Filosofisia tutkimuksia Helsingin yliopistosta
Filosofiska Studier från Helsingfors universitet
Philosophical Studies from the University of Helsinki

Publishers:
Theoretical Philosophy
Philosophy (Swedish)
Social and Moral Philosophy

P.O. Box 24 (Unioninkatu 40A)
00014 University of Helsinki
Finland

Tuomas Vesterinen

Socializing Psychiatric Kinds

A Pluralistic Explanatory Account of the Nature and Classification of Psychopathology

Doctoral dissertation

to be presented for public examination with the permission of the Faculty of Arts of the University of Helsinki, in lecture room U3032, University Main Building, on the 5th of May 2023, at 12 o'clock.

ISBN 978-951-51-9204-2 (paperback)
ISBN 978-951-51-9205-9 (PDF)
ISSN 1458-8331 (series)
Unigrafia, Helsinki 2023

Abstract

This thesis investigates the nature of psychiatric disorders, and to what extent they can form a basis for classification, explanation, and treatment interventions. These questions are important in the light of the "crises of validity" in psychiatry, according to which current diagnostic categories do not pick out real disorders. I address the questions by defending an account of psychiatric disorders that can better accommodate social aspects and non-epistemic values than the symptom-based model of the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the brain disease model of the biomedical approach, including the Research Domain Criteria (RDoC). I concentrate on three ways that psychiatric disorders can differ from prototypical natural kinds, such as biological species and chemical elements. First, the concept of psychiatric disorder or mental disorder is partly value-laden and cannot be defined based only on scientific facts. Second, the objects of psychiatric classifications are interactive human kinds because people with disorders respond to their classifications through looping effects. Third, sociocultural factors can shape disorders in complex ways, which is indicated by their cross-cultural variation. I argue that these challenges can be overcome with a pluralistic explanatory account that grants an explicit role to value-sensitive considerations and social scientific research. Based on this, I argue that particular disorders can in principle support inductive inferences.

My key argument is that although whether a condition is considered pathological is partly a value-laden question, this does not rule out a realist account of particular psychiatric disorders as homeostatic property cluster kinds (HPC). I assert that causal mechanisms responsible for psychiatric kinds can be understood non-metaphysically based on the contrastive-counterfactual and

interventionist theories of explanation. An advantage of this account is that it can address the concern that psychiatric explanations require one to relate heterogeneous causal factors on different levels, such as genetic, neurological, psychological, and social. My novel argument is that mechanistic explanations of scientific kinds have applicability domains over which they account for specific aspects of kinds and warrant inductive inferences. That is, identifying the applicability domain of an explanation spells out the conditions under which the explanation is expected to be reliable and when it can break down. This helps to understand how different discipline approaches, including social and cultural ones, can be complementary and make an explanation of a psychiatric kind more reliable. On the other hand, the account also shows how the complex nature of psychiatric kinds can license re-classifications for different epistemic and non-epistemic purposes. Finally, in the light of the social implications of psychiatric classifications, I suggest a value-sensitive or ameliorative approach to engineering the concept of psychiatric disorder. In conclusion, my account shows that research on the social and cultural factors that shape psychiatric disorders, and weighing the non-epistemic values that influence classificatory practices, can and should inform classificatory choices as well as policy and treatment interventions.

Table of Contents

Acknowledgements

This dissertation is not just the product of armchair philosophizing, it equally originates from discussions, collaborations, travels, interviews, and friendships. For these, I am indebted to numerous people, more than can be singled out here.

My sincere gratitude goes to the external examiners, Professor Dominic Murphy and Professor Rachel Cooper, who have made excellent recommendations for improvements. I would like to express my deepest gratitude to Professor Murphy for agreeing to serve as the opponent. His research in the philosophy of psychiatry has been a major inspiration for this thesis.

Words cannot express my gratitude to my supervisors Professors Gabriel Sandu and Petri Ylikoski. Little did Gabriel know that when he promised to supervise my master's thesis, that he would continue to supervise me until his retirement and beyond. I am grateful for his firm feedback, help whenever I have needed, and humorous yet lucid philosophical company. I am also extremely grateful to my other supervisor, Professor Petri Ylikoski, for suggesting this topic, and providing insightful feedback throughout the process. He has not only helped me to finish this thesis, he has also shown me how to combine social scientific and philosophical interests. I also wish to offer special thanks to the Custos, university lecturer Samuli Reijula, whose work on some of these issues has been a continuous inspiration. I am also grateful to him for arranging the practicalities concerning my thesis defence.

The possibility to travel and discuss with some of the leading people working on these issues has been crucial. I am especially grateful to Professors Raffaella Campaner and Stebastiano Moruzzi for inviting me for a research visit to Bologna and for their great generosity. Professor Campaner introduced me to philosophers and practitioners working on these issues in Italy. I would like to thank especially the philosopher, psychiatrist and psychologist Massimiliano Aragona for taking time to discuss cross-cultural issues related to psychopathology and its treatment.

I am deeply indebted to Professor Alexander Bird for hosting my stay at King's College in London and helping me on a crucial chapter of this dissertation. I would also like to thank Professor Ernie Lepore for hosting my stay at Rutgers,

and Professors Steven Stich and Jerome Wakefield for their warm-hearted hospitality during my stay. I am also grateful to Professor Jean Gayon for inviting me to visit IHPST Université Paris 1 Panthéon-Sorbonne.

I could not have undertaken the last stages of this journey without the RADAR group. I am deeply indebted to Pekka Mäkelä and Raul Hakli for hiring me as a member of their team, which has become my new intellectual home. I wish to thank Dane Gogishin, Olli Niinivaara, and Kamil Mamak, Tomi Kokkonen and Pii Telakivi. I am especially grateful to Tomi Kokkonen for our discussions on natural kinds and mechanistic explanation.

This research was funded by a four-year position at the Doctoral Programme in Philosophy, Arts and Society at the University Helsinki, a grant from the Finnish Cultural Foundation, and finally a position at Academic Project at RADAR. I would also like to thank Docent Emeritus Mark Shackleton for checking the language of my thesis and giving it a final polish.

This endeavour would not have been possible without our philosophy of psychiatry group: Pii Telakivi, Anna Ovaska, Ferdinand Garoff, Sanna Tirkkonen, and Laura Oulanne. Our collaboration gave rise to conferences, discussion, great parties, and friendships.

I am especially grateful to Pii for her friendship, travel companionship, and making everything just a bit more fun. By socializing our thinking we have been able to extend our minds. Moreover, despite our best efforts, Professor Sandu finally received both of our theses "long lost in the mail". I wish to offer my special thanks also to Anna for taking care of almost everything. In addition, I am grateful to Ferdinand for providing "hands-on" knowledge of psychology to bring my often very theoretical approach down to earth. I am also grateful to Sanna for promoting our research.

I am indebted to Antti Sneitz, who first encouraged me with this subject and with whom I taught my first philosophy course to over a hundred students and lived to teach others. Some of those have been at the Milano Università Vita-Salute San Raffaele, for which I am indebted to Carlo Martini, who has kindly invited me to give talks to his psychology and psychiatry students. I am also grateful to Tuukka Tanninen and Jaakko Hirvelä for all their comments and support.

One of the nicest things in academia is that you can meet like-minded people all around the world. My best time in Paris was spent with Diego Fernanders and Renata Arruda, whose philosophical enthusiasm and intellect is humbling. In New

York, I had a great time with Reinier Schuur and Santiago Echeverri. I am especially grateful to Rey for taking me along to the Sydney Winter School and to the cognitive philosophy course at New York University.

Many other people have also contributed: Valtteri Arstila, Joel Backström, Fausto Barbero, Sofia Blanco Sequeiros, Marion Godman, Anton Granvik, Jaakko Hirvelä, Ilmari Hirvonen, Professor Marja-Liisa Honkasalo, Säde Hormio, Maria Hämeen-Anttila, Tero Ijäs, Ragnhild Johrdahl, Professor Timo Kaartinen, Tuukka Kaidesoja, Inkeri Koskinen, Jaakko Kuorikoski, Markus Lammenranta, Kristjan Loorits, Professor Maria Lasonen-Aarnio, Jan-Ivar Lindén, Vili Lähteenmäki, Magdalena Malecka, Caterina Marchionni, Sanna Mattila, Luis Mireles Flores, Professor Uskali Mäki, Michiru Nagatsu, Professor Ilkka Niiniluoto, Hanna Peljo, Professor Filipe Pereira da Silva, Tuomas Pernu, Paavo Pylkkänen, Professor Panu Raatikainen, Mikko Salmela, Matti Sarkia, Päivi Seppälä, Ninni Suni, Professor Tuomas Tahko, Suvi Talja, Niklas Toivakainen, Professor Thomas Wallgren and Hermanni Yli-Tepsa.

I wish to express my special gratitude to Steve Abeni, who acted as my guide to the local villages around Grant-Popo, Benin. I express my thanks also to the Villa Karo Cultural Center. Thanks should also go to Suvi Jaakkola and Ninnu Koskenalho, with whom I organized together a Night of the Arts event on mental illness as a cultural phenomenon. Its popularity showed how relevant these issues are.

I would like to recognize my nearest and dearest. I am thankful to Uncle Ervo for offering his knowledge of medicine, but also his appreciation and encouragement of my philosophical interests. Likewise I wish to thank my cousin Maarit for our discussions over clinical work, and also for showing me that a dissertation is not the only thing in life. My friends Valter and Freddi have shown that satisfyingly heated debates can occur over just about anything, not just philosophy.

It was my parents who set me on this path with their encouragement and example. I'm extremely grateful to my father Ilmari. Because of his anthropological interests, I grew to appreciate that empathy grows by making the strange familiar and the familiar strange. In this vein, I hope this thesis can contribute to humanizing "madness". Ilmari dedicated his life to research and humanism, which were the primary topics of our discussions from early on. But it was not so much the content of our discussions but rather the feeling of how vitally important such questions are. This feeling has passed on also to my sister Terhi, with whom I can

thankfully continue those same discussions. I am also grateful to her for proof reading parts of the dissertation. Words cannot express my gratitude to my mother Päivi for her support and help with this project in every possible way.

Finally, doing philosophy this long has spilled into everything, something that has not gone unnoticed by my family, Salla, Leo, and Aatos. According to Leo, father has too much “hommia” (work to do). Most of what I have written and argued have required a final polish from Salla. Without her emotional and practical support, this demanding endeavour would not have been possible.

