to sidestep the Mismatch Objection while retaining the strong points of a dysfunction-based theory.

In conclusion, I intend for the next 3 chapters to be taken as a unit, as they all explore different branches of the question “what is the nature of mental disorders?” I will be keeping the two interpretations separate until the end of Chapter 4.
Chapter 2: Traditional dimensions in the Disorder Interpretation of the ontological question: “What is the nature of mental disorders?” Making a case for the Harmful Dysfunction Theory

1. Introduction: searching for a general theory of the badness of mental disorders

In this chapter, I will be concerned with the ontological, evaluative interpretation of the question “What is the nature of mental disorders?” As I explained in the introduction to the ensuing 3 chapters, I call this the “Disorder Interpretation,” as the ‘disorder’ in the term ‘mental disorder’ looks evaluative. That is, we think that the presence of a mental disorder indicates that something has, in some way or other, gone wrong. But what exactly secures this negative evaluation? When I dub this interpretation “evaluative,” I do not mean that anyone’s evaluation makes something the case, ie: that there is a disorder just because someone judges there to be. It is just that typically, we think that being mentally ill or having a mental disorder is a bad thing. Those who construct theories of mental disorder from within the framework of the Disorder Interpretation usually attempt to define mental disorders in terms of the nature of this badness/negativity.

To begin our inquiry, we can note that not all problems with the mind are called ‘mental disorders.’ For instance, we make mistakes in reasoning all the time, we go through hard times or “problems in living,” we act immorally. So why are these sorts of problems not considered to be disorders, while certain others are? Why, for instance, is social phobia a mental disorder while shyness and introversion are not? Which kinds of facts or judgments determine when there is actually a disorder? We have already seen some attempts at such definitions. Consider the differences between the following claims:

(a) Mental disorders are faulty brain circuits. These circuits are firing in some abnormal way (where the norm is statistical), and hence, mental disorders are identical with brain disorders. Their badness lies in the fact that neural circuitry is abnormal.

(b) Mental disorders are biological dysfunctions. Their badness lies in the fact that dysfunctional mechanisms do not perform as they are supposed to perform. Those who believe that some functions are observer-independent will say that biological dysfunctions break naturalistic norms.

(c) Mental disorders are conditions of the mind that lead to high levels of disability and distress. Their badness lies in the judgment that this much suffering is too much. These are consequentialist theories based on the harmful outcomes of having a mental disorder. It is this harm that constitutes the badness of the disorder. Hence these theories are sometimes called “Harm-based” theories of mental disorder.

(d) Mental disorders are conditions that have received negative societal evaluation. Their badness lies in the fact that they break social norms.

(e) Mental disorders are not real. Diagnosis does not indicate anything wrong with the individual. All mental disorders can be made sense of with the right amount of understanding. And so on….

The hope in posing the initial question is that we will be able to give some kind of general answer to the question of what makes certain conditions mental disorders. We are wondering if there is a universal theory or definition that captures something distinct about an entire class of entities. One might wonder about the feasibility of such a project. What reason do
we have to think that all of the conditions that fall under the higher-level kind ‘mental disorders’ have **anything** in common, when they are so heterogeneous? For an idea of the difficulty that comes from this heterogeneity, try coming up with a possible commonality between depression and Autism spectrum disorders and ADHD and addiction and schizophrenia. The project indeed looks like a losing battle.

I therefore grant the point that there might be no unifying answer for what all mental disorders have in common. But I should also note that while the current extension of the term can constrain and inform our investigation (for instance, our theorizing probably should not conflict with the DSM about quintessential mental disorders), our theorizing may end up being revisionist. What I mean by ‘revisionist’ is that a suitable theory might tell us that certain conditions that are currently considered to be mental disorders actually are **not** disorders, while conditions that are not currently considered to be mental disorders actually are **disorders**. So it may be that the conditions that end up in the extension of the revised term are not as heterogeneous as we fear. Alternatively, it may turn out that on closer empirical inspection, the current heterogeneous mix of mental disorders really **will** be shown to have some feature in common. I feel that the foregoing possibilities should temper our pessimism about the project of constructing a general account.

But even if no general account is possible, there are reasons to think that a “somewhat general” account is good enough for some purposes. Why do we pose the evaluative, ontological question about mental disorders to begin with? What do we stand to gain from answering it? Besides the obvious increase in knowledge that exploration of this question could yield, there are practical and ethical implications. Certainly, we cannot doubt that the distress associated with what we call ‘mental disorders’ is real. But we want to show something further – that the disorders themselves are real conditions. If mental disorders are recognized as real disorders validated by the natural sciences, then there is an obligation to regard them as real, to provide treatment rather than judgment. One of the greatest threats to this possibility is the charge that all theories about mental disorders are too permeated by social judgments to be scientifically acceptable. I want to give more due to constructivist critiques of Medical models. By ‘constructivist,’ I mean any of a variety of claims that state that mental disorders do not exist until people classify them as such. Presenting the shortcomings of these constructivist critiques will serve as a negative argument in favor of my conclusions in Chapter 1. The benefits to undermining constructivism are first, that mental disorders are then no longer conceptualized as flimsy entities, subject to societal whim, and second, that it will no longer be justified to call people mentally disordered just because they do not fit in to a given society.

In addition to wondering about the value of offering a general theory of mental disorders, we also might wonder what role **philosophy** can possibly play in these discussions. After all, philosophers do not diagnose individuals or perform brain scans to search for neural correlates. Therefore, it might seem irresponsible and arrogant for philosophy to believe itself capable of contributing to a either a general theory or a psychological nosology of mental disorders. But I think that a philosophical analysis can aid psychology and psychiatry in a couple of different ways: for one, it can offer a careful analysis of the work that already exists. As we have seen when trying to make sense of the question “what is the nature of mental disorder?” clarifying what the questions are actually asking (discerning the questions within the questions) is not a trivial task. And furthermore, there really are metaphysical questions on the table, relating to constitution, causation, and the role of context in mental disorders.
My goal for this chapter is to map out the conceptual space of the framework for organizing the theories that purport to answer the question of what makes a condition a mental *disorder*. I will offer three dimensions on which these theories can differ. I have based my choice of dimensions on those that appear most often in the literature, sometimes under different guises. When I initially undertook this investigation, I explored the literature in search of papers where the authors attempted to categorize different *ways* of answering the disorder question rather than offering particular theories themselves. What made this project difficult was the conflicting language that different theorists used (ie: ‘realist’ vs. ‘objectivist’ vs. ‘essentialist’, etc.) to refer to what I eventually took to be the same “dimension.” I attempted to abstract from their word choices to find the common ground. The choice of 3 (rather than 2 or 5 or…) dimensions is not entirely arbitrary, though it also serves a pragmatic purpose of yielding enough combinations for an interesting discussion without multiplying the combinations to an unmanageable level.

In addition to the messiness of the literature that directs itself towards what I am calling the Disorder Interpretation, there is also the problem of which sorts of considerations could decide between theories. After all, why create a framework for organizing the theories, if we do not plan to stake some sort of claim on which is best? But how do we choose which theory is “best”? It seems that we only choose the “best” theory with regard to a prior commitment. I argued in my first chapter for a degree of optimism in the hope of one day having a scientific psychiatric nosology – a classification based in biological etiologies. As an outgrowth of that defense, I will therefore ask: which kind of theory would best fit the goals of psychiatry, as it aims to be properly scientific? For instance, I have alluded many times to the Medical Model’s need for a theory of mental functions and their corresponding dysfunctions. This prior commitment drives me towards defending an objectivist, biological theory. These considerations will motivate my defense of the Harmful-Dysfunction theory of mental disorder, which I will critique in the following chapter. Therefore, I see this current chapter as an attempt to put order onto the heterogeneous literature related to mental disorders, before defending the benefits of the HD theory.

2. 3 dimensions on which to classify theories that answer the Disorder Interpretation of “What is the nature of mental disorder?”

In the introduction to this chapter, I mentioned that certain answers to the Disorder Interpretation of the question “what is the nature of mental disorder?” could validate disorders as real medical conditions that require treatment. Such validation would do two things: (1) destigmatize mental disorders, and (2) accelerate research on empirically-based treatments. Yet, while some theorists of mental disorder are certainly driven to legitimize mental disorders, others want to de-legitimize them, usually by demonstrating that they are somehow constructed entities. For instance, there are those who believe that *any* social etiology whatsoever automatically prevents mental disorders from being natural or objective kinds. An example of this phenomenon might be the preponderance of anorexia and bulimia in contemporary Western cultures, as compared to earlier cultures and other concurrent cultures that do not involve the same beauty norms as our society. If a disorder crops up in only one society, one might argue, how can it possibly be anything other than socially-constructed? Note, interestingly, that we do not do the same thing with physical disorders: if a disease crops up in some remote society, we are likely to think that there is an environmentally-encapsulated pathogen responsible for that disease, rather than thinking that the disease is socially-constructed, and hence not real.
There are a few important things to keep in mind about the terminology and how it can steer us wrong. For one, just because we use evaluative language when we say that mental disorders are *bad* things to have, that something has *gone wrong*, we need not be committed to such badness having its ground *only* in human or societal judgment. The norms at stake could be naturalistic. Nor does evaluation necessarily involve subjectivity. For instance, we could think that there are objective norms on what constitutes the “good life,” and disorders break those norms. So the Disorder Interpretation, in spite of the language that we must use in order to frame it, does not implicitly endorse a key role for human judgments.

Attempts to say what has gone wrong in mental disorders are typically not responses to just one question, but rather, to a series of questions along a few different dimensions. My goal in this section is to clarify this territory. I would like to discuss three different dimensions so that we may then discuss their combinations. These three dimensions will be: 1) Objectivist-Constructivist, 2) Biological–Psychological–Social, and 3) Internalist-Externalist.

2.1 Objectivist-Constructivist

This dimension marks off the most often-discussed dimension related to mental disorder. Questions involving this dimension come in many different forms: are mental disorders real things? Are they observer-independent? Are they natural kinds? Should we be realists or antirealists about mental disorder? Is there an objective fact that determines when someone has become mentally ill, or does society decide? What seems to be at issue here is whether or not mental disorders are conditions that exist independently of our methods of classification. Hence, we ask if the conditions that we call disorders exist as disorders independently of our judgments. I call this dimension Objectivist-Constructivist. The simple rendering of this distinction is that Objectivists believe that what makes a mental disorder a disorder is determined by something factual, while Constructivists think that our judgments/evaluations are the basis of something’s being a disorder. Views can be dually Objectivist and Constructivist if they maintain that facts and judgments are both necessary, but individually insufficient, for the presence of a disorder.

Note that although Constructivists believe that the ultimate arbiter of whether something is a disorder or not is a judgment, this evaluative basis does not mean that constructivists are not at liberty to claim that there are objective facts concerning the biological and psychological states of a person. They just don’t think that these facts on their own could ever make it the case that there is a disorder, without the necessary role of a negative value judgment.

2.2 Biological-Psychological-Social

This dimension is meant to be explanatory of the badness of mental disorders, and it can apply to either Objectivist or Constructivist views. An objectivist who believes that there are facts that determine the presence of a mental disorder will ask about the level at which those facts lie. A constructivist will ask towards which level judgments are made.

Folk psychology will say something like: “mental disorder is a disturbance of thought, experience, or emotion.” Someone who is naive to questions about mental disorder might then ask: what more needs to be figured out when it comes to theorizing about mental disorders – is it not enough to say that one’s psychology is disturbed? After all, we are talking about mental disorders. But this kind of response, if it stops its theorizing here, is circular – to say that someone’s psychology is disturbed is just to reiterate the claim that he is mentally disordered.
What we really need is a theory that accounts for the metaphysical underpinning of such disorders; how are they constituted, and where does their badness lie?

In trying to construct a non-circular account of mental disorder, it is therefore important to ask whether or not mental disorder can be understood in terms of something more fundamental than psychology. My use of the word ‘fundamental’ does not entail that the explanation is given in terms of a “harder” science; it just means that something else besides psychology does the explaining. The two primary candidates here are biology and sociology/society/culture. We might also push a third candidate that involves a normative or ethical conception of what the “good life” is, or what it means for a human being to flourish.⁵⁰

I stated that this dimension is intended to be explanatory of the actual badness of the disorder, meaning that it concerns constitution, not causation. We need to juxtapose this kind of understanding to etiological or symptom-based approaches. Biopsychosocial models of mental disorder have shown time and again that mental disorders have physical, mental, and social causes as well as physical, mental, and social symptoms. Do we thereby define mental disorders on the basis of their causal histories? No. Imagine 3 people who are depressed, one because she endured a brain injury, the other because he lost his partner, and the third because she suffered childhood trauma. These are 3 very different diachronic causal pathways, but these causal pathways all converge on similar pathology. Therefore, it looks like the end-state is constitutive of the disorder, while the causality is secondary. So the question that we are asking here is, what is the ground or basis of the mental disorder, which is constituted by such end-states? We know that something is going wrong, but what kind of thing or property can account for this “going wrong”? What kinds of fact or judgment can make the determination? The question is separate from a consideration of etiology.

Some theorists claim that Biology is the appropriate explanatory level for answering these questions. Biology could make its way into the definition of mental disorder in two different ways: a) the theory might focus on the presence or absence of certain biological markers (ie: faulty circuits). As we saw in Chapter 1, these neural correlates of disorder are not themselves explanatory of disorder without further functional descriptions. So alternately, the role of biology in the underpinning of disorder could be focused on b) biological dysfunctions. It is important to note that the notion of ‘biological dysfunction’ has the benefit of involving a naturalistic evaluative component. Whether or not a mechanism is malfunctioning or not is typically taken to be an objective fact according to norms given by evolutionary theory. It is also important to note that biological functions can be abstracted from their physical realizers. In other words, many kinds of neural states could realize one and the same biological dysfunction.

In contrast to thinking that mental disorders require scientific explanation of some kind, one might instead focus on society as the fundamental explanatory factor. Foucault,⁵¹ Szasz,⁵² and R.D. Lang⁵³ are the best-known proponents of theories that claim that there are no mental disorders at all, just social disapproval of certain behaviors. For these reasons, these thinkers are known as critics of psychiatry. So, one might just think that mental disorders are whatever a particular society deems to be abnormal or deviant. Less extreme social views of mental disorder focus on the context in which disorder arises, and the fact that social norms and customs make a difference to when a certain collection of behaviors is deemed to be a disorder or not (ie: the DSM’s “unless it is a culturally sanctioned response” exclusion). Sometimes the distress that a person is experiencing, though lamentable, is not enough to qualify the person as disordered. Usually, this is because the distress is understandable to the society (“she should be depressed;
her child just died.” “His angry outbursts are unfortunate, but they make sense given that he grew up in such a troubled home.”).

These examples of social constructivism, on which social judgments are the determinants of a condition’s being a disorder, should not be confused with a view on which social facts are constitutive of mental disorders. Remember that the Objectivist-Constructivist distinction applies to explanations given at the social level. For instance, some interpersonal theories of mental disorder take a family or relationship to be the unit and locus of mental disorder, with certain types of observer-independent relational patterns establishing the disorder. Another kind of view that places a strong emphasis on social facts is one that looks at the social consequences of someone’s mental functioning as a constitutive basis for his being mentally ill. Two people have the same symptoms, yet one holds a job and has social ties while the other does not. The negative social consequences of the latter person’s symptoms make it the case that he has a mental disorder while the former person does not.

Thus it is important not to confuse social facts with social judgments, which is yet another reason why the Objectivist-Constructivist dimension must be distinguished from the Biological-Psychological-Social dimension.

A note on ethical theories of mental disorder

I alluded earlier to the possibility of offering an ethical theory of mental disorder. As far as I can see, we have two options here: 1) to focus on positive theories of mental health that tell us what it means for a human being to flourish/live a good life, and to argue that mental disorder is the failure of such flourishing. Or 2) to align mental disorder directly with vice. The latter technique will also have objectivist and constructivist alternatives, depending on how objectivist the theory of vice is purported to be. A constructivist view would likely focus on the relativity of virtue and vice throughout societies, pointing out the danger in conflating the moral and medical.

An objectivist theory of mental disorder focused on virtue/vice might aim to account for all mental disorders, or it might aim to account for just some. An example of the former would be Ancient theories of mental disorder, on which mental health just is moral virtue. Both Plato and Aristotle defended some variant of the view that one has a healthy mind only when the different parts of the soul display a particular kind of unity and work in concert with one another. This unity just is an expression of human virtue, since such a person will act in the right ways at the right times for the right reasons. Lest anyone think that this Ancient view is too antiquated to count in the current debate, we might look at the way that personality disorders exemplify the Aristotelian virtues gone wrong. Recall that Aristotle aligned virtue with the non-arithmetic mean between two vices. So for instance, proper self-regard falls between narcissism and self-denigration, hence Narcissistic Personality Disorder is the manifestation of vice on an Aristotelian view.

When we look at contemporary debates regarding virtues and mental disorder, identity theories (mental disorder is moral vice) are not prominent; in fact, many people feel that ethical failures and mental health failures violate entirely different sets of norms. However, intuitions are not as clear cut with regard to this distinction: is a pedophile mentally ill? One might just say that the pedophile is immoral (some might say that the pedophile is only immoral if he acts on his urges). Other will say that the pedophile has a mental disorder. In that case, is he mentally disordered because he is immoral? As another example, is a sociopath mentally ill? Part of sociopathy is defined in ethical terms, with regard to consistently violating another’s rights. So by definition, the sociopath has a moral failing. Do we thus call him mentally disordered?
Clearly, not all moral failures indicate mental disorders, but it looks like we should probably account for why some do. I argue that an objectivist theory based on psychological and/or biological functions can account for these intuitions if the capacities for moral emotions and ethical reasoning are amongst the functions included in the theory. However, determining which kinds of failures of these functions count as mental disorders rather than immoral exhibitions – and why – will certainly take more work.

I do not aim to tackle the relationship between mental health and virtue/vice head-on. All I contend is that some failures of morality are also mental health failings, and so long as we can give a naturalistic story for our ethical capacities, an objectivist theory based on either psychological or biological explanations will suffice to include these disorders.

2.3 Internalism-Externalism

There is a conflation that we must be wary of when we talk about the Disorder Interpretation of the “what is the nature of mental disorder?” question. We must be cognizant not to confuse an ontological interpretation with a diagnostic interpretation. There is an important distinction between the kinds of facts (or judgments) that make it the case that something is a disorder versus the process used to diagnose a mental disorder (and who decides). Because one of the central parts of this entire debate is the possible role of social judgments in the determination of disorder, it is easy to confuse these two aspects. Therefore, we need to keep in mind the difference between social construction and social influence. I mentioned earlier that almost no one wants to deny the biopsychosocial causal picture of mental disorders. No doubt, mental disorders arise in a social context, and often involve social causes and social symptoms. But societal influence in this way does not mean that the disorders are themselves social constructions. It does not mean that the process of diagnosis is a process of creating disorders, or that social judgment constitutes a mental disorder. After all, there are good reasons to think both that diagnosis can get things wrong and that mental disorders exist whether anyone ever diagnoses them or not. So we need to make sure to keep clear the difference between: 1) the content of the properties that determine whether a condition is a mental disorder, and 2) the origin of the diagnosis of that condition. These may overlap and they may not.

So the internalism-externalism dimension is not intended to refer to something diagnostic, like: who decides? It is not about whether the agent herself or external agents consider her to be in a bad state. While most people with mental disorders suffer, some do not. Therefore, we cannot construct a theory based only upon cases where the individual herself seeks out help for her condition. On the other hand, as we saw with the example of bereavement, a person can suffer and yet not warrant diagnosis. Thus the dimension is not about whether internal or external evaluations are needed to diagnose; it is not a question of perspective.

Rather, this dimension is about whether or not what constitutes a disorder supervenes on the individual, or is instead individuated by and determined by (in the non-causal sense of ‘determination’) relations that the individual bears to the environment or society. These relations could include whether or not there is a function-environment match, what the environmental context is, whether or not one’s behavior is accepted or considered harmful by society’s standards, or whether one’s behavior is probabilistically expected.

A view is internalist if it views mental disorders as supervening on factors intrinsic to an individual, like neurobiological states and the presence of decontextualized symptoms and behaviors. Internalist views can admit that mental disorders have etiologies that are influenced
by external factors, but they will not see those factors as themselves constitutive of the disorder. On internalist views, the disorder’s relationship to the external world will be only causal.

3. Combinations of Dimensions

Now that I have introduced the 3 dimensions on which I am analyzing theories of mental disorder that attempt to answer the Disorder Interpretation, I will present the 12 combinations of these dimensions, and try to fit extant theories into these various combinations. It is important to note a few things about this discussion of combinations. First of all, I run through all of the possibilities for the sake of conceptual clarity, but some of the combinations, even if not logically impossible, are just not plausible. And two, complex views can be built up using these triplets as components, as we will see with the Harmful Dysfunction theory. To go back to the generality question that I discussed in the introduction to this chapter, we might wonder: why couldn’t some mental disorders be objective, while others are constructed? Couldn’t some mental disorders be explained by biology, while others are explained by society or psychology? Can’t some mental disorders be internalist while others are externalist? Couldn’t even one disorder incorporate all of these dimensions?

I answer yes to all of the above questions. But, recall that we are trying to see what we can come up with for a “mostly general” theory of mental disorders, in the hopes of suggesting how all of these conditions may be unified. Perhaps disorders that fall under certain of these dimensions do not have as much to suggest them as being mental disorders as do disorders that fall under other dimensions. I want to explore these possibilities now.

3.1 O-B-I: Objectivist-Biological-Internalist

An OBI theory is one on which the presence of a mental disorder is determined by intrinsic, biological facts about an individual. As an internalist picture, environment and context do not play a constitutive role here; facts about mental disorders supervene on internal biological facts. Medical Models that make the claim that mental disorders are “broken brains” or are brain diseases/lesions are the simplest representatives of this kind of theory. A slightly less stringent Medical Model might view certain neural correlates (neurotransmitters, genes) as being sufficient for the presence of a mental disorder. As I stated in Chapter 1, however, mere presence does not explain disorder without the addition of some relevant notion of ‘function.’ A third set of theories that falls here is the one comprising Biological Dysfunction theories on which either 1) the environment is not actually important to understanding if the function is fulfilled or not, since the function stays stable over an extremely wide range of environmental possibilities, or 2) ‘function’ is defined relative to internal biological systems with no regard for the environment.

3.2 O-B-E: Objectivist-Biological-Externalist

An OBE theory is one on which the important facts with regard to mental disorders are biological facts, but these facts pertain not only to intrinsic features of the individual but also importantly involve factors outside of the individual, like environment, evolutionary history, or context. This combination is best exemplified by Biological Dysfunction views on which ‘dysfunction’ is understood etiologically or historically, with regard to a specific environment in which the function evolved. Disorder arises when there is a function-environment mismatch. A
consequence of this kind of theory is that due to the mismatch, it may be that the environment is more responsible – and more in need of intervention – than the individual.

3.3 C-B-I: Constructivist-Biological-Internalist

Any constructivist view will claim that the presence of a mental disorder is determined by a judgment rather than by facts. In this case, the material for the judgment is provided by internal biological features of an individual. My best guess for what a view like this would look like is one on which the purported biological correlate for a mental disorder is on a spectrum rather than being a categorical phenomenon – for instance, a range of neurotransmitter levels – and we make a diagnostic judgment about where the cut-off is for when someone is considered mentally disordered or not. This view is hard to imagine, because it is difficult to understand the grounds of the judgment without significantly more empirical knowledge about how neural correlates relate to mental symptoms. Any cut-point would likely be entirely arbitrary, or else circularly based on some prior theory of what it means to be biologically normal.

3.4 C-B-E: Constructivist-Biological-Externalist

This view is similar to a CBI one in that it is based on judgments concerning biological factors. However, the relevant biological factors are different here – they involve the environmental context in addition to the individual’s intrinsic properties. I believe that the best way to cognize such a view is to think about the judgments that can be made on function-environment relations. We must imagine that a judgment is made about whether or not a dysfunction is present. This kind of view might not be sustainable because it makes the idea of ‘function’ not only relative to a given environment, but potentially subjective as well.

When a biological dysfunction is based on a mismatch between a function and its environment (ie: a white moth taken from a snowy environment and put into a soot-covered forest now has dysfunctional coloration), the presence of a dysfunction is a relative fact because of the role of the environment. But it is still a fact, still objective. A C-B-E view is one on which there is no fact of the matter on whether a dysfunction exists at all; rather, we make a diagnostic decision. As with C-B-I theories, it is hard to see what the grounds for this value judgment would be if we consider only the biology. On this kind of view, all function-ascriptions would be observer-dependent, negating any hope of a theory of mental functioning based on naturalistic norms.

3.5 O-P-I: Objectivist-Psychological-Internalist

Switching gears, we have now moved into the realm of Psychology’s being the explanatory discipline for what constitutes the badness of a mental disorder. Given that this triplet is objectivist, it claims that mental disorders supervene on intrinsic, psychological facts about an individual. What are these facts? Well, they could be decontextualized psychological symptoms of the kind that the DSM cites (ie: loss of pleasure, hearing voices, manic mood, etc). Again with regard to internalism, it is not that the environment is not allowed to play an important etiological role in someone’s mental disorder; instead, nothing about the environment constitutes the disorder. Thus, the mere presence of psychological symptoms is sufficient for the existence of the disorder. The DSM (without its exclusions clauses) would fall here, if it were taken to be an actual theory.

Another kind of view that could fall under this kind would be one that claimed that mental disorders are psychological dysfunctions. This kind of theory will obviously have to offer
a view about what a “psychological function” is, as distinct from a biological function. Clearly, the level of abstraction will be a further abstraction from just the biological facts, but what does this mean? It might just mean that biological functions are multiply realizable. Or, there could be a fully autonomous psychological level. Or, it could indicate that the brain (a biological entity) is not the bearer of the function so much as the entire person is. This kind of view could begin to make sense of the claim that mental disorder is “in the brain, but not of the brain,” contrary to what certain Medical Models might push.

3.6: O-P-E: Objectivist-Psychological-Externalist

I place the ethical/flourishing views of mental health and illness in the OPE category. The application of Neo-Aristotelian ethical views to mental disorder points to a set of objective psychological capacities that are necessary for mental health. Ethically-inspired theories attempt to come up with a theory of health and then define mental disorders in terms of the absence of health, which is a different tactic than that of offering a direct definition of mental disorder. Starting with Aristotle, virtue ethicists placed the good life on objectivist ground, with its foundation in human nature. Although different theorists may disagree about what the norms of human flourishing are, they all agree that these norms are not subjective. It is also likely that such norms will not be internalist, as to talk about what makes an individual’s life a good (or even merely “decent”) human life necessarily involves the individual’s relationship to, and interaction with, the environment. As George Graham states, our psychological capacities are “enmeshed in the contexts and purposes of our lives.”

As I alluded to towards the end of section 2.2, I put this kind of view in the “psychological” explanatory realm instead of constructing a new “ethical” category, because the best worked-out views of this kind cite basic psychological capacities as the basis of such flourishing. These “basic psychological capacities” might be a subset of a more general set of psychological capacities, those that involve regard for “the ongoing commitment that we persons makes to projects or aspirations that help to give structure, meaning, and purpose to our lives.” In other words, when we ask about the relevant psychological capacities, we need to be sensitive to the reasons for our interest in those capacities. Which kinds of capacities are essential “for lives that we normally or virtually universally want to lead?” This theory looks interest-relative, but its relativity is meant to extend to all human beings, which gives it the appearance of objectivity. It also means that something like evolutionary theory on its own would never be able to provide us the ultimate list of the relevant capacities, for evolutionary theory does not “care” about our goals, or at least not those that are not survival/reproduction-based.

Graham’s list of these fundamental capacities includes: bodily/spatial self-location, historical/temporal self-location, general self/world comprehension, communication, care/commitment/emotional engagement, responsibility for self, recognition of opportunities or affordances, and conscious experience. A theory that asserts that such capacities are the underpinning of mental health is meant to be objective and universal – based fundamentally in human nature - and not dependent on social norms of a particular culture. But it is not dependent on the entities described by the natural sciences, either.

3.7: C-P-I Constructivist-Psychological-Internalist

Again recall that constructivist views base the determination of mental disorder on judgments rather than on facts. In the CPI case, these judgments concern internal psychological properties of a person. A view like this would be one where we looked at a person’s
decontextualized psychological symptoms and made an evaluation about whether they were, say, harmful or bad enough to warrant our calling them pathological. The amount of harm or negative consequences involved with a bout of purported mental illness is a factual matter. The judgment comes when we draw a line between amounts of harm and suffering that indicate mental disorder and those that do not. And this judgment looks to be constructed on the norms of a given culture. For instance, more stoic cultures may say that a certain degree of suffering is actually good for one’s character and just indicates the typical problems in living that any human confronts. As an internalist view, this kind of theory cannot look to context to ask about the justification of the level of psychological distress; certain levels are considered prima facie too much to be considered healthy, without regard to circumstantial factors.

3.8: C-P-E Constructivist-Psychological-Externalist

This kind of view is similar to the CPI view in that mental disorder is determined by the judgments that are made on someone’s psychological symptoms; however, the difference in this case is that we are dealing with an externalist view on which one has to take into account facts external to the individual, like the context within which psychological symptoms have developed. For instance, a judgment about whether someone is clinically depressed or not cannot be made just by looking at the absolute value of the suffering or loss of joy that the person exhibits. We need to understand why those symptoms have come about. For instance, the DSM contains a Bereavement Exclusion Clause on which someone is not clinically depressed if she is showing all the symptoms of depression, yet these symptoms are attributable to the normal grieving process. Therefore, to try to make a judgment based on someone’s internal psychological state alone would be a huge mistake. We cannot tell if someone is disordered or not unless we are able to judge whether or not her suffering is expected, justifiable, or in proportion to her environmental situation. These standards on proportionality appear to be societally-relative.

3.9: O-S-I Objectivist-Social-Internalist

This view is one which may not be a logical possibility. Such a theory would claim that mental disorders supervene on internal, social facts about a person. But how could social facts, relational by definition, be intrinsic? I claim that they cannot. Therefore, I find no reason to go further with discussing this kind of view.

3.10: O-S-E Objectivist-Social-Externalist

On an OSE view, facts about mental disorders supervene on social facts. One possibility for this kind of theory is a Family Systems approach to mental disorder, on which the locus of mental disorder is the family, or some other relational unit. In other words, on this approach, the “individual” that suffers from a mental disorder is not an individual person, but rather, a set of people. Certain kinds of relational patterns are considered intrinsically disordered and pathological, and these patterns constitute mental disorders. For instance, Systems Theorists will look at a family with an alcoholic member and say that the entire family is sick, not just the alcoholic.

Still, an OSE view need not necessarily take a set of people to be the bearer of disorder. We could also take an individual and claim that he or she is mentally disordered on the basis of the social facts that describe that individual; these facts might include the person’s position in society, how she is regarded by society, how well she fits into a society, how successful she is,
how social she is, etc. However, this kind of theory really misses a necessary intrinsic component to mental disorder. It is not as if we can look at the social facts about someone – she is at the poverty line, she has no family and no job – and immediately know anything about her mental health status, except for the fact that she exhibits many risk factors. So these factors do not themselves constitute mental disorder.

3.11 C-S-I: Constructivist-Social-Internalist

Again, we are in the realm of judgments rather than facts now, and the judgments are made concerning internal, social features of individuals. As we saw with OSI views, the notion of properties that are both internal and social is contradictory. Views that are CS are social constructivist views – that is, psychiatry is nothing but the process of judging whether or not someone meets evaluative social norms concerning how she should be and behave. These judgments are made with regard to the individual within society, and therefore cannot be made solely with regard to intrinsic features of a person. So, we shall move on.

3.12 C-S-E: Constructivist-Social-Externalist

A CSE view is a social constructivist theory of mental disorder, and this kind of theory has often served as a harsh critique on the field of abnormal psychiatry. Foucault, Szasz, and Lang all claimed that psychiatric diagnoses were unjustified displays of control by those with more power over those with less power. What is called ‘mental disorder,’ such thinkers claim, is nothing but a judgment of social deviance or social difference.

Of course, a CSE theory need not be as sinister as the above, with regard to the intentions of a society. It may just be that we cannot avoid the fact that societies make judgments about what is normal and what is abnormal. Oftentimes, these decisions are made based on statistical norms alone; but rarity is not an indication of disorder (consider genius-level intelligence), nor is commonality an indication of health (much of a community might experience PTSD in the wake of a terrible wartime experience, for example). While it is true that repeated social problems might be an indication of a disorder, it is equally true that such problems might just be an indication of non-conformity. As Horwitz and Wakefield say, we need to preserve and account for the difference between mental disorders and other “socially disvalued conditions.” A CSE view is the most likely to make the (perhaps ambiguous) claim that “mental disorders are not real.” Actually, most constructivist views will say something of this kind, but at least those that focus their judgments on biological or psychological features take the explanatory virtues of lower-level kinds more seriously.

But why should we think that taking these lower levels seriously is even important? This question leads me into the discussion of how we might decide between these various combinations of dimensions in choosing a theory of mental disorder to endorse.