Helsinki, May 2023

1 Introduction: An Explanatory Account of Psychiatric Kinds

Psychiatric disorders[1] such as schizophrenia, depression, and attention deficit hyperactive disorder (ADHD) are difficult to classify, explain, and treat because there is no clear understanding of what kinds of things they are. Successful classifications in the natural sciences are usually considered to "carve nature at its joints" by arranging the targeted objects based on their naturally shared properties and relations. Therefore, the hope is that if psychiatric disorders resemble chemical elements or biological species, successful psychiatric classifications could be based on their natural structure. Such a natural classification could dispel doubts about medicalization, overdiagnoses, facilitate research, as well as enable efficient diagnoses and treatments. On the other hand, because psychiatric disorders are also socio-cultural phenomena, there is no obvious answer as to how they differ from deviances of social norms. Moreover, psychiatric classifications have ethical and social consequences. They determine which abnormalities are considered genuine disorders, and thereby warrant treatments and exonerate behaviour, but they also inflict stigma, influence self-identities, and alter general conceptions of the normal and the pathological. Hence, the challenges facing psychiatric research and classification are not only empirical, but also involve morally relevant conceptual and philosophical questions about the nature and realness of psychiatric disorders. The purpose of this thesis is to investigate what kinds of scientific objects psychiatric disorders are by employing recent advancements in the philosophical of science, and thereby provide

1 I use the term "psychiatric disorder" synonymously with "mental disorder". Their connotation is different from "mental illness" and "mental disease". "Mental illness" is usually considered to include a cultural interpretation, while "mental disease" has a connotation of resembling somatic diseases. The older terms "madness" and "insanity" are ambiguous because of their colloquial nature.

conceptual tools that can facilitate empirical research and classification as well as policy decisions and treatment interventions.

The conception of psychiatric disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD) has been especially influential and controversial. The conception is based on a pragmatic interpretation of the medical model, according to which psychiatric disorders can be provided symptom-based categorical definitions. However, the symptom-based approach has been criticized for facing a "crisis of validity", according to which the diagnostic categories are not empirically well-funded (Solomon 2022; see Kendler 2021). Arguably, this lack of causal support is exemplified, for instance, by their high rate of comorbidity (same individuals are diagnosed with multiple disorders), culture- and value-dependency, arbitrarily strict definitions, and weak interrater reliability (same cases are diagnosed differently) (see Murphy 2017; Werkhoven 2021). Moreover, since the introduction of manuals, both the number of diagnostic categories, and the people being diagnosed, have increased. In 1952, the DSM-I contained about 100 mental disorders, while the latest version DSM-5 has nearly 300 (APA 2013). Similarly, studies in the early 1980s suggested that one-third of the USA adult population had a diagnosable mental disorder, while in 2013 the figure had risen to approximately half of the population (Cockerham 2017). The growing number of some diagnoses, such as depression and eating disorders throughout the world is also striking (Watters 2011; Pike and Dunne 2015). The suspicion is that these increases are at least partly due to overdiagnoses and medicalization of social problems, instead of genuine increases in mental problems or better diagnostics (Brinkmann 2020). As an example, Horwitz and Wakefield argue in their 2007 book *The Loss of Sadness* that the DSM manual has obscured the distinction between sadness and depression. Despite this argument, the latest ICD-11 (WHO 2022) includes a novel

diagnostic category of prolonged grief. Furthermore, since the manuals are employed for bureaucratic and administrative purposes, they do not only influence research and treatments, but also serve as the bases for policy decisions that have wide-ranging societal consequences.

A commonly held hope in psychiatric research and practice is that an etiological (i.e. causal) approach endorsed by a stronger view of the medical model can overcome the problems of the description-based account of psychiatric disorders. According to this view, psychiatric disorders are discreet biomedical entities situated within the individual, which would make them similar to somatic diseases (Kincaid and Sullivan 2014). Hence, the stronger medical model seems to be committed to essentialism and reductionism, so that although some of the pathological causes can be social and psychological, the disorders themselves are identically realized in the brain. This brain disease view is explicitly endorsed by the Research Domain Criteria (RDoC) programme. The RDoC was launched by the National Institute of Mental Health (NIMH) with the intent that it could eventually replace the DSM's diagnostic categories as the primary bases for research on psychopathological causes (Insel et al. 2010; Kincaid and Sullivan 2014). However, critics argue that such biomedical approaches do not sufficiently concentrate on the experience, social role, and sociocultural context of mental disorders (Kincaid and Sullivan 2014, Luhrman et al. 2016). For example, the cross-cultural variation of psychiatric disorders seems to indicate the substantial role of social factors in disorders. In addition, the strong medical model's "broken brain" view of mental disorders may inflict stigma with its immutable view of psychiatric disorders (Haslam 2014).

In this thesis, I examine two philosophical questions raised by these challenges to classify and explain psychiatric disorders. First, I examine what kinds of scientific objects psychiatric disorders are. More specifically, I ask what kinds of things psychiatric disorders are in

comparison to other scientific or real kinds, such as chemical elements and biological species, on the one hand, and how they resemble and differ from each other and somatic diseases, on the other.[2] The central purpose is to investigate whether a general theory of psychiatric disorders as scientific or real kinds can be formulated. Second, I ask what implications my findings about the nature of psychiatric disorders have for classificatory as well as treatment and policy decisions. I investigate especially the scientific and social implications of the fact that psychiatric classifications do not only sort out kinds of psychopathology, but also kinds of human beings. This means that the second question is also normative. I ask what value-laden choices and responsibilities my account of psychiatric disorders bestows on classificatory and diagnostic practices.

I address these questions by exploring an alternative account to the medical model of psychiatric disorders. I formulate and argue for a value-sensitive and pluralistic explanatory account of psychiatric disorders as scientific or real kinds. To that end, I explore and analyse the philosophical debate over the possibility of conceptualizing and classifying the putative natural kind structure of psychopathology. The question concerning the natural kind nature of psychopathology has been central in the philosophy of psychiatry (see Kincaid and Sullivan 2014). As Serif Tekin (2016: 148) points out, the debate is motivated by the scientific legitimacy of psychiatric research and clinical practice. It is based on the conviction that if an object of research is a natural kind, it is not an arbitrary or value-laden grouping imposed on nature, but is instead an objective cluster empirically found there. This implies that if we can establish that psychiatric disorders can in principle be natural kinds, we can reason that the empirical validity problem is

[2] I use the terms 'real kind' and 'scientific kind' to cover all the kinds that ground relatively robust epistemic projects, such as explanations, predictions and interventions.

solvable, and thereby also address ethical and policy questions. That is, instead of being value-laden social constructs or arbitrary groupings, psychiatric disorders would be objective classificatory and research targets resembling somatic diseases to the extent that their discovery would justify policy decisions and treatment interventions.

However, there are challenges of formulating a natural kind account of psychiatric disorders, even if the most implausible constructivist worries and pragmatic problems related to the classificatory manuals are put aside. According to a growing number of philosophers and researchers, the central scientific challenge with psychiatric disorders is that they are difficult to explain and sort out because they are brought about by the interaction of multiple causal factors on different levels of description (e.g. Sullivan 2014; Kendler 2012, Murphy 2006).[3] This implies that psychopathology may not have a stable and distinctive natural kind structure resembling somatic diseases ready to be discovered (Maung 2016: 16). Alternatively, part of the problem could be that the current concept picks out very different kinds of natural kinds, whose understanding require varying explanations from biological, psychological, and social factors. On the other hand, even if it were possible to modify our current diagnostic categories to pick out stable and scientifically relevant natural kinds, Cooper (2005) and Beebee and Sabbarton-Leary (2010b) argue that whether we label them as natural kinds of psychopathology, rather than just biological or psychological kinds, is still at least partly a value-laden choice. Based on these challenges, the purpose of my thesis is to examine whether a pluralistic explanatory account of psychiatric kinds

[3] The notion of "level" is used in various ways in the philosophy of science (e.g. Eronen 2013, 2015; Levy 2013; Kokkonen 2021; Kuorikoski 2009). I employ it mostly to describe levels in mechanistic explanation without making ontological commitments (see Craver 2007).

can accommodate values and disciplinary relative-explanatory approaches, including social scientific explanations.

I am not simply asking whether we can carve psychopathology at its joints. In practice, I am also asking how we can classify psychopathology responsibly in view of its sociocultural nature. Psychiatric classification reflects, but also reciprocally affects, the social context where disorders emerge, develop, and are experienced. This is important because psychiatry is an applied science guided by both moral and practical goals and by epistemic ones (cf. Zachar 2014b). These factors are not only relevant for clinical practice, but also for classification and research. The reason is that as research objects, humans are active social agents, and therefore care and react to how they are studied and classified. As the philosopher Ian Hacking (1995a: 21) points out, psychiatric disorders are subject to looping effects: "People classified in a certain way tend to conform to or grow into the ways that they are described: but they also evolve in their own ways, so that the classifications and descriptions have to be constantly revised." In fact, sociocultural and institutional factors influence all disorders and even make possible some conditions to the extent that they are not timeless and universal (Young 1995, Hacking 1998). Consequently, psychiatric disorders can be considered kinds of humans or human kinds in addition to natural or scientific kinds. The implication is that cultural and social factors related to psychiatric disorders bestow more responsibility on psychiatric research and classificatory practice than what is involved in, for instance, classifications of biological species, chemical elements, or somatic disease. Based on this, I investigate how non-epistemic values could be incorporated more explicitly– value-sensitively – into the classification and specification of psychiatric disorders.

1.1 The Nature and Realness of Psychiatric Disorders

The first question of this thesis concerns the nature of psychiatric disorders as kinds of scientific objects. The problems related to current classifications and explanations of psychiatric disorders raise two interrelated philosophical questions about their nature.

First, can the concept of psychiatric disorder or mental disorder be provided a definition based on scientific facts or is it irreversibly value-laden? That is, what, if anything, unites all the conditions that fall under the heading of psychiatric disorder? Second, can particular psychiatric disorders be in principle natural or real kinds that resemble chemical elements, biological species, or somatic diseases? Although these questions are interrelated, there is a need to clarify to what extent a condition's status as a putative dysfunction, i.e. an internal psychological or biological system that is not working as it is supposed to do, is related to it being a particular scientific kind, i.e. a non-arbitrary grouping or distinction in the world. This further question is also relevant because specific psychiatric conditions may differ substantially from each other, even if their status as psychiatric disorders were justified in the light of shared nature or objective values. So far, the nature of the concept of psychiatric disorder and nature of individual psychiatric conditions have largely been examined separately. I address this gap by examining the underling conceptual assumptions in the discussion over how to define psychiatric disorder, as well as the conceptual and causal relations between the concept and the kinds it purportedly applies to.

The philosophical debate over the nature of psychiatric disorders has largely relied on analysing the general concept of psychiatric disorder or mental disorder and the role conceptual analysis plays in determining what the disorders are. The debate is motivated by providing an objective definition of the concept to overcome constructivist worries raised by the validity problem of current and past

diagnostic categories. Naturalists, such as Christopher Boorse (1976), argue that the concept of psychiatric disorders can be given a definition based on scientific facts, whereas normativists, such as Rachel Cooper (2005), claim that such definitions are irreversibly value-laden. On the other hand, Jerome Wakefield (1992a) has presented a hybrid account that incorporates both value choices concering harmfulness, and scientific facts about biological dysfunctions.

However, the discussion has increasingly been criticized for relying too greatly on conceptual analysis, which is arguably based on conflicted and context-dependent lay and professional intuitions (e.g. Murphy 2006, Lemoine 2013, Schwartz 2014*).* In other words, according to this interpretation, it is taken for granted in the discussion that by analysing how the term "psychiatric disorder" is generally used, the nature of disorders it purportedly refers to can be revealed, and thereby we can come to terms about how to define the general concept. Arguably, the problem is that the participants thereby also implicitly allot conceptual analysis the task of constraining and interpreting empirical discoveries by determining what the conditions for genuine psychiatric disorder are. This criticism has been especially propounded by Dominic Murphy (2006), who does not only point out the failings of conceptual analysis, but also argues that current diagnostic categories are saturated by unspecified folk-psychological and theoretical preconceptions. At the extreme, the present diagnostic categories have been called "epistemic prisons" that severely hinder empirical research (Hyman 2010).

Based on this criticism, I make a distinction between top-down and bottom-up methods to define the concept of psychiatric disorder and determine the nature of psychiatric disorders. Top-down method relies on conceptual analysis and aims to understand psychiatric disorders in their totality in the hope that once such a definition is reached, it can be employed to make a decision whether a particular kinds is a disorder.

Following others, I argue that Boorse and Wakefield's attempts to naturalize the concept can be understood as top-down approaches because they rely, at least partly, on conceptual analysis to determine the correct definition. Likewise, I explore how normativist approaches rely on discovering shared intuitions concerning how the concept is employed and thereby make judgements about what it means to be a disorder. Subsequently, I examine alternative bottom-up approaches, which aim to be methodologically naturalistic by arguing that empirical research alone can determine the true nature of psychiatric disorders. I point out that they are open-ended as to which disorders should be understood as disorders, and instead concentrate on more particular disorders, or even smaller units of analysis. As examples of bottom-up approaches, I analyse Murphy's (2006) disease model and the RDoC program. I also argue that contextual models, such as Broadbent's (2020) recent non-objective naturalism as well as Canguilheim (1991) and Hacking's (1995) more socially oriented accounts, can be understood as bottom-up approaches. While Murphy argues that psychiatry should ideally merge with cognitive neuroscience, the contextual accounts concentrate on the reasons why we employ the concept of psychiatric disorder.

The purpose of this discussion is to explore the possibility of formulating a bottom-up value-sensitive account of psychiatric concepts, which would grant an explicit role to value-laden decisions over whether particular conditions are psychiatric disorders, and how their boundaries are determined (see Aftab 2019, Bueter 2019, Gagné-Julien 2021). I want to investigate especially whether the values related to the concept of disorder, and psychiatric classification in general, could be studied and made more explicit, and their conceptual and causal interactions with empirical research and the conditions themselves should be assessed, and if deemed necessary, controlled. Such an account would acknowledge that because there can be genuinely different

ways to classify psychopathology, classificatory choices need to rely on ethical and pragmatic considerations.

Given these problems of providing a naturalistic account of the superordinate concept of psychiatric disorder, some philosophers have nonetheless argued that it does not imply that particular psychiatric disorders are mere social constructs imposed on nature. Rather, they have examined the possibility that only the general concept may not pick out a natural kind, but particular psychiatric disorders may nonetheless be natural kinds. This means that although the superordinate concept is currently conflicted and normatively loaded, individual conditions so labelled can nonetheless in principle be scientifically relevant value-free natural kinds (Cooper 2005, Murphy 2006, Beebee and Sabbarton-Leary 2010b). This would mean that at least some of the kinds labelled as disorders can be natural phenomena, either non-pathological neurocognitive or psychological conditions, or their different kinds of malfunctions. That is, labelling them as illnesses or disorders would be an additional normative projection, not something found in their nature. In other words, although they would not be *natural kinds of psychopathology*, they could still be genuine *biological or psychological joints of nature*. This view can be contrasted with an exclusive social constructivist account of psychiatric disorders, endorsed by some of the anti-psychiatric accounts, and thereby represents a weak realist account of psychiatric disorders or psychiatric realism.

A realist approach to particular psychiatric kinds can be coupled with Richard Boyd's (1999a) ecumenical thesis of natural kinds as homeostatic property clusters (HPC thesis). It has become the most prominent theory of natural or real kinds in the special sciences, including psychiatry. According to the HPC thesis, a natural or real kind consists of a property cluster and a homeostatic causal mechanism or mechanisms that are responsible for the systematic and regular

clustering of the properties. Crucially, replacing explanations based on essences and ironclad laws of nature with exception-permitting mechanisms, supports an ecumenical and domain-specific account of scientific kinds, which can still support sound inductive inferences. Consequently, if some psychiatric disorders are property cluster kinds, they can in principle be as causally potent and inductively rich as biological species or somatic diseases.

However, the discussion over natural kinds in psychiatry differs from similar discussions concerning more prototypical examples in the literature, such as biological species. The reason is that unlike in biology, there is no robust classificatory and inductive success in psychiatry that would call for natural kind explanations. Rather, prototypical natural kinds mentioned in the literature have been invoked as examples of objective classificatory targets to the extent that if psychiatric disorders resemble them, psychiatry could be placed on a par with natural sciences. Nevertheless, it has been a challenge to formulate a scientifically relevant ecumenical thesis of scientific kinds, and to maintain, based on empirical evidence, that such an account is applicable to research and diagnostic objects in psychiatry. On the one hand, the naturalistic and ecumenical views of natural kinds, such as the HPC view, which could potentially accommodate the multi-causal structure of putative psychiatric kinds, are criticized for their conventionalism and explanatory irrelevance (Craver 2009, Slater 2015). On the other hand, some argue that such an account means that socially constructed kinds can be real kinds as well (Boyd 2010, Mallon 2016, Khalidi 2013, Guala 2016, Godman 2020). This means that the status of psychiatric disorders as inductively rich kinds would not by itself differentiate them from socially constructed kinds such as gender or race or even institutional kinds such as money. Therefore, the notion of natural kinds could not be employed to make the distinction between natural and social phenomena as some have hoped. Because of this

watering down, Hacking (2007a), for example, argues that the concept of natural kind has lost its purpose.

The challenges of providing a natural kind account of psychiatric disorders have been addressed by pragmatism and eliminativism. Peter Zachar (2014b: 154) argues for a pragmatic solution, according to which economic and socio-political priorities, for instance, are relevant in determining how to classify psychiatric disorders. This view seems to be committed to an anti-realist view of psychiatric disorders, which is exemplified by the non-causal approach of the DSM and the ICD classification manuals (see Zachar 2000, Kendler et al. 2011: 1146). Another suggestion has been to give up the categorical approach to psychopathology. Sullivan (2014) and Tabb (2019), for example, point out that the complexity of psychiatric problems makes the kind-oriented philosophical approach currently unattainable, while Hasslam (2014) argues that it can even be socially harmful. Instead of a kind-oriented discussion, Tabb (2019) has recently appealed for more diverse non-categorical approaches inspired by the development of precision medicine and the Research Domain Criteria (RDoC) project. She argues that such approaches offer a better ground for scientific research, policy decision, and treatments. However, settling on a pragmatic account, or giving up on the categorical project altogether, have their own problems. The pragmatic approach does not seem to differ significantly from the descriptive approach of the classificatory manuals, which, as mentioned above, has been criticized on the grounds of comorbidity, lack of validity, and the medicalization of social problems. The challenge with the non-categorical approach, on the other hand, is that since it does not define mental disorders, it does not offer an explicit way to distinguish psychopathology from normality. Critics such as Wakefield (see Kincaid et al. 2021) argue that this could contribute to delegitimizing psychiatry as a medical science.

I discuss and examine the possibility of answering these challenges by providing a realist account of psychiatric disorders with sound – albeit limited – inductive power. My aim is to provide a moderate pluralistic explanatory account of scientific and psychiatric kinds that can facilitate interaction between discipline-related approaches to their explanation and nature. To this end, I formulate and defend an applicability domain approach to address the challenges of employing the HPC view in psychiatry. In particular, I explore how the contrastive-counterfactual theory of causal explanation (Ylikoski 2001, Woodward 2003), which has become the dominant explanatory account in the special sciences, can be employed to address the challenges of employing the HPC view of natural kinds. For my purposes, this theory comes with two benefits. First, it meshes well with the mechanistic account of scientific or real kinds (Pöyhönen 2013b) and enables a causal approach to intentional behaviour. The reason is that the theory is metaphysically non-committal, and is therefore suitable to describe in epistemic terms the interaction of various causal factors that are responsible for the properties of psychiatric kinds. Althought the account does not resolve the mind-body problem, it does help to understand how psychological and physiological factors can be connected to explain psychiatric phenomena. Second, according to the theory, reliable explanations are domain-relative in virtue of tracking specific invariance-relations, rather than being based on universal iron-clad laws of nature. This offers an epistemic way to understand the limited stability and scientific relevance of psychiatric kinds (see Pöyhönen 2013a). Consequently, the theory can potentially overcome some of the problems related to the descriptive and biomedical approaches to psychiatric disorders. Classifications of psychiatric disorders could be based on permissive causal mechanistic explanations, rather than on unreliable descriptions or impoverished biological explanations (Ylikoski & Pöyhönen 2013). In short, its permissive approach

to causal and mechanistic explanation can be employed to specify the interaction of various causes and factors that realize psychiatric disorders, and thereby identify the applicability domain within which kind concepts can be expected to support inductive inferences.

I reflect and compare the non-metaphysical account of psychiatric kinds to various reductionist and pluralistic explanatory approaches in psychiatry. In psychiatry and clinical psychology, there are multiple competing explanatory approaches to psychopathology. A common strategy to address this pluralism has been to attempt to formulate an integrative approach to facilitate "multi-level" explanations of psychiatric disorders. However, integrative views, such as those endorsed by Kendler (2005), Mitchell (2013), and Schaffner (2006), have not been very fruitful in practice. Therefore, I explore a less demanding approach to explanatory progress based on moderate or interactive explanatory pluralism, endorsed especially by Helen Longino (2013), according to which explanatory success is not necessarily measured by integration. Rather, the merits of different explanatory approaches can also be analysed based on what they can offer to the overall understanding of a phenomenon.