4. Choosing a theory based on combinations of dimensions: A prior commitment to the Medical Model should drive our decision-making

I have attempted to put some order on the set of theories that answer what I am calling the “Disorder Interpretation” of the question “What is the nature of mental disorders?” As a reminder, the Disorder Interpretation involves the observation that when someone has a mental disorder, somehow something has gone wrong with her – she is worse off than she otherwise
would be. The philosophical issue here is whether this negativity associated with mental disorder is something objective or otherwise naturalistic, or if it instead depends on an evaluative judgment (or perhaps both). In the preceding section, I tried to create a logical space in which to fit theories that answer this question, based on 3 dimensions. The challenges that ensue from the construction of this logical space are two-fold: 1) fitting previous theories into this space. I have performed this task with a handful of theories in the prior section, but more complex theories, as we shall see shortly with regard to the Harmful Dysfunction theory, require more resources than just one triplet provides. And 2) now that we have elucidated the triplets of combinations of dimensions, what sort of guidelines are there to help us decide which kind of theory we should endorse?

I find this question to be a very difficult methodological question. Ordinarily, when we try to decide between theories, we look at explanatory virtues like: how many explananda are accounted for, considerations of unity, alignment with truth, ability to support inductive generalizations, predictive power, etc. We can use intuitions to guide us – for example, I have mentioned that one might be tempted to use DSM diagnoses as constraints on what should and should not be included under the general term ‘mental disorder.’ I do not think that this move is generally advisable, given that the DSM is a mostly-athetoretical diagnostic tool based on descriptions rather than etiologies or constitutive claims. Thus, it looks as if we can only answer the question “which kind of theory is best?” if we have some prior goal in mind that we can apply to the question “best for what?”

As I have repeatedly stated, one of the reasons that this debate arises to begin with is that people are concerned about the “reality” of mental disorders. As such, many frame the reality question in terms of an analogy to physical disorders, assuming that most people question neither the existence of nor the badness of physical disorders. So they will ask: are mental disorders as real as physical disorders? Are they the same kind of thing as physical disorders? Are mental disorders diseases? Again, the various ways that the question is posed complicates the matter. In Chapter 1, I defended a Medical Model of mental disorder and claimed that Medical Models are unified by two features: 1) the central role of the brain, and 2) the availability of causal explanations for mental disorders. But these features on their own do not give us an answer to the current question relating to the badness of mental disorders; causal explanations need not be at a particular level of explanation, and the brain can be important at the same time as judgments about the brain and its functions can be important as well. So a defense of the Medical Model does not automatically insure the truth of objectivism, for example.

In fact, one might use the following observation to support constructivist theories of mental disorders: the classification system for mental disorders has always shown a degree of instability. Social norms about which sorts of conditions we disvalue have also shown this kind of instability. Therefore, the paralleling of these types of instability speaks on the side of a constructed understanding of the negativity associated with mental disorders.

However, an objectivist would respond to such a supposition with the claim that the instability of diagnostic systems is not an inherent feature of them; rather, it is exactly what we would expect from a system that has not yet managed to get things right. It is exactly what we would expect from a scientific field that is moving slowly towards better and more accurate explanations.

Again, when we ask which kind of theory is best, we need to qualify “best” with an answer to “for what?” Because I defended a Medical Model of mental disorders on independent grounds in Chapter 1, I’d like to ask which kind of theory (according to the dimensions that I
have given in this chapter) works best with a Medical Model. I argue that a prior commitment to the Medical Model does in fact give us reasons to endorse an objectivist theory.

4.1 Scientific considerations and objectivist intuition

I made the methodological point that in order to choose the “best theory” out of our choices, we need to defer to a prior goal. I am therefore going to make this investigation into the best theory of the Disorder Interpretation a conditional one. I will ask: if psychiatry is to fit into the natural sciences, if psychiatry is to be a branch of medicine, what kind of answer to the Disorder Interpretation must we endorse? I did mention earlier that the desire for psychiatry to be a scientific discipline is certainly not a reason to think that it is such a discipline. But empirical research has provided a foundation for the claim beyond mere desire. There are just too many neuroscientific studies and genetic studies bearing fruitful information for us to ignore the possibility. It is also worth noting that many people who write about mental disorder take this conditional for granted. For example, when Horwitz and Wakefield discuss the kind of theory of mental disorder for which we should aim, they write that “adequate accounts must not only distinguish disorder from social values but also explain in what ways disorders are real medical ailments that represent, at least in part, some objective problem in individual functioning.”

Because an Objectivist view grounds mental disorders in facts and because the natural sciences are built on facts, an acceptable scientific theory of mental disorders (and one that validates their existence) is going to have to have at least an Objectivist component. This means that the ground of the badness of disorders will itself have to lie in some kind of fact. We need to be careful to realize that constructivists do recognize the existence of scientific facts, too. Constructivism about mental disorder is not at all antithetical to materialism about the mind. To say that mental disorders have physical realizers is to say something that all materialists would agree with, including constructivist materialists. As Graham states: “the sheer physicality or neurophysical basis of a disorder does not necessarily mean that mental disorders are diseases,” where I take it he is using the word ‘disease’ to represent some faulty physiological process. A constructivist might concede that there are objective neural correlates for psychiatric symptoms without granting that those correlates are bad things or mechanisms that are “going wrong,” or anything other than useful constructs.

So the results of neuroscientific and genetic empirical studies, while providing us with some reason to think that mental disorders are observer-independent, do not get us all the way to a fully objectivist theory of the Disorder Interpretation. The reason that they fail to do so is because a naturally-occurring substrate need not automatically come with a designation of badness. Constructivists might claim that of these naturally-occurring states, only thus-and-so set is the one that contains disorders, on the basis of judgments made. A fully objectivist theory would have to provide some reason for thinking that there are naturalistic norms – observer-independent norms - that objectively determine whether something is functioning or malfunctioning. I will discuss this issue more in the following chapter.

4.2 If physical disorders are constructivist, then we cannot expect that constructivism alone could invalidate mental disorders

Objectivists say more than materialism does about the relationships between mental disorders and the brain. Citing neural correlates or genes is no explanation of why a condition is considered to be a mental disorder. Thus, as I showed in Chapter 1, Medical Models that rely on identity claims about brain circuits individuated via anatomical considerations fail to be properly
explanatory. I have argued that Medical Models can overcome this failure by presenting a theory of brain *functions* and then claiming that mental disorders are brain dysfunctions. But some constructivists claim that even this move does not take us where we need to be. They claim that the cited dysfunctions are not the locus of the *badness* of mental disorders; badness requires yet a further judgment. The constructivist would then say: because *physical* dysfunctions do not require a further judgment in order to be bad conditions, mental disorders are sufficiently *unlike* physical disorders. This lack of likeness warrants rejecting an objectivist Medical Model of mental disorders. Importantly, this manner of argumentation depends on the intuitive claim that physical disorders or diseases can be given a fully objectivist rendering. Depending on one’s previous commitments, this assumption can ground different conclusions about the Disorder Interpretation. Here is the formalized argument, which I argue is unsound:

$P_1$: The badness of physical diseases and disorders are accounted for by observer-independent facts. Thus, their badness is intrinsic to them and does not require further judgment.

$P_2$: Mental disorders are just physical disorders.

$C$: Thus, mental disorders must be objectivist.

I claim that this argument is unsound because premise 1 might just be false. We might think that to claim that a purely objectivist model of *physical* disorders (ie: their badness supervenes on objective physical facts) is wrong, thereby moving all of medicine into at least partially constructivist territory. It is just not clear that a *physical* condition can qualify as a disorder without an accompanying judgment related to its negative effects on a person. Consider infertility – infertility is a biological dysfunction, as it lowers the fitness value of its bearer. But if the bearer does not want children, can we really say that her infertility is a bad condition? The question on the table is: regardless of whether the purported disorder is mental or physical, is something’s being a disorder *ever* just a matter of mere fact? Constructivists about all disorders say no, that although there are of course always objective conditions that individuals are in, these conditions cannot be disorders or illnesses until they are judged to be bad things. Indeed, the literature on ‘disease’ is rife with just such observations. For example, oftentimes the word ‘disease’ is used to point to an objective physical condition, while ‘illness’ is used to refer to a disease that is judged to be a bad thing to have.

Thus, a proper theory of the badness of mental disorders might have both objectivist and constructivist components. The addition of the constructivism is not a reason to believe that these disorders are not a part of natural science. Consider the following argument:

$P_1$: Physical disorders are clear-cut cases of subject matter that falls under the scope of explanation by the natural sciences.

$P_2$: An account of the badness of physical disorders might need to rely on our evaluations about when a condition is bad.

$C$: Therefore, a constructivist component – or a judgment of value – is not enough to make a subject fall outside the scope of the natural sciences.

Mental disorders might be different from physical diseases, but this difference could be due to the different kinds of physical correlates and etiologies involved, rather than the latter’s being objective while the former is constructed. If physical disorders can have a constructivist component while still remaining scientific kinds, then so can mental disorders. Therefore, I conclude that mental disorders must have some objective component, and are permitted to have a constructivist component.
4.3 A theory of mental disorders must account for the significant role of harm, disability, and negative consequences; but this imperative does not necessarily mean that harm is constitutive of a disorder.

In the previous sections (and in the prior chapter), I discussed the observation that strong Medical Models of mental disorder do not provide a suitable answer to the Disorder Interpretation of the question: “What is the nature of mental disorders?” unless they also offer a theory of how biological substrates relate to the relevant functions. This observation dovetails with my claim that if the goal is to make psychiatry more like other sciences, it must have an objectivist component. Thus, an Objectivist view that provides a substantive answer to the Disorder Interpretation will need to be grounded in facts that go beyond the mere presence of a physical correlate. We need a fact that can be explanatory of badness, and the obvious concept to turn to is that of ‘Function.’

Opponents to Medical or Objectivist models might immediately point out that in my naturalistic theorizing, I am focused on the wrong thing. What is bad about mental disorders, such people argue, is just that they have negative consequences. People suffer great harms when they have mental disorders, whether they are capable of realizing that or not. Those who I will from now on refer to as “Harm Theorists” claim that such harm and suffering is constitutive of mental disorders; without these features, a condition cannot be a disorder. Such Harm Theories stretch across all dimensions: one can undergo biological, psychological, or social harms. One can undergo harm in relation to external circumstances or the harm can be more internal or endogenous, “without cause.” Harm and distress can be understood both objectively and constructively. On the objective end, it might just be a fact whether someone is harmed or suffering distress or not; on the other hand, we make evaluations about which kinds of harm, and which amounts of them, are considered unacceptable for mental health. People undergo much distress all the time without being mentally disordered. And some people with mental disorders might not experience much distress at all (narcissists, sociopaths, those experiencing a manic episode, etc.).

This constructivist rendering of harm is, however, vulnerable to charges of subjectivity. If the badness of mental disorders is taken to be constituted by disability, suffering, harm, or negative consequences, and these constitutive factors all arise from judgments, then mental disorders need not be anchored by anything objective. As a result, some Harm-based views of mental disorder could count virtually any mental condition as a disorder, given how ‘harm’ is construed. This observation is what drove anti-psychiatry theorists to argue that there really are no such things as mental disorders, absent judgments that those in power make.

Unfortunately, a unilateral focus on harm will not be enough for our purposes, especially when we are exploring the possibilities for a science of mental disorders. Harm is a consequence of being mentally disordered rather than being what constitutes the disorder. Therefore, what we need to do is to make explaining the harms caused by mental disorders a constraint on theory construction.

Dysfunction-based models of disease and mental disorder are naturalistic, in that they locate the explanatory factor of what makes something a disorder in natural, objective facts: whether an organ or body system is malfunctioning or not. What constitutes or determines disorder, on this kind of view, is the presence of the dysfunction. I will save a more in-depth discussion of how we should understand ‘function’ for my next chapter. But for now, my goal is to show that if we are going to offer a suitable Objectivist answer to the Disorder Interpretation,
it looks like we will need to make an appeal to biological (or psychological) dysfunctions. I make this claim because facts that concern functions are the best (and only?) candidates that could underlie the evaluative component of ‘disorder’ without appeal to an additional judgment. That is, there is something bad about dysfunctions whether or not someone ever recognizes that fact.

But what remains to be seen is whether dysfunctions have the right kind of “badness” to explain mental disorders. Some theorists claim that when we look at dysfunctions, we find many examples that although not biologically optimal, do not necessarily look like conditions that we would consider bad; that is, they are not necessarily states that “one ought not to be in.” In the rare case, this could be because they are good states to be in, but all that we really need is to show that they are neutral states to be in. For example, we might wonder if a dysfunctional stress system should really count as a mental disorder if an individual can funnel that stress into successful work. Remember also the case of infertility. Though a biological or psychological dysfunction indicates that something has gone wrong, this “going wrong” might not itself cause harm to an individual. Therefore, it is not immediately obvious that one is worse off when one has a dysfunction.

These kinds of examples – purported dysfunctions without disorder – do not meet the constraint of explaining the harmfulness of mental disorders. Thus, they need to be rejected or altered. When we are trying to understand if a dysfunction intuitively qualifies as a disorder or not, we usually look to that individual’s context. Harm makes its way back into the analysis when we consider a dysfunction to be bad when it involves a poor fit between an individual and her environment, where “poor fit” is evidenced by negative consequences. This is an externalist view, because the responsibility for the harmfulness of the condition may not be based solely in the dysfunctional mechanism within the individual. On the contrary, the problem could lie with the particular environment, or with the individual’s interactions with this environment.

When we combine the considerations in this section and the previous one, we reach the following conclusions: 1) we should support theories that have an objectivist “anchor” in a dysfunctions, 2) constructivist components do not render a subject unfit for scientific explanation, and 3) only a subset of dysfunctions might actually be disorders – those that cause harm. These considerations lead us to consider combination theories. I did say early on that the components of the dimensions need not exclude one another. That is, we do not necessarily need to pick either objectivism or constructivism. Let’s assume that a necessary condition on having a mental disorder is that one has an objective dysfunction of some kind. This condition protects against too much subjectivity. From the vantage point of this necessary condition, we can ask whether or not this Objectivist component is sufficient to explain mental disorders, or whether it must be supplemented by a constructivist component. The observation that we can come up with examples of dysfunctions that are not harmful, and where this lack of harm is itself what underlies their not being disorders, speaks in favor of a joint O-C view.

Thus, we have two choices: either we 1) admit that theories of disease, disorder, and illness will have to involve both an objective and a constructed component, that is, both facts and judgments. Or, 2) we attempt to account for harm without the addition of any constructivist tools. For this latter possibility to be an option, we would have to somehow tie the concept of ‘dysfunction’ necessarily to ‘harm.’ This task would require a theory on which the relevant mechanisms could not malfunction without necessarily causing distress or harm. Because of the difficulty of this task, I accept option (1) for now.
4.4 Internalism vs. Externalism, and Biology vs. Psychology: why we can postpone decisions on these dimensions at the moment

I have just made the case that if psychiatry is to be properly scientific and if it is to be a branch of medicine, then it will have to include an objectivist component. Then, I noted that mental disorders involve more than just this objective component, and that a constraint on a good theory is that it accounts for harm and negative consequences. Thus, a proper answer to the Disorder Interpretation of the question “What is the nature of mental disorders?” will be objectivist, and will likely turn out to be dually objectivist and constructivist. But what about the other two dimensions that I mapped out? Is there a reason to favor some of these over the others?

Let’s consider the Internalist-Externalist dimension. One reason to favor externalism is that it looks like an individual’s behavior and experience is the result of dynamic interactions with the environment. People become mentally disordered within a context, and the matching between one’s behavior/experience and one’s context is important with regard to whether he is functioning well or not. For example, if we have two people who are acting severely depressed and we are told that one just lost a child while the other got a promotion at work, we are inclined to think that the former is not mentally disordered while the latter is, although the two are internally identical in the relevant neurological ways. The benefit of externalist views is that they capture this interplay with the environment. Our functioning is intimately tied to our encounters with the world.

The benefit of internalist views, on the other hand, is that they more fully capture the intuition that with mental disorders, something is going wrong with the individual. On externalist views, an individual-in-an-environment is disordered; if you counterfactually took that individual out of that environment, he may be disordered in his new environment, or he may no longer be. This would not just be a matter of the diagnosis changing; if context partially constitutes disorder, and you take the context away, then the disorder could possibly go away as well, even though nothing about the individual changes. It might strike us as irresponsible to allow the presence of a disorder to be so counterfactually flimsy. So I believe that there are strong reasons that speak in favor of either Externalist or Internalism, and when I talk about the Harmful Dysfunction theory of mental disorder, I will talk about the variants along this spectrum.

With regard to Biological-Psychological-Social, I believe that although etiologies of mental disorder will no doubt involve social factors, and although interventions can and should target the social environment, we should not think of mental disorders as being socially-constituted. Mental disorders are neither socially constructed (this would leave objectivity out completely) nor conditions that (generally) apply to units greater than the individual.

Things get more difficult when we consider the Biological vs. Psychological, with the notion of dysfunctions in mind. First, it is not clear that biological functions and psychological functions are even two different species of function. Perhaps they are two levels of description of the same function (computational vs. implementational, for example). Or, as in the “flourishing” type models, can we discuss psychological functions autonomously from biological functions? Or is the difference to do with the bearer, where biological dysfunctions are brain dysfunctions, and psychological functions are functions of an individual person? Or is this distinction between the brain and the individual a faulty dichotomy, since proper theories of brain functions essentially make reference to an individual at the organism/person level? Again, we may be unable to offer a general account; some of the dysfunctions involved in mental disorders might be biological while others are psychological. Thus, although I have whittled down the preference
to Biological or Psychological, these two possibilities will be pitted against one another more fully when I discuss functions in more depth in the following chapter.

5. Making a case for the Harmful Dysfunction theory of mental disorder

Is there a theory of mental disorder that has both O and C components, can support either E or I variants, and stays within B and P explanatory frameworks? The best contender for such a theory is Wakefield’s Harmful-Dysfunction (HD) theory of mental disorder. Wakefield writes that a mental disorder has the following 2 components:

1) “The condition results from the inability of some internal mechanism to perform its natural function.” And,
2) “The condition causes some harm or deprivation of benefit to the person as judged by the standards of the person’s culture.”

Component (1) is taken to be a matter of objective facts – a condition is a dysfunction regardless of how, or even if, we classify it. Component (2) is constructivist and makes overt reference to cultural judgments of harm. Cultures may disagree about whether or not a dysfunction is harmful at all, and if so, how much of a given manifestation of a dysfunction is enough to qualify it as such.

The Harmful Dysfunction view of mental disorder deserves a much deeper treatment than this short paragraph offers, which is why I spend the next chapter exploring its implications in more detail. We have already seen the benefits of the HD theory throughout this chapter: its objectivist component anchors it to the natural sciences, the notion of ‘dysfunction’ is explanatory in a way that reference to mere neural correlates is not, and it can account for the role of harm. However, the theory also faces a number of problems that threaten its viability. It is to these problems that I now turn in my third chapter.
Chapter 3: Problems with the Harmful Dysfunction theory and the need for a looser conception of ‘function’: introducing Genetically-open Functions

1. Introduction: questioning the concept of ‘function’

My goal in this chapter is to determine the defects of the Harmful Dysfunction theory of mental disorders (HD) and then see if those defects can be corrected. In particular, I demonstrate that the theory needs to offer a more detailed understanding of the relevant notion of ‘function’ that it employs. In Chapter 2, I demonstrated that a theory of mental disorders that fits the constraints of a Medical Model must have an objectivist component. Dysfunction-based theories are more substantive in telling us what mental disorders are than are theories that focus solely on the harmful or distressing consequences of mental disorders. These points should be justification enough for why I am now focused on the HD theory. Though we have seen reasons to defend the HD theory, we must be aware of the serious problems that it faces. This chapter focuses on one of these problems, which I introduce as the Mismatch Objection in section (2). This objection claims that the current HD formulation is predicated on an improper or incomplete notion of the relevant functions. I will spend the remainder of the chapter focused on how a dysfunction-based theorist can get around the Mismatch Objection.

Recall that a dysfunction-based theory of mental disorder emphasizes functions that have gone wrong in some way. For simplicity’s sake, from here on out, I will refer to the purported functions that are relevant to mental disorders as “psychological functions.” My use of this term does not come with an ontological commitment to a special kind of function yet; it just picks out the supposed functions that are important in a discussion of mental disorders. Purported examples of psychological mechanisms that can malfunction are: threat-detectors, self-monitoring of body movements, self-attribution systems, loss-response mechanisms, other-minds systems, etc. Dysfunction theorists believe that psychological dysfunctions and mental disorders are either identical to one another, or that the latter are a set of the former. But what exactly do we mean by ‘function,’ when it comes to mental disorders? And which particular functions are involved in mental disorders? Are the examples given above accurate examples of psychological functions? Furthermore, are psychological functions relevantly different from the biological functions of the brain?

I would like to be clear that it is not my goal to offer a positive general theory of functions. Rather, my goal is the negative project of arguing that the understanding of ‘function’ implicit in Wakefield’s work is incomplete and inadequate as it stands. In particular, I think that Wakefield fails to consider that some (though far from all) psychological functions might be importantly different from the noncontroversial functions of biology. Examples of such noncontroversial functions are: the function of the heart is to pump the blood, the function of glucagon is to raise the concentration of glucose in the blood, etc.

Generally speaking, there are two distinct notions of ‘function’ that arise in the philosophical and scientific literature: (a) a biological notion (used in the biological sciences) and (b) an evolutionary notion (more often used in philosophy).65,66,67 According to the biological notion, a function of X is what X does to contribute to the workings of a complex system. On this view, a function is merely a capacity, or a causal power, of X. What we take the function of X to be could therefore change depending on which capacities and systems we find scientifically important or interesting; the function is observer-relative. However, this biological notion of function does not capture the intuition that a function of some trait X is more than just...
something that X does. There is teleology involved: if X’s function is Y, then X is for Y. But how is such teleology naturalistically-explained? According to the theory of natural selection, Y counts as a function of X if Y raised the fitness of organisms with trait X, explaining X’s continued presence in the population. This is an adaptationist or evolutionary notion of ‘function,’ and it is intended to be observer-independent. We can take Larry Wright’s formulation of evolutionary functions to be characteristic of this notion. He writes that:

“The function of X is Z means
(a) X is there because it does Z,
(b) Z is a consequence (or result) of X’s being there.”

In section 3, I will discuss these biological and evolutionary concepts of ‘function’ in more detail, especially with an eye to how they each understand the environment-function relation.

It is important to remember that ‘function’ – and the related concept ‘fitness’ – both apply only relative to an organism (or population) in an environment (or set of environments). In other words, you cannot just pick a trait and ask (with no further information) – what is its function, what is its fitness? So, a function is only understood relative to an environment. This relativity explains my emphasis in this chapter on a function’s ‘reference-environment’: the function is defined and identified as what it is with reference to a certain environment. But, then, what determines the relevant reference-environment for a function? And how much is allowed to change in a given environment before it stops counting as the same reference-environment? Later on, I will show that given differences in the properties of the function-bearers at hand, the kind of reference-environment can differ: some will be narrow (very little can change before the reference-environment no longer exists, and the function – defined with reference to that environment- is obsolete), while some will be more “wide” (the identity of the reference-environment is preserved across variations in many properties, and the function can be scrutinized in environments other than the one in which it arose).

To restate an earlier point that I made: my primary goal is one of critique. I mainly want to clarify what is left out of Wakefield’s understanding of ‘function.’ However, in section (4), I explore a line of thought that I believe naturally arises from the discussion of the Mismatch Objection to the HD theory. I raise the possibility that there might be a set of functions that are unlike those that are characteristic of our biology. My exploratory claim will be that the former are more “loosely” related to their reference-environments than are the latter (“loose” in the sense that their reference-environments are large sets of possible environments that can undergo much change while still remaining relevant to the function at hand). I call these special kinds of functions “genetically-open functions.” I claim that these open functions are not relevantly similar to the stereotypical functions of biological kinds (like organs and metabolic pathways in the brain). However, because this is an exploratory line, I certainly do not want to make the positive claim that all psychological functions are genetically-open.

Genetically-open functions have two key features that mark this departure from standard functions of our biology:

1) The actual implementation of the function is not fulfilled until one is able to calibrate to local conditions, those being the conditions of the current environment. However, because that environment is constantly changing, proper functioning is judged with regard to what an organism would do, given various environmental changes. Proper functioning is therefore marked by a disposition to be able to accurately calibrate to local conditions, not by the actual
calibration to any one particular environment. Thus, the reference-environment for a genetically-open function is a set of potential environments.

2) The outcomes of genetically-open functions are not genetically-reducible, meaning that they are not limited to combinations of “building blocks” that are found in the species’ genome. The endpoints of these functions tend to be complex behavioral, affective, cognitive, or motivational responses that are nonetheless not reducible to behavioral, affective, cognitive, or motivational components that are coded in our genetics. I will explain this irreducibility more by juxtaposing it to an example of a genetically-reducible but somewhat “open” function – that of the acquired immune system.

I end section 4 (and the present chapter) by raising, and attempting to dissolve, an objection to the notion of ‘genetically-open function.’ The objection claims that the notion of an ‘open function’ has no content and has subverted the notion of ‘function’ to an unacceptable degree. In response to this charge of subversion, we can do one of two things: we can reject a componential analysis of psychological functions and ask a more holistic question: what is the function of the human psyche or the human brain, or maybe, what is the function of a human being? Or, alternatively, we can maintain a componential analysis and try to find a middle ground between loosening a function and making it completely open. For this chapter’s purposes, I embrace the latter strategy.

However, before I can explore these strategies, we need to understand the motivation behind exploring the notion of ‘function’ at hand. And this move requires us to delve into the objections one might make against the Harmful Dysfunction theory of mental disorders.

2. An objection to the Harmful Dysfunction theory of mental disorder: The Mismatch Objection

Let’s rehearse the definition of the Harmful Dysfunction theory. Wakefield writes that a mental disorder has the following 2 necessary, and jointly sufficient, components:

1) “The condition results from the inability of some internal mechanism to perform its natural function.” And,
2) “The condition causes some harm or deprivation of benefit to the person as judged by the standards of the person’s culture.”

Component (1) is taken to be a matter of objective facts – a condition is a dysfunction regardless of how, or if, we classify it. Component (2) is what I have been calling “constructivist:” it is based on an evaluation or judgment of someone’s condition, relative to norms that are operative in a society at the time. In the prior chapter, I discussed why one might reasonably defend the HD theory, mainly focusing on the fact that the necessary condition of the presence of a dysfunction anchors disorders to something independent of human values. Regardless of whether or not a human evaluation then judges the dysfunction to be harmful, this anchoring means that not any condition whatsoever can count as a mental disorder. This type of view also lends itself nicely to the intuition that something is going wrong within an individual who is mentally disordered. It is not just that she has found herself in a hostile society. In some way, she (or her brain) is not functioning properly.

In spite of these points in its favor, the HD theory is not without its problems. For one thing, some opponents to the theory do not think that the concept of ‘dysfunction’ is at all relevant to mental disorders. Some people think that evolutionary biology has nothing to with
mental disorders because the disvaluing of mental disorders does not have to do with their impacts on biological fitness. Similarly, others believe that biology is an improper explanatory level for mental disorders, which we usually characterize as arising at the level of the person. I have discussed a variety of these anti-dysfunction viewpoints in the previous chapter. However, as I am ultimately looking to defend a dysfunction-based theory of mental disorders, I will focus on objections whose resolutions will allow us to refine the HD theory and make it more defensible. I will be focused on one objection in particular, which I call the Mismatch Objection. In the next section, I will lay the groundwork for getting around this objection, via a discussion of the relation between function, fitness, and environment.

I characterize the objection as follows:

**The Mismatch Objection**: Wakefield’s use of the term “natural functions” links psychological functions too closely to the Environment of Evolutionary Adaptedness, making purported cases of mental disorders vulnerable to charges of mere mismatches between a mechanism and the present environment.

Note the way that the dysfunction condition is stated in Wakefield’s HD formulation: “The condition results from the inability of some internal mechanism to perform its natural function” (emphasis mine). What is a “natural function?” My interpretation is that Wakefield’s use of the term “natural function” aligns with the earlier etiological/evolutionary definition of ‘function’ given by Wright: a natural function is an adaptation, and a “characteristic c is an adaptation for doing a task t in a population if and only if members of the population now have c because, ancestrally, there was selection for having c and c conferred a fitness advantage because it performed task t.” Therefore, we might call c the function-bearer and t its function. C’s function is t (c is for t) when doing t contributed to the survival of organisms with c in the past. The function is explained naturalistically and is hence observer-independent (this is why this component is objectivist). However, this reference to the past could mark the undoing of the application of such a theory to psychopathology.

All traits originally evolved during a certain time period in a certain environment. Biologists refer to this environment as the Environment of Evolutionary Adaptedness (EEA). Richters and Hinshaw refer to Tooby and Cosmides when they define the EEA as a “composite reference to the evolutionary history of environmental conditions and selection pressures responsible for shaping biological designs…the EEA provides a rough guide to the range of environments in which [organisms] are suited by design to function adaptively.” The EEA can differ depending on the species or trait with which we are concerned. However, for the majority of human traits, the EEA probably references a period thousands of years ago, when human life was very different than it is now. The problem, contend theorists against HD, comes when we mistakenly evaluate some function that evolved in environment-EEA with regard to how it performs in environment-N. Thus, the charge goes, the environment in which purported psychological capacities/functions evolved is very different from the modern world in which mental disorders arise and are diagnosed.

The Mismatch Objection capitalizes on this fact when it claims that “pathological” symptoms, like heightened anxiety and pessimistic cognitive patterns, were very likely adaptations with high fitness values in the EEA. When placed in current environments, their fitness values plummet, but not because there is something internally wrong with an individual or with these mechanisms. Rather, the phenotype-environment interaction as a unit is what is at fault, with the environment bearing most of the responsibility for this faultiness. It is as if we took a smoke detector, placed it near an oven that was prone to smoking, and then claimed that
the detector was dysfunctional because it went off constantly.\textsuperscript{72} The analogy here is supposed to be that the anxiety-system is a threat-detector. The fact that someone has high levels of anxiety in contemporary times is \textit{not} a problem with the person herself, but rather, an indication that the environment is just full of a great many threats that were not around when the detector originally evolved. In fact, on such a picture, the detector of an anxious person is functioning very well. However, someone with bad anxiety problems is \textit{not} mentally well. So something must be wrong with this account.

Thus, the objection goes, judging a mechanism on what it was supposed to do in an obsolete environment does not cohere with an intuitive vision of mental disorders. It seems obvious that functionality relative to the distant EEA should have no bearing on whether a person is mentally disordered now or not. So, we need to keep in mind that a proper reconstruction of a Dysfunction-based theory of mental disorders will bear the constraint of accounting for ‘function’ in terms that go beyond a fit with the presently-irrelevant EEA.

With the problem now introduced, I would like to go on to talk about strategies for overcoming the Mismatch Objection, strategies that are mostly focused around offering a novel notion of the kind of function that is at stake in mental disorders. My aim is, as stated, to introduce an amended Dysfunction-based theory that avoids the objections against the HD theory while still retaining its strongest points.

3. Function, Fitness, Environment

3.1 Important concepts

A theory of functions has to account for the role of the \textit{environment} (either biological, social, or both; and historical vs. current vs. future) with regard to understanding functions and their corresponding dysfunctions. We have seen how questions of the following kind arise: are functions defined relative to a certain environment? If so, which matters more, the present environment, or the environment in which the function originated? Which features of a given environment matter in the designation of a “reference-environment?” Do psychological functions have a different relation to the environment than do standard biological functions? If a purported dysfunction is understandable in the context in which it arises, does this understanding negate its dysfunctional nature? Thus, the environment comes up in a myriad of ways in discussions of mental disorder and/or psychological dysfunction. What we say about the role of the environment will end up impacting constitutive claims about mental disorders (like if they are constituted by external factors in addition to internal ones) as well as diagnostic considerations (like the role that context plays).

The theory of natural selection states the following: “Natural selection is differential reproduction due to differential \textit{fitness} (or differential adaptedness) within a common selective environment.”\textsuperscript{73} Hence, fitness is defined as adaptedness. Adaptedness has to do with how well an organism can respond to the demands in its environment, allowing it to then survive and reproduce. Thus, ‘fitness’ is a \textit{relational} concept. You cannot just ask what the fitness of trait \(c\) is; you must ask what the fitness of trait \(c\) is in environment \(n\).

Note that I say the fitness of \textit{trait} \(c\). Many people erroneously believe that individual organisms themselves are the bearers of a fitness value. But it is traits – or phenotypes (the physical and behavioral characteristics of an organism) – that are adaptations, hence it is phenotypes that have fitness values. It is in this way – phenotypes being adaptations with high fitness values – that we can introduce the notion of an evolutionary function-bearer. A function-
bearer is some trait or phenotype that is adapted to the relevant environment. And the function of that function-bearer is whatever the function-bearer does that is causally responsible for its increased fitness. Evolutionary/etiological/historical/selected-effects theories of function view functions as “past effects that explain the current presence of the function-bearer.” In other words, functions exist as they do now because they served as adaptations in the environment in which they evolved. This is the “evolutionary notion” of ‘function’ that I introduced at the beginning of this chapter.

As a simple example of such an adaptation, let’s imagine land organisms that had rudimentary lungs and those who did not. Having or not having lungs is a phenotypic trait. Because of the demands in the EEA- for example, the need to acquire oxygen to drive energy-heavy metabolic reactions- those who had lungs had higher viability and higher reproductive rates than those without lungs. Therefore, those with lungs were more fit, more adapted to their environments. Because oxygen-fixation was selected-for, because it was an adaptation, it gains the status of a function. Hence, the function of the lungs (in the EEA) was to capture oxygen. When the lungs fail to capture oxygen, then they have malfunctioned. Because the notion of ‘function’ is intertwined with that of ‘fitness,’ and ‘fitness’ is relational, then ‘function’ is relational too. It only makes sense to talk about something’s being a function in relation to environment.

3.2 Different kinds of fit to environment; wide vs. narrow reference-environments

Thus, on an evolutionary understanding, ‘fitness,’ ‘adaptation,’ and ‘function’ are concepts that concern the relation between a phenotype and a particular environment; they do not supervene on intrinsic facts about an organism. We can understand a function relative only to a particular environment (some say that this environment can only be the EEA), and from here on out, we will call this relevant environment the “reference-environment.” Ultimately, I am interested in the question of what the reference-environment is for psychological functions.

We can feel the strength of this relation between functions and their reference-environments by imagining lifting an organism from its current environment and placing it into various other environments. Such a move could alter the fitness value and adaptational status of its phenotypes immediately, even though all of the organism’s intrinsic properties remain constant. As an example for clarification, consider the pigmentation of moths. Consider a white moth that lives in a snowy environment. The moth’s pigment is an adaptation that raises its fitness in this environment: the white color effectively camouflages the moth from predators. The pigment has the function of providing camouflage, because providing camouflage was what was responsible for the increased fitness of white pigmentation in the evolutionary history of the organism. But take this moth out of its current environment, and place it in a forest whose leaves are covered with black soot from a nearby factory, and the pigment’s fitness value will be low relative to this environment. Thus, there is no universal survival value or fitness level associated with any given trait, devoid of context.

One might think that there are exceptions to this principle. And in fact, it can appear – deceptively - as if certain biological mechanisms or phenotypes are just prima facie adaptive, adaptive in all environments. This masking of the relational nature of biological mechanisms seems to me to be the product of the ubiquity of these particular mechanisms, combined with the facility with which we can overlook features of the environment that have been fairly stable since the beginning of evolutionary time. Such stable features may include the presence of water, oxygen, gravity, and so on. Examples of the mechanisms that I have in mind are functions that show very little variation within or across species: organ systems or species-specific
morphological features, like vertebrates having legs. Traits like being able to pump blood (if one has blood) or being able to locomote look to have universal adaptive benefit, leading us to conclude that they have environmentally-independent high fitness values. This supposed universality of fitness benefit makes it easy to overlook the fact that one could imagine environments where even if the environment did not lower fitness values for these phenotypes, it might at least make seemingly beneficial traits adaptationally neutral or irrelevant. If, for example, there arose environments where movement across land was no longer needed, this environment would make the possession of limbs adaptively irrelevant.

I claim that the reason that we overlook the relationality-to-environment of these kinds of functions is because the reference-environment is widely defined. By “wide,” I mean that the identity of the reference-environment is preserved across variations in many properties over time. In other words, for wide reference-environments, most of the changing environmental conditions, or the micro-environments that have cropped up through evolutionary time, do not change the selection pressures directed on the function at hand. Since the properties that define what I am calling a “wide” reference environment have not changed through evolutionary time, the reference-environment has subsisted through that time period. If psychological functions had this kind of reference-environment, then the kinds of example that the Mismatch Objection is predicated on would not arise. The Mismatch Objection requires there to be a significant change in the relevant environment between the EEA and now; it requires that the original reference-environment be lost.

Compare this situation with a trait whose function is understood with regard to a “narrow” reference-environment. For narrow reference-environments, very few environmental properties can change before the reference-environment no longer exists. Thus, on the opposite side of traits that look universally adaptive, we have phenotypes that are very closely linked to specific environments – fixed camouflage patterns, for instance, or beaks that are sized and shaped to match a particular food source. For these phenotypes, almost all possible alternative environments would involve a decrease in fitness because those counterfactuals would involve a loss of the original reference-environment to which the function was closely adapted. If psychological functions were like this, most people’s mental health statuses would have been extremely unstable over time. We should not be thinking of mental disorders as the kinds of things whose statuses as mental disorders can change from environment to environment.

Therefore, to foreshadow, in order to present a justified evolutionary account of mental disorders as dysfunctions, we are going to need a theory that allows for functions to be defined with regard to reference-environments that are neither wide nor narrow. The important point to recognize here is that the function would need to have some capacity to genuinely function or malfunction outside of its original reference-environment. When I say “genuine,” I mean that the malfunction would not just be blamed on environmental change; something would be going wrong with the mechanism itself.

3.3 Returning to the Mismatch Objection: the distinction between evolutionary functions and current functions

But the Mismatch Objection looks to negate the possibility of our judging the functionality of a function in any environment other than the one in which it originally evolved. The Mismatch Objection says that if functionality is a relation between a function and an environment, then apparent malfunction might be due entirely to the environment. Such a view would have the unfortunate consequence of yielding many false-positive diagnoses of mental disorders. Remember, the smoke alarms that keep going off 24/7 have nothing wrong with them if they
happen to be placed next to a smoking oven. The same, one might say, goes for someone with 
generalized anxiety in today’s over-stimulating world. I will continue to use the example of 
anxiety as I explore the options for responding to this objection.

The way that I see it, we can respond to the Mismatch Objection in 4 different ways:

<table>
<thead>
<tr>
<th>Response to Mismatch Objection</th>
<th>Dysfunction?</th>
<th>Disorder?</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>It is a mismatch</td>
<td>Yes, but not internal</td>
<td>No</td>
</tr>
<tr>
<td>(ii)</td>
<td>Non-dysfunction theory</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>(iiia)</td>
<td>Current role theory</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>(iiib)</td>
<td>Genetically-open functions</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The first (i) response is to align oneself with the smoke-detector analogy and just say that a 
dysfunction is indeed present, insofar as the function-environment unit is not working properly. 
However, the mismatch between the environment and the mechanism is to blame for the 
dysfunction. Thus, the person himself does not have a mental disorder; he is not internally 
dysfunctional.

In response to strategy (i), I claim that it is reasonable to conclude that a theory of mental 
disorders has failed if it yields the result that those with crippling generalized anxiety are 
mentally well. We do think that people have clinical anxiety disorders, in spite of the 
contributions of today’s high-paced, stress-filled society. One way of honoring these intuitions is 
to give up a commitment to a dysfunction-based theory of mental disorders. Thus, (ii), we might 
conclude that the presence of a dysfunction is actually not a necessary condition on something’s 
being a mental disorder after all. This response would assert that the people who are the 
exemplars of the Mismatch Objection are mentally disordered, but that they are disordered in the 
absence of being dysfunctional. However, this option leaves us without the anchor in objectivism 
provided by the emphasis on dysfunctions. So as someone defending a dysfunction-based theory 
of mental disorders, this option is not open to me.

So how can we insure that the anxious person in this example does have a psychological 
dysfunction, even with the apparent mismatch lurking in the background? Remember that earlier 
I said that my aim is to explore the possibility of being able to judge the functionality of a 
mechanism outside of the environment in which that mechanism originally evolved. To 
elucidate such a possibility, let’s talk about an alternate way for a smoke alarm to malfunction, 
besides the environment’s being at fault, ie: the constantly smoking oven. This alternative 
possibility is to cognize the dysfunction as being due to the function (or internal mechanism) side 
of the function-environment unit. Imagine that the alarm is over-sensitive and goes off for small 
triggers that are not threatening (ie: a person smoking a cigarette, or a person blowing out a 
candle). If we think about a person having a “threat detector” like this, then it would seem like 
we would want to say that something is going wrong within a person who experiences crippling 
anxiety over small triggers.

Richters and Hinshaw interpret Wakefield as arguing in just this way when they quote him 
saying that natural selection could have “occurred for a certain adaptive range of intensities [of 
fear and anxiety responses] but not for responses that are so intense as to be maladaptive.” 
Furthermore, they write that “a natural function of anxiety mechanisms is to respond to danger 
with intensity levels that are roughly proportionate to the degree of danger present.” Thus, the 
anxiety mechanism that evolved in the EEA is not necessarily ill-fitted to the current 
environment, at least not at the population level. And we can see this fact by observing that
although everyone in contemporary societies faces a slew of stressful situations, not everyone has anxiety attacks.

Now, on the traditional evolutionary view of functions, a mechanism cannot be counted as playing out its function unless it is in its original environment. Hence, it cannot malfunction either, without being in its original environment. But we have just seen a counterexample to this claim. For the person with anxiety based in an oversensitive threat-detection mechanism, we can justifiably claim that this mechanism is dysfunctional in the current environment, which is different from the mechanism’s EEA. The dysfunction is internal to the individual and hence blame falls on the “function” side of the “function-environment” relation. In other words, we can evaluate the functionality of the mechanism in the current environment. Therefore, in order to preserve a dysfunction-based theory of mental disorders, what we need is a theory that allows us to decouple the formulation of a mechanism’s function from the environment in which that function evolved. We need a theory on which the reference-environment of a function is not necessarily fixed to the EEA. In the table above, this move is (iii), and I propose two alternate ways to implement it.

The first way (iiia) is to adopt a “systemic” notion of ‘function,’ where what matters to defining a function is its current role in a system (this is the “Biological” notion of ‘function’ that I referred to at the start of this chapter). The second way (iiib) is to make ‘function’ relative to a suitable range of environments, a range that encapsulates historical and present environments, as well as potential future environments. For this second option, malfunctions occur not because of actual function-environment mismatches; rather, they occur because certain functions are plastic and have flexibility-of-response-to-environment built into them, not just incidentally, but as part of the function itself. Let’s first explore option (iiia).

You might ask: why are we even talking about evolution? Why does it matter what a phenotype originally evolved to do? Why does fitness, with its emphasis on reproduction, matter now, especially to psychopathology? Admittedly, there is no obvious reason why “mechanisms selected for inclusive fitness advantages” should translate in any straightforward way into psychological well-being or freedom from symptoms of anxiety, depression, or social maladaptiveness as judged by contemporary standards.”78 In other words, if mental disorders are diagnosed on the basis of someone’s mental health now, why would evolutionary considerations even be relevant? If you are in this camp, you will claim that we can answer the Mismatch Objection by claiming that these people do in fact have mental disorders because their minds are in some way dysfunctional. It is just that the notion of ‘function’ has to be changed such that it is defined with reference to the current environment rather than to the EEA.

As stated in the introduction to this chapter, there is a distinction made between two independent concepts of ‘function’: the evolutionary and the biological. While philosophers have been primarily concerned with the former, biologists who actually study organisms, including neurobiologists and psychiatrists who study brains, tend to be concerned with the latter.79 So let’s look at this different notion of ‘function.’ I ultimately argue that we should reject it.

3.4 A new notion of function: systemic/current role/a-historical

The 3 expressions after the colon above are meant to be synonyms of the same general theory of functions, on which the function of some mechanism/phenotype is its current role in an overall system. This notion of functions is the one that I have referred to earlier as “Biological.” In this section, I will first explain this concept, pointing out how it differs from the “Evolutionary” understanding of functions. Then, I will present a couple of objections against
this theory, namely that it is overly liberal and cannot account for teleology. Though I will explore ways in which the proponent of a-historic functions might try to get around these worries, I will ultimately argue that these objections overwhelm the strengths of the theory.

3.41 What is an a-historical function?
A-historical functions break the link between a function and its EEA. Instead, the reference-environment is always taken to be the current environment. This concept of ‘function’ does not appeal at all to evolutionary processes, fitness, or adaptations. Therefore, its function-attributions have no claim to being observer-independent. Rather, an a-historical function of some function-bearer X is just X’s “contribution to a capacity of a complex system.” This means that what actually counts as a mechanism’s function can change with changes in environment. For instance, the initial function of feathers in the EEA was warmth, but now feathers have the function of being used for flight.

This a-historical notion of ‘function’ is typically based in pragmatic concerns, namely, why scientists are studying a particular system – that is, which capacities of the system interest the scientists, and what the roles of parts of the system are in effecting that capacity. Privilege is not afforded to capacities on the basis of evolutionary or adaptational considerations, as ‘function’ is now observer-relative rather than observer-independent. Thus, when we ask about someone’s mental functioning, we wonder how well she is doing with regard to her current context. We are not asking how she would fare in viability and reproductive success back on the plains of the EEA. Clearly, there is something right about this move; it is not as if psychiatrists do evolutionary analyses on their patients, and if they did, we might wonder what the point of such an activity was. On the other hand, some mental disorders have enabled people to become rich and powerful, in spite of – or maybe even because of – obsessions, paranoia, and misconceptions of grandeur. So “thriving” on its own does not entail that one is in a state of good mental health.

3.42 Problems with the notion of a-historical functions
My objective in this section is to raise some problems with the notion of a-historical functions. I do so in order to suggest that we might want to solve the Mismatch Objection in some alternate way, thereby laying the groundwork for my theory of genetically-open functions. Therefore, I simply raise and explain these objections, rather than delving into how an a-historical functions theorist would respond. I grant that the dialect relating to each of these points could be pushed further, but doing so is beyond the scope of this chapter.

(a) The problem of liberalism
The a-historical use of ‘function’ is typically deployed on the side of a dysfunction-based theory of mental disorders. It allows for malfunctions to occur even if the environment has changed, since the current environment (rather than the EEA) is all that matters when characterizing a function. The definition that I provided for the systemic notion of function uses the phrase “contribution to a capacity of a complex system” (emphasis mine). The use of the indefinite article “a” here is not accidental, as the entire point of moving away from a historical function view is to be able to make function-attributions that are not necessary linked to evolution, but have instead to do with current causal roles.

The trouble with focusing on causal roles as the underpinning of functions is that all purported function-bearers (organs, mechanisms, behaviors, etc.) have multiple causal roles. Which one(s) we focus on depends on which capacities we happen to be interested in. Choice or
judgment is made first about which systems are important, and then which capacities of the components of those systems are relevant, with nothing to constrain the choice of systems or capacities except for one’s interests. This lack of constraint leads to liberal lists of function-bearers, as well as of functions: any system that does anything at all will have component parts with functions, and anything that any one of those parts does can be its function.

Sober gives an example of such liberalism, or permissiveness, when he writes: “The heart has a given weight. It contributes to the overall capacity of the organism to tip the scale at some number of pounds. Yet, it seems strange to say that a function of the heart is to weigh what it does. The worry is that the distinction between function and mere effect seems to get lost.” This liberalism conflicts with the teleological aspect of functions, the fact that there is a difference between effects that c is for, vs. accidental effects of c. The liberalism objection claims that the a-historical function theory provides no principled way to distinguish between the two. In the case of the heart, the cited “function” looks trivial.

If we prefer to have an objectivist theory, with an anchor in the biological sciences and with ontological commitments to real kinds of function-bearers with real functions, then this definition of ‘function’ looks thin indeed. I claim that it is not up for the task set for it. Remember, evolutionary functions explain the normative aspect of functions by positing biological norms that are independent of human interests or conceptual schemes. So we come back to the following constraint on a theory of functions: it must account for teleology. A function is not just what a mechanism does; it is what a mechanism should do. We should therefore not accept a theory that allows any effect or capacity to be considered a function.

Just a note: someone who supports a theory of mental disorders based on a-historical functions is well-placed to explain why our ideas of which conditions are mental disorders have changed so much over time. It is because which psychological functions we take seriously have changed so much over time. And thus, some conditions that really were mental disorders in the past are not anymore, and some conditions that were not mental disorders in the past now are. I think that this is an erroneous way to view the changing landscape of mental disorders, and I think that it is another way of embracing social constructionism about mental disorders, couched in biological theory. The reason that some mental disorders have dropped out of the DSM (ie: homosexuality) or have been added is because the turn towards Medical Models is allowing our classification schemes to align more accurately with an observer-independent reality.

(b) Liberalism and the choice of system

But the problem of liberalism does not stop with the choice of capacity, or effect, of a function-bearer. Choosing a system on which to focus is also problematic. The evolutionary notion of ‘function’ makes an implicit reference to organisms – function-bearers have functions because they were good for the survival and reproduction of organisms. The fact that a current-role theorists cannot make a similar move is a point against the theory. Kitcher is one philosopher who claims that current-role theorists are vulnerable to liberalism when choosing a particular system on which to focus. He uses the following example: “Any complex system can be subject to functional analysis. Thus we can identify…the ‘functions’ of mutant DNA sequences in the formation of tumors.” The negative outcome of the liberalism here is the choice of a system that is harmful to human beings. For tumors, mutant DNA obviously plays an important functional role. But this DNA does not have a function for the organism itself. In fact, it harms the organism. Thus, liberalism leads to functions that are trivial or even worse, harmful.
Both of these results are unacceptable if we believe that functions that are realized in human organisms are to the benefit of those organisms.

A-historical theorists might amend their theories to account for this point about systems by adding some kind of “organism-constraint” on their methodology. They could claim that the only effects of those components that count as functions are those that are tied to the utility, survival, or flourishing of organismal systems. Therefore, identifying current functions is not just an “anything goes” kind of process. A commitment to the integrity and survival of the organism drives the choice of relevant systems and effects. Peter Godfrey-Smith reiterates this point when he writes that “biologists apparently reserve ‘function’ for activities or powers which are, in some intuitive sense, helpful and constructive.”

First, let me note something strange about this outcome: with enough imagination, mental disorders themselves could be allowed to play functional roles for the benefit of the organism. Then it could be that mental disorders are not dysfunctions after all, and are rather, adaptive responses to challenging situations. For example, we could view depression as an “alarm system” that alerts a person to the fact that something is going wrong in her life. Or we might view repression as a beneficial defense-mechanism against truths that would otherwise overwhelm someone’s psychological stability. Or, we might consider the delusions of schizophrenia to be quite functional in that they take high volumes of disparate sensory data and make a (admittedly delusional) holistic story from them, saving the person from an even more confusing array of stimulation.

Something has clearly gone wrong here if we have concluded that mental disorders are good for a person, but what precisely has gone wrong? I maintain that just because a phenotype can be interpreted in a way to make it look beneficial or functional, that does not mean that on overall balance, it is not after all a bad thing to have. Some mental disorders – or more likely, symptoms of mental disorders - might indeed be said to loosely serve functions to the benefit of the person, but this is typically because something has already gone wrong. One does not need to utilize defense mechanisms like repression if she is mentally healthy. The brain does not need to scramble to come up with a “story” for unifying odd sensory data unless a malfunction has already occurred. Therefore, even if we were to grant the claim that some mental disorders perform functions, the implementation of these functions would be contingent on the presence of a prior dysfunction.

I will not spend time analyzing this move towards an “organism-constraint” on the part of the current-role theorist, but I will point out that solving the problem of liberalism via an appeal to “the good of the organism” is a strange tactic for a position that was originally created to avoid discussing evolutionary considerations like fitness or adaptation. This way of “solving” the problems of liberalism appears to me to end up just reminding us of why we needed an evolutionary notion of function to begin with.

(c) A thin conception of ‘validity’

Another danger with current-role methodology is that without careful precision about what actually counts as a function-bearer/component part of the brain, one could start to make general claims about what the brain is for based on which kinds of disorders show up in the clinic. Imagine that someone comes to the psychiatrist’s office in distress over the time he wastes watching television, and he is subsequently diagnosed with Binge-Watching-Disorder. Now, clearly, there were no televisions in the EEA, but that does not matter to the person’s functioning in the here-and-now. A theorist might say, “well, one of the current functions of the brain is to make sure that we properly regulate the time that we spend watching TV.” This scenario might
sound absurd (and if it does not, you can imagine more absurd examples), but current-role methodology allows for it. However, how we do we even know if a supposed effect of the brain is a real effect? Rather than a realistic conception of validity (“a diagnostic category is valid only if it represents a real entity”), we now have, at most, a utility-based conception of the validity of psychiatric constructs. That is, a construct (ie: “generalized anxiety disorder”) would count as validated (without ontological commitments) so long as its individuation was helpful in diagnosis and treatment.

In other words, we can carve the functional space of the brain (or psyche) up in whichever ways we deem most useful to treat people, no matter how seemingly absurd. While it is debatable whether diagnosing these behaviors as disorders stretches the meaning of ‘disorder’ beyond anything that we should accept, there is much to suggest that doing so is actually useful, insofar as there is a well-defined syndrome for which treatment programs can be created. If diagnosing these disorders ultimately helps people in a statistically-significant way, perhaps this success is all that we need for validating these constructs and calling them dysfunctions. We do not need to be committed to there actually being such distinct mechanisms somewhere in the brain.

Unfortunately, this thin notion of ‘validation’ negates the reason for turning to dysfunction-based accounts of mental disorder to begin with. In aiming for an objectivist theory of mental disorders, the hope was that science would provide us with the relevant list of “psychological functions,” grounded in something (maybe systems neuroscience or cognitive neuroscience) besides utility for a particular research program. I argue that we cannot attain such a benefit if we just focus on how a system or mechanism happens to function biologically in the current environment. Moreover, our choice of functions would then be based on subjective interests, and therefore, the space of possible dysfunctions would be constrained by those choices. This means that if our interests were different, or if the conditions that disturbed people in our society were other than what they are, we would alight on an alternative set of relevant functions of the psyche. And this outcome does not line up with the explanatory frameworks of the natural sciences.

(d) A-historical functions, summing up:

Remember what an appeal to a-historical functions was supposed to do for us: it was supposed to allow us to get around the Mismatch Objection by defining functions in terms of a non-fixed reference-environment (namely, whatever the current environment happens to be) rather than in terms of the EEA. However, my aim in this present chapter is to preserve the benefits of the HD theory by qualifying the theory in light of the objections raised against it. Chief among these benefits is the scientific explanatory structure that the HD theory provides. Dysfunction-based theories of mental disorder that utilize an a-historical notion of ‘function’ lose this crucial feature, as they are prone to liberalism and offer no account of teleology. Moreover, some of the attempts to get around these problems involve deferring to evolutionary theory, which is precisely what the theory was constructed so as not to have to do.

So, what do we do now? Recall that the general strategy for solving the Mismatch Objection, given by (iii), was to allow a function’s reference-environment to be something other than the EEA. Rather than making the current environment the reference-environment, as (iiiia) did, (iiiib) makes the reference-environment a set of possible environments. This set include both the EEA and the current environment, among other environment. The ability to fit to the members of this set, however, is a capacity of a function that itself has an evolutionary history. That is, the
flexibility itself was selected-for, in that it is part of the original function to adapt to whatever environment is encountered. This move restores a teleological account of functions. I call functions that adhere to this formulation “genetically-open functions,” and I claim that they can save Dysfunction Theories from the Mismatch Objection without falling into liberalism. I will now go on to explain genetically-open functions in more detail.

4. A new solution to the Mismatch Objection: Genetically-closed vs. genetically-open functions

The aforementioned considerations set up the problem that we now face if we still want to rest a theory of mental disorders on dysfunctions. Our constraints are: 1) to preserve some notion of evolutionary functions, in order to avoid both the liberalism and subjectivity that arise with a purely a-historical theory of functions. This goal means that we need to look beyond how a mechanism works in the current environment. 2) However, we also must avoid relating psychological functions too closely to the EEA, in order to avoid the Mismatch Objection. Nonetheless 3), we still need to maintain that adaptation and fitness are relational to a reference-environment.

In this section, I push for the recognition of a novel type of function that I believe will help us to meet these constraints. I call this novel type of function a “genetically-open function.” Genetically-open functions have two characteristic features: 1) they are programmed to undergo ontogenetic (experience-dependent) development in the fulfillment of their functions. Therefore, their reference-environments are sets of potential environments that the organism may encounter. And 2), their end states are not genetically-reducible. Genetically-open functions are possible because of the unique plasticity of the brain. This plasticity allows for changes in the nature and strengths of synaptic connections between neurons in response to experience. Therefore, all genetically-open functions are functions of the brain. However, not all functions of the brain are genetically-open, and I make no commitment to the claim that all psychological functions are genetically-open. I just suggest that some psychological functions are genetically-open.

Most generally, we can say that genetically-open functions confer flexibility of the implementation of one’s function onto the function-bearer. They are thus dispositions to respond appropriately to whatever environment happens to be encountered (think about the earlier discussion of proportionality/appropriate response). The reference-environment for a genetically-open function is a set of possible environments, but not just because being adapted to the EEA happens to allow one to be adapted to other environments in addition. Instead, the reference-environment is a set of environments because future modifications of the function relative to the local environment are constitutive of a genetically-open function.

In this section, I will first discuss how brain plasticity provides for this more open notion of ‘function’ in the current context. I will then provide a close reading of a paper by Richters and Hinshaw, from which I derive my usage of “genetically-open.” In particular, I spend time exploring what it means to not be genetically-reducible. I will end the section by confronting and answering an objection that claims that open functions are content-less and subvert the very notion of ‘function.’

4.1 Evolution of mentality

Recall that my notion of “genetically-open functions” applies only to brain functions and not to other organs, as the notion is dependent on neural plasticity. Let’s now discuss a plausible
evolutionary history for such plasticity. If evolution favors traits that endow an organism with increased chances of viability and greater reproductive success, then it would have behooved species to develop mechanisms that would allow them to respond to a range of environmental contingencies. Consider simple creatures who have fixed responses to a type of stimuli – for instance, a bacterium that has a tactile encounter with something in its path, and depending on the few properties of that thing that it can detect, either responds by moving towards the stimulus or away from it. Even in “higher” animals, we see instances of fixed-action-patterns, where a “fixed-action-pattern” is defined as “an instinctive behavioral sequence that is relatively invariant within the species and almost inevitably runs to completion.” These kinds of functions – and their outcomes - are therefore hardwired into the organism. Once they get triggered, a particular behavioral suite unfolds without the organism’s being able to stop or change this behavior partway through.

As an example, consider the egg-retrieval behavior of the graylag goose:

“Like many ground-nesting birds, if an egg becomes displaced from the nest, the greylag rolls it back to the nest with its beak. The sight of the displaced egg is the sign stimulus and elicits the egg-retrieval behaviour. If the egg is removed from the goose during the performance of egg-rolling, the bird often continues with the behavior, pulling its head back as if an imaginary egg is still being maneuvered by the underside of its beak. The greylag will also attempt to retrieve other egg-shaped objects, such as a golf ball, door knob, or even a model egg too large to have possibly been laid by the goose itself.”

The greylag goose is in the unfortunate position of not being able to change the way that its egg-retrieval function is enacted when the environment changes (ie: when the egg is no longer present). My point here is just to highlight the fact that functions that could avoid this kind of automaticity have survival value. A more functional egg-retrieval mechanism, for example, would allow the goose to monitor its environment, to receive feedback, and to halt or change its response if it turns out to be inappropriate or no longer necessary.

Thus, I make the following claim about evolution: *ceteris paribus*, functions that allow one to override simple instinctual responses and react in flexible ways to the environment have high fitness values. (Note: this does not mean that organisms could do without the instinctual responses; it simply means that supplementing them with higher, regulatory systems or more “thoughtful” systems would generally be a good thing). So, most processes that allow an organism or a system to “pull back” and take stock of the local environment are likely to have survival value.

We might even go so far as to claim that this flexibility of response or behaviors is why the brain (or the higher cortex) evolved to begin with. I recognize that the “just-so story” method of theorizing about why certain structures or behaviors evolved – not least of all making this kind of claim about the psyche or brain as a whole – is rarely justified. For instance, psychological mechanisms and consciousness might have just been side-effects for selection of generally larger brains for some reason other than the benefits of consciousness. Still, one could claim that the survival value of having a brain – and of having a human brain vs. a mammalian brain vs. a reptilian brain – is so obvious that it is just an axiom to state that psychological mechanisms have generally been selected-for.

Keeping in mind what I was saying just a bit ago about “pulling back” from any given environment, we have to consider the survival value of higher cortical features that allow
planning, foresight, and detection of environmental features that go beyond those lent by the sensory organs (for example, the ability to detect value relative to needs, that is, emotional valence). In contrast to the other organs, “the human brain…contributes to survival by adapting to rapidly changing characteristics of physical, social, and cultural environments.” Thus, I believe that our psychology has an evolutionary history, and I want to contend that the selection of psychological functions had to do with their adaptive value in providing a level of flexibility-of-response heretofore unavailable. From this contention, I will now provide a characterization of genetically-open functions, and clarify the features that distinguish them from the ordinary biological functions with which we are familiar.

4.2 Richters and Hinshaw: natural functions vs. evolutionary functions, and the introduction of openness

I am about to argue that some psychological functions are relevantly different from ordinary biological functions. This consideration is intertwined with the related thesis that non-brain-based biological functions are genetically-closed, while some psychological functions of brains are genetically-open. I borrow this distinction from a paper by Richters and Hinshaw (hereafter referred to as the singular “RH”), entitled “The Abduction of Disorder in Psychiatry.” The overall aim of RH’s paper is to argue against a dysfunction-based theory of mental disorders. In my exposition of their work, I will not explore this aspect of their paper. Instead, I will be concerned with their work insofar as it provides us with a characterization of a “genetically-open function.”

To start exploring RH’s understanding of this concept, once again recall Wakefield’s initial formulation of HD: one of the necessary conditions for being a mental disorder is that “the condition results from the inability of some internal mechanism to perform its natural function, wherein a natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mechanism.” I interpret Wakefield as intending for ‘natural function’ to refer to what the function was initially adapted for, which is just our familiar concept of an “evolutionary function,” adapted to the EEA. Thus, for Wakefield, ‘natural function’ = ‘evolutionary function.’ This is the notion of ‘natural function’ that Wright’s work offered.

However, RH makes the point that these two concepts are actually distinct and can come apart. RH believes that Wakefield relies on the following faulty premise in formulating his HD theory: the use of the concept “mental disorder is a straightforward extension to mental processes of the same general concept of disorder used in physical medicine.” I interpret RH here to be saying that although the natural functions of physical systems are the same as the evolutionary functions of these physical systems, this same equality does not hold when it comes to some brain functions. Again, the reason that these two concepts can come apart in the case of the brain is because of the brain’s plasticity. For RH, some natural functions of the brain are more than just evolutionary functions, because they have more causal history associated with them than just what their evolutionary pasts confer. Thus, some natural functions of the brain are evolutionary functions (they were selected-for), which have then been further modified by ontogenetic development to fit into some new environment. So a natural function is the evolutionary function as it is manifested when it contacts a given environment.

“Ontogenetic development” refers to changes that occur in a mechanism within the lifetime of an organism. Some evolutionary functions are therefore “genetically predisposed to continue evolving during ontogenetic time in response to environmental experiences.” I believe
that this predisposition is what the term ‘genetically programmed’ refers to (the genetics has a developmental plan built into it; there is no observer-relative “program” being enacted). Thus, an encounter with the current environment completes the mechanism’s development. The technical way of stating this process is to say that the functional mechanism becomes “calibrated to local conditions.” In other words, while evolution created the mechanism, which is coded in the genome, there was some “openness” of design in allowing the final touches of the function to occur only once it is in the relevant environment. This openness is part of the function; as I mentioned earlier on, this is a different situation from one in which an evolutionary function (that undergoes no further ontogenetic changes from its original design) just happens to do ok in the present environment.

Let me offer a simple physical example of the difference between (a) just happening to fit with environments other than the EEA and (b) actually calibrating to those new environments. Imagine an organism on a planet that is only hit with rays of light between 500-600 nm. Let’s say that because of some biological “overshooting,” this organism evolves with the same ability to detect light as us (between 400 and 700 nm.). The organism is clearly fit to its EEA. Let’s then suppose that the planetary environment changes, and light from 450-650 nm. starts to hit the planet. Then, the organism is fit to that environment as well, and to other environments that might occur within its range. Thus, the evolutionary function of the organism’s “eyes” is well-suited to non-EEA environments, because it is operative over a range of wavelengths. However, that range of detectable wavelengths stays the same throughout environmental changes. Compare this situation with one on which a creature’s eyes are equipped with the ability to calibrate to whatever local conditions happen to be found (maybe it has a “sliding” range of wavelengths that it can detect, say a relative difference of 500 nm between the lower and higher thresholds). Thus, if the light situation should change, with mainly ultraviolet wavelengths composing the rays hitting the surface of the planet, the organism’s brain will “rework” the visual system, or “slide” the range to best represent those wavelengths. And this can keep on happening as the environment changes. Both of these mechanisms allow for some flexibility. But only the second function is partially constituted by the ability to flexibly calibrate its range.

Now let’s apply this same kind of example to one involving psychological functioning related to mental disorders. Consider the earlier discussion concerning anxiety and proportionality. On this current view that distinguishes evolutionary functions from natural functions, the “threshold” of anxiety responses – that level above which physiological responses will be deployed – will be determined only once the mechanism can “sense” the kind of environment with which it is dealing. In fact, this view allow this threshold to undergo further changes as the environment changes. Moreover, being able to change the threshold with changes in environment would be part of the realization of that function. For instance, in a highly stressful environment, negative feedback will work to “heighten” the anxiety threshold. The high stress becomes “normal” and anything above that is taken to be threatening. So the person does not become crippling anxious even in a stressful environment.

This kind of “calibration” would explain why the anxiety system is not like a smoke detector, and why it is possible for a threat-detection system to malfunction when it is constantly activated even in a highly threatening environment. The threat-detection system is not intended to have just one set-point, like the smoke detector. It would count as malfunctioning because it was not properly sensitive enough to its environment to be able to calibrate itself to that environment. As a matter of fact, most people do not suffer from anxiety even with the stress of contemporary living. Empirical evidence like this suggests that certain mechanisms that were
selected-for in the EEA nonetheless are capable of being functionally evaluated in environments other than the EEA. *Because* of their intrinsic plasticity, the mechanisms themselves are held accountable for failing to conform to unforeseen environments. So to reiterate, RH couches this discussion of genetically-open functions in terms of how the brain and other biological organs differ from one another. This difference concerns both how functions are realized and how much flexibility or openness marks them. They believe that Wakefield has made a mistake with his premise that all brain functions are just like other biological functions.