1.2 Classification of Psychiatric Disorders and its Consequences

The second question of the thesis is about what consequences my theory of psychiatric kinds has on classification and diagnostic interventions as well as the responsibility it bestows on those practices. Psychiatry lies in between the natural and human sciences because it not only deals with natural phenomena, it is also concerned with human experience and social behaviour. Hence, there is a need to consider social and ethical aspects related to the classification on psychopathology. The social

nature of humans as research objects, as well as the sociocultural causes of psychopathology, raise questions about reactivity, cross-cultural validity, and responsibility related to classification and explanation. Although there have been studies on each of these questions separately, their interrelatedness has not been sufficiently investigated. I address this gap by exploring how my theory of scientific kinds can account for psychiatric disorders as social as well as natural and psychological phenomena.

I explore the social nature of psychiatric disorders based on the claim that they differ from other kinds, such as biological species, because they are interactive human kinds. Ian Hacking (1986, 1995b) argues that at least some psychiatric disorders are not natural kinds, but are instead interactive human kinds because their classification induce looping effects, thereby making them difficult to track and explain. The looping effect describes the interaction between classifications and the targeted kinds of people or human kinds that purportedly share behaviour and traits. The idea is that classificatory practices induce reactions from the members of the human kind by enabling new intentional ways of being and acting. Tracking these changes requires revisions in the original classification, which may in turn lead to further changes in the members of the kind. Consequently, the interaction between the classification and the affected members of the human kind creates a feedback loop that renders that kind a moving target. According to Hacking, this classificatory instability generated by the looping effect distinguishes the human sciences from the natural sciences. In particular, the interactive human kinds studied by the human sciences do not support the robust explanations, predictions, and interventions (i.e. epistemic projects) that the natural kinds picked out by the natural sciences do.

Hacking's description of the looping effect has instigated a debate over whether human kinds can be given a realistic interpretation. Critics

such as Cooper (2004) and Khalidi (2010) argue for a realist view on the basis that some prototypical natural kinds are also subject to the looping effect, such as domesticated animals and disease entities. Moreover, Murphy (2006) asserts that looping effects can stabilize human kinds, whereas Mallon (2016) argues that in general our knowledge can keep up with their rate of change. On the other hand, Laimann (2018) and Allen (2018) have recently defended Hacking's position. They argue that although some biological kinds are subject to the looping effect, interactive human kinds differ from interactive biological kinds because their classificatory-induced reactions are difficult to explain and predict. Allen associates the problem with classificatory-induced intentional reactions being ontologically anomalous, whereas Laimann associates it with the complex social interactions that underlie our inability to uncover mechanisms for patterns of change and stability. In sum, the looping debate is based on the dichotomy of whether interactive human kinds are real kinds that support robust epistemic projects.

However, the discussion has mainly concentrated on whether looping effects are problematic for epistemic projects. My aim is to identify how knowledge of looping effects can supplement the explanations of interactive human kinds, such as psychiatric kinds. The central challenge with Hacking's account of human kinds and the looping effect is that he seems to maintain that those kinds require explicit conceptions and intentional reactions to them. However, this account in general does not seem to apply to most psychiatric disorders as such because they have existed before we created their classification and conceptualization. Nevertheless, I wish to explore how the looping effect has consequences that make psychiatric disorders as human kinds difficult to classify and explain.

I also explore whether my applicability domain account can explain cross-cultural variation of psychiatric disorders. I examine social

construction and sociocultural explanations of psychiatric disorders, and their implications for a realist account of psychiatric kinds. To this end, I compare my applicability domain account to the medical model, and especially the Research Domain Criteria (RDoC) research program, according to which psychiatric disorders are brain diseases. The cross-cultural variation of psychiatric disorders raises the question of to what extent they are socio-culturally caused and constructed. The anthropologist Arthur Kleinman (1991) has argued that the diagnostic categories found in classificatory manuals are culture-specific, and thereby their universal application would be a "category fallacy". Anthropologists call the effects of cultural meanings on psychopathology "idioms of distress", which describe culturally sanctioned and conceptually enabled ways to enact symptoms of psychiatric disorders (Kirmayer 2001). Moreover, according to recent studies, cognitive processes that were previously taken to be universally shared may in fact be unique to WEIRD (Western, Educated, Industrialized, Rich and Democratic) people (Henrich et al. 2010). To address these claims, I explore how my non-metaphysical approach to explanation can accommodate sociocultural causation and construction as an explanatory approach to account for the cross-cultural variation of psychiatric disorders. As case studies, I investigate how the interaction between cognitive and sociocultural factors shape and realize culture-bound syndromes (i.e. cultural syndromes). I also raise the possibility that some psychiatric disorders, especially culture-bound syndromes, are co-constructed by cognitive and sociocultural factors to the extent that locating them on the social level, rather than in our brains or minds, may be a more fruitful explanation.

Finally, I investigate how my findings about the nature of psychiatric disorders, and its implications for their classification, relate to recent discussion on conceptual engineering. According to conceptual engineering, we should aim to refine scientific concepts to

accommodate their scientific and practical purposes, rather than try to discover their purported meanings through descriptive conceptual analysis. Conceptual engineering as explication has recently been employed by Schwartz (2014) and Griffiths and Matthewson (2018) to argue that the concept of mental disorder should be refined to accommodate scientific aims and discoveries. Although I am sympathetic towards this approach, I also wish to explore whether conceptual engineering applied to psychiatric concepts can be combined with my pluralist value-sensitivity account. For this reason, I employ Sally Haslanger's (2012) ameliorative approach to conceptual engineering, according to which some social kind or type concepts, such as "race" and "gender", should be refined to accommodate morally and politically motivated goals, instead of putatively discovered scientific facts. In particular, based on my discoveries concerning the social consequences of the looping effect, I will investigate whether that knowledge could inform classificatory decisions relating to psychopathology. In particular, I follow Ludwig (2016) and Brigandt (2020), who have lately argued that conceptual and classificatory decisions about psychiatric kinds should be informed by non-epistemic values, such as how to balance the right for treatment with the stigma associated with diagnoses.

1.3 Methodology

This thesis is an investigation in the philosophy of psychiatry. I examine issues concerning the nature and classification of psychopathology by applying concepts and theories developed in the philosophy of science. I concentrate especially on causal and mechanistic explanation as well as theories of natural kinds. This thesis is also partly based on interdisciplinary research. In addition to having participated in seminars

and workshops in philosophy, psychiatry, and anthropology, I have conducted interviews in Italy and Benin. In Italy, mostly at NIHMP (the National Institute for Health, Migration, and Poverty) in Rome in 2017 and 2019, I interviewed professionals about their community-based approach to diagnoses and the application of DSM categories. I have also conducted fieldwork in Benin, West Africa, in 2020, where I studied *vodun* beliefs related to mental illness. The aim was to understand the impact of culture on the perception and treatment of mental disorders. However, this thesis is a philosophical rather than an empirical examination, and these interviews and my fieldwork have functioned mostly to ensure that the philosophical questions and arguments I examine in this thesis are empirically relevant.

My philosophical approach is naturalistic. I do not claim to provide guidelines on how empirical research on psychopathology should be conducted or what results would be conceptually acceptable. Rather, I believe that conceptual questions over psychopathology should be settled in communication with scientific findings. This is because intuitions are, as recent studies have demonstrated, relative to culture, profession, and identity. Relying on armchair intuitions alone would be articularly dubious when examining concepts employed by psychiatry and anthropology, which in the past have both been accused of ethnocentrism, racism, and gender bias. Nevertheless, I believe that philosophical and conceptual analysis that is supported by empirical research can serve an important role in settling muddled conceptual issues in psychiatry. As Kendler (2016) points out, scientific research on psychopathology is broad, requires diverse disciplinary-related approaches, and has wide-ranging social, economic, and ethical consequences. Moreover, Murphy (2006) has rightly pointed out that a great deal of confusion in psychiatry stems from folk psychology. Many current diagnostic categories have roots in folk psychology and reflect preconceptions about what disorders should be like. Just as the category

of superlunary objects in the geocentric model hindered the development of physics (Kuhn 1970, Griffiths 1997), many current diagnostic categories based on folk psychology may restrain and mislead empirical research in psychiatry. Hence, conceptual clarity is required to settle folk-psychological commitments and to facilitate successful comparisons between various disciplinary-related explanations of psychiatric disorders. In short, this thesis is based on the convinced that philosophy can in part function as a medicine for the conceptual confusion in psychiatry.

1.4 Overview of the Thesis

In the second chapter**,** I examine the ontological commitments made by classifications and explanations of psychiatric disorders, as well as the philosophical debate about whether the concept can be given a definition based on scientific facts. I begin by explicating the notion of psychiatric realism as a naturalistic and discipline-related thesis concerning scientific realism in psychiatry. I then examine the descriptive and pragmatic approach of the DSM and ICD classification manuals and the criticism they have as received. I also provide a brief overview of different explanatory approaches to psychiatric disorders. The chapter is, however, mostly dedicated to the philosophical discussion about the concept of psychiatric disorder. For this reason, I distinguish between conceptual and methodological discussions. In the conceptual discussion I examine attempts to provide a natural definition of psychiatric disorder and arguments that support a value-based account. I analyse and criticize the most prominent naturalistic theories of mental disorder, Christopher Boorse's (1976) biostatical theory and Wakefield's (1992a) harmful-dysfunction theory, based on their

inability to provide a naturalistic account of dysfunction. As a representative example of normativists accounts of the concept of psychiatric disorder, I examine Cooper (2005, 2020) view, according to which the concept is inherently value laden. I point out that her account seems to have dubious relativistic consequences. Finally, I follow Cooper and Murphy by arguing that the superordinate concept of psychiatric disorder is value laden and open-ended, although the conditions that fall within its extension can be real kinds.

The second discussion concerns the nature of conceptual analysis in determining what psychiatric disorders are. I make a distinction between top-down and bottom-up approaches to defining psychiatric disorder. The naturalist views by Boorse and Wakefield largely rely on conceptual analysis to provide an account of what disorders are in their totality and are thereby top-down. Murphy, however, has especially defended a naturalistic bottom-up approach, according to which the right method to uncover the nature of psychiatric disorders is not conceptual analysis but empirical research.