Before we move on, I want to anticipate a potential misunderstanding about ontogenetic development. It would be a mistake to understand the point about “calibration to local conditions” as meaning that the realization of a function requires mere experience, just *an* environmental influence. This understanding would make the concept of ontogenetic development trivial *since all* biological functions unfold in time via development and experience in *some* environment or other. Organ systems, the paradigm examples of biological function-bearers, do not enact their functions without the proper environment, and indeed some of their malfunctions are the result of poor environmental conditions (ie: the effects on the heart of in-utero nicotine concentrations). Even the most stalwart of genetic mutations, those that will be expressed with near-100% probability (sickle cell anemia, long expansions of the Huntingtons Disease gene) still need *an* environment in order to come about. But what I have just been talking about in terms of biological functions is just the *background condition* of requiring *an* environment for the function to be realized at all. With genetically-open functions, fitting to the particular environment at hand is considered *constitutive* of the realization of the function (ie: the “anxiety detector’s” threshold being set only once the general safety-level of the environment is detected).

4.3 Genetically-closed vs. genetically-open designs

In the last section, I explained how the notions of ‘natural function’ and ‘evolutionary function’ can come apart in the case of some brain functions. This explanation was based on the claim that some brain functions have, as *part* of their functions, the requirement of calibrating themselves to local environmental conditions. Thus, the realization of natural functions requires more history than does the realization of evolutionary functions. I want to now move on from the distinction between natural and evolutionary functions towards the elucidation of the central distinction at hand: genetically-open vs. genetically-closed designs. My ultimate claim will be that some psychological functions are genetically-open, which allows us to get around the Mismatch Objection without having to accept an a-historical theory of functions.

In the introduction to this section, I wrote that genetically-open functions have two characteristic features, the latter of which is not shared by any genetically-closed functions: 1) they are genetically programmed to undergo ontogenetic development in the fulfillment of their functions, and 2) their end states are not genetically-reducible. I have just explained characteristic (1) by demonstrating how the fulfillment of some natural functions requires more causal history than just the evolutionary history. Now I want to look at (2). It will be helpful to talk about the “end states” of genetically-open functions by juxtaposing them to the end states of genetically-closed functions. This move will require us to understand genetically-closed functions. To begin doing so, I will quote RH at length on genetically-closed functions, and then offer multiple interpretations of what he means by this term. I will embolden these various interpretations and reject them until we end up with an acceptable interpretation.
About genetically-closed functions, RH says:

The “functioning designs [of most biological structures] were fine-tuned by selection pressures for reliability in producing and maintaining specific beneficial end states in unmodifiable ways under those conditions. The heart, for example, is designed to perform the unique function of pumping blood and to respond adaptively and efficiently in this capacity to changing demands on the body. The evolutionary cost of this efficiency, however, is inflexibility in the heart’s functioning design. It is genetically closed in the sense that it cannot devise novel structures (other than those encoded in its genotype) to improve pumping efficiency in response to novel environmental conditions, and it is incapable of performing functions other than those for which it is genetically designed. The structural components of the heart and their integrated performance characteristics contribute to blood pumping in ways that are completely determined by a genetically closed blueprint…the heart is the biological equivalent of a function-specific machine, genetically programmed to produce and maintain particular end states in particular ways.”

Here is a list of important, and potentially ambiguous, concepts that arise in this selection: “end states,” “unmodifiable (ways or end states),” “inflexibility,” “cannot devise novel structures,” “genetically closed blueprint,” and “function-specific machine.” Let’s now look into these concepts and see if we can interpret them in a way that aids us in figuring out exactly what is meant by “genetically-closed.” First, consider the line: genetically-closed functions “produce and maintain particular end states in particular ways.” These lines indicate to me that enacting a function involves reaching an end state in some way. So there is some kind of process that culminates in an end state. When we consider the heart, the end state is pumping blood. The “way” is via contractions of a particular kind of muscle (which is made of certain kinds of cells, has 4 chambers, and so on). We might interpret RH as saying that, in the case of the heart (and for genetically-closed functions, generally speaking), neither the end state nor the way that that end state is accomplished can be changed for a given function. For example, the heart cannot come up with a new way to pump blood and it cannot have a function other than to pump blood.

As a result of this analysis, maybe we should conclude that the hallmark of genetically-closed functions is that both the end state and the way to that end state are fixed. However, as I shall now show, RH himself offers an example of a genetically-closed function with modifiable end states: the acquired immune system. This example will thus demonstrate that the property of having fixed end states cannot itself be definitional of a function’s being genetically-closed. Exploring a more complex genetically-closed function will yield insight into the work that ‘genetic’ is doing in the terms ‘genetically open/closed functions.’

The function of the acquired immune system is to detect and fight off pathogens in the local environment. While the innate immune system responds non-specifically to threats via inflammation and other automatic processes, the acquired immune system’s function is to fight off specific pathogens by deploying antibodies targeted to the antigens of those pathogens. The reason that biologists call this system “acquired” is because the needed antibodies cannot be determined until an organism encounters the antigens of local pathogens, which evolve rapidly and might be entirely novel entities. Thus, the system must have the disposition to flexibly respond to unanticipated and unpredictable pathogens.
Why, then, is the acquired immune system considered genetically-closed? Our earlier attempt at a definition for ‘genetically-closed’ (“both the end state and the way to that end state are fixed”) does not apply to this case. After all, the process is itself predicated on unpredictability and flexibility. The system does not “know” in advance which pathogens it will encounter, so it needs a mechanism that is disposed to respond to whatever it might encounter. Therefore, the actual end states of the function – the antibodies themselves – are modifiable (changeable, given experience) and furthermore, their possibilities are infinite. This process coheres with the processes of “calibration to local conditions” and “ontogenetic development” that I discussed in the prior section. But the acquired immune system is genetically-closed. Therefore, it cannot be ontogenetic development – or mere unpredictability or flexibility of end states - on its own that marks off genetically-open functions from genetically-closed functions. Although ontogenetic development is part of a function’s being genetically-open, it is clearly not the whole story, which is why the formulation requires a second condition.

So the acquired immune system is genetically-closed in spite of exhibiting ontogenetic development and modifiable end states. What it does not seem to exhibit is a modifiable way by which those end states are produced, so perhaps this feature is the hallmark of being genetically-closed. Recall that I had mentioned the difference between (a) end states and (b) the manner/way in which those end states are produced. Even in a system as seemingly open as the acquired immune system, there is actually no flexibility in how these unforeseeable end states are produced. The acquired immune system has a fixed number of genes coding for various proteins, and then those genes are used in new combinations to create novel antibodies once the local pathogens have been encountered. It is true that there are an infinite number of possible antibodies that the organism might end up constructing. But, it is also true that the possibilities are all there in the original genetics, due to a finite number of protein “building-blocks.” One might be tempted therefore to say that being genetically-closed requires a fixed way of getting to end states, even if those end states themselves are modifiable.

Although this move brings us closer to an acceptable definition of ‘genetically-closed,’ I claim that this formulation is not yet complete. What we need is a proper explanation for why the “way” of getting to the end states must be fixed. This explanation is that the end states are the results of a process that comes from a “genetically closed blueprint.” But, what is a “blueprint?” I claim that having a genetically-closed blueprint means having genetically-reducible end states, though I must be careful to explain how I am understanding ‘reduction’ in this context.

So, what kind of reducibility is at issue? I am calling the end states of the acquired immune system “genetically reducible,” even though the environment plays a large role in determining which of the infinite dispositional possibilities for a design are actualized. This designation looks to go against a traditional usage of ‘reductionism,’ defined in terms of deduction or predictability. On this understanding of ‘reduction,’ if something is genetically-reducible, it should be predictable from the genetics alone. But I have just shown that, at least in the case of the acquired immune system, the roles of ontogenetic development and environmental variables keep us from being able to predict the final end states (the actual antibodies) from the genetics alone. To make this kind of irreducibility clearer, imagine identical twins with identical genomes. The identity of the genomes does not insure that there will be identity of antibodies in the future. And thus, one cannot derive the reduced entity (the antibody) merely from the reducing entity (the genotype). Thus, it might look like a mistake to even use the term ‘genetically-reducible.’ After all, sometimes “A is reduced to B” means that “A is fully...
explainable in terms of B.” But the presence and creation of particular antibodies is not fully explainable in terms of the genetics – appeal to environmental encounters are also integral.

However, my usage of the term ‘genetically-reducible’ has more to do with the constraints that the genetics impose on the possible end states than it has to do with predictability. It is more like a componential reduction. Although we cannot derive the final antibodies from a description of just the genome, the antibodies are nothing but different combinations of components (proteins) that each correspond to a gene. No additional components can be made, at least not without the long process of mutation-driven evolutionary change that allows for them (and this process certainly does not occur in the lifetime of a single organism). Thus, I take this to be the second feature of being a genetically-closed function: the function is genetically-reducible, in that the end states themselves contain nothing over and above the concrete components coded for in the genome. Genetically-closed functions may have modifiable end states, but these end states are genetically-reducible because the way to get to the end states is unmodifiable. I take it that this reductive, componential account is what RH was driving at when he wrote that genetically-closed functions “cannot devise novel structures (other than those encoded in its genotype).”

4.31 Genetically-open functions

How is the situation different with genetically-open functions? We can answer this question by investigating the end states of genetically-open functions. How can an end state not be genetically-reducible? With the acquired immune system, we saw that once the genetics are set in place, nothing but new combinations of basic genetically-coded components are possible. If these genetically-coded, but infinite, antibodies are the end states of the acquired immune system, what are the end states of genetically-open psychological functions? Our best answer to this question is that the end states of psychological functions are behaviors and responses. And my claim is that complex human behaviors-especially with regard to the cognitive, affective, rational, and motivational systems that go wrong in mental disorders-are just not built from genetically-coded components of “smaller” behaviors. Although everything that we end up doing is influenced by our genetics, the components of the behaviors are not always in the genetics. In the case of genetically-open functions, the genetics gives us capacities for behaviors, not constraints.

One might retort to my claim: but some behaviors are genetically-reducible (or at least biologically reducible)! For instance, we know that in response to an anxiety-provoking stimulus, the lower brain areas function to produce one of a confined set of biologically-reducible behaviors, ie: fight or flight or freeze. I call these behaviors “biologically-reducible” because although there is not a gene for flight per se, still, the pathway that leads to this behavior is an identifiable, pre-existing component of the biological mechanism for responding to threats. It is genetically-shared among all members of our species.

My immediate response to this objection is to point out that I am not claiming that no behavior can be the end state of a genetically-closed function; what I am saying is that some psychological functions are open functions because their end states are genetically-irreducible, complex behaviors. As it turns out, human beings do not respond to anxiety with only a set of biologically-coded behaviors. Instead of fighting, fleeing, or freezing, we might cry, or take a drug, or talk to a therapist, or “talk ourselves down” via some form of cognitive reasoning, or
implement any number of further strategies to deal with (or try to avoid) the experience of anxiety. I argue that the opening of such behavioral possibilities again has to do with the brain plasticity made possible by the evolution of higher brain areas.

RH highlights this point when he writes that

“the evolved systems of many higher-level brain systems do not contribute to adaptive functioning by producing and maintaining fixed, task-specific endstates…Their modifiable endstates, as it were, are subservient to the brain’s genetic equivalent of a policy instruction: to interact with the environment in ways that maximize the body’s chances of survival and inclusive fitness.”

I have already discussed why the end states of genetically-open psychological functions are modifiable, which is because there are an infinity of non-componential behavioral responses that we might employ in light of an environmental challenge. The term “task-specific” bears further explanation, though. Interestingly, after the quote given above, RH states that “moreover, the brain’s ability to maintain this ultimate end state – by formulating, evaluating, selecting, implementing, and revising its behavioral strategies based on environmental contingencies – is critically dependent on the openness of its functions design” (emphasis mine). I have italicized “ultimate end state” because I believe that this is the proper term to juxtapose to “task-specific end state.” Presumably, our basic biological functions can drive us towards multiple conflicting behaviors at once (thirst drives me to seek out water, fright of predators keeps me from doing so, and so on). The ultimate behavior that someone exhibits comes about as a result of weighing and decision-making (whether this process is conscious or subconscious, personal or subpersonal).

My interpretation of RH’s focus on “ultimate outcome” is that RH takes this ability to “formulate” and “evaluate,” etc., as the keystone to genetically-open functions. Higher mental processes allow one to pull away from the immediate responses to an environment in order to assess the situation in more detail. They allow us to deliberate, to evaluate the rationality of our responses, to change our behaviors. These abilities are not “task-specific.” Rather, they are used in the service of various specific tasks.

For instance, we do not always feel anger and then lash out; oftentimes, we feel anger, and try to interpret that anger so as to make the best choice about our behavior. I claim that once we realize the importance of higher brain functions to psychological functions, we are required to “zoom out” from lower brain areas in order to find the relevant function-bearers. The regulatory and integrative functions of the higher cortical areas are integral to the functioning of the lower mechanisms. Together, they contribute to a task-specific function. It seems that clinical anxiety becomes a problem not only when one experiences anxious feelings that are non-proportional to the stimulus, but also when higher regulation of this system becomes ineffective. The anxiety takes over without a person’s being able to do anything about it. So on such a view, it would not be right to say that the function-bearer of our threat-detector system is just a component of the limbic system; rather, cortical areas partially constitute the function-bearer as well. This consideration links up with the speculation that I offered in Chapter 1, that frontal-subcortical circuits (that “vertically” link the regulatory/rational cortex with the subcortical areas) might in fact at least partially constitute psychological function-bearers. So, again, perhaps the relevant “anxiety system” includes not just the limbic regions, but the higher regulatory regions as well.

Some disorders involve undoubtedly strange symptoms (such as the voices heard by schizophrenics, or the feeling in Capgras Syndrome that one’s loved ones have been replaced
with physically indistinguishable imposters, or the belief that one is dead that is held in those with Cotard’s delusion\textsuperscript{95}). However, part of the strangeness of these disorders has to do with the reaction or interpretation that a person has to his or her symptoms. The schizophrenic person takes at face value that the voices are real. The Capgras and Cotard’s sufferers do not question their own rationality or perceptual reliability. Thus, it appears that for some disorders, more is malfunctioning than just some function that performs a specific task; a more global, regulatory problem of reflectivity or rationality is occurring as well.

4.32 How genetically-open functions solve the Mismatch Objection

Now that we understand what genetically-open psychological functions are, let’s return to investigating how they can help us to sidestep the Mismatch Objection. This objection involved examples where the function-environment unit breaks down, supposedly because the mechanism is not in its proper reference-environment, not because there is actually an internal dysfunction. The environment is to blame, and therefore there can be no dysfunctions in the current environment. In response to this objection, strategy (iii) was to come up with a notion of ‘function’ that divorces a function from being characterized in relation to only the EEA. Now, we must ask: what is the reference-environment for a genetically-open function, if not the EEA? Recall that genetically-open functions have a more complicated causal history than do other functions. They have an evolutionary function, and part of that function is to exhibit experience-dependent ontogenetic development in environments other than their EEA. Thus, a genetically-open function malfunctions when it fails to perform its \textit{natural} function, which involves the disposition to respond flexibly to local conditions (ie: setting up a proper “threshold”) in enacting X (where X is the evolutionary function, say, threat detection). Thus, the reference-environment for a genetically-open function is a range of possible environments. The functionality, or lack thereof, of the function is judged with regard to its \textit{disposition} to fit to at least most of the environments in this range.

This rendering of the reference-environment allows us to claim that when a dysfunction occurs, the function side of the “function-environment” unit is to blame. In other words, we can conceive of the dysfunction as \textit{internal} to the individual, a problem with her dispositions. The fact that the organism is no longer in the EEA cannot be deployed as an excusing factor. This move brings us towards an \textit{internalist} theory that nonetheless retains the individual-environment fit as being important to the account.

Thus to conclude: genetically-open functions 1) are programmed to exhibit ontogenetic development and modifiability of end states, and 2) these end states are not genetically-reducible. Only the brain has the necessary plasticity to underlie such functions, so only brain functions can be genetically-open, though not \textit{all} brain functions are genetically-open. The genetically-open functions of the brain are those that couple specific tasks with higher functioning: capacities for “interpreting, evaluating, and learning from experience, and for generating, implementing, and revising behavioral strategies…”\textsuperscript{96} These higher functions can operate on their own, or they can be linked to lower brain processes, in a regulatory or integrative capacity. Many (but not all) psychological functions are genetically-open. Because of the flexibility of these genetically-open functions, their reference-environments are a set of possible environments to which they would become fitted if they were functioning properly. Dysfunctions of genetically-open functions are internal, occur when the function loses its plastic capacity to fit to a new environment, and hence do not fall prey to the Mismatch Objection.

4.4 An incomplete theory
I concede that my claims about genetically-open psychological functions are speculative and at this point, incomplete. But I do think that fruitful work could come from trying to fill in the details of the theory. I suggested the notion originally because it allows a Dysfunction theorist to get around the Mismatch Objection without falling into liberal accounts of functions. But I also suggest it because I believe that it does justice to the intuition that our psychological functioning is complex and interactive, stretching across subpersonal and personal levels. Modularity theorists and some cognitive scientists postulate that the mind is composed of entities that are analogous to “organs,” which are likely multiply realized by the brain. We can imagine that each organ has a function, and hence can malfunction, and that these malfunctions are what mental disorders are. Clearly, I do not disagree that mental disorders are dysfunctions, but I do think that the “functions” that go wrong are not always so narrowly individuated. For instance, in clinical depression, a person’s functioning seems to have fallen apart in very many different ways. If the neural correlate studies are right, and if the neural correlates of depression are globally distributed and heterogeneous, then it seems unlikely that this dysfunction occurs in anything like an “organ.” It seems unlikely that there is any such organ or any associated specific role, like merely being responsible for the hedonic tone of experience.

An acceptable Dysfunction-based theory of mental disorders is going to have to offer a way of individuating the relevant psychological functions. This constraint reaches all the way back to what we were discussing in Chapter 1 about how neural circuits might map onto brain functions, and moreover, what those brain functions should be taken to be. I admit that it is a weakness of my work that I do not offer a list of genetically-open psychological functions here. I provide only a brief account of how the notion might apply to one particular disorder (Generalized Anxiety Disorder). This problem is not unique to my theory of psychological functions, however. The epistemic situations pertaining to the identification and individuation of functions are different for biology and psychology. When it comes to biological functions, it is relatively easy to identify the most basic function-bearers. Gross anatomy suggests basic divisions to us. One can point to a certain organ (ie: the heart) and ask “what is this thing’s function?” Of course, there are more complicated biological function-bearers – metabolic pathways and so on – but nonetheless, these function-bearers are concrete biological entities or molecules. When it comes to psychology, the relevant physical substrate is no doubt the brain. But unlike the rest of biology, the partitioning of the brain into functional parts (or distinct function-bearers) is still in a state of relative infancy. Thus, my only retort to the incompleteness of my theory is to emphasize that dysfunction-based theorists who do not make reference to genetically-open functions are also far from being successful in offering a method of individuation for psychological functions.

4.5 Objection: are “open functions” really functions?

I would like to develop and respond to an objection against the notion of genetically-open functions, namely the one that says that ‘open function’ subverts the very notion of ‘function.’ When we think of functions, we typically think of specific tasks. Open a toolkit and you can identify a determinate function for each tool; in each case, you can say “the function of X is to Y.” On the other hand, if we say that “the function of X is to be flexible and do whatever is required in environment C,” we might think that this formulation has no content whatsoever. It is as if we had a tool in our toolbox that magically would conform itself to whatever needed to be done at any given moment in any given environment. It seems misleading to say that this tool
would have a function. Not to mention, it is hard to imagine what kinds of properties would underlie the miraculous nature of such a tool, since presumably some causal mechanisms would have to underlie its universal functional capabilities. Likewise, when RH claims that genetically-open functions exhibit a “variety and range of functioning capabilities” with no “fixed task-specific endpoints,” this description seems so open-ended that one has to wonder what even one of these functions would look like.

For example, consider language. What is the function of language? Although language likely has an evolutionary past, meaning that it was likely selected-for, many would want to say that there is no one function of language anymore. Language does many things for us – we use it to communicate, to entertain certain cognitions, to plan, to deliberate, and so on. Language is thus instrumental. It is used for many ends. If something is used for almost everything, then saying that it has a function looks to distort the meaning of ‘function,’ since there is no specific function to point to. The formulation “the function of language is to play whatever role an organism needs it to play,” looks quite vapid.

My first response to this objection is to yield to the example being employed and admit that perhaps language does not have a function. But I would then claim that this example is not characteristic of all psychological functions. I would say that genetically-open psychological functions are not instrumental in the same way that language is. One might immediately disagree with me and say that if anything exhibits an instrumentality of function, it is consciousness. Consciousness is used in the service of all kinds of goals, and it also adds functions (at least in the current-role, observer-dependent sense) to other organs. For instance, perhaps the evolutionary function of the hand was to manipulate tools, but conscious intentions now allow my hands to take on many other functions, including writing philosophical theories on pieces of paper.

But maybe consciousness, even if it is a part of our psychology, is not itself a psychological function-bearer. Perhaps it really just makes instrumental contributions to other, more specific functions. In spite of the remarks about both language and consciousness, I believe that the formulation of a psychological function does involve more content than just “the function of X is to do whatever is required in environment C.” Psychological functions will make reference to something contentful, as I showed with “the function of the anxiety system is to monitor and respond to threats.” This formulation has specific content because it mentions threats.

The key to understanding why genetically-open functions do not fall prey to this objection about instrumentality is to recognize that the flexibility involved in psychological functions is a flexibility of response, not a flexibility of function itself. Just because a function is open does not mean that the function changes (the anxiety system is not all of a sudden going to start controlling our responses to losses instead of monitoring threats, for example). Nor does being open mean that anything at all is possible in the fulfillment of the function. As RH writes: “mental and psychological processes are characterized by only limited degrees of design openness; all aspects of their development, functioning, and experience-based modifiability are to varying degrees determined, shaped, and constrained by genetically closed guidance mechanisms of the brain.” Thus, open functions are not like all-purpose magical tools.

Still, someone might push for the instrumentality objection in a slightly different manner by picking up on what I was saying at the end of the last section about the widened grain of the function-bearers of psychological functions. Recall that I suggested that the biological circuits underlying psychological functions likely stretch between cortical and subcortical regions, accommodating the contributions of higher-level processes like regulation and integration. An
objecor might argue – why not widen this grain indefinitely? Perhaps there is only one open function, which is the function of the psyche/mind/brain as a whole. This function is, as RH stated, “to interact with the environment in ways that maximize the body’s chances of survival and inclusive fitness.”

There are two moves to make in response to a claim like this: 1) the first response is to claim that even if the genetically-open functions of the higher brain are very much integrated and interconnected with both one another and lower brain systems, nonetheless, systems neuroscience will be able to perform a componential analysis of this complicated structure. Individuating and untangling the proper components might prove to be harder than it is with the rest of biology. But still, the fact remains that the brain is ultimately subject to decomposition and localization, just like any other functional system.

Alternatively, someone might argue that 2) either the brain, or the mind, or both, are holistic entities that do not admit of componential analysis. On this kind of view, it might indeed be improper to use the terms ‘function’ and ‘dysfunction’ with regard to this holistic entity, as there would just be one global “mode of being” whose failings would not be decomposable into parts. Philosophers who would want to push for some kind of phenomenological understanding of mental disorder actually might accept this option. Still, recall that it was my goal to show how a theory of mental disorders could fit into the framework of the natural sciences. The natural sciences involve componential analysis. If the function-bearer of genetically-open psychological functions is just the whole brain, or the whole person, we lose this scientific constraint. How would we ever begin to come up with what a function of “the person” is? Thus, I go with option (1), a move that once again makes it clear that a completed dysfunction-based theory will have to provide a methodology for the individuation or decomposition of psychological functions.

4.6 Concluding remarks

At the end of the last chapter, I gave reasons for thinking that the HD theory of mental disorders is a good theory to endorse. In this chapter, I explored problems with the HD theory, with an eye towards using those problems to strengthen the theory. I believe that we can get around these problems via a proper understanding of the genetically-open nature of many psychological functions. In the future, the HD theory will have to offer an individuation scheme and a nosology of psychological functions (and hence mental disorders), but the fact that it cannot yet do this is an indication of the limitations of our current epistemic situation rather than anything endemic to the theory itself.

Having come to the end of the discussion of the Evaluative/Disorder interpretation of the “what is the nature of mental disorder?” question, I will now go on to explore the Structural interpretation.
Chapter 4: The structural interpretation of the “what is the nature of mental disorder?” question: defending the claim that mental disorders are homeostatic property cluster kinds

1. Introduction: Causal Structure

In my “Introduction to Chapters 2-4,” I noted an important distinction between two different interpretations of the question “what is the nature of mental disorder?” I have discussed one of those interpretations – the Disorder interpretation – at length in the prior 2 chapters. That interpretation concerned itself with exploring the badness associated with mental disorders. Within the discussion of that interpretation, I defended a Dysfunction-based theory of mental disorders, and pursued an exploratory claim that many psychological functions are genetically-open. But in addition to the disorder interpretation of the question, there is also a structural interpretation of the question “What is the nature of mental disorders?” This interpretation will be the focus of the current chapter. In this interpretation, we are concerned with the question of what kinds of things mental disorders are, in particular, what their causal structures are. How do their various etiological factors and symptoms causally relate to one another? And how might answering this question help us to construct a nosology, or classification system, of mental disorders?

Remember that when I began this work, I brought up the topic of validity, as it relates to psychiatric constructs. How accurately do these constructs portray the observer-independent world? I use the word ‘constructs’ because the mere use of diagnostic categories does not itself commit one to any ontology beyond one including symptoms. But, recall that I stated that a traditional constraint on the validity of psychiatric constructs is the realist conception that “a diagnostic category is valid only if it represents a real entity.” In fact, I have claimed that a scientific conception of mental disorders will have to show that disorders are somehow more than just constructs. I do believe that mental disorders can meet the traditional constraint on validity, as I believe that they have systematic observer-independent causal structures. So, now I turn to a discussion of what it means for something to be a “real entity” in this way. I start this task by turning to the literature on natural kinds.

As I showed in Chapter 1, Strong Medical and Disease models of mental disorders employ a reductive notion of ‘realness’ – an entity is real insofar as it can be identified with a kind at the genetic or neurobiological level. On this view, “real” mental disorders are natural kinds, and natural kinds require an identifiable fundamental and robust biological cause, where ‘cause’ is understood constitutively. The mental disorder just is that biological entity, which is then causally responsible for the symptoms associated with that mental disorder.

However, a few considerations should make us skeptical of whether or not this is an accurate picture of the structure and scope of mental disorders. For one, the science of mental disorders is in its very early stages, and although there are many empirical studies that confirm correlations between psychiatric constructs and neurobiological and/or molecular processes, not even one clear identification has been made between the two levels. We cannot know if this result is due to faults in our science, or if there really are no such identities to be discovered. I demonstrated in Chapter 1 that recent work on faulty brain circuits does give us reason to believe that perhaps some identifications will be made (like “OCD_handwashingvariety = faulty FSC circuitn”). Still, the fact that brain circuits can become faulty in multiple ways should make us question whether these identifications should be viewed just like the simpler identifications characteristic
of currently-accepted type-identities (table salt = NaCl, HIV = infection with a particular \textit{Lentivirus}, etc.). Perhaps their complexity and potential multiple realizations means that reductions are impossible.

Regardless of such considerations, Disease Model theorists remain optimistic about finding identities between mental disorders and kinds at the neurobiological or molecular levels. For them, the ideal is still to find physical etiological factors for all mental disorders; indeed, this is just the goal that prompted theorists to try to construct – and thus far unfortunately, fail to actually construct – an etiologically-based DSM 5. We see a similar kind of aim in the RDoC (Research Domain Criteria). NIMH states that the project “is an attempt to create a new kind of taxonomy for mental disorders by bringing the power of modern research approaches in genetics, neuroscience, and behavioral science to the problem of mental illness.”\textsuperscript{99} Clearly, the theorists working on these initiatives remain hopeful about finding biological causes of mental disorders, which will serve to both causally explain \textit{and} validate the existence of such disorders. However, this kind of thinking implicitly accepts a traditional picture of natural kinds, on which all properties of a kind radiate out from a common core cause, or essence. I maintain that this acceptance biases us towards a particularly constrained theory of what can be real or natural.

In this chapter, I take a critical stance on essentialist frameworks and attempt to demonstrate that they have gotten the causal structure of mental disorders wrong. The claims that I will defend are: 1) mental disorders are homeostatic property cluster kinds (HPCs). 2) HPC kinds are still natural kinds, even though they do not have essences. And 3) mental disorders are networks, not entities.

In section (2) of this chapter, I will investigate the concept of “natural kinds,” with an emphasis on essentialist natural kinds. I will juxtapose essentialist natural kinds to various types of non-natural kinds (ie: syndromes, family resemblances). In section (3), I will explore ways in which we can maintain the virtues of natural kinds without being essentialists. In particular, I will discuss paranatural kinds, those kinds defined by an \textit{absence} of a particular natural kind. Although I conclude that some mental disorders may be paranatural kinds, I argue that a theory of paranatural kinds does not yet provide us with a clear picture of the causal structures of mental disorders. This consideration leads me to an exposition of homeostatic property cluster kinds and their features. For simplicity’s sake, I then discuss the non-psychiatric example of a biological species in order to make the HPC concept clear. In (4), I apply the concept of HPC kinds to psychiatric constructs, and confront places where the theory has holes – particularly with regard to the nature of the supposed homeostatic mechanisms involved in these kinds. Section (5) will serve as a defense of the HPC view. In this section, I will present the benefits of holding such a view. These benefits involve being able to account for aspects of mental disorders that are not as well explained via more traditional views like biological essentialism. Finally, section (6) is a conclusion that introduces an exploratory line about the possibility of bringing together the Disorder and Structural interpretations. My suggestion is that some of the features of HPC kinds can explain why mental disorders are bad things to have.

2. Natural kinds vs. Non-natural kinds

The primary question that I am asking is: what kinds of causal structures do mental disorders have? What are the linkages like between different symptoms and etiological factors? In the background of these structural questions is the issue of reality – are mental disorders naturally-occurring? Are they natural kinds? We can generally think of natural kinds in the following way:
“To say that a kind is natural is to say that it corresponds to a grouping or ordering that does not depend on humans.” In other words, the similarity between members of a natural kind is not conventionally determined by human choices or pragmatic interests; the grouping of natural kinds is observer-independent. Because the unity of natural kinds comes from the world, natural kinds play important roles in natural laws, inductive inferences, and predictions. These scientific virtues are, I presume, part of the reason that philosophers of psychiatry aim to defend the existence of natural kinds in psychiatry.

So, when one asks if mental disorders are “real,” the most likely interpretation of this question is that one is asking whether or not mental disorders are natural kinds. A corollary of this question is whether mental disorders have the same causal structure as physical disorders, with the assumption that physical disorders are natural kinds. As we saw in Chapter 2, when mental disorders fail to exhibit the reductive structures characteristic of quintessential physical disorders, one (wrongly) infers that mental disorders are not real because they are not natural kinds. My argument will be that people who conclude in this manner are confusing the category of essentialist natural kinds with the more general category of natural kinds. I aim to show that natural kinds need not be essentialist, and in fact, many mental disorders (as well as many physical disorders) are non-essentialist natural kinds.

### 2.1 Essentialist natural kinds have essences, from which radiate all characteristic properties

So, what is an essentialist natural kind? Essentialist natural kinds have relatively simple causal structures. In the following diagram, E stands for “essence” and P for “characteristic properties.” E is both the constitutive cause of the kind under discussion and the diachronic cause of the properties, and the P’s are the properties. We can imagine far more P’s than just 3, and we can imagine further causal chains extending from the P’s.

![Diagram of essentialist natural kinds]

Note some important features of kinds with essentialist causal structure: 1) there is one cause cited (E), which is taken to be definitional – both necessary and sufficient for the kind, 2) all of the characteristic properties are effects of this one common cause, 3) the properties are not in direct causal relations with one another, and 4) the essence is ontologically separable from its characteristic properties; in other words, even if the kind fails to diachronically cause its characteristic properties, nevertheless E still exists.

Chemical elements are the quintessential essentialist natural kinds; for instance, hydrogen is all and only those things that have one proton; having one proton is hydrogen’s essence, and having one proton causes hydrogen to exhibit other features (a certain mass, certain bonding properties, etc.). Or, going up a level from the atomic to the molecular, consider water, whose essence is the H\(_2\)O molecule. The properties of the hydrogens, the oxygen, and their bonds are responsible for all the characteristic properties of water: its boiling point, its freezing point, its surface tension, and so on. Being H\(_2\)O is both necessary and sufficient for being water. In non-standard conditions, water may not boil at 100 degrees C, but this does mean that it is not water.
Physical diseases supposedly have this essentialist structure too, though their E’s are either specific genes or infectious agents. Consider tuberculosis. The *tubercle bacillus* bacterium is the essence of the disease and it is causally responsible for all of the symptoms (cough, fever, night sweats, etc.). However, with different background conditions than what are typical, the bacterium might infect a person *without* causing symptoms (such a person would still have Tb, he would just be carrier); on the other hand, a similar symptom set might manifest even without the TB bacteria. The absence of the essence automatically renders this symptom set *not* a case of TB. This independence is what I had in mind when I stated that essences are ontologically separable from their characteristic properties. The hope, for essentialist theorists, is that mental disorders will turn out to have causal structures like germ-based or genetic diseases.

When we ask of something, “is it real?” essentialist natural kind theories have an advantage. The essence is something that we can point to as an existing, concrete entity, and its causal powers are further pieces of evidence for its existence. This “pointing to” links with the concept of validity; if we can just find *something* specific from which radiate all the symptoms of a mental disorder, then we could validate the existence of what a diagnostic construct describes.

Unfortunately, there is no good reason to think that the majority of mental disorders are essentialist natural kinds. As I have already stated, the empirical evidence has failed to yield type-reductions. Although this failure is not a reason to conclude that such essences will never be found, it is a reason to explore other possibilities. In addition to the lack of good evidence for essences, essentialists run into the problem that the symptoms of mental disorders look to be in direct causal contact with one another. As I stated earlier in a footnote, direct causal contact between symptoms may not actually be disallowed by an essentialist theory. Still, this kind of causal contact might make us wonder if some of the symptoms are *initially* caused by other symptoms rather than by the purported essence. For example, the rumination characteristic of clinical depression appears to have a direct causal impact on suicidal ideation. Racing thoughts cause insomnia. Insomnia causes motor retardation. This is a different topology than one on which, say, serotonin deficiency independently leads to rumination, suicidal ideation, racing thoughts, and insomnia. In addition to there being causal interactions between symptoms, symptoms of mental disorders also appear able to “reach back” and have a causal impact on their purported causes as well (rumination causes suicidal ideation, which then causes one to ruminate even more). So, if mental disorders have a causal structure not captured by the concept of an essentialist natural kind, what kind of causal structure do they have? And does their failure to be essentialist impugn their ability to be natural kinds?

I would like to focus on two contrast classes for essentialist natural kinds: 1) non-natural kinds and 2) natural kinds that are not essentialist. For this latter category, I have two frameworks in mind: (i) paranatural kinds (which are not technically natural kinds, but are nonetheless grounded in natural kinds) and (ii) homeostatic cluster property kinds. I will start with the first contrast class, as it makes a more overarching distinction, that between natural and non-natural kinds.

2.2 Non-natural kinds: social constructions, syndromes, and family resemblances do not appeal to categories that exist outside of human classification

I already presented one strategy for making the distinction between natural and non-natural kinds in Chapter 2: social construction. Social constructivists with regard to mental disorders feel that if mental disorders require an evaluation of badness for their existence as mental disorders, then this normativity cannot be provided by the natural world; all judgments of
badness are observer-relative. Hence, mental disorders are not natural kinds, although the conditions upon which the judgments are made might be.

Another way to elucidate the distinction between natural and non-natural kinds is by deploying the physical-disorder paradigm’s distinction between a disease and a syndrome. Diseases are those ailments with a known etiology (based on genetics or infectious disease particles, as seen in the prior sub-section with the example of Tb). Syndromes, on the other hand, are just collections of symptoms with no known underlying cause. As an example, consider CFS (chronic fatigue syndrome), which is marked by symptoms like persistent fatigue, malaise, and joint pain. Of course, it could be the case that a cause (or essence) will someday be found and that CFS will attain the status of a ‘disease.’ But as of yet, the cause of CFS is not known – perhaps because CFS is nothing but a collection of symptoms. Furthermore, the choice of which of those symptoms “belongs” to the syndrome is not set in stone. If we have symptoms A-E, one medical professional might say that the syndrome includes A, B, C, D, and E, while another draws the boundaries just around A, D, and E; and so on. To be an instance of syndrome, one does not need to have all of the symptoms associated with that syndrome.

As I stated in my introduction, one might be pessimistic about the reductive goal of discovering essences for mental disorders, because she believes that there are no such things to be found. This outcome would attain if there were nothing more fundamental underlying the symptoms of a disorder, no naturally-grounded explanatory factor for the unity of the instances of that kind. Instead, the collection of symptoms would be contingently brought together on the basis of utility, according to some human purpose (say, ease of diagnosis). The symptoms would either tend to co-occur without explanation, or else individual cases of the syndrome would be similar enough according to human judgment to be lumped together for pragmatic reasons. For example, scientists might just have an interest in those people who express both fatigue and joint pain, so they call this set of symptoms by a certain name. Because syndromes are nothing over and above (or nothing more fundamental than) their symptoms, a syndrome cannot exist independently of its symptoms. If the symptoms are not present, there is no syndrome, since the syndrome just is the collection of symptoms. Note how different this outcome is from the “independence” between essences and their symptoms that marked the “disease” Tb example.

A third way of talking about non-natural kinds besides either social construction or the syndrome analogy is via Wittgenstein’s notion of a “family resemblance.” Concepts that have a “family resemblance” definition have no necessary and sufficient conditions on membership; rather, the entities that fall into the extension of the word are related the way in which members of a family are related – A may share a property with B, B may share a property with C, while A and C share no properties at all. Applying this framework to mental disorders would mean that tokens of one category of mental disorder (ie: clinical depression) have been grouped together conventionally on the basis of a family resemblance, and the degrees of resemblance necessary to fall under the same term are based on our own pragmatic interests in any given case.

We may view the current DSM as operating in the “syndromatic” manner, when it comes to categories of mental disorders – someone is diagnosed with a mental disorder if he or she meets a certain lower threshold on a number of listed symptoms. In most of the cases, none of the symptoms is individually necessary to the diagnosis. Thus, even in the DSM, we do not see a commitment to essentialist natural kinds (though, that is not to say that that was not the goal of the creators of the DSM; indeed it was their goal). Thus, one can describe the typical profile of a mental disorder without being committed to the disorder’s status as an essentialist natural kind.
As even further evidence against mental disorders being essentialist natural kinds, what we see described by the DSM are a whole lot of disorders with overlapping symptoms. Anxiety and depression have many symptoms in common, as do depression and bipolar disorder, and so on. So there is a whole mass of symptoms, stuck together in various ways, and perhaps different theoreticians would carve these symptoms up into disorders differently, or not carve them at all. In any event, a morass of symptoms with overlapping groupings delineating supposed separate disorders is not the ideal picture for an essentialist natural-kinds theorist.

3. Natural kinds: alternatives to essentialism

My primary goal in this chapter is to push the claim that mental disorders are homeostatic property cluster kinds, a type of natural kind that does not require us to posit biological essences. Thus, I am arguing that the concepts ‘natural kind’ and ‘essentialism’ can come apart. The HPC theory is an alternative theory within a framework of natural kinds.

One might argue at this point that not all hope is lost yet for essentialism, because we can account for the problems that crop up in essentialism by deferring to complex essences. A theorist who believes that mental disorders are essentialist natural kinds may claim that the entities on the neurobiological side of the “mental disorder = neurobiological kind” identity are not simple. Perhaps there are necessary and sufficient conditions that provide essentialist definitions for mental disorders, but these essences are not entities like atoms, molecules, infectious organisms, individual genes, etc. Perhaps they are entire metabolic pathways or combinations of genes.

While I grant that this outcome is a possibility, it is likely that these supposed “complex essences” would deviate too far from the traditional meaning of “essence.” Being complex, they would have parts, and as a result of having these parts, they might even be multiply realizable. It might be that one of these components of the essence ($E_1$) necessarily causes some of the characteristic effects of the kind, while another component ($E_2$) necessarily causes other characteristic effects of the kind. Can an essence be “$E_1 + E_2$”? Alternatively, can an essence be disjunctive, as multiple realizability would entail ($E_a$ or $E_b$ or $E_c$)? What if each component was multiply realizable? The possibilities for such causal structures do not do justice to the notion of essences as necessary-and-sufficient/definitional for the kind. Note that I am not claiming that the causal structures of mental disorders do not have these kind of conjunctive and disjunctive components – in fact, they probably do. Rather, I am claiming that “complex essence” is not the proper way to understand them, and that such complexity deviates too far from the restrictions on an essentialist theory.

Our cause of accounting for mental disorders as natural kinds is not lost, however; even if mental disorders do not have essences, we need not conclude that they are therefore not natural kinds. Instead of contrasting essentialist kinds with non-natural kinds (as I did in section 2 of this chapter), we might stay within the realm of natural kinds and juxtapose an essentialist model of natural kinds with alternative understandings of natural kinds. I will now explore such options.

3.1 Option 1: Paranatural kinds are marked by the absence of natural kinds

Bortolotti, Broome, and Sabbarton-Leary refer to the work of Roy Sorenson when they write that “a paranatural kind is parasitic upon some natural kind or other; it is an ‘absence defined by a natural kind.’” For examples of natural kind/paranatural kind pairings, they offer the following two examples: light/shadow, heat/cold. While light is a definite item of our ontology (photons), a shadow is the absence of photons. Likewise, while heat is a definite item
of our ontology (molecular motion), cold is nothing but the absence of this kind of motion. Thus, strictly speaking, the claim that mental disorders are paranatural kinds cannot be a claim that mental disorders are natural kinds. But I include this theory here because the theory inherits most of the virtues of a theory on which mental disorders are natural kinds.

Now, why do we postulate natural kinds to begin with? Recall my earlier statement that "to say that a kind is natural is to say that it corresponds to a grouping or ordering that does not depend on humans." Natural kinds are therefore parts of the "furniture of reality." They can figure in laws, they are projectible, and they ground inductive inferences, yielding high predictive power. Bortolotti et al.’s point is that since the characterization of any paranatural kind is derivative of the characterization of the natural kind of which it is an absence, then the paranatural kind also has the sort of unity that is needed to ground the explanatory benefits listed above. So, we gain the virtues of positing natural kinds without having to expand our ontologies.

How does this general theorizing about paranatural kinds apply to mental disorders? First, it is important to recognize that Bortolotti et al. do not claim that no mental disorder is a natural kind. Rather, they say that the natural kinds of psychiatry are only "those sorts of causes that have a biological basis, which are independent items of our ontology." What is an "independent item of our ontology?" We can explore the meaning of this phrase first within the realm of physical medicine. Take a kidney – there are many ways for a kidney to fail. It might be that the kidney contains a tumor. A tumor is considered an "independent item of our ontology." It is a novel addition to the natural kind, kidney. On the other hand, failures of metabolic pathways that are an expected part of kidney functioning do not count as "independent items of our ontology." These failures are just ways for what the kidney is already doing to go wrong. The functional mechanism is the natural kind, and the failures compose a paranatural kind – the absence of the kind. Even though a paranatural kind is not an independent item of our ontology, it is not strictly speaking nothing at all. What exists is a dysfunctional mechanism, which dovetails nicely with the previous chapter’s claims that mental disorders are dysfunctions.

On this picture, natural kinds of mental disorders would be tumors, or lesions, or some additional system (one that healthy people would lack) that was responsible for symptoms (this would be the case, for instance, if rather than having some kind of malfunctioning sensory system, a person with schizophrenia actually had a distinct kind of sensory system that produced vivid hallucinations).

But there is a problem with an account of mental disorders as paranatural kinds, a problem that will lead us into a discussion of homeostatic property cluster kinds. The problem is this: a theory of paranatural kinds fails to provide any further explanation for the unification of cases of a particular mental disorder beyond the fact that they are all absences of the same natural kind. Remember that instead of paranatural kinds being “entire pathological mechanisms,” they are “deviations resulting in slight qualitative changes to normal function.” Their abilities to mimic natural kinds in their inductive potential and projectability have to do with their negative qualities (not being that natural kind). Still, isn’t there reason to think that these deviations from normal functioning have more – something positive - in common besides this shared absence? Usually, the individual tokens of a given type of mental disorder share a distinct set of symptoms, and perhaps even etiological factors. The para/natural kinds theorists do not talk overtly about the causal structures of paranatural kinds, perhaps because they find it odd to talk about the causal structure of absences. They define mental disorders by negation rather than offering a positive theory. But indeed, mental disorders do have causal structure.
So, it is not that I am arguing that a paranatural kinds theory of mental disorders is wrong. In fact, I believe that it is a good candidate to be linked with a Dysfunction-theory of mental disorders. It is just that it fails to provide an account of the positive features of mental disorders.

Thus, I want to change the topic slightly and ask: what are the causal structures of mental disorders? I now introduce HPC kinds, which are non-essentialist natural kinds.

3.2 Option 2: HPC kinds are non-essentialist natural kinds composed of a cluster of properties and a mechanism that holds those properties together

The alternative natural kind to essentialism is the homeostatic (or mechanistic) cluster property kind (HPC or MPC kind). An HPC theory recognizes that not having necessary and sufficient conditions on something (i.e., the necessary presence of the essence; the necessary presence of characteristic properties, *ceteris paribus*) need not impugn that thing’s status as a natural kind. It need not leave a disorder with the status of “syndrome,” or something whose existence as a disorder is forever dependent upon value judgments or conventional decisions. On the contrary, a “homeostatic property cluster kind” has a definition that is “determined by the members of a cluster of often co-occurring properties, and by the mechanisms that bring about their co-occurrence.” This definition has 2 distinct components: 1) the cluster of properties, and 2) the mechanism. The relevant mechanisms are processes that make the continued co-presence of the properties more likely. The mechanism is something in the world, so the clustering is nomologically grounded *a posteriori* in a phenomenon that occurs independently of the construction of conceptual schemes. This kind of grounding is juxtaposed to the conceptual *a priori* grounding of non-natural kinds, where the categories that we come up with analytically entail certain properties.

Consider the following diagrams, from Ahn and Kim. In their studies, clinicians were asked to surmise about the causal structures of both clinical depression and anorexia nervosa, and these figures are the *composites* of those results. The boldfaced boxes contain symptoms in the DSM, while arrows represent purported causal relationships. Numbers above the arrows indicate relative strengths of the causal links.
Granted, in pushing what they call “Causal Category Theory,” Ahn and Kim have a slightly different goal than I do. They are interested in the empirical question of how mental health clinicians think about mental disorder. They believe that to learn a mental disorder concept is to learn a causal structure. They aim to show that clinicians exhibit causal thinking, and are thereby ontologically committed to more causal structure than what is indicated by mere correlations between symptoms (which is all that essences would commit us to). Of course, the fact that clinicians conceptualize causal structure in a given way is not necessarily an indication that the causal structure is actually that way.

And that is precisely where I want to push further. I want to defend a claim that mental disorders are kinds that have causal structures that are more complex than essentialist causal structures. What the above sorts of diagrams do for us is to make it intuitively plausible that essentialist causal structures misrepresent the causal structure of mental disorders. What might not be intuitively shared by all readers, however, is the belief that the diagrams support a conclusion that this causal structure is a homeostatic cluster. What exactly does this term mean?

I stated at the opening of this section that HPC kinds have two components. The first component is a cluster of properties, none of which is individually necessary for the existence of an instance of that kind. So there is no essence, just the collection of properties. The properties co-occur with a high degree of probability and they nomologically make one another’s presence more likely. This “making more likely” is a cluster-level probabilistic property, corresponding to the unit and not to any one symptom. In other words, not all of the symptoms make all of the other symptoms’ presences more likely. In fact, most symptoms are not causally connected to all others, and moreover, the strengths of the causal connections within a cluster are quite variable. Thus, individual tokens of a kind may causally differ markedly from one another. One might charge that this account of property clusters is imprecise because it does not set a lower bound (in terms of connectedness or probability levels of causal linkages) on a property’s being part of the cluster. In addition, we might wonder: what is ruled out? Couldn’t we get any causal structure to fit this picture?

Ruben raises this problem of imprecision when he writes the following:

"[there is the] question of just how much more likely need the presence of some of the properties make the presence of the others in order for there to be a homeostatic clustering of properties. Would any increase in likelihood, however slight, suffice for a group of properties to count as homeostatically clustered? Although Boyd does not explicitly set a lower bound on how much of an increase in likelihood is required for a group of properties to count as homeostatically clustered, he does make it clear that, where HPC kinds are concerned, the homeostatic clustering of properties should be ‘causally important.’ This suggests that, at least with respect to HPC kinds, we should expect that the presence of some properties in the relevant group raises to a considerable degree the likelihood of the others being present. (Indeed, if this were not so, then it would be hard to see how HPC kinds could fulfill their role in facilitating reliable inductive inference). Of course, even if this is accepted, we are still left with a good deal of imprecision in the account of property homeostasis.”

While it is beyond the scope of this chapter to explore the possibilities for dealing with the problem of imprecision, I bring it up to show where future work on a theory of HPC kinds might focus. It is also possible that the imprecision will end up being an explanatory virtue that allows us to account for variability, fuzzy boundaries, and continua in accounts of mental disorder. However, for now, I acknowledge that if the theory cannot come up with a list of causal
structures that are ruled out as not being HPC kinds, then it will just include too much to be a tenable theory.

To return to the exposition of HPC kinds, I have just been discussing how one of the two features of HPC kinds is the property cluster. The account of the property cluster might appear to overlap with the family resemblance or syndrome concepts, which have no ontological commitments beyond the presence of the correlated properties themselves. But there is indeed a significant difference between HPC kinds and non-natural kinds: HPC’s have a deeper explanatory factor underlying the unity/co-occurrence of their properties. This explanatory factor is taken to be causal rather than conceptual, nomological rather than logical, something that is discoverable a posteriori. This explanatory factor is referred to as a “homeostatic mechanism,” and it is the second characteristic of HPC kinds.

“Homeostasis” refers to something’s ability to remain stable over time. Thus a homeostatic mechanism is any mechanism that preserves the stability of the clustering of a set of properties. It could do this either by directly producing, underlying, or otherwise sustaining the causal linkages between the properties and their causes. Therefore, the kind-ness of HPC kinds lies in “stable patterns of complex interaction.” The distinction between etiological factors and symptoms is artificial on this theory; a more proper distinction is that between central (those properties with more – and stronger – causal connections to other properties) and peripheral properties.

We can view HPC kinds as a compromise position between essentialist natural kinds and non-natural kinds. HPC kinds are like syndromes in that they do not point to a singular essence or to a set of necessary and sufficient conditions that must be met in order for something to be a member of that kind. However, they are more like essentialist natural kinds than syndromes in that they exist independently of human classification. The explanation for their clustering makes appeal to natural causal mechanisms. This appeal allows us to secure the existence of natural kinds without a reduction to, or a privileging of, the biological level. Since the similarities between instances of a kind are therefore explained without recourse to human objectives, HPC kinds can play the scientific roles given to natural kinds.

5.1 An example of a homeostatic property cluster kind: biological species

In the philosophical literature, the primary example of an HPC kind is a biological species. In this case, the cluster of properties would be the typical morphological, behavioral, and genetic characteristics of the species. Think about the diversity of the members of a species, and you will note that not all members will share all of their properties; indeed, there may not be even one property that they all share. Membership in a species can be more central or peripheral depending on the amount of overlap in these properties. So why group these animals into one kind? Do biologists just make judgments about family resemblances? No.

The proper answer to why members of a species are grouped together lies in the homeostatic mechanism responsible for the property clustering. The stable sharing of a large number of properties between members of a species is not an accident nor is it a human creation; it is in nature. There is a homeostatic mechanism for this stability: the species’ genome, and mechanisms of descent. Interbreeding allows for continual exchange and phenotypic manifestations of an “imperfectly-shared” set of features. One might argue that a shared genome is a shared essence, but the fact that any species-genome is full of so many polymorphisms
(different versions of genes) shows that what is shared is at best something abstract rather than something concrete, something complex rather than something simple.

HPC kinds have some interesting features, which will become relevant to our discussion of mental disorders. First, a) cases of indeterminacy are permitted. With regard to biological species, this situation in fact occurs during the transition from one species to another. In a given case, we might not have an answer as to whether an individual belongs to a particular species or not, especially if that individual shares only a small portion of its properties and/or genome with the more quintessential members of that species (those who share in more of the cluster). Still, a theory of HPC kinds allows for this indeterminacy because sharing in only a small set of the property cluster still qualifies one for membership in that kind. Still, its smaller number of typical properties means that it can perhaps be a member of another species as well, so long as it shares in a portion of that property cluster as well. Under essentialism, the essence is either present or absent; there is no in-between.

Second, b) the relevant homeostatic mechanisms may actually be external to the individual entities. I briefly said that the homeostatic mechanism responsible for the clustering of properties representative of a species is either the genome or interbreeding, or both. But there is more to be said here. An internal component of this mechanism is the species’ genome. The species’ genome is responsible both for the stability and the variety of species traits throughout succeeding generations. However, one might add that another component of this mechanism is that the members of the species live in similar environmental conditions with similar selective pressures, and that interbreeding between species is relatively low. This means that the environment also plays a role in maintaining the homeostasis of the species’ characteristic properties, rendering the relevant mechanism both internal and external. Thus, external conditions can “induce” a real clustering of properties that is stable enough to play a role in natural laws. I will return to this notion of external homeostatic mechanisms in my discussion of mental disorders as HPC kinds.

4. Are mental disorders homeostatic property cluster kinds?

In this section, I would like to advance the claim that types of mental disorders (ie: clinical depression, schizophrenia, etc.) are homeostatic property cluster kinds, and as such, are neither constructs of symptoms nor entities, but rather, are causal systems or networks. Remember that HPC kinds have two components: 1) a property cluster and 2) a mechanism. Let us look at each of them in turn, with regard to mental disorders.

4.1 The property cluster does not distinguish between causes and symptoms

As regards the property cluster for a certain mental disorder, we should view it as containing both symptoms and etiological factors; or rather, we should no longer make this distinction. The inclusion of the causes alongside the symptoms is non-traditional, when we consider the unidirectional path from cause to symptoms with essentialist kinds. What we need to keep in mind here is that this is a causal network. Not all of its causal chains are unidirectional. So, while it is possible that a particular causal chain is set-off with the presence of a given etiological factor, that factor may cause symptoms that then come back around and affect that initial causal
factor. Then this causal factor becomes what we would have previously called a “symptom.” Is rumination a cause or a symptom of depression? The most accurate answer is likely “both.”

As for the actual content of the members of a cluster, these are properties of a disorder that show up across cases. They can come from all levels of description (chemical, biological, neurobiological, psychological, social, etc.). Remember that each token case of an HPC kind need not express all of the properties in the cluster, and indeed there is no one property that is necessary for membership. So the identity of the cluster abstracts away from particular cases – and is a composite of them - in order to include a range of more central and more peripheral characteristic signs and symptoms.

The word ‘central’ refers to properties that are more enmeshed in the causal network. This centrality means that central properties either 1) have causal connections to a relatively large number of other properties in the network or 2) that their causal connections to other properties are strong/they raise the probability of those symptoms to a higher degree, or (3) both (1) and (2). If we look at the causal diagram for anorexia, “lives in an industrialized society” is a central property (it causes 3 other properties), while “need for control” is more peripheral (because it has only one causal arrow emanating from it). Central properties explain why we were tempted to believe that certain symptoms were necessary and sufficient conditions on having a particular mental disorder. Because of their central placement in the causal network, it is highly likely that they will be present in any given token of the disorder. But, this very same point also explains why there can be cases of a disorder that fail to show these supposed necessary and sufficient conditions. The part of the causal structure that is activated in a particular case might just not include that central symptom. Yet as a result of partaking in even a small part of the causal network, such a case really does count as a case of, say, anorexia.

Of course, there is the further question: how much of the causal structure has to be activated in order to count as an instantiation of the kind? We will get to that question, but for now, it is just interesting to note that if we accept an HPC theory of mental disorders, then diagnosis becomes a two-step process. As Borsboom writes: diagnosis includes “(a) whether the causal system of symptoms is at all activated in a person, and (b) where in the causal sequence a person is located at the time of the diagnostic interview.”

When I talk about the benefits of adhering to this kind of theory, I will return to this two-step diagnostic process, and describe how it can account for both categorical and dimensional intuitions about mental disorders.

4.2 The homeostatic mechanism may not be the same in all token cases of a type of mental disorder

Recall that HPC kinds are not family resemblance concepts, for the reason that their clustering is causal rather than conceptual. In some way, the world itself is responsible for the fact that these properties tend to show up together with a high degree of probability. The homeostatic mechanism is whatever is responsible for this clustering. This observer-independent causal network is what makes mental disorders fit subjects for the natural sciences. In the case of biological species, this mechanism was presented as being the genome, combined with external environmental pressures. The question that I will now explore is what we should take the mechanism to be in the case of mental disorders. Interestingly, Ruben interprets Boyd (perhaps the originator of the HPC concept) as giving the following definition of HPC kinds:

“A family of properties, F, is homeostatically clustered if and only if either
i. (under appropriate conditions) the presence of some of the properties in F makes more likely the presence of the other properties in F, or
ii. there exist mechanisms or processes that make more likely the continued presence of the properties in F, or
iii. both i and ii.”

The reason that I say that this formulation is interesting is that it does not actually require the presence of an independent homeostatic mechanism in order for something to be a cluster-kind. It might be that in some manner of speaking, the properties themselves + their causal relations (in other words, the entire causal network) is itself the mechanism. In fact, I will end up supporting a claim much like this conclusion. But first, I want to look independently at reasons why a search for a more concrete mechanism may be foolhardy.

The question that I am asking is, with regard to mental disorders, are the property cluster and the homeostatic mechanism two different things, as is the case with biological species? If so, and if we imagine the mechanism as causing and sustaining the clustering, then I see us falling into territory that is similar to the “complex essences” picture discussed in in the opening to section (3) of this chapter. On this picture, the homeostatic mechanism is something more fundamental than the property-cluster, something underlying it. It is hard to imagine this kind of picture being able to avoid reductionism, or if not reductionism, then privileging of lower-level explanations. Even though the biological species picture admits the importance of the environment, nonetheless, the basis for the mechanism is still the genome. Thus, while there are no type-reductions here, like “species-characteristic feature a1 reduces to gene A,” there is still a deference to lower-level explanation. What the kind really is is the mechanism, which is still a chemical or biological entity.

Thus, we might imagine that although clinical depression fails to have a simple neurobiological essence as once was hoped (ie: serotonin deficiency), it could be that there is a complex, widely-realized neural processes that extends over multiple parts of the brain and is responsible for the co-occurrence of the properties in the cluster. Again, this would not be a traditional type-reduction, especially since such a large network would certainly be multiply realized. Still, we could say something like “depression really just is this set of neurobiological stuff happening.” This outcome would be explanatorily reductive without being a type-reduction because we would likely see the lack of necessity of any one non-functional property of the biological realizers. The symptoms would then again become secondary to the underlying state, and the symptom/cause divide would be opened up again.

In their paper, Kendler, Zachar, and Craver claim that some HPC kinds can take on this form, where all instances of a kind share an “underlying state” that holds the property cluster together. They offer the following figure, with this caption:

“…we have a series of causes (C1 to C4) that interact with each other to produce an underlying state (US) that in turn leads to the individual clinical features (F1 to F5). These causal processes could be psychological or biological. These clinical features in turn could causally interact with each other.”
On a picture like this, we might be tempted to identify the disorder with either US or the set of c’s. However, in order to support an HPC theory, we need not view mental disorders as clusters of properties being held in homeostasis by any kind of further mechanism. Some cases of depression have lowered left hemisphere activity, some show decreased hippocampal volume, some have decreased frontal cortical tone, and so on. The explanatory reductionist would conclude that these biological correlates are the “c’s” in the diagram above, that together lead to a complex multiply-realizable state that underlies the property cluster by causing its individual properties to manifest. However, I think that a better solution to this neurobiological complexity is to actually put these individual biological correlates (the c’s and US’s) into the property cluster itself. After all, for the moment, neural correlates are just correlates. Nothing is indicated about causality. It could be that what we take to be lower-level, fundamental causes are themselves caused by symptoms at a higher level (for all we scientifically know at the moment, it could be that rumination itself directly causes hippocampal degradation).

Any attempt to find a complex homeostatic mechanism that underlies each type of mental disorder would inevitably lead us to embrace a strong version of the Medical/Disease model, a model that we have seen is untenable. This move involves a tacit acceptance of the reductive assumption that something more fundamental underlies mental disorders beyond the symptoms themselves. Although symptoms are not necessarily rendered epiphenomenal (look back at the caption’s statement that “these clinical features in turn could causally interact with each other”), they are secondary in that they depend on the underlying states for their existence and maintenance. Thus, interventions would be best targeted towards the US rather than towards the symptoms. However, what if symptoms themselves constitute disorders? If that is the case, then we need a picture on which symptoms are primary targets for, and loci of, causal interventions.

One way that we might try to do justice to both the causal power of symptoms and to the existence of a homeostatic mechanism is to claim that in a token case of the disorder, one or a couple of the symptoms are the mechanism; for example, perhaps a central symptom could play the role of homeostatic mechanism. We have already seen that not all cases of a particular mental disorder need to share a designated central symptom, so one might think that this is not a good hypothesis. However, no one ever said that the unity of the kind, say “depression,” had to be based on uniformity of mechanism rather than overlapping cluster properties. That is, in each token case of depression, the central symptom sustaining the causal network could be distinct. For instance, one person’s depression could be homeostatically maintained by something external – say, ongoing trauma (like domestic violence, or living in wartime conditions), while
another person’s depression could be homeostatically maintained by something neurological (low serotonin) or something psychological (depressed mood). We could come up with any number of mechanisms, from all levels of description (biological, psychological, social). The figure given by Ahn and Kim, remember, is a composite of causal structures hypothesized by different theorists. Similar diagrams have been offered by Boorsboom that are composites of the causal structures of actual cases of the disorder. Thus, even if something like “depressed mood” is the most central property in the majority of cases (and hence the best contender for a property-based homeostatic mechanism), some individual cases that made up the composite may have had distinct central properties.

What these possibilities bring up for me is the question of unity: what is responsible for all of the cases of disorder X being instances of that particular disorder? In this section, I have shown that the answer “having an identical homeostatic mechanism” is false. So where is the unity, if in neither an essence nor an identical homeostatic mechanism?

4.3 The unity of a mental disorder kind is nothing but a shared causal structure

What unifies all of the instances of an essentialist natural kind is the necessary and sufficient presence of the relevant essence. What unifies (some) homeostatic property cluster kinds is the sharing of a homeostatic mechanism that is responsible for the clustering of a certain set of properties. In the biological species case, we have seen that reference to this mechanism (the genome) is partially reductive, because there is a dependence on something that is “underlying.” Every member of the species has this genome (even though the genome is multiply realized through different alleles of different genes).

However, I have discussed why this kind of partial reduction does not apply to mental disorders, unless we understand them as having complex neural/genetic identities. I suppose that I am expressing skepticism with the claim that states that all cases of say, depression, are cases of depression because they exhibit homeostatic mechanism D. This type of view looks essentialist, and one might also be tempted to think of the mechanism as something that causes all of the clustered properties. However, the mechanism is not causally responsible for the presence of the properties themselves; rather, the mechanism is responsible for their clustering, for their making one another more likely, for their maintenance and possible growth via emergence of more peripheral properties.

Take a case of depression’s being homeostatically-maintained by ongoing trauma in the form of domestic violence. It could of course be the case that the domestic violence may have directly caused low mood or insomnia or suicidal ideation. But the violence need not necessarily have had this initial causal role. Instead, the domestic violence keeps the symptoms of depression stably entwined. For instance, perhaps one’s symptoms of negative cognitions start to wane as she takes on a new job outside the home. The extinction of these thoughts leads to a reduction in rumination, and therefore, a reduction in insomnia. However, her terror at her home situation is a separate path to insomnia. Her lack of sleep then causes low mood, which then causes rumination and negative cognitions. The environment serves as the homeostatic mechanism, even though it is not the direct cause of all of the symptoms and their causal relations. As the homeostatic mechanism, the home-life holds these symptoms together.

Of course, homeostatic mechanisms can be either internal or external, or some combination of the two. My point is that I do not think that the unity of a kind like “depression” need rely on the identity of the homeostatic mechanism between cases, because the mechanism can differ from person to person. What the various cases of a given disorder then seem to really
have in common is that they share in the same cluster of properties. This is why I cited Ruben’s line before, that what makes something a homeostatic property cluster kind might just be “i. (under appropriate conditions) the presence of some of the properties in F makes more likely the presence of the other properties in F, or…” where the other side of that disjunction is the homeostatic mechanism. In cases that exemplify (i.), we could say that there is nothing over and above the cluster of properties, and that the cluster is the mechanism. In other words, it is the entire network of causal linkages between the properties of the cluster that is responsible for the clustering. Because this is just how things nomologically are, there is no further explanatory “why” question to ask about why the properties are clustered in nature like this; they just are. This situation is different from the one that we encountered with regard to family resemblances, where human interests decide which property combinations are interesting to consider together, without ontological commitments to real causal linkages between those properties.

In fact, Kendler, Zachar, and Craver give us a diagram of this kind of HPC kind (the kind where the cluster itself is the mechanism), and they caption it with the following: “One possibility for a property cluster kind in which individual clinical features (labeled F1 to F5) are causally inter-related to one another. There is no underlying essence that is responsible for the clustering of the symptoms. For example, if the disorder is major depression, suicidal ideation (F2) might be caused by both depression mood (F1) and feelings of guilt (F4).”

4.4 As HPC kinds, mental disorders are causal networks, not the entities postulated by latent construct theory

Recall that one of the larger background issues of this discussion about the causal structure of mental disorders is the question of naturalness, reality, or validity. If mental disorders are HPC kinds, what does this structure say about whether they are naturally-occurring or not? What does it say about where lines can be drawn between different disorders, on the one hand, and between health and disorder on the other? This is an apropos time to discuss the distinction between “discrete categories” and “dimensional continua.” Categorical views take our diagnostic constructs to refer to entities with clear-cut boundaries in nature. You either have clinical depression, or you do not. You have either anxiety or clinical depression, and though you could have both, you would not have a combination depression-anxiety, because those are two different entities. On a dimensional view, diagnostic constructs are on a spectrum, without definitive boundaries. On one end, you definitely do not have clinical depression and on the other you definitely do, but in the middle, it is somewhat unclear. A dimensional view does not necessarily preclude that kind from being a natural kind; it just means that nature might not “decide” in mid-range cases. As a physical example of such dimensionality, consider
hypertension. While high blood pressure really is a condition that naturally occurs, there is no way of specifying (outside of human interests) when someone switches over from low to normal to hypertensive.