Based on this distinction, I argue that the applicability of Cooper's normativism is limited because of its reliance on descriptive conceptual analysis. On the other hand, I suggest that Murphy's bottom-up naturalism has an explanatorily unnecessary strong ontological commitment to psychiatric disorders as neurological problems. Further, I explore the possibility that the concept of psychiatric disorder has been contextually embedded in our evolution (Broadbent 2020) and sociocultural practices (Canguilhem 1991, Hacking 1995b). Based on this, I argue that the concept needs to be open-ended to cover all the disorders we want, and will want, it to apply to. To this end, I present my own approach – value-sensitive pluralism – according to which the values associated with psychiatric disorder should be examined, and if deemed necessary, revised in a value-sensitive fashion based on their social impact. However, I also argue that psychiatric disorders can be

real in different ways, and that they can be re-classified differently for different epistemic and non-epistemic purposes.

In the third chapter, I examine whether particular psychiatric disorders can in principle be inductively rich and scientifically relevant real kinds. I begin by arguing against essentialist views of natural kinds in favour of naturalist and ecumenical approaches. Following others, I argue that the homeostatic property cluster theory of natural kinds (HPC view) matches how psychiatric disorders are employed to make inductive inferences in psychiatry (e.g. Kendler et al. 2011). However, I assert that the problems of a purely mechanistic interpretation of the HPC view, endorsed especially by Kendler et al. (2011), lead to the individuation problem, as suggested by Craver (2009). I assert that the problem lies in the purely mechanistic interpretation of the HPC view, which raises some of the same problems that the essentialist interpretation had.

In the fourth chapter, I formulate and defend my own account. I begin my providing a an account of mechanistic explanation in psychiatry based on the contrastive-counterfactual theory of explanation (Woodward 2003; Ylikoski 2001). Thereafter, I argue that mechanistic explanations of human kinds have *applicability domains* over which they reliably account for aspects of the kinds in counterfactual situations. The applicability domain is better when it can account for more aspects of the kind in a wider range of alternative situations. This means that explanations of a homeostatic property cluster kind fall on dimensions of goodness described by the applicability domain. The idea is that a better mechanistic explanation of a kind enables more secure domain-relative projections (i.e. generalizations and predictions) based on that kind-concept. Moreover, an explanation ideally spells out its applicability domain because human kinds such as psychiatric kinds support limited epistemic projects. In this case, classificatory projects that apply the explanation do not exceed the limits of their own

applicability. That is, identifying the applicability domain of an explanation spells out the conditions under which the explanation is expected to be reliable and when it can breakdown. This helps to understand how disciplinary explanations, including social and cultural ones, can be complimentary and make an explanation of a psychiatric kind more reliable. On the other hand, the account also shows how different epistemic and non-epistemic interests can license re-classifications of disorders for research, clinical and pragmatic purposes. Thereafter, I compare my account to other approaches to explanatory pluralism. In particular, I point out some of the shortcomings of biomedical approaches to explanatory integration. In contrast, I suggest that my non-metaphysical applicability domain account retains the benefits of causal and mechanistic explanation, while foregoing the problems of more exclusive ontological approaches.

In the fifth chapter, I investigate the instability of psychiatric disorders that is generated by their interactive nature and susceptibility to looping effects. Ian Hacking uses the looping effect to describe how classificatory practices in the human sciences, including psychiatry, interact with the people classified. While arguably this interaction renders the affected human kinds unstable and hence different from natural kinds, realists argue that some prototypical natural kinds are also interactive and human kinds in general are stable enough to support explanations and predictions. I defend a more fine-grained realist interpretation of psychiatric kinds as interactive human kinds by arguing for an explanatory domain account of the looping effect. First, I argue that knowledge of the feedback mechanisms that mediate the looping effect can supplement and help to identify the applicability domain by which a kind and its property variations are reliably explainable. Second, by applying this account to cross-cultural case studies of psychiatric disorders, I distinguish between congruent feedback mechanisms that explain matches between classifications and

kinds, and incongruent feedback mechanisms that explain mismatches. For example, congruent mechanisms maintain Western auditory experiences in schizophrenia, whereas exporting diagnostic labels inflicts incongruence by influencing local experiences. Knowledge of the mechanisms can strengthen explanatory domains, and thereby facilitate classificatory adjustments and possible interventions on psychiatric disorders.

In the sixth chapter, I compare my applicability domain account with the medical model based on their ability to explain the cross-cultural variation of psychiatric disorders. I argue that the medical model – both the ecumenical pragmatic approach and the stronger biological approach – cannot account sufficiently well for cross-cultural variation of psychiatric disorders and sociocultural causation. In contrast, I argue that my approach can offer a heuristic means to facilitate interaction between various discipline-related explanatory approaches, including sociocultural explanations, based on its a non-metaphysical approach to psychiatric kinds. Further, I assert that because cross-cultural cognitive variation cannot be ruled out, neither can the possibility that psychiatric disorders even on a lower level of cognition vary across cultures. In support of this argument, I examine culture-bound syndromes. I argue that sociocultural causation falls on a continuum, and knowledge of social factors can rule out exclusive social construction. More specifically, I argue that the relevance of sociocultural factors depends on how the explanandum phenomenon is specified – that is, which aspects of a psychiatric disorder we are interested in. Based on this, I argue that sometimes a more powerful explanation is offered by construing a psychiatric problem as constituted by a dynamic social process, rather than locating it within the individual. Finally, I make some remarks opposing semantic arguments that support biological accounts of psychiatric

disorder, on the one hand, and accounts that seem to support incommensurability and antirealist interpretation of psychiatric progress, on the other.

In the seventh chapter, I present my conclusions and reflect on them in the light of recent research on conceptual engineering. I argue that further research should be conducted to establish a value-sensitive methodology for the conceptual engineering of psychiatric concepts. I point out that the pluralistic nature of psychiatric explanations, and the fact that psychiatric kinds can be classified in different way, suggest that those classifications should be conducted in a value-sensitive manner. I assert that the feedback mechanisms that medicate the looping effects of psychiatric disorders should not only be identified to supplement the epistemic aims of explanatory models and classifications, but also to evaluate these models and the classifications that rely on them, based on their anticipated social outcomes. The reason is that if conceptions and models of psychiatric disorders become embedded into social reality, they may influence the health and social life of those classified. Especially in borderline cases, the choice whether to pathologize a condition, or which alternative explanatory model to apply, should be weighed against its potential social consequences. Consequently, empirical research on the consequences of alternative ways to model and specify psychiatric kinds should inform value-sensitive choices on how to apply those models and classifications.

2 What is Psychiatric Disorder?

There are competing explanations and classifications of psychopathology that have different commitments to its nature. The concept of psychiatric disorder or mental disorder is, at least currently, relevant to all of them. It makes a commitment to the type of general theory of psychopathology that is required. The concept provides information about the attributes that purportedly bind different psychiatric disorders together, and thereby channels research on their classification and explanation. Moreover, the concept of psychiatric disorder is not only important for academic reasons, but also because it has social, political, and ethical consequences. For example, it underlies decisions over who is entitled to treatment, channels funding for research and clinical practice, and influences lay conceptions of the normal and pathological.

In this chapter, I examine how the question over the nature and realness of psychiatric disorders can be approached and what role the nature of the concept of psychiatric disorder plays in these approaches. I begin by pointing out that rather than the general questions of scientific realism, psychiatric research, classification, and practice raise more specific discipline related conceptual questions about the nature of psychiatric disorder. In particular, rather than asking whether psychiatric disorders are real, it is more fruitful to ask in what sense they are real, and how different explanatory approaches could be combined to understand their nature. Based on this, I elaborate a naturalistic interpretation of scientific realism in psychiatry – psychiatric realism – and compare it to pragmatist and the social constructivist approaches to psychiatric disorders. Subsequently, based on these distinctions, I provide a brief historical account of the concept of psychiatric disorder and why it remains contested. I then examine the definition of psychiatric

disorder in the classification manuals as well as various explanatory models of psychopathology.

Most of the chapter, however, is dedicated to the philosophical debate over the concept of psychiatric disorder. I examine whether the concept of psychiatric disorder can be given a definition based on scientific facts, or whether it is irreversibly value laden. I start by distinguishing between two ways the concept has been discussed in the philosophical literature. Although recently different ways to approach the discussion have been offered (see Wilkinson 2023), this discussion is important for understanding what kinds of things the disorders are, and what is the right method to discover them. The first discussion is between naturalist and normativist (or constructivist) approaches. While naturalists argue that the concept can be defined based at least partly on objective scientific facts, constructivists argue that the concept is irreversibly value laden. As representative examples of naturalism, I compare Christopher Boorse's biostatistical theory and Jerome Wakefield's harmful dysfunction theory. As an example of a recent normativist account, I analyse Cooper's account. I argue that while these traditional naturalistic approaches cannot provide an interpretation of dysfunction based on biological or psychological scientific facts alone, normativism cannot avoid relativism and thereby it stands against our realist intuition about psychiatric disorders.

The second discussion concerns the most appropriate the correct methodology to discover the nature of psychiatric disorders. I distinguish between top-down approaches that rely heavily on conceptual analysis to determine the nature of psychopathology, and bottom-up approaches that are convinced that the task should be left to empirical research. As examples of this approach, I scrutinize Murphy's neurocognitive approach, as well as contextual approaches by Broadbent, Canguilhem and Hacking. While Murphy's neurologically oriented approach would mean large overhauls in the current diagnostic categories,

the contextual approaches are based on providing accounts of why we have the concept of psychiatric disorder in the first place. After pointing out some of the challenges this accounts face, I put forward my own bottom-up value-sensitive approach, according to which the value-laden process of determining psychiatric disorders should be paid more attention. The underlying concepts and classifications of psychiatric disorders need to be studied and balanced with empirical discoveries.

2.1 Psychiatric Realism and the Concept of Psychiatric Disorder

Although there is no general agreement about the nature of psychiatric disorders, most researchers and practitioners are convinced that they exist independently from their conceptualizations, classifications, and research. This view can be understood as a central tenet of psychiatric realism, according to which real psychiatric disorders do exist. In the following, I will provide a conceptual and historical overview of the debate surrounding psychiatric realism, and briefly situate my own account in relation to it.