The traditional picture of both categorical and continua-based diagnostic constructs views these constructs as making reference to latent entities or structures. When something is a latent construct, we do not have direct evidence of it; instead, we see its indicators, which are, in the case of mental disorders, the symptoms. We infer from the symptoms that the latent construct (the disorder) exists. Thus, psychiatrists infer from the presence of certain symptoms that someone has this unobservable underlying thing/entity, that we call “depression.” Again, this way of thinking tends towards essentialism and a one-cause/many-effects picture. But what if there was no such latent construct? In fact, what if there was no construct at all, just the cluster of symptoms? Cursory consideration of this possibility might lead one to say, “yes, but then mental disorders are just syndromes – and syndromes cannot be validated in the same way that diseases can be (via future neurobiological or genetic identifications), so this kind of theory takes us back to where we were – questioning if mental disorders are real, and yet unable to defend a claim that they are.”

Of course, we have already seen that this skeptical answer is far too hasty. In fact, an HPC theory allows us to validate the existence of mental disorders without deference to a latent construct. There is no construct; there is just the cluster, holding itself together in virtue of its causal linkages. On this kind of view, a given mental disorder, like clinical depression, is not an entity – it is an entire network. I started this chapter by claiming that while we typically do not doubt the reality of symptoms associated with mental disorders, many people do doubt the existence of real illnesses that underlie those symptoms. The kind of view that I am offering goes a step further and says that in not being able to doubt the existence of the symptoms – and in the overt recognition of the co-occurrence of, and direct causal relations between, these symptoms (provided by statistical scientific studies) – that one does not need more in order to validate the disorder. The disorder just is this clustering of symptoms, it just is the causal network in which these properties figure and make one another more likely. Thus, as Boorsboom writes, latent variable models “view correlations between indicator variables as spurious,” because after all, the explanation for the correlations is the single cause – the latent variable – causing each indicator individually. However, on a causal network or HPC theory, the relation between the latent construct and the indicators is mereological, and the construct is validated by the existence of its parts (the symptoms). The construct is multiple realizable, in a manner of speaking, because an instance of it just requires an activation of some part of the relevant composite causal network.

So, we are to imagine all of the possible symptoms (biological, psychological, social, etc.) of all categories of mental disorders in a large causal network, with “clumping” occurring to indicate different observer-independent networks. With such clumpings, we would see various central symptoms (again, those with more – and stronger – causal connections) closer to the middle of clusters, with peripheral symptoms on the outside, but causal arrows going in all manner of directions. This clumping – which is part of the natural causal structure of the world – is what is responsible for the individuation of mental disorders, and so it is not up to us. Thus, mental disorders are natural kinds.
5. Why think that mental disorders are HPC kinds? What benefits do we gain from acceptance of this theory?

As we shall now see, the theory that mental disorders are HPC kinds both (1) gives us a tremendous amount of explanatory power and (2) accounts for obvious features of mental disorders better than essentialist theories do. I now turn to elucidating these benefits of the theory. What follows is thus a list of reasons to defend an HPC theory of mental disorders against either biological essentialism or various non-natural-kind theories:

(a) HPC kinds respect the autonomy of the mental

Strong Disease Model theorists push a claim that mental disorders are just like physical disorders, except perhaps that they have mental symptoms. But they take these symptoms to be the epiphenomenal results of some underlying, more fundamental biological process. Theories that view all mental symptoms as indicators of latent constructs make this same error. On the contrary, other theorists argue that the ‘mental’ in the term ‘mental disorder’ is not just an idle label for a category of symptoms. Rather, the use of the term ‘mental’ refers to the fact that the mental symptoms themselves help to cause and sustain the disorder. Recall the mereology remark that I made just a bit ago: certain mental symptoms constitute cases of a mental disorder, meaning that they participate outright in the causal network. They are not epiphenomenal. In fact, in many individual cases of mental disorders, mental symptoms are relatively central (someone will say something like “if I could just get rid of this pervasive feeling of dread, I feel like all my other symptoms would relent.”) In other words, in many cases, the symptoms that do the most to maintain homeostasis of the symptom-set are mental. On an essentialist theory, the essence (taken to be a neurobiological kind) could exist without the manifestation of any characteristic symptoms, including the mental ones that look so central to each disorder.

Although it is beyond the present scope for me to go too much into moral responsibility and autonomy, I believe that this autonomy of the mental allows for a middle-ground between blaming the patient and seeing the patient as a passive victim. While it is true that patients who suffer from mental disorders often feel a lack of volition – likely due to the causal network and its sustainers taking over – the accessible mental symptoms, with their real causal powers, may serve as places where a therapist – or oneself - can intervene “above” the neurobiological level. In the next chapter, I will delve further into the question of the interventional autonomy of the mental. But for now, let us just note that an HPC theory takes the mental seriously as a locus of causal power.

(b) The HPC model is more accurate with regard to symptom-symptom causal relations and heterogeneous, complicated etiologies

This point piggybacks off of what was presented in (a) about the autonomy of the mental. The point about the causal power of mental symptoms can be applied to all properties in the cluster. Once again, the distinction between etiological factors and symptoms break down, as all of the properties are likely to play both roles in some capacity or other. The properties are interdependent. This point speaks in favor of an HPC theory mostly because of the observable evidence about mental disorders with regard to how they present and then unfold. As examples, consider: insomnia causes tiredness, guilt causes suicidal ideation, phobias cause avoidance behavior and avoidance behavior causes habituation to staying inside one’s home. For instance, even if genes cause initial disordered thinking in a schizophrenic person, this does not make the disordered thinking epiphenomenal. The disordered thinking might directly cause
familial strife, leading to increased cortisol levels, and further dysregulation of the expression of the original gene. Barring cases of sudden brain trauma or a surgery, it is usually very difficult to locate the cause of a person’s mental distress, or even the time at which the distress began. More often than not, the initial causation is more complicated, multifactorial, and gradual than such examples allow. Furthermore, the genesis of the disorder tends to be different with every individual case. The HPC theory can explain these features by deferring to the notion of a causal network rather than a causal chain (from essence to epiphenomenal symptoms).

To be fair, biological essentialism does not necessitate the epiphenomenalism of symptoms, just because those symptoms radiate out on independent paths from the essence. This kind of structure on its own certainly does not prevent those symptoms from then interacting with one another in symptom-symptom causal interactions. But to get to that point, the essence has to initially be present, which is what I am arguing against. My point is that most any symptom can trigger the presence and growth of the causal network.

(c) HPC kinds allow for variable members

Recall that according to HPC theory, every type of mental disorder is marked by a characteristic set of properties that make one another’s presence more likely. However, the individuation of a given category of disorder is made on the basis of having some of this cluster/causal network “activated,” meaning that no one property is necessary in order to have a case of the disorder. What this means is that the cluster is multiply realizable. If we imagine the composite network for a given disorder represented topographically on a piece of paper, one person might partake in the causal structure indicated on the left side of the page, while another might partake in a portion at the bottom of the page. They are partaking in the overall same causal structure (because the network is held together homeostatically); they are just partaking in different parts of it.

This is why, for example, cases of depression can differ so markedly from one another: one person is rageful, cannot sleep, has no appetite, cries for no reasons, while another person is despondent and sleeps and eats all the time. These people have likely “entered into” the causal network via different routes. As time goes on, their cases may worsen and resemble one another more, as they make their ways deeper into the network. Also, tokens of types of mental disorders might also be heterogeneous because their central sustaining properties are distinct, or because they are maintained via different internal or external homeostatic mechanisms. In order to offer this same high degree of explanatory power, an essentialist theory will need to offer a reason for why all sufferers of a disorder, sharing a common essence, look so markedly different in their manifestations of the disorder.

(d) HPC kinds allow for fuzzy boundaries between health and disorder

I noted that the HPC theory will claim that in anyone suffering from a particular disorder, the relevant causal network has been “activated.” I also stated that on this theory, diagnosis is a two-step process that includes evaluating “(a) whether the causal system of symptoms is at all activated in a person, and (b) where in the causal sequence a person is located at the time of the diagnostic interview.” But, how much of the causal structure needs to be activated before someone has that disorder? The answer is: we do not know, perhaps because there is no such answer. Or, if there is an answer, it will come after careful empirical studies of a particular category of mental disorder, studies that attempts to find and highlight the symptoms that tend to be central in instances of that disorder. Sometimes someone seems more depressed than healthy
people do but still does not seem to have depression; HPC theory explains this scenario by its claim that one can partake more or less in the overall causal network. Depending on how much of – and which parts of - the relevant causal structure is present, individual cases of a given disorder will vary from more to less severe.

Again, it is not necessarily the case that essentialism cannot account for fuzzy boundaries between health and disorder. For instance, hypertension has a concrete biological cause, but sometimes it is indeterminate whether someone is hypertensive or not. Essentialism could place tokens of a disorder on a continuum as well. It is just that a theory of HPC kinds has a better answer to why this spectrum exists. The essentialist would most likely have to defer to background conditions that made a single essence more or less causally powerful. For instance, the essentialist might attempt to explain a low-grade case of schizophrenia (ie: a person who hears voices only every now and then) by giving a story about why a gene was not fully active in that case (perhaps because other genes are required for the essence to reach its fullest fruition). The HPC theorist’s explanation is far simpler and is just that a person can partake in more or less of a problematic causal network

(e) HPC kinds explain co-occurring disorders

Certain mental disorders tend to show up in the same person with a high degree of probability. For instance, clinical depression and anxiety tend to co-occur. So do depression and substance-use disorders, as well as anxiety and substance-use disorders. If every kind of mental disorder was characterized by its own essence or its own underlying mechanism, such co-occurrences would be hard to explain. For instance, if there were neurobiological essences, we would need some empirical theory about why two essences were likely to show up together (gene linkages, perhaps). However, a theory like HPC, which makes the property cluster paramount, more easily accounts for comorbidity. Consider again the entire pathological “symptom space,” with clusters/bunches representing what we call different disorders. Some clusters might be closer to one another in this space on account of bridging symptoms, symptoms that are constituents of both disorders. Not only do these bridging symptoms pull the disorders close together, but they also serve as pathways from one disorder into the other. Therefore, symptoms from different clusters are “literally intertwined with one another.”

Borsboom and Cramer employed network analysis techniques (on the composite causal space of categories of mental disorders) in order to identify the 3 bridge symptoms between depression and anxiety: fatigue, difficulty sleeping, and difficulty concentrating. Let’s consider a person who suffers from depression. She experiences low mood and a loss of interest in daily activities. As this condition progresses, she finds that she has difficulty sleeping. This symptom provides an “entrance” into the anxiety symptom cluster via a direct route of causal connections. Her difficulty sleeping causes her to become on-edge.

So, how do we diagnose this patient? Does she have both depression and anxiety? Does she just have one of these disorders, qualified with an adjective relating to the other (depressive anxiety or anxious depression)? Or does she have a depression-anxiety hybrid? Such a decision might just be a matter of convention. But the conventional decision would not change the ontological facts about the causal pathways between these symptoms, and the degree to which the causal networks are being activated.

Even if narrow categories of mental disorders (where we have different diagnoses for, say, gradations and type of depression) have diagnostic and/or interventional utility, perhaps widening the diagnostic categories would more properly exhibit the real structure of these
symptoms. Borsboom and Cramer call the kind of larger cluster that occurs between substance-use disorders, depression, and anxiety a “giant component” or a “small world.” While it is true that the small world is composed of 3 clusters, the small world itself has unity via a highly interconnected and large causal network, which is relatively far away in “symptom space” from other disorders. When we map the total causal symptom space of mental disorders, we find “points of rarity,” where clusters are relatively cut off from the encroachment of symptoms outside of that network. These “distances” can explain both the presence of co-occurring disorders, as well as why some disorders rarely ever show up together.

(f) HPC kinds explain vulnerability and worsening of conditions

Theorizing about mental disorders goes beyond diagnosis and treatment. It also involves prevention, both with regard to who is vulnerable to getting a mental disorder to begin with, and then how a particular individual can avoid relapse. HPC theory can help us with these aspects of disorders, with its emphasis on causal links between symptoms. Remember that this chapter began by posing what I called the Structural interpretation of the nature of mental disorders question. HPC theory claims that mental disorders are constituted by homeostatic property clusters. But a structure is not just its properties; it is also their relations. And more can be said about their relations beyond “so-and-so properties are causally connected.” The connection strength between certain symptoms matters. If two symptoms are strongly connected, then the presence of one is very likely to cause the presence of the other. Recall that “central properties” were defined disjunctively: either they are connected to many other properties or their causal connections yield relatively higher probabilities of the presence of their effects (or both).

Borsboom and Cramer define risk structurally in the following two ways: 1) risky structure involves symptoms that have strong connections. They liken this kind of structure to dominoes that are placed closed together; hence it is a dispositional property. If one symptom in a strongly-connected network is triggered, it is likely to activate much more of the network (one domino falling will knock down a slew of others). And 2) central symptoms – those with more causal connections – have a stronger causal influence on the entirety of a network. It may be that central symptoms are different for each individual’s causal network, but regardless, it is important to identify what these symptoms are, and to put the most interventional power into staving them off. If insomnia is only a peripheral symptoms of someone’s depression, then a bad night of sleep is not a big deal; but if humiliating experiences are a central symptom, then it is important to focus therapy here, towards building tools to keep the effects of such experiences from escalating and progressing.

Recall, though, that an HPC theory does not privilege any level of explanation, and that the properties from the cluster can come from any and all levels. Borsboom and Cramer make a provocative suggestion about the role of genes in risk and vulnerability. They suppose that perhaps the causal connection strengths between symptoms are determined genetically. Examples of genetically-influenced vulnerabilities might be how much fatigue a person experiences as a result of insomnia, or how much the typical stressors of daily life lead to negative affective states. If one person is genetically predisposed to be extremely affected by daily stress while another is not, this predisposition could explain the former person’s increased vulnerability to developing depression.

Note that an HPC theory can also account for the chronic nature of some mental disorders. Perhaps someone’s mental disorder goes into “remission” when enough of the causal structure disappears. However, the disappearance of a significant portion of some of the causal
structure does not mean that more of the causal structure might not lie in wait, latent and dispositional. For instance, take the genetically-based vulnerability to being affected by certain experiences in extreme ways. This trait does not go away just because someone’s depression gets better. It just becomes de-activated. But its continued presence means that the person might have to be continuously vigilant about keeping any causes of activation of that system at bay, lest they set off a new snowballing of the network.

(g) A theory of mental disorders as HPC kinds is non-reductive yet naturalistic

While essentialism privileges the biological, chemical, and physical levels of description, HPC theory does no such privileging. On an HPC theory, properties are more or less central depending on their roles in the causal network, not whether they are biological vs. psychological vs. social, etc. Moreover, this lack of reduction does not make mental disorders into constructed kinds of things. The property cluster and its network of causal relations are absolutely real things, whose structure exists independently of either human classification or recognition of this structure. The clustering of the properties is nomologically-determined, and hence HPC kinds have all of the features that make natural kinds so integral to science, including the ability to support inductive inferences.

(h) Essentialist natural kinds are a variant of HPC kinds

But, one might argue, couldn't it be that some mental disorders are not HPC kinds? Might we find particular mental disorders that do have the required essences? To this question I answer “Yes.” As heterogeneous as the various mental disorders are, it is likely that they exhibit all kinds of causal structures. As I stated earlier, the topological possibilities of causal structures are infinite. Nonetheless, I claim that the majority of these structures will be cluster kinds. This claim does not commit me to saying that the biological level is not important (it may turn out empirically that every mental disorder is marked by central symptoms that come from the biological or genetic levels). Nor does it commit me to saying that no mental disorders have biological essences. In fact, biological essences are HPC kinds too, albeit with a far simpler causal network – a common cause and a series of unidirectional causal chains, held in homeostasis by the presence of the essence. So essential kinds are HPC kinds, just taken to a certain extreme. It could be that apparent mental disorders with fundamental biological explanations – ie: the mental problems associated with Huntington’s Disease – are essentialist kinds. Thus, this picture of mental disorders does not conflict with the possibility of discovering that some mental disorders have essences after all. Essences are just another type of HPC kind.

(i) HPC theory allows for symptom-based therapy consistent with interventionist views of causation

An HPC theory takes symptoms to be constitutive of a disorder, rather than manifestations of an underlying cause. Such an outlook affects what we take “treatment” to be. On an essentialist theory, one would think that a proper target for treatment would be the “underlying cause,” with biological substrates given precedence, and thus one might focus on physical interventions – medications, ECT. Although therapists would obviously try to alleviate the painful symptoms, they might not consider someone “treated” until the underlying cause was dealt with. On the other hand, if the symptoms constitute the disorder, then interventions that
directly target those symptoms are actually more efficacious in treating the disorder than they have been given credit for.

While essentialist theories are in line with mechanistic views of causation, an HPC theory aligns itself more with interventionist views of causation. An interventionist view of causation asserts that Y is a cause of X when a surgical intervention on Y has a causal impact on X. The benefit of this kind of view of causation is that the targets of interventions need not have carefully worked-out mechanisms associated with them. Therefore, there are many more plausible targets of interventions – possibly every property in the cluster. When I say “plausible,” I mean that we intervene directly on that target to effect change, without thinking that such targeting will really work indirectly, after “trickling down” to a lower-level.

6. An exploratory conclusion: bringing together the Disorder and Structural interpretations of the “what is the nature of mental disorder?” question

Recall that when one poses the question, “What is the nature of mental disorders?” she could mean one of two things: On the Disorder interpretation, she is asking what makes mental disorders bad things. On the Structural Interpretation, conversely, she is asking what the causal structure of mental disorders are – how their symptoms and causes relate – and perhaps also, whether or not they are natural kinds. I wrote in my introduction to Chapters 2-4 that “whichever kind(s) of structure(s) that we land on for mental disorders, the structure on its own does not necessarily imply anything evaluative. Of the structure given above [a diagram of an essentialist natural kind], for example, it makes no sense to ask whether it is good or bad. The kinds of things that exhibit this structure might themselves be good or bad according to some perspective, but the structures cannot be good or bad.” It might just be that the conditions that are mental disorders are natural kinds, but that these conditions are not mental disorders until further judgments are made (this would be an objectivist-constructivist hybrid, and the Harmful Dysfunction theory is one theory that builds itself from this framework). Thus, goodness or badness cannot be read off of a structure devoid of content.

In this section, I would like to explore if there are possible ways to bring the Disorder and Structural interpretations together. Because goodness or badness is not a property of structures themselves, I will not be suggesting that having an HPC structure is sufficient for something’s being a mental disorder. All sorts of things - species, metals, emotions - are candidate HPC kinds. Structure is not of the right logical type to play a role in such a sufficiency claim. Nor am I suggesting that having HPC structure is necessarily a bad thing. In fact, I imagine that good conditions exhibit this kind of structure as well. We can imagine this kind of scenario with regard to good health, for example, or even one’s personal goals. A good diet is likely a central property of what might be a causal network representing “good health.” Once one achieves a good diet, other health-related properties arise concurrently at a rapid rate: one’s cholesterol drops, one loses weight, which causes one to move easier, which causes one to exercise, and lose more weight, which motivates one to be better at his diet, and so on. Once a large part of the cluster is activated, or once a central property is activated, it is difficult to extinguish the network, because so many different pathways are holding the “net” in place. So, this kind of homeostasis can be a good thing.

Because of their homeostatic natures, HPC kinds are robust networks when much of their causal structure has been activated. Because of their feedback loops and the multiple interconnections between properties, once a crucial degree of structure is achieved, it becomes very difficult to make that structure fall apart. This is all well and good in the case of good health
and goal-setting, but it is a terrible thing when it comes to mental disorders. If mental disorders were essentialist natural kinds, and if we had knowledge of what their essences were, we could come up with interventions targeted to wipe out those essences. Doing so would extinguish all of the properties caused by that essence.

However, for HPC kinds, the situation is nothing like this. If treatment helps to extinguish peripheral symptoms of a case of mental disorder, it is possible that relief in these symptoms will have only a minimal impact on the activity of the network as a whole. If treatment is aimed at central symptoms, it could be that one causal pathway to/from those central symptoms is blocked, while many others remain open. Remember, we are talking about homeostasis here – the symptoms make one another more likely. I claim that homeostasis is closely linked to another phenomenon, self-reinforcement. For a self-reinforcing network, the system will shift so as to maintain itself in the face of any intervention. In other words, if the causal network begins to fall apart, the network will “work” to get itself back into balance. Over time, as the network grows, it takes on a “life of its own,” because the network becomes less and less sensitive to external perturbances (external to the network, not to the person).

A suggestive claim is that when a mental disorder’s HPC network has a self-reinforcing “life of its own,” this condition is intrinsically bad for the individual. It is bad enough that the cluster involves unpleasant symptoms, but furthermore, the self-reinforcing nature of the network makes it extremely difficult to obliterate those symptoms. Typically, one does not want to be in this condition, and furthermore, one feels powerless to get out of the condition. As I said, small attempts to extinguish a particular symptom or “unwind” a particular causal pathway tend to be unsuccessful because of the many linkages in the network as a whole. And even worse, it could be that unsuccessful attempts to combat the network serve as further input into growing and sustaining the network. This self-reinforcement is due to the peculiar mental ontology of mental disorders – the fact that mental symptoms are non-reductive causal nodes in the network. For instance, if one tries to overcome her negative cognitions about herself but fails, this failure might actually bolster that very symptom. Or notoriously, rumination looks to beget rumination; the more one tries to stop thinking about her depression or her compulsive thoughts, the more such thoughts assail her and grow out of control.123

It is interesting to me that the end of my prior chapter on genetically-open functions ended with an emphasis on the importance of a certain kind of flexibility in one’s mental life, while this chapter has ended with an emphasis on a loss of flexibility through losing control of one’s self. While I do not pursue this path further, I want to suggest that it is worth thinking through the following thesis: (some) mental disorders are dysfunctions of genetically-open functions. Genetically-open functions involve a disposition to flexibly respond to the world; hence, their dysfunctions involve a loss of flexibility. When an agent loses flexibility in her mental life, this loss is intrinsically bad for her, as the homeostatic causal structure of her disorder becomes self-reinforcing and takes on a life of its own. Therefore, the dysfunctions involved in some mental disorders are already bad without the need for an additional judgment of “harmful.” If we were able to defend a thesis like this, we could actually make a Dysfunction theory of mental disorders fully objectivist. Again, I do not go further with this thesis here; but it is something to think about when it comes to the implications of the theories that I have been defending.

Another potential way to draw the Disorder and Structural interpretations together is to claim that mental disorders are indeed paranatural kinds – absences of properly functioning mechanisms. This claim is similar to calling mental disorders dysfunctions and claiming that there are many ways for each particular function to go wrong. But as I stated earlier, being an
absence of a natural kind does not mean that there is nothing there; in fact, the absence of a well-functioning mechanisms is best described in terms of a cluster of properties held in place by some homeostatic mechanism, which may or may not be nothing beyond those properties and their relations.

My main intent in this chapter was to support the related claims that mental disorders are networks, not entities, and that they are HPC kinds. The exploratory aim of this last part of the chapter was to explore a possible link between the supposedly independent Disorder and Structural interpretations. Throughout the chapter, I made reference to interventions on mental disorders, and how interventional strategies might change depending on which structural theories we accept. An HPC approach to mental disorders, contrary to an essentialist approach, opens up the possibilities for successful interventions in a non-reductive framework. I now turn to the question of whether psychotherapy – and other “mental” forms of treatment - has any ineliminable role when we accept a weak Medical Model of mental disorders in a materialist framework.
1. Introduction: 2 treatments for depression, and the materialism-dualism tension

When it comes to the treatment of clinical depression, we must note 2 empirical claims that are supported by multiple studies: 1) Antidepressant medication (ADM) and psychotherapy (PT) are equally effective in treating depression, and 2) of those treated with either ADM or PT, those treated with psychotherapy tend to stay well for longer and to suffer fewer relapses after treatment is terminated. In conjunction, these claims appear to support a conclusion that not only is PT an effective treatment for depression, but also that it provides benefits that ADM does not. My goal in this paper is to defend the ineliminability of psychotherapy from a completed science for the treatment of depression. My use of the term “completed science” indicates that this ineliminability is not merely pragmatic. It is not just that PT does certain things that we cannot yet accomplish through medical methods. Rather, my point is the stronger claim that it is not possible for PT to ever be eliminated by solely physical methods as a treatment for clinical depression, because no merely physical method could ever duplicate the results of PT.

I should state at the outset that I make this argument while operating within a materialist framework, which might sound strange, as I am defending the integrity of a therapy that is typically couched in mentalistic terms. From here, one might claim that if PT is actually efficacious, then this efficacy must support a dualistic picture of the world, on which physical symptoms of depression are treated via physical interventions, and mental symptoms are treated by mental interventions. But we need not accept such a conclusion. My view unifies both primary treatments of depression via an appeal to neurobiology. In sections 3 and 4, I introduce an empirical hypothesis on which general categories of both depressive symptomology and interventions can be mapped onto distinct brain areas. Respectively, these are the limbic system and the prefrontal cortex. My argument that all interventions on depression work by changing the brain in systematic ways does not lend itself to dualism. The first part of my paper aims to present an empirically-supported hypothesis of the systematically different ways in which ADM and PT act on the brain.

Although I emphasize the importance of the brain for both ADM and PT, I do not thereby align myself with the materialist who supports the collapse of psychiatry into just another branch of physical medicine. Those who defend such a move might pose the question: if PT changes the brain via the indirect route of discussing one’s problematic emotions and cognitions, then wouldn’t it make more sense to find a way to bypass this psychotherapeutic medium and instead change the brain directly? I answer no, and my reasoning has to do with the manner in which PT is performed.

In section 5, I claim that PT exhibits interventional autonomy. Interventional autonomy is comprised of 2 features: 1) the global nature/holism of the substrates of psychotherapeutic interventions, and 2) path-dependence. The former has to do with psychologists being able to change a large and distributed process of the brain in one neat set of moves. The latter refers to the importance of the accompanying changes that are dependent on the process of PT. I argue that these changes are those that are responsible for PT’s increased resilience against relapse when compared with ADM. We cannot just drop these features out and ensure the same result.
Unfortunately, both features of interventional autonomy fall prey to a common objection about the current limits of medicine: if all of the changes wrought by PT are materially-realized, why couldn’t there be a brain surgery that led to precisely these same physical states? I use a thought-experiment involving fine-grained “brain-tweaking surgery” to explore this move. The paper then proceeds in section 6 with my defense of the interventional autonomy of PT in the face of this objection. I discuss the question of what the success-criterion for recovery from depression should be, rejecting normative and epistemic approaches. This discussion is what ultimately marks this paper as philosophical or conceptual, rather than merely empirical. I argue that even when we adhere to a results-based criterion, the path-dependence of PT is still crucial to recovery. Thus, I show that even if all of the neurobiological results of PT could be duplicated via physical means, still the person who underwent such an intervention would not be well-prepared to function in the world.

I end the paper in section 7, by raising a further objection that perhaps my rendering of the thought-experiment unjustly stacks the deck in my favor, because on it, only inessential features of the thought-experiment turn out to be objectionable. I overcome this objection by making certain constitutive claims about both clinical depression and recovery, namely, that both involve not just brains, but ways of being in the world. As a learning process, PT (and recovery) is necessary experiential. These moves are meant to defend the ineliminability of psychotherapy as a treatment for clinical depression, even as psychiatry becomes increasingly scientific.

2. A caveat about the deployment of empirical studies in this work

I start this paper by noting some claims that I state are “supported by multiple studies.” It would be unfair of me not to admit that another person might read these studies – as well as additional studies – and come to alternate conclusions. After all, the results of any scientific study are consistent with multiple interpretations, and these interpretations can change if new sets of relevant data are introduced. I wrote the present chapter because I feel that defending the cited empirical claims is justifiable in light of the evidence at hand. However, one might be skeptical about getting on board with my project to begin with, if she thinks that evidence points to any of the following conclusions: (i) neither PT nor ADM is particularly successful in treating depression, (ii) either PT or ADM is more efficacious than the other, (iii) when either treatment is successful, we do not know why, hence there is no systematicity in interventional success, or (iv) there is no indication that PT and ADM act directly on different parts of the brain. I draw numerous conclusions from empirical studies, and an opponent could disagree with me on any or all these points.

For instance, Valerie Hardcastle writes that “it is widely known and accepted that the combination of pharmacotherapy and talking therapy is the most effective treatment for depression.”\textsuperscript{140} She cites Cuijpers\textsuperscript{141} and Hollon.\textsuperscript{142} As we will see in section 4 when I discuss the synergistic effects of the two types of therapies, I do not necessarily disagree with her point. Nonetheless, though she herself is not committed to the following conclusion, her point is consistent with a view on which PT and ADM do the same exact thing, thus we should use both so that one gets a “double dose” of the treatment. Hardcastle also disagrees with my optimistic assessment that distinct kinds of interventions act systematically on distinct kinds of symptoms/brain areas. She writes, of “best practice recommendations,” that
“We cannot tell which type of therapy, if any, will be helpful on the basis of presenting symptoms or patient history. Some patients with low mood, but no cognitive distortions, for example, might be very responsive to psychotherapy, but get no relief from SSRI’s...It is my contention that there is no evidence that pharmacotherapy works via subcortical mechanisms while psychotherapy works via cortical ones. In fact, we really have no good explanations for why either therapy works, just as we have no detailed and widely accepted explanation for what depression actually is.”

I invite skeptical readers to review the scientific studies that I have cited. Together, I think that these studies provide a provocative background on which to start asking about the efficacy and mechanisms of psychotherapy. Of course, it is possible that I am wrong. However, even if there are competing hypotheses about the outcomes of these studies (including the hypothesis that these studies give us no insight into what is going on neurobiologically when we treat mental disorders), I do think that my empirical hypothesis is plausible, given the data, and that is reason enough to take the philosophical work that results from it seriously.

3. All clinical depression symptoms and interventions can be explained by neuroscience: introducing the higher-lower distinction

Recall that I am arguing from a materialist standpoint: even if neuroscience can fully explain the efficacy of PT, there are still reasons to maintain that PT is ineliminable. It is important to note that I do not claim that psychotherapy is either necessary or ineliminable in all individual cases of depression, just that it is necessary and ineliminable in the treatment of the disorder as a kind. In this section, I aim to argue for the truth of the conditional that opened this paragraph. I show that depressive symptoms and their interventions can be systematically explained by appeal to neuroscience, through the deployment of a “higher-lower” distinction. My goal in highlighting this distinction is to use it in an empirical hypothesis about the neurobiological mechanisms behind both ADM and PT. This hypothesis is that, in spite of being equally effective in treating depression, these two interventions act on distinct psychological and neurobiological substrates, and therefore, cannot be substituted for one another.

The following table attempts to show that there is a way to distinguish higher vs. lower depression symptoms, map these higher vs. lower psychological symptoms onto malfunctioning higher vs. lower brain areas, and eventually, to distinct kinds of interventions. Note that the following table expresses general correlations, not reductive identities.

<table>
<thead>
<tr>
<th>Categories of Symptoms and/or functions</th>
<th>Lower</th>
<th>Higher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood, vegetative, cognitive/implicit schemas, automatic emotional responses</td>
<td>Emotional regulation and inhibition, explicit cognition, reflective thinking, schema construction</td>
<td></td>
</tr>
<tr>
<td>Subcortical, limbic system, amygdala</td>
<td>Prefrontal cortex</td>
<td></td>
</tr>
<tr>
<td>ADM</td>
<td>Psychotherapy</td>
<td></td>
</tr>
</tbody>
</table>

In non-pathological cases, the functional architecture of the relevant higher and lower brain areas is well-supported in the biopsychology literature. When we look at the pathological cases, we get a breakdown into 3 categories of symptoms/neural correlates.
1) **Lower:** Depressed people tend to demonstrate low mood and the inability to regulate emotions. Their cognition is dominated by the activity of negative schemas and cognitive distortions. Empirical studies support the neural realization of these symptoms in lower, subcortical brain regions. Empirical data supports the theory that moods are dependent on neurotransmitter concentrations in these regions, particularly serotonin and norepinephrine. Low concentrations of these neurotransmitters are correlated with heightened limbic activity, which indicates that one’s threshold for experiencing negative emotion is lowered, as well as sensitized to more stimuli. More limbic activity is therefore associated with more phenomenological distress.

2) **Higher1:** Higher brain areas are responsible for emotional regulation of the limbic system, via inhibitory cortical-limbic connections. Emotional regulation makes negative emotions less likely to initially arise. Furthermore, when these emotions do arise, a person is better able to handle them. Regulation also dampens the activation of negative cognitive schemas.

3) **Higher2:** The cortex is also the locus for active reflective capabilities, problem-solving, and the creation of new thought-patterns and schemas, via cortical-cortical connections. An important aspect of these connections, that I will get to later, is that they involve specific content.

With regard to both of these “higher” functions, those with depression tend to show decreased cortical activity and decreased modulation of limbic regions by the cortex.

On the basis of the previous divisions, I claim that improvements in both higher functions are responsible for the increased resilience of PT when compared to ADM, and that Higher2 in particular underlies what I will say about interventional autonomy and path-dependence. But before I can make these points clearer, I must elaborate on the concept of a ‘schema.’ The way in which schemas supposedly play a role in PT is essential to my claims about the ineliminability of PT. Despite this centrality, I do not take negative schematic content to be constitutive of clinical depression; I do not claim that – by definition - all people with depression have negative schemas. In section 3, I will explain how my empirical hypothesis fares with those cases that are not so constituted. But to defend PT from elimination, it is enough for my purposes that a large number of depressed patients exhibit negative schematic content. Recall that I am interested in what can be said for PT with regard to treatment of depression as a kind, not a token.

Beck, an early proponent of cognitive theories of depression, emphasized the role of “maladaptive schemas” in the pathology of depression. Clark et al. interpret Beck when they write that schemas are “enduring structural representations of human experience…that direct the identification, interpretation, categorization, and evaluation of experience.” Hence, schemas are implicit – preconscious – networks of beliefs and procedural/behavioral frameworks. The content of schemas becomes filled in through developmental time, via associational learning and repetition. At a neurobiological level, learning occurs as a result of neural wiring via synchronous firing. The colloquial neuroscience phrase is: “those neurons that fire together, wire together.” As an example of the genesis of such a schema, imagine that a depressed person visits his mother and interprets her as disapproving of him. Neural pathways representing the thoughts, “Here is my mother,” and “My mother thinks that I worthless,” fire at the same time, which strengthens their connection, making them more likely to fire together in the future. Every subsequent joint firing makes them again more likely to fire simultaneously, until the mere thought of this person’s mother can trigger very strong depressive cognitions in him. Proneness
to this kind of thinking can lead to a positive feedback loop, where more and more contingent factors can get wired into the schema.\textsuperscript{156}

I believe that we can infer two important and related points from these considerations: 1) The decreased limbic activity witnessed after use of ADMs may mean that negative mood and emotional reactivity are dampened, but this outcome is still consistent with the deeper schematic structure remaining unchanged, and 2) neither heightened mood (Lower) nor general emotional regulation (Higher\textsubscript{1}) involve \textit{specific content}. However, those with depression suffer from cognitive distortions and negative schemas that involve personalized and idiosyncratic content, a fact that is fundamental to my argument.\textsuperscript{157} It is fundamental because on my hypothesis, only PT can directly change the idiosyncratic content of schemas. Changing this content is what underlies the \textit{resilience} against depressive relapses.

Let me explain this key point about resilience. With the functions of the higher brain in mind, we might cognize the goals of PT as follows: 1) boost activity of Higher\textsubscript{1}: One will learn to regulate negative emotions and to make implicit negative schemas explicit in order to reflect on them and make them less active. 2) Boost activity of Higher\textsubscript{2}: Build new, positive schemas by the repetition of certain experiences, behaviors, interpretations, and reflections. To use our earlier example, a therapist would get the patient to process the situation with his mother, to make his implicit schemas explicit, and then to challenge them with new interpretations. The patient then needs to go out into the world and spend time with his mother so that he can couple the neuronal pathways coding for the experience of her with pathways coding for more rational interpretations of her behavior.\textsuperscript{158} Thus, neurobiological change is effected through consciously ignoring\textsuperscript{159} old schemas and creating new ones. Because PT creates new brain structures, the existence of the pathways representing new schemas will outlast the end of PT treatment. On the contrary, soon after ADM is terminated, serotonin concentrations tend to go back to pre-treatment levels and de-activated schemas become active again, without any positive compensatory schemas having been built. These observations explain the resilience of PT when compared to ADM.

4. Empirical hypothesis and explanatory power: ADM and PT function differently because their initial sites of action are distinct; however, some of their ultimate results will be similar due to bottom-up and top-down processes

Recall that my empirical hypothesis, based on the distinction between “higher” and “lower,” is that ADM primarily targets mood and emotional reactivity (and the limbic system), while PT primarily targets cognitive and emotional regulation and inhibition (and the prefrontal cortex). The qualification of “primarily” is significant, and is juxtaposed to “ultimately.”\textsuperscript{160} In order to explain this qualification, I introduce another empirical claim, in addition to those that I presented in my introduction: (3) Patients who undergo effective monotherapy (only ADM or only PT) typically will show many of the same long-term brain changes as one another.\textsuperscript{161,162,163} Thus, those who are treated with depression via either form of therapy have \textit{both} less overall limbic activity and increased cortical activity. This claim is consistent with the competing theory that states that ADM and PT are redundant therapies that act via the same mechanism to enact the same changes. However, I support the alternate interpretation that PT and ADM work via different mechanisms to enact different changes. And I believe that the empirical data supports this interpretation, as it demonstrates that even though PT and ADM lead to many of the same brain changes, there are systematic differences between ADM and PT’s effects both with regard to a) final results and b) timeline of changes.\textsuperscript{164,165,166,167,168}
In this section, I extend my empirical hypothesis and argue that the third empirical claim stated above is more accurate when it is reformed to state that the ultimate brain changes associated with both types of therapies are similar (not the same). I maintain that there is a global neural correlate of depression that involves subcortical regions, cortical regions, and their interactions. Any intervention might change a part of this correlate, or the whole. I argue that depending on where an intervention primarily acts, we will see similar but different results, and different timelines. Thus, if ADM acts on the limbic system, the treated limbic system might go on to change the PFC via bottom-up mechanisms, leading ultimately to heightened cortical activity in addition to the lowered limbic activity. Similarly, PT might initially act on the cortex, which then might change the limbic system via a top-down mechanism, again ultimately leading to lowered limbic activity in addition to the heightened cortical activity.

Each of these primary changes only tends to make the secondary changes possible, not inevitable. This hypothesis not only explains the neural data on brain changes and timelines, but it also can account for the decreased relapse rate in those treated with psychotherapy vs. ADM. Furthermore, by considering these interactions between the higher and lower levels, I will show how my empirical hypothesis has high explanatory power when it comes to understanding the multiple presentations of depression.

4.1 Bottom-up interactions: heightened mood as an enabling condition on cognitive change

I will start by investigating what I call “bottom-up” interactions. My empirical hypothesis can explain the following phenomena: patients who do not respond to PT, patients with a severe presentation of depression who seem to require medication, and patients who “plateau” in their recovery after a period of time on ADM’s. As we saw in the prior chapter on HPC kinds, manifestations of clinical depression are heterogeneous. Earlier in this chapter, I stated that some depressed people may not have negative schemas or cognitive symptoms, and indeed, some cases of depression look entirely mood/lower-level based. Some studies have shown that these kinds of severe cases of depression have a much better chance of being successfully treated via ADM rather than via psychotherapy. In such cases, mood can be so low that positive thoughts and overt efforts to “pull oneself up by the bootstraps” have no impact. In fact, I argue that this situation occurs when the limbic system is so hyperactive that even moderate cortical inhibition enacted via one’s will would not have much impact. Worse, failed attempts to get better through one’s individual efforts would only create, or provide reinforcement of, negative schemas related to self-efficacy.

Cases like this are likely to be those that do not respond to PT, at least not until after one has been treated with ADM. These considerations indicate that we can view a treated mood as an enabling condition on cortical changes. Such enabling is an example of a bottom-up process. Although a better mood makes hedonically congruent mental states more likely to occur, the change in mood does not itself create strengthened cortical-limbic regulation or positive schemas. Thus, the discussion of the “bottom-up” impact of mood not only explains why some severely depressed people cannot be helped through PT alone, it also explains why some depressed patients just need medication in order to improve.

Furthermore, it explains cases (i) where one “plateaus” (stops improving) after a certain amount of time on ADM’s as well as those cases that I cited at the beginning of this paper, where (ii) a person treated with ADM exhibits changes that make her look like she underwent PT as well. In both of these cases, as I hypothesized, ADM acts on the subcortical and limbic regions,
elevating mood, which then frees a person up to challenge her schemas from a reflective standpoint. But since mood is not identical to schemas, heightened mood on its own does not automatically lead to the acquisition of positive schemas nor to the ability to challenge one’s negative cognitions; a person has to have the right kinds of experiences to build up the neural pathways in the cortex that underlie subcortical inhibition and regulation.

If a person does not reflect on his schemas, or have the right kinds of experiences to change these schemas, then although his limbic system is less active, when it does become active, he has no way to cope. His mood changes will lead to improvement, and then he will stop improving, unless he is pro-active in building up his schemas and practicing his inhibition. Interventions at the lower level work bottom-up to effect the upper level, but not without the additional step of independently increasing activity at the higher level. Typically, a person needs professional help to increase activity at the higher level, but it must be the case that some people just encounter the right kinds of experiences, or else perform CBT on themselves (without necessarily being aware of that). Both of these scenarios can explain “spontaneous remissions.”

4.2 Top-down interactions: from emotional regulation to mood improvement and resilience

The same kind of logic as we saw with ADM pertains to the timeline of PT on its own. With PT, it has been demonstrated that cognitive and motivational changes (as well as cortical changes) precede vegetative/mood changes (as well as subcortical changes). This timeline is opposite of that seen with ADM, even if many patients have similar ultimate outcomes. So, I claim that higher brain changes can work top-down to elevate mood, not because they themselves are actively raising mood, but rather, because the person has acquired the ability to regulate the lower level processes such that negative emotions a) do not arise as often (as a result of direct inhibition), and b) when they do, the person has learned skills to keep these emotions from escalating.

My hypothesis about PT and ADM working via different mechanisms can explain the observation that PT takes longer to have an impact than does ADM. My explanation for this fact has to do with the different properties of the direct neural targets of these different interventions. Changing neurotransmitter concentrations (and the concomitant upregulation of receptors) takes relatively less time than the process of extinguishing negative schemas and then building new positive schemas (via neural networks) at the higher level. The latter processes are the result of associational learning and practice, which require repetition and the proper experiences as opportunities for such repetition. To use our earlier example again, a therapist would get the patient to process the situation with his mother, to make his implicit schemas explicit, and then to challenge them with new interpretations. The patient must then go out into the world and spend time with his mother so that he can couple the neurons coding for the experience with her together with neurons coding for neutral feelings or more rational interpretations.

The payoff for this extra time is the resilience of these network changes. Neural pathways that have been honed require much non-use to become “extinct.” Repetition of experiences over some time frame is necessarily to build the cortical tone that is responsible for emotional regulation and inhibition, as well as to both build and strengthen positive cognitive schemas. Each negative slip-up will reinforce the old schemas. Thus, building positive schemas that become part of one’s “default” cognitive framework is not something that happens right away. You cannot “flip a switch” and create tools of resilience, the way that you can change mood. This inability accounts both for why so many patients find psychotherapy difficult and why the
changes wrought by it are more resilient than those wrought by ADMs. Psychotherapy is a learning process, wherein one strengthens neural pathways that aid in reassessment, reinterpretation, emotional inhibition, and reflection rather than automatic responses/behavior.

In this section, I have supported the following two claims: 1) My empirical hypothesis (about the higher-lower distinction) does not conflict with the evidence that oftentimes, the ultimate changes of PT and ADM, as monotherapies, are the same, and 2) the hypothesis has high explanatory power with regard to the different presentation of depression. In particular, it offers a neurobiological explanation for why those treated with PT alone tend to show increased resilience and lowered relapse rates when compared with those who are treated with ADM alone.

The next question to ask is: do these points about resilience really indicate something special about psychotherapy? From here on, we enter the second part of this paper, where I focus wholly on PT and drop the subject of ADM and interventions on lower brain areas. That is, from here on out, I compare PT not to ADM but rather, to purported neurobiological equivalents of PT. Now, we start the real philosophical work.

5. Interventional autonomy of psychotherapy – imperiled by counterfactual medical interventions?

The philosophically significant work comes when we ask: if we accept the empirical hypothesis that I presented in (3) and (4), what are the consequences for the ontological status of PT? In these final sections of my paper, I highlight what I feel is the most significant ramification of PT’s having effects that ADM does not: the interventional autonomy of PT. As I mentioned in the introduction to this chapter, “interventional autonomy” is comprised of 2 features: 1) the global nature/holism of the substrates of psychotherapeutic interventions, and 2) path-dependence.

Let’s first consider the importance of holism to PT’s autonomy: The total neural substrate underlying all of the maladaptive cognitive patterns, emotional dysregulation, and weakened cortical skills, etc., in someone’s case of clinical depression likely spans a vast and interconnected area of the brain. Therapists can neatly get into contact with a large proportion of this realizer by getting the person to think overtly about the ways that his beliefs relate, or by getting him into a depression-inducing situation via individualized triggers. For instance, we can trigger a huge schema of worthlessness just by mentioning a few scenarios that went into strengthening this system. The entirety of the schema can be stimulated by stimulating just a part of it. Then, this global stimulation allows a therapist (and the patient) explicit access to the totality of the schema, in order to challenge and change it. When I said earlier that psychotherapy is way of changing the brain via the “medium” of talking about one’s mental states or via the “indirect route” of exploiting psychological properties to gain access to the relevant parts of the brain, this method is what I was talking about.

Because the realizer is widely-distributed over perhaps the entirety of the brain, and because schematic content is not given in the language of synaptic connections, neuroscientists cannot look at a set of neural networks and discern which ones are coding for negative vs. positive schematic content. It looks like if a neurologist wants to change this vast network, she is going to have to work piecemeal, synapse by synapse, only after she solves the problem of finding normative information (about which areas are functional or dysfunctional) in the neurobiology itself. A neurobiologist cannot look at neurons and isolate only those synaptic connections that code for depressive schemas or cognitive distortions – the meaning cannot be
read off of the neurobiology. As rTMS intervention attests to, at this point in time, neuroscientists are able to crudely increase or decrease only general activity in the PFC.\textsuperscript{178} Therefore, because of the idiosyncratic nature of any particular case of depression – because of the specific content of the cognitive symptoms – it looks like focusing on psychologically-accessible content is the only way to target all that needs to be changed at one time. When a cognitive-behavioral therapist gets a patient to think overtly about the ways in which her beliefs and emotions relate, to challenge her cognitions, to have new experiences, to construct new schemas, etc., many of the relevant neural networks are stimulated simultaneously, and are therefore available for rewiring as a unit. Being able to target such a vast set of networks as a unit in one fell-swoop is a significant virtue of PT, and one that appears to be dependent on the autonomy of the mental properties themselves.

Unfortunately, this defense of PT does not give me what I need, because it does not establish PT as a necessary type of intervention for depression. Instead, it establishes only that its results are necessary for relief from depression. We can imagine a counterfactual situation in which science has progressed to the point where negative and positive content can be read off of neurons and their connections. In any event, we are to imagine that future neuroscientists can do a “brain-tweaking surgery” to get someone’s brain into the exact global state that it would be in following successful PT. The long process of PT would be replaced by a surgical intervention that would directly change the brain. As a result, the “psychological properties give access to the holism of realizers” feature of interventional autonomy would no longer carry any weight in saving PT from elimination. The psychological states that were once used to gain access to the totalizer are no longer necessary for such access. Because all of the results of interventions on mental disorders can be explained via neurobiological mechanisms, everything that PT could do would now be done by neurological interventions alone.

At first, I imagined that the introduction of the second feature of interventional autonomy (“path-dependence”) would be answer enough to this objection about the brain-tweaking surgery. Thus, I thought that we could distinguish the neural end-states of PT from the total end-states of PT. Additional changes to the brain accompany the process of PT, like the acquisition of skills that add to one’s resilience: learning how to take care of oneself and avoid stress, building up an arsenal of positive examples of self-efficacy, and so on. It seems that this experiential process of learning, of doing the hard work, is what distinguishes the long-term changes of PT from the short-term reversible changes of ADM. However, I soon realized that the brain-tweaking surgery objection can be used on this consideration as well. Because my view is materialist, all of these accompanying changes can be couched in neural terms as well, and hence these results too could be duplicated in the counterfactual situation.

It will not work for me to simply bite the bullet and make a normative judgment that it is just better for someone to go through the hard work of PT herself. In what way would this outcome be better than having the brain-tweaking surgery? I would need an argument that the hard work has a special kind of value. But such a stance looks ill-placed when it comes to treatment of mental anguish. Instead, we can make headway on this problem by investigating the success-criterion that we place on depression treatment. Since clinical depression is oftentimes a chronic disease, I claim that recovery should involve treatment not just of this bout of depression, but resilience against further recurrence and relapse as well. With this criterion in mind, what could be wrong with the brain-tweaking surgery? After all, a person awakens from this surgery with her faulty brain areas reset into healthy states, and along with these changes, she also has skills of resilience that leave her well-poised to function in the world. Saying that
she did not get this way of her own accord and that therefore this is not a desirable state to be in seems cruel and without justification.

I grant this point. Nonetheless, I still claim that the importance of path-dependence is maintained and that it is actually better for a person to have PT than it is for him to have the brain-tweaking surgery. After briefly exploring whether this judgment is an epistemic one – and deciding no - I explain the basis for my argument, which is entirely results-based. Because the brains of the people undergoing these interventions (PT vs. brain-tweaking surgery) are assumed to be precisely the same as one another at the end, the difference must be externally-based. I maintain that this difference has to do with the veridicality of the patient’s memories and beliefs regarding her interactions with other people and the world.

6. Path-dependence and the importance of veridicality

Implicit in my theory of the interventional autonomy and necessity of psychotherapy in the treatment of depression is a certain theory of what successful recovery from depression is. I make this success criterion as explicit as possible in this section. Doing so will help me to present an argument for why I disagree with the claim that it does not matter how someone gets into a healthy brain state, so long as he gets there. This argument is built on the importance of path-dependence and is a defense of the claim that the counterfactual brain-tweaking surgery would leave a depressed person worse-off than he would be than if he had had PT; and worse-off based on results, not on the basis of our judgments. In other words, I will not argue for a normative claim that there is an implicit value to doing the “hard work” of psychotherapy, as opposed to having the “quick fix” of surgery. Depression is a condition marked by considerable suffering, so refusing to countenance a treatment because of this kind of normative judgment is unnecessarily inhumane. This is not to say that offering a theory of recovery is itself not a normative endeavor; it is – it involves making evaluations on the standards of recovery.

Some might claim that depression is successfully treated when one has relief from major symptoms. Biological essentialists would claim that depression is successfully treated once its cause is removed. On the contrary, my theory involves an emphasis on resilience and protection against relapse. Depression is, in many cases, a chronic illness. Empirical studies show that many of the faulty neural pathways involved in depression remain present, but inactive, when someone is in remission, but respond with renewed activation under stress or other biological conditions. The process of kindling can occur, under which every recurrence of depression becomes more severe than the last and leaves one increasingly vulnerable to the next recurrence. Hence, treating an individual’s depression in the long run must involve not just treating this bout of depression, but the additional component of preventing recurrences. I have tried to show that ADM fails to provide this kind of protection, which supports the claim that some intervention that produces the same results as psychotherapy is ineliminable.

But in order to defend the claim that this needed intervention can only be psychotherapy, more is required of me. I need to argue for a conclusion on which counterfactual higher-level physical therapies would fail to accomplish what psychotherapy accomplishes. Actually, it might be more accurate to say that I must show that even if these therapies have the same results as psychotherapy, that we have reason(s) not to count them as being successful. I argue that we do have such reasons, and that these have to do with the fact that the results of these interventions are not achieved in the required way. This point is about path-dependency; I
disagree with the claim that it does not matter how someone gets into a healthy brain state, so long as he gets there. There is more to restored health than a healthy brain.

6.1 Deception

Recall that the question that I am dealing with is this: if a medical intervention can get a person into a state of mental health— a total brain state that offers tools of resilience against future relapse— then is there any reason not to count this person as being recovered? My answer is yes, that this person does not qualify as being recovered, because the causal history of his reaching this state is problematic. The person is worse-off than he would be had he gone through PT to get to this same end-state. But worse-off in what way? My first attempt to answer this question is epistemic: the person is a victim of deception. Ultimately, I do not think that this reason is strong enough to defend path-dependency. However, this point will lead me into a deeper discussion of the veridicality of memories and then to the strongest argument that I can offer for path-dependency.

In the psychotherapy case, we imagine someone with depression building up an arsenal of experiences that underlie a new schema of herself as an efficacious, worthwhile individual. This person forces herself to do things that she would normally avoid, she reframes her automatic cognitive distortions, and she recognizes her ability to achieve goals— at first small ones, and then larger ones. The actual experiences of these accomplishments— along with her memories of them, and the consolidation of the memories into positive schemas/beliefs— is what leads to her heightened feelings of self-worth. When she has a negative emotion, she can fall back on all of these memories, learning, and hard work. Following the counterfactual brain-tweaking surgery, she would have all of the same schematic content and memories, leaving her well-placed to not relapse. But the memories would be false; the experiences never happened. Furthermore, one might wonder about whether the person’s entire set of dispositional mental states would be consistent with one another. How are the memories of self-efficacy integrated with the person’s actual memories from the last year, for example?

Thus, one might argue: the causal history, or path, to get to the final end-state of recovery from depression matters insofar as it should not include deception or inconsistency. Implicit in this claim is a premise that deception is somehow contrary to mental health. But we must admit that a degree of deception, and inconsistency, is not problematic in and of itself; in fact, deception and inconsistency are a part of all of our mental lives. No one’s entire belief-set is consistent, and nearly all memories of events that actually occurred are usually false in one way or another. If mental wellness can include deception, then we are not entitled to automatically fault the person in the brain-surgery case for being deceived.

In fact, it could be that minor amounts of deception can actually aid mental wellness. I find evidence for this claim in the empirically-supported hypothesis of “Depressive Realism,” which maintains that depressed people perceive the world- and themselves— more accurately than do non-depressed people. Mentally healthy people, on the other hand, tend to make positively-biased appraisals of the world and of themselves. However, such skewed beliefs give a person the confidence necessary to pursue goals, while the depressed person is more likely to become derailed by the realistic assessments. A similar phenomenon of a moderate amount of deception working in one’s favor is the example of a person who has high self-esteem “for no reason.” This person has not necessarily accomplished his goals or had positive interactions with people, he is not as good at certain things as he believes that he is, etc. We thus might claim that this person is deceived about himself. However, does this deception make the person mentally
unwell? We might indeed claim that he is very mentally healthy (if the self-esteem does not become extreme and spill over into narcissism). This person is likely to keep pursuing goals and living happily, as a result of his high self-esteem. Thus, if we do not fault this kind of person for being deceived, then we cannot automatically fault the person in the brain-surgery case for being deceived in the same way.

Thus, it cannot be that deception (and/or inconsistency) in itself is the problem here. If it were, we would be expecting something that is not even achievable by non-depressed people to set the standard for recovery from depression. Deception on its own has given us no reason to think that the person who had the surgery is any less well-off with regard to getting out into the world and living a life without the handicaps that come with depression. Thus, we need a further reason besides deception to support the constraint of path-dependence on recovery. So either the kind and/or degree of deception/inconsistency are at fault, or something beyond the deception itself is the problem. I’ll argue in the latter manner. In the counterfactual scenario, the problem is not just that someone’s memories are not perfectly veridical; rather, the problem is that the memories are complete fabrications. Furthermore, they are fabrications that underlie the precise component of mental life that is at fault in the disorder. And finally, these non-veridical memories are likely to lead to negative behavioral and introspective outcomes.

6.2 Veridicality principle regarding memories

The principle that I support in this section is: the possession of completely fabricated memories discounts a person from qualifying as being recovered from depression, no matter how healthy of an occurrent state those memories put her in. This point is more than epistemic. Recall that we are trying to come up with success-criteria on recovery from depression, a disorder that is characterized by high levels of mental and physical suffering. As stated, to hold back any intervention that could put someone in an end-state that would prevent such suffering appears inhumane. Thus, rather than talk about our evaluations about what counts as recovery, we should confine ourselves to looking at results/effects. And unfortunately, fabricated memories lead to undesirable results. In this way, I argue that the actual having of particular experiences is necessary for recovery from depression.¹⁸¹

Let’s imagine the example of the man and his mother again. In the case where the man participates in PT, he works hard to reframe his cognitive interpretations of certain experiences with his mother. When he spends time with her, he uses tools to interact with her in less pathological ways, building up a better relationship through time. As he moves forward, he falls back on these previous memories of overcoming obstacles to help him continue to set goals and overcome yet more obstacles. In the brain-surgery case, the man has all of these memories, but all of them are fabrications. Certainly, many of them concern only himself and what has supposedly gone on introspectively. But many of these memories involve ways in which he has interacted with other people. And this relational feature of the memories is what ultimately undermines the brain-tweaking surgery and upholds the importance of path-dependence of higher-level interventions.

Notably, any memories that have to do with work done alongside or with another person will be denied by the other person involved. The man will ask his mother to recall positive interactions between the two of them. His mother will be bewildered and tell him that such occurrences never happened. Is the man equipped to deal with this kind of dissonance? I think
not. As he moves through the world, more and more of his memories will be questioned by others who do not share them. These are the very memories that underlie his resilience, and once they are called into question, his resilience will begin to crumble. If depression keeps a patient from living because of all the increased worry and rumination in his head, this brain surgery will not help much, since now the person will be in a state of confusion over why his experience does not cohere with that of others.

Thus, it is all well and good for a person to have an intrinsic state of mental health, but once one tries to navigate the world in such a state, it will quickly become apparent via interactions with others whether the person engaged in PT or had the brain surgery. My claim is that the brain-tweaking surgery changes are not sustainable, as they will lead to further states of confusion and dysfunctionality in the world. Hence, my claim that causal history matters is not an epistemic judgment, nor is it a normative judgment about what criteria one needs to meet in order to be considered mentally healthy. Rather, it is ultimately based on how well someone is able to navigate the world – including the social world – following higher-level interventions on depression. Although it looked like the person who had the brain-tweaking surgery was well-placed to live a healthy life because he was in possession of tools of resilience, we can see that unless he got there via actually undergoing the relevant experiences, he is still at a disadvantage compared to the person treated with psychotherapy.

6.3 Objection: perhaps the distress of non-veridical memories will just become an unpleasant side-effect of brain-tweaking surgery

Someone raised an interesting objection to my results-based defense of path-dependency when he said, “Well, so sometimes the person’s memories are not going to line up with the memories of other people. But so what? If this becomes a medical procedure that a lot of people have, then this ‘mismatching of memories’ will just be a common side-effect of the surgery. We can imagine that someone will be called-out by his mother for citing a false memory and contrary to his resilience becoming imperiled by this challenge, the person will simply say ‘Oh yes, that’s that brain surgery side effect again. No big deal. This isn’t a real memory. Wow, I’m feeling better from that awful depression that I had before.’”

My response to this objection has a lot in common with my response to the larger objection that is coming in the following section. It has to do with what constitutes depression (and other mental disorders), and what concomitantly constitutes recovery and/or health. When we medically intervene on someone’s condition, there is typically a certain “fit” to the treatment: “person-to-world” rather than “world-to-person.” Of course, there are objections to this fit; for instance, we might send someone with Tb to a warm and humid climate, or we might move an alcoholic to a substance-free recovery center. In those situations, we are changing the environments around these people in the hopes that their disorders will then wane. This discussion recalls the Mismatch Objection in my third chapter on mental disorders as possible dysfunctions. But if the disorder is *within* an individual, one might wonder about how well we are actually treating the person by focusing our interventions on a world-to-person fit. It was suggested to me that we could actually perform a global accommodation of the brain surgery by getting all of the people in the person’s life “in on” the non-veridicality, and then getting them to “play along” with any false memories. Besides the effort involved in such a task, this move seems wrong. But, why?

One is reminded of Nozick’s “experience machine.” Nozick imagines humans hooking themselves up to experience-machines, which can mimic any experience that the person wants.
He poses the question: would we hook ourselves up to such machines (to experience unending pleasure, for example), and if not, why not? Although Nozick feels that we would not hook ourselves up to these machines, he does not offer explicit reasons why not, except to say that we prefer to actually do things and to be in contact with reality. In a similar way, I believe that someone’s mental health status has to do with her situatedness in the world, her way of being in the world. Therefore, it is not the normative goal of therapy to adjust the world to the person. Rather, the normative goal of therapy is to adjust the person to the world. Perhaps I come up against my earlier claim that our assessment of the thought-experiment should be merely results-focused, and not based on normative judgments about what a person should do to get healthy; if suffering is alleviating, what else matters? But if we support the global accommodation scenario of the thought-experiment, then I do not see why we should not just go a step further and treat someone’s depression by putting her in a pleasurable experience-machine.

Now that I have discussed a more extreme version of the thought-experiment, let me explore an objection based on a weaker version of the thought-experiment. This discussion will bring me back to these considerations about the normativity of treatment/recovery.

7. Objection: I unfairly stack the deck by the way that I set up the thought-experiment
7.1 A mitigated thought-experiment

One might object to my whole project by claiming that my thought-experiment is either absurd or irrelevant. The fact is that we just do not have any idea right now about what kinds of brain interventions will or will not be effective in treating mental disorders. Similarly, one might object to my interpretation of the empirical data and claim that the evidence does not support a claim that pharmacotherapy systematically works via subcortical mechanisms while PT works via cortical mechanisms. If this objection is correct, then it means that it is just wrong to assume that cortical surgery will target higher-level symptoms like cognitive schemas.

While I grant the legitimacy of these kinds of objections that hit at the fundamental structure of my project, I want to spend time exploring a more specific objection that arises within my construction of the thought-experiment itself. This objection claims that I make a dialectically weak move in the way that I set the experiment up, insuring that the outcome will turn out in my favor. Here is a diagram of my “brain tweaking” thought-experiment, and how it supposedly supports the ineliminability of PT:

**Internalism:**
- Brain-tweaking surgery $\rightarrow$ Brain$_x$
- Non-veridical memories

**Path-dependence:**
- PT/“hard work”/experiences $\rightarrow$ Brain$_x$
- Veridical Memories

Brain$_x$ = the global internal state of the brain, which is the same (as denoted by the x) across the two interventions. In this case, the fact that both processes end in Brain$_x$ means that the question becomes one of whether or not recovery supervenes on the intrinsic global state of the brain. I argue that it does not, because of the detrimental impact of non-veridical memories on the stability of recovery.

But now let’s imagine the thought-experiment differently. This alternative rendering comes about in response to an objection that claims that I add elements to the brain surgery that are inessential to it; and if it is only these inessential features that are objectionable, then I should reconstruct the thought-experiment to exclude them. This exploration will lead us to a deeper discussion of the concept of “recovery.”
So, one might ask: are any particular memories actually essential to recovery from depression? I argued that continued experiences of oneself as capable of implementing new cognitive and behavioral strategies strengthen one’s ability to implement these strategies in the future. These experiences also lead to the heightened self-esteem that gives one resilience against depressive relapse. When one is unsure, one falls back on her memories of these times, and they strengthen her. Now, if I make these memories necessarily psychotherapy-involving, then that means that the brain-tweaking surgery will always lead to non-veridical memories. This move is unfair not only because it leads automatically to my conclusion (brain-tweaking surgery could never, by necessity, be equivalent in results to PT), but moreover, because it does not take into account the possibility of spontaneous remission. Some people who suffer from mental disorders get better on their own, without any treatment. Presumably, these people are in some healthy intrinsically global brain state, one that lacks memories of PT. Thus, PT-specific memories cannot be essential to recovery.

Now, it may be that these people are only temporarily experiencing an alleviation of symptoms, and that they have no resilience, because they did not build up the necessary arsenal of skills. On the other hand, I argue that if they do have resilience in the face of further challenges to their mental health, it is likely because they luckily found themselves having the right kinds of experiences to rewire their cognitive schemas (in this example, I do not have in mind people whose cases of depression are more biologically-based, in being localized to subcortical mechanisms). What the objection against me really is asking is: is there a way to give a person resilience (via brain-surgery) without thereby giving them false beliefs as well? What we need to do to accommodate this concern is to change the surgery so as to give the person less. No veridical memories are passed to a person during the surgery. Here is a diagram for the argument that one can gain resilience via a brain surgery that does not involve non-veridical memories:

**Internalism:**

\[
\text{Brain-tweaking surgery} \rightarrow \text{Brain}_z \rightarrow \text{Brain}_x \text{ vs. } \text{PT/“hard work”/experiences} \rightarrow \text{Brain}_z \rightarrow \text{Brain}_y
\]

Brain\(_x\), Brain\(_y\), and Brain\(_z\) all denote global internal states of the brain. Brain\(_z\) is an intermediate state that involves tools of resilience that have no specific content (they do not involve specific memories). Brain\(_z\) is shared by those who underwent either intervention, immediately following that intervention. Brain\(_x\) and Brain\(_y\) are the brain results that come about after those skills of resilience have been put into action in the actual world. They do involve memories, but because the people’s actual experiences (post-intervention) determine these memories, Brain\(_x\) and Brain\(_y\) are not identical to another. Still, both only contain veridical memories.

Recall that one of my primary aims in this paper is to demonstrate that PT is more resilient than ADM therapy; it leaves a person with stronger dispositions to avoid relapse. I push this claim in service of defending the ineliminability of autonomous mental routes of treatment for mental disorders. What this objection is really driving at is the question: what constitutes resilience? The question asks what is going on with Brain\(_z\). Is this brain really resilient? What do the person who went through PT and the one who went through the non-memory-based brain-tweaking surgery have in common? Furthermore, what don’t they have in common that could ground my claim that path-dependency is still important?

I claim that Brain\(_z\) is a “healthy” brain insofar as it involves improvements in Higher\(_1\) capabilities like emotional regulation. Still, I claim that changes in Higher\(_1\) are not enough on
their own to yield the resilience necessary to avoid relapse. The path from Brain_z to Brain_y occurs with the help of PT. One subjects herself to certain experiences, and actively works to rework her cognitions. Through time, these experiences slowly and step-wise build up a new self-image for her. The process is reinforcing because it is a process of learning; earlier changes set the stage for later changes, as one remembers: “I did this before, I can do it again.” On the other hand, what does the path from Brain_z to Brain_x look like? Brain_z gives patients a disposition to regulate their emotions, but it does not on its own prime them or motivate them or teach them how to go beyond this inhibitory process to a positive process of schema construction. Thus, one’s being in Brain_z following the non-memory-based version of the brain surgery does not ensure that one will ever reach Brain_x.