As a general view, scientific realism can be divided into ontological, semantic and epistemic stances (e.g. Psillos 1999 xix; Niiniluoto 2019). The ontological stance is that the world consists of a mind-independent natural kind structure; the semantic stance is that the theoretical terms and concepts of our best theories and models refer to those unobservable natural kinds; the epistemic stance is that those theories and models are well confirmed and approximately true. Consequently, scientific realism about psychopathology would mean that mental disorders are natural kinds; psychiatric classificatory terms refer to them and our best explanatory models and theories about them are well confirmed. However, the debate over the nature of

psychopathology is largely discipline-relative, and therefore the general questions of scientific realism as such are not relevant in the philosophy of psychiatry. Part of the reason is that psychiatric disorders differ from prototypical natural kinds, such as chemical elements and biological species. They are mind-dependent and fuzzy, and their identifications may be value-dependent. Moreover, it is generally admitted that current classifications and explanations are mostly incorrect and unwarranted (see e.g. Kendler 2021). Hence, instead of the traditional realism debate, scientifically relevant debate over the nature of psychiatric disorders has concentrated on questions that have arisen in the empirical research and classification of psychiatry.

Instead of the traditional realism debate, a scientifically relevant debate over the realism of psychiatric disorders, or psychiatric realism, has concentrated on two questions. The first challenge has been to determine the nature of the concept of psychiatric disorder. This discussion is important because the meaning of the concept of psychiatric disorder is a theory of the nature of the objects it putatively refers to. In the discussion, realists commonly hold that the concept can be given an objective and naturalistic definition based on biological or psychological facts. This would mean that psychiatric disorders are objective scientific entities, thereby implying that psychiatric science can resemble somatic medicine. The second challenge concerns the correct means of determining the nature of psychiatric disorders. This is an important question for two reasons. One reason is that it is unclear as to what extent the nature of psychiatric disorder, if there indeed exists such a thing, can be determined by analysing common intuitions associated with the concept. If our current folk conceptions and intuitions – which also seem to play a role in psychiatric research and classification – are incorrect, psychiatric disorders, in the sense commonly assumed, do not strictly speaking exist. This does not mean, however, that the many conditions we label as psychiatric disorders would not exist, only

that they may not be the kinds of objects they are conceptualized to be. The second reason is that, in the light of historical and cross-cultural variations in both our perceptions and the conditions themselves, it is not clear whether we have identified psychiatric disorders correctly to begin with. If these assumptions are correct, instead of relying on conceptual analysis to determine what disorders are, we should be more revisionist concerning their conceptualization and identification. This would imply that the task should be left to empirical research.

Psychiatric realism can be contrasted with social constructivism and instrumentalism. Minimally, a realist interpretation of psychiatric disorders maintains that although they can be influenced by their conceptualizations, classifications, and explanations, they are nonetheless not constituted by them. In the past, the most influential challenge to psychiatric realism has been social constructivism, which can be divided into social construction about concepts and kinds (Bell 2014). Social constructivism about concepts is an antirealist and nominalist view of psychiatric disorders. It is based on the idea that psychiatric disorders are nothing but social or subjective beliefs about concepts that are imposed on patients and society at large. Such an antirealist account of psychopathology was maintained by some anti-psychiatrists. They argued that psychiatry in general, and the concept of psychopathology in particular, are employed for social control and to label social deviance.

More moderate views on the social construction of psychiatric disorders, some of them compatible with weak psychiatric realism, have been supported by conceptualists or by discourse arguments about kinds of psychopathology. These views rely on historical and cross-cultural variations of the conceptions associated with the concept of "psychiatric disorder" and the conditions so labelled. They maintain that the concept of psychopathology is value laden and depends on interests to the extent that those beliefs and values influence the

conditions and patterns of behaviour labelled as pathological, and may even, to a certain extent, constitute them. Nevertheless, these accounts are not necessarily antirealist, because they can agree that psychiatric disorders are real joints or conditions in the world, and not just social deviances or subjective projections, although they are shaped by sociocultural processes.

On the other hand, current classificatory approaches are based on a weak medical model view of psychiatric disorders that is instrumentalist or pragmatist in its orientation. Instrumentalists and realists alike may answer strong social constructivists that although our current theories and models are not perfect, they do nonetheless enable limited prediction and interventions. Their difference is that while realists argue that we should seek to uncover the underlying causal structure of psychiatric disorders, pragmatists claim that disorder delineations are necessarily determined by pragmatic and ethical considerations. Nevertheless, the argument of pragmatists is that this is not a problem as long as the pragmatic categories can support predictions and medical interventions.

2.1.1 A Brief History of the Concept of Psychiatric Disorder

The question over the nature of psychiatric disorders, and the correct means to determine it, has played an important role in the history of psychiatry. Modern psychiatry has gone through different stages in its approach to the nature of mental disorders. The father of modern psychiatry, Emil Kraepelin (1896), endorsed a reductionist and realist account of psychiatric disorders. He believed that mental disorders are "natural disease entities", that is, neurobiological brain diseases that are discrete and discoverable (Pietikäinen 2013). He believed that these neuropathological entities could either be detected directly, through

their distinct aetiologies, or by analysing the unique symptom clusters they generate (Bental 2003: 15). Although during Kraepelin's time it seemed that only symptom-based determinations were possible, he was convinced that those determinations would eventually converge with their discovered aetiology and neuropathology, hence forming a natural psychiatric classification (Murphy 2014: 107). In other words, Kraepelin argued that mental disorders are akin to somatic diseases generated by bacteria entities, and are therefore natural kinds that can be defined based on sufficient and natural conditions. This view was validated to some degree in the beginning of the twentieth century, when the spirochete bacterium was found to be responsible for syphilis by means of invading the central nervous system (see e.g. Bolton 2013: 446). Eventually this discovery led to the discovery of a cure by penicillin, fuelling yet to be fulfilled optimism over finding similar causes for other mental diseases.

The antipsychiatry critique of the 1960s represents a philosophically anti-realist and social constructivist view of psychiatric kinds. According to the critique, mainstream psychiatry labels behaviour and traits as pathological that are merely socially deviant and morally questionable in the eyes of society. In other words, anti-psychiatrists argued, for slightly different reason, that the concept of mental disease is thoroughly value laden and figures in power relations, instead of picking out genuine mental disease (see Bolton 2013: 440).[4] Hungarian-American psychiatrist Thomas Szasz's argument for anti-realism is especially influential and philosophically succinct. Szasz's view can be understood as a form of error theory (Wrigley 2007). According to him, although the sentence "someone is mentally ill" is truth-apt, its referent is empty. That is, although it is meaningful and

[4] Anti-psychiatry was influenced by Foucault (1992, 2006), Goffman (1961), Laing (1987), Rosenhan (1973), Thomas Szasz (1999), and Thomas Scheff (1999).

natural to ask whether someone is mentally ill, the only legitimate diseases are physical, and since mental disease is by definition mental, such things cannot exist. Conversely, if mental illnesses were real diseases, they would be brain diseases, not mental problems. Therefore, the conditions that we call mental illnesses are merely problems of living, or myths, that the current society deems problematic for social and political reasons (Szasz 1999, 1961: 113).

Szasz (1961) compared the concept of mental disorder with the proposed mental disorder called *drapetomania*. In 1851 the physician Samuel Cartwright claimed to have identified *drapetomania* as the common cause for slaves' urge to escape.[5] On the other hand, if a slave was able to escape, the disorder turned into *dysaesthesia aesthiopis*, a mental disease caused by the inability to cope with freedom (Zachar 2014b). Cartwright believed that he had an ideal remedy at hand: "With the advantages of proper medical advice, strictly followed, this troublesome practice that many negroes have of running away, can be almost entirely prevented" (Cartwright: 1851/2004: 34). While it is hard to determine whether this proposed diagnostic category had any practical influence, insanity diagnoses have been used to justify coercive measures. But Szasz (1961) makes a stronger argument. He argues that the behaviour and traits that are labelled as mental disorders even nowadays are no different from *drapetomania*. They are merely problems of living, which, based on our social norms, we tend to label as mental problems.

The concept of psychiatric disorder (i.e. mental disorder) gained its modern definition and societal status when it became important for the Diagnostic and Statistical Manual of Mental Disorders (DSM) in the 1970s due to pressure from anti-psychiatrists on the one hand, and gay-

[5] It seems natural for any slave to attempt escaping, but perhaps the term was meant to explain the difference between individuals who escaped and those who did not.

rights movement on the other. Homosexuality had been included in the DSM-II as it had been considered a mental disorder at least from the 19th century onwards (Cooper 2005). However, the gay movement in the 1960s campaigned that homosexuality was not a mental disorder, and labelling it as such was a form of social repression that aimed to subdue social deviance, making the label "homosexuality" in essence similar to *drapetomania.* To defend the realism of mental disorders, and the diagnostic category decisions made so far, Robert Spitzer, who eventually became the chairman of the DSM-III committee, argued for a neutral position (Cooper 2005: 85). He maintained that homosexuality should be considered a mental disorder only in the event that it caused distress or disability to the individual (Cooper 2005: 85). Thus, individuals who do not experience distress or disability do not have a mental disorder, whereas individuals that do, should be labelled as having a mental disorder. In 1973, homosexuality was replaced by the condition "Sexual Orientation Disorder".[6] Later, Spitzer played a pivotal role in how the definitions of mental disorder were formulated.

The anti-psychiatric movement is exemplified by the famous Rosenhan study (Science 1973). Rosenhan sent actors to mental hospitals in the USA. These actors, who claimed to hear "empty" sounds, and "hollow thuds", exhibited no other symptoms. All the actors were hospitalized and medicated, and although they ceased reporting about voices, and acted normally, they were not discharged from the hospitals. The situation only became more dire once they admitted to having acted, so much so that some remained hospitalized up to a few months. When Rosenhan exposed the experiment, one hospital asked him to send more actors, claiming that they could detect them. After a few months, the hospital proudly claimed to have

[6] The American Psychiatric Association's finally excluded homosexuality from the Diagnostic and Statistical Manual of Mental Disorder in 1980.

discovered 41 actors. Rosenhan answered to the embarrassment of the whole psychiatric system that he had not sent any.

Because of these manifest problems with psychiatric diagnoses, the American Psychiatric Association formed a task force to come up with an appropriate definition of mental disorder. The aim was to render classifications valid and reliable. Reliable in the sense that the same symptoms would be diagnosed in the same way, to counter the Rosenham problem. Valid in the sense that the classification would capture what they are meant to capture, in other words, to exclude non-disorders such as drapetomania and homosexuality. The upshot of the task force was to define mental disorders based on descriptions of symptoms, hence rendering the definitions atheoretical by extracting causes (with a few exceptions).