So if having specific memories of oneself having been efficacious are constitutive of resilience, then it looks like I did not unnecessarily stack the deck in my favor when I constructed the thought-experiment as I did. As it turns out, what we have dropped out of the brain surgery in response to this objection is not inessential to one’s recovery. The objectionable aspect of the thought-experiment – the creation of non-veridical memories – is crucial to the resilience that we are looking to preserve.

7.2 A note on why path-dependence is unique to mental disorders, and what this means for the concept of ‘resilience’

I am attempting to state that there is something unique about some treatments for mental disorders like depression, such that they exhibit autonomy from the physical level. My reasoning for this claim is that even though the mind supervenes on the brain, recovery from depression in some cases necessarily proceeds via psychological, experiential routes. These psychological routes cannot be duplicated by physical procedures, even if such procedures duplicate the relevant end-states of the psychological routes.

However, I can imagine someone applying my argument about path-dependency to different levels of interventions on physical disorders, thereby negating my claim that the autonomy rests on something unique to mental disorder interventions. So, I ask: do higher-level interventions on physical disorders express interventional autonomy as well? I answer “no,” based on crucial differences in the constitutional natures of mental and physical disorders.

Let’s look at how the argument for the path-dependency/interventional autonomy of physical disorders would supposedly go, via the example of heart disease. Let’s imagine someone with heart disease, and let us consider “lower” and “higher” realizers of this disease. “Lower” and “higher” refer to substrates that are more local and global, respectively. The lower realizer would be the clogged arteries in the heart, while the higher realizers would be more holistic metabolic problems (high cholesterol, high cortisol levels, etc.). We then can imagine two categories of intervention for someone with heart disease, one that targets the lower level and one that targets the higher level. The former would be the clogged arteries in the heart, while the higher realizers would be more holistic metabolic problems (high cholesterol, high cortisol levels, etc.). We then can imagine two categories of intervention for someone with heart disease, one that targets the lower level and one that targets the higher level. The former would be bypass surgery, where the clogged arteries are rerouted into clean arteries. The latter would be “lifestyle changes,” like changing one’s diet, starting an exercise regimen, and lowering stress levels.

The objection that I am thinking of with regard to my claim that path-dependency applies only to interventions on mental disorders would proceed as follows: one could say: “Look, interventions for physical disorders also involve interventional autonomy of the higher level. There is no way to change everything that lifestyle changes alter, except via the path of actually enacting those changes. That is, even if we were to produce (via sophisticated surgery) all of the
physical end-states of lifestyle changes (the person has brain states representing dispositions to eat healthy food, his body is in-shape, etc.), this result would not count as recovery from heart disease.

It must be emphasized that we are discussing higher-level interventions now (PT in the case of depression and lifestyle changes in the case of heart disease), and we are comparing the non-physical (mediated by something not describable in physical terms) vs. physical paths (counterfactual surgeries) to get to those changes. The defender of path-dependency would maintain that some of these higher level changes for the heart disease patient are path-dependent, and that the person who gets in shape slowly through time and toil will have other changes that are not captured merely by the physical “exercise-and-diet-end-state.” For example, the person who actively worked out will have created certain strategies for extending these changes into the future – eating well in spite of temptation, forcing oneself to work out, knowing how to pace oneself on a long run, etc. There is a resilience inherent in someone who has “done the work herself,” whereas someone who finds herself suddenly in good shape (defined merely by being in a particular global physical state) might quickly reverse this state because she does not have the tools to preserve it.

As we did in the psychotherapy case, the retort to this counterfactual would be to claim that all that we have produced in the counterfactual is the “exercise-and-diet-end-state” and not the “exercise-and-diet-total-end-products” state. We would need to extend the surgery to cover these accompanying end-states. The correct counterfactual would make the person’s brain pathways and relevant auxiliary physical states identical to those of the person who got in shape on her own. The person who says that there is nothing unique about the mental case would continue to push back and say that still, this person should just not be counted as being recovered. But why would we claim such a thing?

My view is that we would stop this process of pushing the counterfactual further at this point, and end the dialogue. This person would count as having resilience against further recurrences of her heart disease, because she would have the protective mechanisms in place. Sure, she might have false memories of having trained for a marathon, but though this deception might be epistemically bad and lead to arguments with people who know that she did not in fact train for a marathon, such deception does not negate her status as recovered from her heart disease. It does not really matter how this person got into this state; all that matters is that she is better and poised to stay better.

This is the crucial part of my argument, and it will hopefully tie together everything that I have been saying about the constitution of mental disorder and mental health/recovery. In section 6.1, when I discussed the problem of fabricating memories, I stated that the memories created by the brain-tweaking surgery “are fabrications that underlie the precise component of mental life that is dysfunctional in the disorder.” What I meant with this statement was that many individual cases of depression are partially constituted by cognitive distortions relating to the self – one’s efficacy, one’s worth. These symptoms are not just side-effects of the disorder but are significant components of the disorder itself (perhaps they are “central properties,” in the HPC language given in Chapter 4). Physical disorders, on the other hand, are not so constituted. Even though treatment of physical disorders might involve some “higher-level” interventions related to lifestyle changes, duplicating just the results of these lifestyle changes without having taken the experiential path to get there will suffice to count someone as being recovered from the physical disease. This is because physical disorders are defined with reference to their physical properties.
Let me be clear: discovering fabricated memories in one’s treatment would always be disruptive, whether it is that one finds out that he did not actually do the work to get his body into this amazing shape or whether he discovers that he did not build his self-esteem slowly through setting small goals and restructuring his cognitions and responding differently in relationships. The difference is that the knowledge is the former case is unlikely on its own to lead to relapse of heart disease, whereas the knowledge in the latter case will very likely have an impact on the progress of one’s recovery from depression. Learning that one did not do the actual work to get oneself into good shape might be disillusioning, but it should not cause us to say that someone is not healthy after all. If someone has a healthy heart and healthy dispositions, he just is physically recovered.

On the contrary, I claim that learning of the deception in the depression case will very likely have an impact on the progress of one’s recovery from depression, starting a harmful cycling back into low self-esteem. Since depression is an illness that crucially involves negative cognitions and low self-esteem/self-worth, it is significant that the person actually experiences the events that lead to the construction of improved self-image. The symptoms of depression are self-reinforcing. They tend to beget one another. If I feel bad about myself, this might cause me to ruminate, which may cause me to feel even worse about myself. This self-reinforcement (unique to mental disorders) is what underlies my claim that path-dependency matters for mental disorders, but plays no role in recovery from physical disorders.

In this section, I have made claims about the constitution of both mental and physical disorders and the constitution of recovery. In the following section, I conclude this chapter by explicitly discussing the meaning of ‘recovery.’

7.3 What recovery means

At the beginning of section 6, I stated that offering a theory of recovery is a normative endeavor because it involves making evaluations on the standards by which we designate someone as “recovered.” What I have done in this chapter is to make a series of claims about what it means to be recovered from depression. I have argued for a certain conception of depression that involves negative schematic content involving oneself, others, the world, and the future. If these schemas are constitutive of (at least some) cases of depression, then recovery will, by definition, have to involve the eradication of these schemas. Furthermore, because depression is often a chronic condition, recovery must involve resilience against relapses.

Merely duplicating a healthy brain in an unhealthy patient is not enough on its own to constitute recovery. So, recovery does not supervene on the brain. My argument for this claim is based on neither value judgments nor epistemic judgments about deception simpliciter. Rather, it is built from considerations of the ill effects of non-veridical memories on how a patient is disposed to be in the world. I argue that not only will the patient’s social relations be strained, but furthermore, discovery of the non-veridicality will erode the supposed resilience. Thus, even if we ignore moral and epistemic values entirely and focus only on results, it turns out that a person who has achieved PT results without undergoing PT is worse off than the person who actually did undergo PT to achieve PT results. When medical professionals treat someone with depression, they are aiming to help that person to function again in the world. They are aiming to fit that person to the world, not the world to the person. The importance of path-dependency preserves this methodological goal.
Thus, I have shown that when it comes to recovery from depression, psychotherapy exhibits interventional autonomy from the physical level. This interventional autonomy means that the causal path to recovery matters to the end results. This phenomenon is unique to mental disorders. Path-dependence is a results-based consideration that preserves an ineliminable place for psychotherapy even in a fully medicalized psychiatry.
Conclusion:

This conclusion will serve two purposes: first, I want to restate my main claims and demonstrate that they are consistent with one another. And second, I would like to discuss the kinds of issues that remain unsettled in my work, or rather, those lines of inquiry that are opened by the project.

(i) Concerns of consistency

The thesis that I have aimed to defend in this work is that the lack of reduction between mental disorders and neurobiological kinds does not entail that mental disorders are not explainable by the natural sciences. The failure of reduction is due to the heterogeneity and complexity of the neurological correlates and the causal networks that constitute mental disorders. The failure of reduction is not indicative of the need for some extra-scientific element in an account of mental disorder (like an appeal to dualism, or to social constructivism), nor is it evidence that mental disorders are not as real, as natural, or as valid as physical disorders are. All the failure of reduction tells us is that explanations of mental disorders are unlikely to appeal to robust and fundamental explanatory factors. Instead, explanations will be multi-level.

Any appearance of inconsistency in my claims can be attributed to confusing a scientific account of some subject matter with a reductive account of that subject matter. For instance, I argued that a scientific account of mental disorder must meet two constraints, namely that it can provide an account of the objective causal structures of mental disorders and that it references the brain as playing a systematic role in those disorders. While empirical research increasingly demonstrates the correlations between brain circuits and mental disorders, these neurobiological kinds are likely so widely-distributed and heterogeneously-realized that, except in rare cases, they are not likely to be 1:1 identifiable with mental disorders. This irreducibility does not show that mental disorders are not natural kinds, however. As I showed in Chapter 4, they are just not essentialist natural kinds. Rather, we should view types of mental disorders as homeostatic property cluster kinds, with their symptoms and etiological factors forming a causal network. None of the causal factors on its own is individually necessary for the presence of the disorder, and symptoms can be more or less “central,” depending on how many symptoms they are causally connected to, and how strong those causal linkages are. Some HPC kinds have a homeostatic mechanism that maintains the causal structure, though I showed how there need not be a homeostatic mechanism in addition to the causal network itself. HPC kinds are natural kinds because their causal structures are observer-independent. The connections between symptoms are nomologically-based and discoverable a posteriori, neither knowable a priori nor dependent on our conceptual schemes.

So that is one positive claim that I made about mental disorders: they are HPC kinds. I made another positive claim: mental disorders are dysfunctions, or more precisely, they are brain dysfunctions. This dysfunctionality is what makes a mental disorder a bad condition to be in; one breaks a natural norm. I claim that the particular kind of dysfunction that afflicts many of those with mental disorders is the failing of what I call a “genetically-open function.” This kind of dysfunction involves a loss of flexibility with regard to the environment, an inability to calibrate to local conditions. This flexibility is part of the evolutionary history of the mechanism, not just something that the mechanism happens to exhibit across a range of environments.

How do we combine all of these claims? One misguided option would be to state that faulty brain circuits are the homeostatic mechanisms that hold the cluster properties of a mental
disorder in place. I call this theory “misguided” because I think that it is implicitly reductive – if the circuit “underlies” the cluster, then perhaps the disorder is nothing but the circuit, and maybe this is a kind of borderline essentialism, albeit a complex one. On such a picture, mental disorders would be identical to complex neurobiological kinds. However, as I stated in Chapter 4, I believe that the physical (or neurobiological) symptoms/etiological factors of mental disorders should be “taken up” from being considered “underlying” and put into the cluster itself, on par with variables from other levels. Some brain state will play a crucial role in the cluster, but it may not be exactly the same state across every token of the disorder. This rendering shows us how the brain can play a crucial role in explanations (by being centrally involved in the observer-independent causal structure) without disorders thereby reducing to neurological kinds. It turns out that variables from other levels – psychological, social, environmental – play important causal roles as well.

In Chapter 4, I brought up paranatural kinds, which are kinds marked by the absence of a natural kind. How might we link up HPC kinds and dysfunctions with paranatural kinds? The following move is just a suggestion, but it could be that mental disorders are brain dysfunctions, and dysfunctions are deviations from normal functioning (understood from an evolutionary perspective). But there is not just one way to malfunction; in fact, the physical correlates of any one dysfunction are quite varied, and accompanied by symptoms from non-physical levels. In order words, the dysfunction is multiply realized at the neurobiological level. Furthermore, the dysfunction is the absence of the natural kind (the natural kind being the well-functioning mechanism) but it is not an absence simpliciter. The dysfunction has its own causal structure, which is that of an HPC kind. A person is diagnosed based on which cluster(s)’ causal structure she partakes in, and how much of it she partakes in/how deeply into the cluster she is.

(ii) Remaining questions and open lines of inquiry

In addition to my positive claims, I also presented some exploratory claims throughout this project. I want to rehearse those exploratory claims again here to demonstrate what I believe to be some fruitful research topics coming out of my work.

(a) If mental disorders are dysfunctions, then a theory of these dysfunctions requires a worked-out notion of ‘function.’ I argued that the understanding of ‘function’ given via Wakefield’s Harmful Dysfunction theory was inadequate. I claimed that his notion allowed for mental disorders to be nothing but mismatches between evolutionarily well-functioning mechanisms and novel environments. To deal with this objection, I introduced the idea of a ‘genetically-open function,’ and I stated that such a function has two features: it undergoes ontogenetic development and calibration to local conditions during its instantiation, and it is not genetically reducible. However, this theory needs to be filled out much more. What is a proper individuation scheme for genetically-open functions? How many are there? Does such a view lead inevitability to the claim that a human being and/or the psyche has just one open function? Furthermore, we may ask, in individual cases, what are the neural correlates of these functions? I argued that because of the complexity of behaviors that result from the activity of genetically-open functions, their realizers probably span from more primitive to more reflective/regulatory parts of the brain. I would be interested to see if this suggestion is borne out by empirical studies.
(b) I made a distinction between the Disorder (what makes a mental disorder *bad?*) and Structural (what is the causal structure of a mental disorder?) interpretations of the “what is the nature of mental disorder?” question. When I introduced this distinction, I noted that these are separate interpretations because structure is just formal and cannot itself be either good or bad. However, at the end of Chapter 4, I attempted to bring the two interpretations together. This is another place where I think that more work could be done. I suggested that the homeostatic clustering of causal factors of a given mental disorder may become self-reinforcing and take on a “life of its own.” The more causal structure that gets activated, the harder it is to de-activate any particular part of the network. If mental disorders have these kinds of structures, they would allow us to explain risk factors, vulnerability, the chronic nature of mental disorders, and why people feel so out of control when they are mentally ill, why they are not able to just pull themselves out of the disorder.

I think that this hypothesis about the link between these two interpretations requires elucidation. What is most intriguing to me is that I believe that finding this link would be a way to make dysfunctions *intrinsically* harmful, thereby negating the need for a *judgment* of “harmful,” or any other constructivist constraint on some condition’s being a mental disorder. The dysfunctions would be intrinsically harmful because they would involve a loss of personal control.

(c) The next two points have to do with the internalism-externalism dimension, and the question of whether or not mental disorders supervene on facts internal to the individual or not. Recall that at the end of Chapter 2, I claimed that the most important constraint for a Medical Model was that a theory of mental disorders has some objectivist anchor. As a result, I left the other two dimensions (Internalist-Externalist, Biological-Psychological-Social) somewhat to the side. On the one hand, having an internalist theory would cohere with the idea that there is really something wrong with the individual suffering from the mental disorder, rather than there being something wrong with the environment, or even with there being a mismatch with the environment. Genetically-open functions may allow us to acknowledge the importance of a reference-environment in the characterization of a function, while nonetheless still allowing for the dysfunction to be internal.

Because the reference-environment for a genetically-open function is a set of possible environments that one may encounter, a well-functioning genetically-open mechanism is disposed to flexibly respond to a range of environments. The suggestive line of inquiry here would be that the dysfunctions that are mental disorders are marked by the loss of an *internal disposition* to respond flexibly to whatever environment happens to be encountered. This view would have the counterintuitive implication that a person could appear to be well-functioning in his current environment, yet still be mentally disordered on account of his not having the disposition to respond to the changing world in flexible ways. Mental disorder, on such an account, would be marked by a certain kind of dispositional *rigidity*.

(d) On the other hand, we might argue that my work has demonstrated that mental disorders should be construed in an externalist way – that is, as *not* supervening merely on facts about the individual. In my 5th chapter, I argued for the ineliminability of mental forms of treatment, by appeal to path-dependency. An implication of this view is that recovery is not an internalist concept. Recovery from a mental disorder supervenes on facts about
causal history and facts about one’s relationship to other people (like, for instance, whether one shares veridical memories with other people). The question that such an observation raises for me is: is it possible for a theory of recovery to be externalist while the concomitant theory of disorder is internalist? If I claim that recovery is about getting a person to fit to the world in a certain way, that recovery is a certain way of being in the world, then mustn’t it also be the case that being disordered is a failure to fit, that it is also a certain way of being (or failing to be) in the world? And therefore, isn’t the world, or environment, in some way necessary to its characterization? So, these remarks in (d) and (e), about internalism and externalism, require further work.

(e) Finally, there is the issue of the Biological-Psychological-Social dimension, which I characterized in my second chapter as “explanatory.” I have claimed that mental disorders are dysfunctions, brain dysfunctions. But are these functions really best characterized as “brain” dysfunctions? After all, mental disorders afflict the whole person, not just his brain. In chapter 3, I had mentioned that I was using the term ‘psychological functions’ merely to refer to the functions that supposedly go wrong in mental disorders. But now we can start to see how the term ‘psychological functions’ could lead us to ontological commitments about a special kind of function, or rather, a special kind of function-bearer. In arguing that we should not conceptualize mental disorders as brain disorders, Natalia Banner writes:

“What it is that identifies a person as suffering from a mental disorder in the first place? In some loose sense, in psychiatric disorders the person’s self is affected…Thus the primary locus of disorder is among mental relationships, within a person’s own psychological life, and in interpersonal relationships. In other words, the correct level of analysis for picking out that something is wrong, is the person, the self. The manifestation of the illness occurs at the level of observable behaviour and felt experience. Indeed, in no other medical discipline is the subjective experience of the patient quite so crucial to diagnosis and the aims of treatment….Whatever the aetiology of the condition or its causal pathways in the brain – whether it is genetic, the product of gene-environment interactions, psychological, social or spiritual causes – the attribution of ‘disorder’ applies to the thoughts, feelings and behaviours of the person. Mental disorders are not mental disorders because of an assumed mental cause, but rather because they identify problems in the person’s mental life and relationships.”

By this reasoning, one is tempted to claim that the function-bearer of the malfunctioning mechanisms involved in mental disorders is the person, rather than brain components. Towards the end of my discussion on genetically-open functions, I made mention of potential neurobiological realizers that stretch between lower task-specific functions and higher-level regulation of those functions, regulation that oftentimes involves reflection and deliberation. Someone might think that such reflective processes are not actually processes of the brain, but rather, processes of the self, or person. As a result, the claim that the biological level is the proper level for explaining the badness of mental disorders would be wrong. Rational norms, rather than naturalistic ones, would be broken. Indeed, Banner advocates for a version of psychological realism. She claims that with a mental disorder, “the ‘organ’ affected is the person, within his environment. It would be a fallacy, she says, to suggest that the brain is the locus of disorder.”

I disagree with this claim. I disagree because I believe that once we focus on brain dysfunctions from an evolutionary biological perspective, we can see that asking what parts of the brain are for will end up automatically making reference to a person-in-an-environment. Again, I think that a failure to see this point derives from a conflation
between brain anatomy and brain function. A neural function can be to the benefit of an entire organism without that organism’s being the bearer of that function. Thus, I do not think that we need to utilize an explanatory level beyond the biological, as the correct (evolutionary) notion of biological functions can account for both the role of the environment and the impact on the whole organism. Thus again, the question that we return to is whether such a notion of brain dysfunctions requires externalist commitments.

(iii) Closing remarks

The five points just offered are questions that are opened up by, and that remain unanswered within, this dissertation. If I have done my job properly, what is not open for discussion is the claim that mental disorders are properly naturalistic and fit for study by the natural sciences, even though they remain unreduced to lower-level kinds. The same is true in spite of some of their mental interventions being ineliminable regardless of how the science progresses from here on out. These conclusions will hopefully do something to bridge the gap between those who feel that mental disorders refer to nothing at all and those who feel that mental disorders are nothing more than simple brain diseases.

Moreover, I hope that the discussion of homeostatic property cluster kinds and reinforcement between properties will help bridge the divide between the competing perspectives on the sufferer of mental disorder- as someone who is a passive victim vs. as someone who is weak-willed. The fact is, certain etiological factors and symptoms may be “reachable” via interventions that one performs on oneself; however, if these factors are deeply embedded in a homeostatic network, individual efforts may prove to be futile. These conclusions should allow us to continue on the path of not blaming people for their mental disorders, and yet remain optimistic that the causal structures can be targeted in many ways at many levels, restoring at least some personal power back to the individual.
Endnotes


3 Banzato et al, p. 49.


8 Insel (2010), 44.

9 Ibid.

10 The use of ‘loop’ and ‘pathway’ might strike you as just as problematic as ‘brain circuit.’ I will shortly discuss what we can say about the ontology of brain circuits, pathways, loops, etc.

11 Thomas Insel (January 2013), TED talk: http://www.ted.com/talks/thomas_insel_toward_a_new_understanding_of_mental_illness

12 Brainfacts.org

13 Janice Wood, “How your brain circuits become miswired,” From Cell, as seen at psychcentral.com


15 Middleton et al, 47.


18 Insel (2010), 44. (emphasis mine)


20 Note: the sub-circuits point would be more compelling if all of the various types of OCD subtypes were associated with different regions of the OF cortex, because the prior consideration actually involves more than just one of the 5-7 general categories of FSC circuits.

21 Insel (2010), 47.


23 Technically speaking, this hate circuit is an FSC circuit that extends from the superior frontal gyrus of the cortex to the subcortical insula and putamen.


25 Tao et al. (2013).

26 Murphy, Dominic, Psychiatry in the scientific image, MIT Press: 2006.

27 “Localization” in this context just means that one finds where a particular function is realized in the brain; “localization” should not be confused with a “local” or narrow realization of a function in one spatially continuous part of the brain – scientists can localize a function to the brain even if the function is globally distributed across the brain.


29 Ibid.

30 Ibid., 127.
mental disorders, but that

Psychiatry


Disorders: a conceptual taxonomy,” the categories might be constructed even though t

comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus,

and hence that the

essentialist view lends itself to the “category” dimension, since the necessary

pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view is a strong natural

questions as it pertains to individual categories of disorders, but it is worth saying something about them here.  An

and Categories

dimensions that could possibly be parsed out of this division: Objectivism

Social

Questions -

whether or not a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Dominic Murphy’s model of mental disorder is revisionist in this way in that it widens the extension of ‘mental disorders’ to include many conditions that we would normally think of as being only physical disorders (blindness, for example).

This simple binary may hide the complexity of the dimension. For instance, Zachar and Kendler come up with 3 dimensions that could possibly be parsed out of this division: Objectivism-Evaluativism, Essentialism-Nominalism, and Categories-Continua. The latter two of these three dimensions relate more to the Structural Interpretation of the question as it pertains to individual categories of disorders, but it is worth saying something about them here. An

essentialist view is a strong natural-kinds view that states that a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Dominic Murphy’s model of mental disorder is revisionist in this way in that it widens the extension of ‘mental disorders’ to include many conditions that we would normally think of as being only physical disorders (blindness, for example).

This simple binary may hide the complexity of the dimension. For instance, Zachar and Kendler come up with 3 dimensions that could possibly be parsed out of this division: Objectivism-Evaluativism, Essentialism-Nominalism, and Categories-Continua. The latter two of these three dimensions relate more to the Structural Interpretation of the question as it pertains to individual categories of disorders, but it is worth saying something about them here. An

essentialist view is a strong natural-kinds view that states that a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Dominic Murphy’s model of mental disorder is revisionist in this way in that it widens the extension of ‘mental disorders’ to include many conditions that we would normally think of as being only physical disorders (blindness, for example).

This simple binary may hide the complexity of the dimension. For instance, Zachar and Kendler come up with 3 dimensions that could possibly be parsed out of this division: Objectivism-Evaluativism, Essentialism-Nominalism, and Categories-Continua. The latter two of these three dimensions relate more to the Structural Interpretation of the question as it pertains to individual categories of disorders, but it is worth saying something about them here. An

essentialist view is a strong natural-kinds view that states that a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Dominic Murphy’s model of mental disorder is revisionist in this way in that it widens the extension of ‘mental disorders’ to include many conditions that we would normally think of as being only physical disorders (blindness, for example).

This simple binary may hide the complexity of the dimension. For instance, Zachar and Kendler come up with 3 dimensions that could possibly be parsed out of this division: Objectivism-Evaluativism, Essentialism-Nominalism, and Categories-Continua. The latter two of these three dimensions relate more to the Structural Interpretation of the question as it pertains to individual categories of disorders, but it is worth saying something about them here. An

essentialist view is a strong natural-kinds view that states that a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Dominic Murphy’s model of mental disorder is revisionist in this way in that it widens the extension of ‘mental disorders’ to include many conditions that we would normally think of as being only physical disorders (blindness, for example).

This simple binary may hide the complexity of the dimension. For instance, Zachar and Kendler come up with 3 dimensions that could possibly be parsed out of this division: Objectivism-Evaluativism, Essentialism-Nominalism, and Categories-Continua. The latter two of these three dimensions relate more to the Structural Interpretation of the question as it pertains to individual categories of disorders, but it is worth saying something about them here. An

essentialist view is a strong natural-kinds view that states that a given class of mental disorders contains members that all have a common essence marked off by necessary and sufficient conditions. A nominalist view is more social-constructivist, in that it claims that nature does not provide divisions between mental disorders, but that we pull certain conditions together on the basis of pragmatic goals and judgments of similarity. Obviously, an

essentialist view lends itself to the “category” dimension, since the necessary-and-sufficient-conditions provide a sharp cut-off point between those who are disordered and those who are not. The “continua” dimension, on the other hand, might be essentialist or nominalist. It could be that disorders fall onto a spectrum of more to less severe cases, and hence that the boundaries between both 1) different disorders and 2) disorder and health, are fuzzy. When it comes to diagnosis, we may be forced to make somewhat arbitrary decisions about where to draw these lines. Thus, the categories might be constructed even though the underlying continua are not. Kendler and Zachar, “Psychiatric Disorders: a conceptual taxonomy,” American Journal of Psychiatry, 164:4, April 2007, 557-565.


Graham, 138.

Ibid.


Horwitz and Wakefield (2007), 15.

Ibid., 53.


Wouters (2005).


“The etiological,” “historical,” “selected-effects,” and “evolutionary” theories of function are meant to be synonymous with one another; I will use the term “evolutionary function” as shorthand for this collection of terms from this point onward.

Richters and Hinshaw (1999).

The example is influenced by Darwin’s theories on the peppered moth during the Industrial Revolution in Britain.

Richters and Hinshaw, 441.

Richters and Hinshaw, 442.

Wouters (2005).

The example is influenced by Darwin’s theories on the peppered moth during the Industrial Revolution in Britain.

Richters and Hinshaw.

Richters and Hinshaw, 442.


This may not be the best example to exemplify current role theories of functions, since one might imagine that feathers were actually selected-for flight by the process of natural selection (in a later EEA than the EEA that “selected” feathers for warmth). In that sense, this function is not observer-independent.

Sober, 87.


Godfrey-Smith, 347.

It is true that the functions (understood in a “current role” manner) of other organs have a kind of “plasticity,” since they can be usurped for various other purposes (usually by conscious intentions – I can make one of the functions of my hand to throw a baseball). However, the plasticity of the brain is intrinsic to the kind of thing that it is. All of the functions that we consciously impose on other organs comes from this original plasticity in the brain.


Richters and Hinshaw, 440.

Recall that I am using the term ‘psychological function’ to refer to those functions that supposedly malfunction in mental disorders; the term itself does not make ontological commitments that these functions are necessarily unique in kind when compared with the rest of the biological functions.


Richters and Hinshaw, 439.

Richters and Hinshaw, 440.

Ibid.

Ibid.
Descriptions of Capgras Syndrome and Cotard’s Delusion can be found in any abnormal psychology text.

Richters and Hinshaw, 441.

Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


Ibid.


This hypothesis was strongly informed by Klaus Grawe’s *Neuropsychotherapy: how the neurosciences inform effective psychotherapy*, London: Lawrence Erlbaum Associates, 2007.

In Chapter 1, I make the argument that scientific explanations of mental disorders need not involve reductions. This table is an example of that point.


ADM’ here refers to current antidepressant medications that operate via changing neurotransmitter systems.
The actual story is more complicated than this— for example, because the right cortex deals more with avoidance and negative emotion, its activity is increased with those in depression. The left cortex deals more with approach and positive emotion, so its activity is decreased in those with depression. But the cortex also actively inhibits lower brain levels, so overall, its lowered activity is associated with psychological distress. See Frewen et al., “Neuroimaging studies of psychological interventions for mood and anxiety disorders: Empirical and methodological review,” Clinical Psychology Review 28 (2008) 228–246.


It is worth noting that there is a debate about the relationship between schemas themselves and schematic content. Some psychologists view these schemas as separate entities from the dysfunctional thoughts and emotions to which they give rise. See Dozois et al., “Changes in Self-Schema Structure in Cognitive Therapy for Major Depressive Disorder: A Randomized Clinical Trial,” Journal of Consulting and Clinical Psychology 2009, Vol. 77, No. 6, 1078–1088.

We all operate with hundreds of schemas, only fractions of which may be active at any time. My hypothesis with regard to depression is that schematic content becomes more and more negative, and more individual schemas link together to create a much larger negative schema, which dominates the majority of a person’s contact with the world.

It has been suggested that the success of rTMS — repetitive transcranial magnetic stimulation — undermines my argument. rTMS acts to increase activity in the cortex, often in very specific regions. However, these regions are not yet specific enough to get at very particular schematic contents.

Interestingly, I believe that this behavioral step shows us why mere insight alone does not move therapy along; insight just allows a person to entertain certain cognitions rather than providing experiential evidence to actually bolster those cognitions and weave them into schemas. Schemas must be well-ingrained in the brain to play a role.

I use the word ‘dampen’ rather than saying that the negative schemas have been destroyed or have disappeared, because it looks like all schemas might endure indefinitely, if only in dispositional form. What makes particular schemas active/more likely to be used is how often they are currently being used. We can think of the metaphor of trails through a forest — some get grown over, though they still exist. The ones that are most regularly used are cleared out and therefore the most likely to be used again. Source: Grawe, 2007.


Clark (2010).


Clark (2010)


This claim applies to depression as a type of mental disorder, not to token cases of depression.

I have in mind here just brute neurotransmitters deficiencies, or perhaps “biological clock” malfunctioning, which is often the case with SAD (seasonal affective disorder).

DeRubeis et al. (1999).

As I have stated in footnote 150, I am unable to provide a philosophical discussion of moods in this paper. However, the theory from which I am operating claims that moods are hedonic frameworks for structuring other mental states. They are like pairs of “tinted spectacles,” coloring every other mental state, and they also make congruent cognitions and schemas more likely to arise and to linger in consciousness.

Goldapple.

See footnote 158.

This is not entirely right, as studies have shown that no network ever really becomes “extinct.” They just become so weak as to hardly ever be used, like a railroad track that has had grass grow up all over it.


It has been said that during rTMS, “Stimulation is highly focused and interests specific regions of the cerebral cortex.” However, the focusing has to do with which areas of the cortex are known to perform certain functions rather than with where certain person-specific schemas are realized. (source: http://www.depression-guide.com/rtms.htm)


These experiences need not be the direct results of psychotherapy that takes place in a structured setting. Sometimes, people perform cognitive-behavioral therapy on themselves, and other times, people just happen to encounter the right kinds of situations to rewire their networks.

It was suggested to me by John Campbell that if I wanted to push the epistemic line as to why the false memories are problematic, I might adopt the view of knowledge that Timothy Williamson espouses in Knowledge and its Limits, Oxford University Press 2000. Williamson adopts a non-reductionist view of knowledge, where knowledge is its own unitary mental state and does not involve internal (true beliefs) plus external factors. Williamson argues that knowledge is more causally efficacious than true belief is when he states that “reference to states of knowing is essential to the power of causal explanation” (63). In this paper, I do not align myself with such an externalist view of knowledge, since I point to the importance of the phenomenological distress involved in being confronted with situations where one’s memories continually fail to align with others’ memories; this is an internalist consideration. However, were we to consider cases of socially-isolated individuals who might never learn of their deception, and we were then to ask what reasons might speak in favor of my view, I suspect that I might defer to a theory like Williamson’s. I might then claim that knowledge (in the form of true memories) is resilient in a way that true beliefs are not, because it makes the proper epistemic contact with the past. The true memories’ causal relation to the past buffers these memories against disintegration and replacement.

This point was suggested by Joseph-Kassman Tod during my Wollheim Society presentation, 10/24/14, “The ineliminability of psychotherapy as a treatment for depression: path-dependency and resilience.”

This point was made by Andrew King during my Philosophy Forum talk, 10/30/14, “The ineliminability of psychotherapy as a treatment for depression: path-dependency and resilience.”

Hardcastle.

This suggestion came primarily from Adam Bradley and Jim Hutchinson, in relation to my 10/24/14 Wollheim Society presentation.


Ibid.