In the course of this thesis, I argue against both the Kraepelian and the Szaszian views. This is not surprising since neither of the approaches are endorsed as such anymore. However, the definition of mental disorder in the DSM is neo-Kraepelian (see Walter 2013). That is, mental disorder is defined based on – yet to be discovered – discrete dysfunctions that can be identified based on manifest and unique symptom clusters. As a result, the DSM is in theory realistic about mental disorder (the distinction between psychopathology and normal health is based on dysfunctions), but in practice instrumentalist about particular disorders (they are defined by symptom delineations). The Kraepelian view is also apparent in the Research Domain Criteria (RDoC) approach to mental disorders, which ultimately considers mental disorders to be brain diseases (see Chapter 6). On the other hand, RDoC also agrees with Szasz to the extent that genuine disorders are not mental problems. Szasz's arguments can also be found in the normative views of mental disorders to the extent that they hold social values to play a definitive role in determining how psychiatric disorders are defined. My approach aims to find a middle ground between these

approaches in relation to the nature of particular psychiatric disorders. They are neither determined by biological causes nor are they merely social constructs.

2.1.2 Definitions and Validity in Classification Manuals

The most influential definition of mental disorder (i.e. psychiatric disorder) can be found in psychiatric classification manuals. However, they have been faced with "crises of validity", according to which the diagnostic categories do not define real disorders. I first provide an overview of the manuals, and thereafter provide an overview of the criticism they have received.

There are two prominent classification manuals of mental disorders that are used for definitions, classifications, and diagnoses of mental disorders. The more prominent is the Diagnostic and Statistical Manual of Mental Disorders (DSM*)* produced by the American Psychiatric Association (APA). The other one is the international Classification of Mental and Behavioural Disorders*,* which is part of the International Classification of Disease and Related Health Problems (ICD) produced by the World Health Organization. Both have gone through numerable revisions, with the result that the number of conditions has increased. The manuals are similar enough to warrant considering them together. Although I will mostly concentrate on the DSM, my arguments also apply to the ICD.

The third edition of the diagnostic and statistical manual of mental disorders, the DSM-III, published in 1980, has gone through a substantial overhaul, so that the classification is rendered atheoretical by describing symptoms, not causes. Thereby, unlike taxonomies in physical medicine, the manual does not define mental disorder with

reference to causes.[7] The descriptive approach to classification is the result of aiming at neutrality in the theoretical conflict between psychoanalysis and biological psychiatry, on the one hand, and wanting to answer the gay movement and subdue the anti-psychiatry movement, on the other.

The manual has gone through three revisions since then, the latest edition being DSM-5, but the central descriptive approach has remained. The manual conceptualizes mental disorders as symptoms that cluster together to form syndromes. The symptoms are mostly disjunctive (although some necessary) lists of psychological states and behaviour described in the terminology of clinical phenomenology. The DSM provides a definition of mental disorder that is meant to cover a myriad of conditions, instead of imposing strict restrictions on what those conditions are. These are, for instance, mood and anxiety disorder, schizophrenia, other psychotic disorders, adjustment disorders, somatoform disorders, dementias, attention deficit hyperactivity disorder, conduct disorder, autism, personality disorder, and many more (Bolton 2008). The definition has retained the main ideas of the DSM-III of Spitzer and Endicott (1978) The current version DSM-5 and ICD-10 provide the following definitions of mental disorder:

"*A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour that reflects a dysfunction in the psychological, biological, or developmental processes underlining mental functioning. Mental disorders are usually associated with significant distress or disability in the social, occupational, or other important activities. An expectable or culturally approved response to a common stressor or loss, such as*

[7] Notable exceptions are PTSD and "psychosis due to substance abuse" (see Bolton 2008).

the death of a loved one, is not a mental disorder. Socially deviant behaviour (e.g. political, religious, or sexual) and conflicts that are primarily between the individual and society are not mental disorders unless the deviance or conflict results from a dysfunction in the individual, as described above" (APA 2013: 20).

"*The term 'disorder' is used throughout the classification, so as to avoid even greater problems inherent in the use of terms such as 'disease' and 'illness'. 'Disorder' is not an exact term, but it is used here to imply the existence of a clinically recognizable set of symptoms or behaviour associated in most cases with distress and with interference with personal functions. Social deviance or conflict alone, without personal dysfunction, should not be included in mental disorder as defined here*." (World Health Organization 1992: 5).

According to these definitions, mental disorder consists of *dysfunction* and *harmfulness* described as distress and/or disability.[8] The conditions were already present in the way Spitzer et al. (1978) first defended the definition of mental disorder in the DSM-III, which is essentially the same as that of the DSM-5.[9] These conditions are operationalized so that dysfunction is a "clinically significant disturbance in the individual" that is related to "personal distress or impairment in social, occupational, or other areas" – this is the operationalized definition of harmfulness (Bolton 2008). Operationalizations by themselves,

[8] The definition therefore resembles Wakefield's (1992a, 1992b) harmful dysfunction definition. The distinction between harmfulness and dysfunction, and the definition overall, was incorporated into the DSM-III as a compromise when homosexuality was removed from the list of syndromes (Cooper 2005: 83). Spitzer argued that homosexuality is not a mental disorder unless it causes distress and impairment.

[9] The DSM-5 edition does not consider harmfulness to be necessary anymore, whereas in the previous edition since DSM-III the conditions were sufficient and necessary (see Cooper 2015).

however, are not enough to distinguish the disorders from social normality, as acknowledged by the DSM definition. This is because in many cases the same behaviour and psychological states can be symptomatic for normal as well as abnormal factors. For instance, in some cases, the symptoms of major depression and bereavement are indistinguishable (Bolton 2008). To overcome this problem the manual has the clause that the symptoms should not be "expected or culturally approved". In brief, as Wakefield (1992a) points out, the definition of mental disorders is an "unexpected distress or disability", where the unexpected condition stands for dysfunction, and the distress or disability condition stands for harmfulness.

The DSM categories were aimed at being reliable and valid. Whereas reliability is the measure of how coherently the same diagnosis is made, validity is the measure of how well the diagnosis measures what it is supposed to measure. The operationalizations in the DSM are based on construct validity, which measures the success of classification by relating observational features with other observations (Murphy 2006: 218). This was intended to render diagnoses reliable, meaning that different clinicians diagnose patients with the same syndromes, and crucially, that the categories offer commensurable frameworks for research. Especially intra-rater reliability (consistency in the same individual's repeated measurements) and inter-rater reliability (consistency in the different individuals' repeated measurements) have been the aims (Kincaid 2014: 3). These goals were largely achieved by rendering the defining features observational, and forgoing theoretical and causal assumption about the syndromes (Bolton 2008).

However, there is a tension in DSM between reliability and validity. The problem is that validity may not be reached by rendering the conditions as reliable as possible through operationalization

(Murphy 2014).[10] Although the categories might measure validly their constructs, the syndromes, they might not detect real mental disorders that putatively underlie these symptom clusters.[11] Part of the reason is that the language of clinical phenomenology found in the DSM is closely linked with the norms of folk psychology, in some parts, and with unspecified theories of abnormalcy, in other parts (Murphy 2006: 221). That is, according to Murphy (2006: 221), the terminology of clinical phenomenology is full of concepts that are not operationalized, such as the internal psychological states of patients, and other unidentified theoretical commitments (Murphy 2013). This implies that the distinction between mental disorders and normalcy in the manuals is influenced by social norms rather than being purely based on medical and scientific norms (cf. Bolton 2008). In other words, the observable symptoms clusters may not offer the means to demarcate mental disorders from normality, on the one hand, or from each other, on the other.

Bolton (2008) argues that one might try defend the DSM manual by arguing that psychiatric disorders are implicitly recognized, and therefore that there is no need to rely on a theory or model to distinguish normal behaviour and psychology. This would mean that mental disorders are those things that patients bring to the clinic. The implication is that our folk-psychological view of normality underlies, as it should do, classifications of psychopathology. In general, this is the underlying idea behind the DSM classification manual (and ICD), clinical work, and folk conceptions of mental disorders (ibid.).

However, the DSM is commonly criticized for pathologizing normal psychological states and behaviour such as normal grief

[10] Solomon has called this "Crises of Validity" (Presentation 22.10.2021).
[11] Even this is questionable because Spizer and Endicott (1978) intended the definition to capture dysfunction, not merely its operationalization. However, Haslam (2013: 988) argues that they should not be understood to provide an ontological account.

(Horwitz and Wakefield and 2007) and hyperactive children, just as run-away slaves and homosexuality were pathologized before. The history of psychiatry and DSM is full of diagnostic categories that are not considered real anymore. Moreover, the current categories have implausibly strict definitions (e.g. major ADHD requires five symptoms of inattention or hyperactivity that have lasted at least 6 months, see APA 2013). The categories have also been criticized for comorbidity by sharing the same features in the symptom profiles (e.g. schizophrenia and bipolar disorder are associated with the same cognitive deficits) (see Poland and von Eckardt 2013: 747, Haslman 2013). The idea is that comorbidity indicates deeper causal differences that have not been properly distinguished (Solomon 2022). In other words, based on this criticism, it is doubtful whether factual abnormalcy can be just read off from one's behaviour or psychological reports, without reference to causes or theories (Bolton 2008).

The problem with the validity of the DSM categories harks back to the empiricist programme of scientific explanation and progress, in particular to Hempel's view of scientific progress. The idea of providing an operational definition of mental disorders originates with Hempel's (1965) famous paper on scientific taxonomy (Murphy 2006: 220, Kendler 1975, see Fulford et al. 2013: 6). Apparently, Hempel's influence was indirect but fundamental (Schaffner and Tabb 2015: 213). Hempel argued that acquiring a mature scientific taxonomy has two stages. In the initial stage, terms are defined by techniques of measurement and observation (e.g. "harder" mineral is defined by the operation that scratches one mineral but not another, Hempel 1961: 10-11: Parnas and Bovet 2015: 194). However, the mature stage in classification is reached when theoretical terms refer to theoretically postulated entities that are embedded in theories. This idea was picked up by psychiatrists to the extent that psychiatry was considered to be at the descriptive rather than the etiological (causal theoretical) stage in

its development. Accordingly, Hempel believed that psychiatric taxonomy would later reach a mature causal theoretical stage when its terms would denote theoretical entities.

However, Hempel's view of scientific progress is underwritten by the logical empiricist view that theoretical terms should be translated into observational language and that scientific explanation is based on exceptionless laws of nature (for the DN model, see Chapter 4). The logical empiricist programme, however, faced severe criticism, so that critics of scientific progress turned to anti-realism, whereas truth-based scientific progress came to be defended by scientific realists (see Psillos 1999: 11). Especially Thomas Kuhn (1996) and Paul Feyerabend (1965) argued, relying on the same descriptive semantic theory as the empiricists, that there is no real scientific progress because scientific terms in consecutive scientific theories are incommensurable. Scientific realism, on the other hand, has been defended by abandoning semantic descriptivism and the deductive-nomological theory of explanation. Scientific realists typically argue that theoretical terms refer to unobservable natural kinds, and more recently, that explanation, particularly in the special sciences, is based on causation and mechanistic models, not exceptionless laws.

The realist view of scientific progress has implications for psychiatric taxonomy. The construct validity of the DSM can be understood as making a commitment to logical empiricism so that both of Hempel's stages are permeated by descriptivism (Haslam 2013). This means that in the first stage, theoretical terms receive their meaning through observations, and in the latter stage they receive their meaning in connection with other theoretical terms in the natural law-based model of explanation (DN model). However, if the requirements of descriptivism and the law- based model of explanation are abandoned, there is no need to assume that the concepts of psychiatric disorders could not figure in exception-ridden explanations, i.e. in the

second stage. According to this view of psychiatric realism, there can be incremental progress that is based on causal and mechanistic explanations of psychiatric disorders understood as real scientific kinds.

In the light of this, to address the validity problem of the classification manuals, we would need to uncover the underlying causes of the disorders. Murphy (2013), for instance, argues for psychiatric realism, namely that classifications should pick out underlying natural kinds, not merely directly observable symptoms. This would mean that observational features should be validated by measuring them against their underlying causes (p. 162) (see also Haslam 2013).

I will next provide an overview of different theories of psychopathology. After, I explore the philosophical challenges of providing an account according to which all disorders share similar underlying natural features, and that besides validity, pragmatic and ethical questions may also be concerned.

2.1.3 Psychiatric Disorder in Explanatory Theories

There are numerous psychological and psychiatric explanatory theories of psychiatric disorders, most of which can be divided in the following way (adapted from Bolton 2008).

1. Folk-Psychological Models
    a. Psychodynamic Models
    b. Phenomenological approaches
2. Disease and Dysfunction models
    a. Boorse's biostatistical model
    b. Wakefield's harmful-dysfunction model
    c. Biomedical disease model
3. Normativist Models

    a. Anti-psychiatry Model
    b. Harmful Models
4. Mixed models
    a. Biopsychosocial model
    b. Network model

The theories can be distinguished based on whether they are realist about mental disorders. Folk-psychological models (1), disease models (2), and mixed models (3) propose a distinction between normality and disorder, that is, between social norms and scientific facts, and are therefore realist about mental disorders. They agree that an objective and scientific distinction between mental disorders and normalcy can be drawn, but uncovering it requires at least some research and is not merely a surface feature that can be operationalized based on folk psychology (Bolton 2008). The normativist theories instead argue either that there are no such things as mental disorders or that there may be real psychological or neurological conditions that we label, based on our values, as mental illnesses. The former approach was endorsed by anti-psychiatrists, especially Szasz (1999), according to whom mental disease is used for labelling social deviance. In contrast, normativist approaches, such as Rachel Cooper's (2005) (see next section), consider the superordinate concept of mental disorder to be value-laden, while the conditions it picks out can in principle be real conditions distinguished by their harmfulness.

Realist theories depart in their commitment to what the problem is, and thereby how social norms and medical or scientific norms can be distinguished. According to the psychodynamic model, mental disorder is a matter of breakdown of meaningful connection in mental life (Bolton 2008). Consequently, it holds that mental disorders can be detected by a disruption in meaningful connection in the patient's inner and social life, and that the disruption is either due to mental or

neurological breakdown. Crucially, rationality in intentional behaviour functions as a demarcation criterion between what is normal and what is not. In other words, the rationality requirement adheres to the intuitive way we tend to detect when someone is not thinking and behaving as expected. Karl Jaspers (1963) was an early proponent of this approach to mental problems (Bolton 2008: 9). He argues that physiologically based mental disorders are not meaningful, whereas some others are meaningful. In addition, Freud argues that "symptoms" have functions in the sense that they are normal reactions to abnormal circumstances.[12] The phenomenological approach can be similarly understood. Parnas (2008), for instance, defends the role of consciousness in psychiatric classification, so that the signs of a disorder should be contextualized in the semantic web of the inflicted individual's life (see also Schaffner and Tabb 2015: 217). A strong version of the meaningful connection view holds that only the folk-psychological approach can provide us with the correct definition of mental disorder (Graham 2009). In brief, the rational-based approaches do not only provide an account of psychiatric disorders as irrationality, but also an account of their identification.

The disease model approaches rely, instead, on brute scientific facts (Bolton 2008). The strong interpretation of the medical model holds that mental disorders are brain diseases on the neurological level, and that their explanation requires the methods of the natural sciences. The neo-Kraepelin view is an example of this approach. Boorse's (1977) biostatistical account and Wakefield's (1992a) harmful-dysfunction theory also endorse a naturalistic approach to mental disorders based on biological dysfunctions (see next section). However, unlike the strong medical model, the biostatic and harmful-dysfunction are at least partly committed to conceptual analysis providing the

[12] According to DSM-5, PTSD is a normal reaction to abnormal circumstances.

correct theory of mental disorder. There are also mixed models. The biopsychosocial model, for instance, emphasized the need for pluralistic explanations, including the need to understand illness experience (Sullivan 2017). However, the biopsychosocial model may be compatible with the medical model because it does not rule out that there could be underlying diseases that are separate from the illness experiences (see Murphy 2006: 114). Further, Borsboom (2017) has presented a network model, according to which psychiatric problems can be construed as causal networks of symptoms. The idea is that psychiatric symptoms need not be produced by an underlying disease-mechanism, but can rather be bound together by mutual interaction to produce the disorder.

In this chapter, I will mostly concentrate on dysfunction (2), disease (3), and normativist models (4) because they represent alternative philosophical interpretations of psychiatric disorders as real scientific kinds. I will also provide an overview of explanatory models that concentrate on the reasons why we have concept(s) of disorder.

2.2 Philosophical Approaches to the Concept of Psychiatric Disorder

The philosophical debate over the nature of psychiatric disorders has largely relied on analysing the concept of mental disorder and its role in determining what mental disorders are. While some philosophers argue that a general concept is not needed (e.g. Ereshefsky 2009), I primarily concentrate on the arguments according to which the concept plays a relevant role in research and clinical practice. There are two general approaches to the concept. Naturalists argue that the concept of psychiatric disorders can be given a definition based on scientific facts, whereas normativists claim that such definitions are value-laden social

constructs. Therefore, the debate over the nature of psychiatric disorders asks whether the distinction between normalcy and pathology is inalienably based on social norms or can be partly or wholly based on scientific facts. Recently, different ways to frame the debate have been offered (e.g. Broadbent 2019; Amoretti and Lalumera 2021, Wilkinson 2023). Nevertheless, I concentrate on the distinction between naturalism and normativism when presenting a framework to analyse other approaches.

I begin by introducing the approaches of the prominent naturalists, Christoffer Boorse and Jereome Wakefield, and the objections they have faced. I then reflect on Rachel Cooper's normativist approach, and point out some of its problems. Next, I criticize the debate for largely relying on conceptual analysis and for trying to find out what disorders are in their totality – an approach in which top-down analyses are made. In contrast, I introduce bottom-up approaches to psychiatric disorders, taking a more empirical approach to determining what psychiatric disorders are. Finally, based on the arguments I have presented, I propose a value-sensitive bottom-up approach to define the concept of psychiatric disorder.

2.2.1 Naturalism

The Biostatistical Theory of Disease

Christopher Boorse's (1975, 1976, 1977) biostatistical theory of disease (BST) is the most prominent naturalistic approach to defining disease. Although Boorse's view has changed since he first formulated it in the 1970s, he has consistently maintained that a disease is an internal state which reduces normal functional abilities (see Cooper 2005). The BST account holds that health and disease are value-free theoretical concepts

that can be defined based on biological facts (1997: 4). Boorse provides four criteria for disease.

1. The reference class is a natural class of organisms of uniform functional design; specifically, an age group of a sex of a species.
2. A normal function of a part or process within members of the reference class is a statistically typical contribution by it to their individual survival and reproduction.
3. Health in a member of the reference class is normal functional ability: the readiness of each internal part to perform all its normal functions on typical occasions with at least typical efficiency.
4. A disease is a type of internal state which impairs health, i.e. reduces one or more functional abilities below typical efficiency.

(Boorse 1977: 555).

According to these criteria, disease is a dysfunction in a system or subsystem of the body, such as in parts of the mind, nervous system, or inner organs (Cooper 2002: 264). But because typical functional abilities vary, Boorse (1976) utilizes the idea of a reference class to qualify statistical normalcy within species, sex, age groups, and in some cases biological ethnicity. Hence, although the ability to reproduce and survive decreases in aging, elderly people should not be considered to have a disease if those abilities are not lower than their own reference class composed of people of the same age and sex. Moreover, in the BST account, the disease of a system or subsystem is based on a causal-role account of biological function, according to which the function is what the subsystem contributes to the goal of the larger system. In other words, the function of a subsystem is defined as a non-historical contribution to the larger system's current needs, purposes, and goals (Boorse 1976). These factors, in turn, are structured based on survival

and reproduction. Therefore, a disease is a system or subsystem that hinders individuals' normal functional abilities to survive and reproduce below that of their reference class. For example, an eye that cannot dilate is not contributing to the larger system's, (e.g. the individual human's) ability to reproduce and survive (cf. Cooper 2005: 15). In such a case, according to Boorse's account, the eye should be considered to be dysfunctional and to have a disease.

Where psychiatric disorders are concerned, the BST account does not require disease to be physiological. Boorse uses a computer analogy to emphasize that just as software problem can run on perfectly functioning hardware, mental problems can be instantiated by neurologically perfect brains (Boorse 1977, Papineau 1994, Wakefield 2017, Jefferson 2018, Kingma 2013: 366). Moreover, this leads naturally to the conclusion that mental disorders can be multiply realized by our genes and brains so that one type of mental disorder can be realized by many different brain structures (Jefferson 2018).

However, the BST account has been rightly subjected to two types of criticism in particular. The first criticism is based on analysing the disease conditions that fall in or outside the extension of the concept of mental disorder in the light of the BST account. Arguably, the statistical approach by the BST account runs against our intuitions by potentially excluding conditions that may become common, for example epidemics (e.g. Covid-19), from being diseases, while including statistically rare but normal conditions, such as homosexuality, as diseases. Critics propounding the epidemic argument ask us to imagine a situation where a disease becomes so widespread that it would be statistically normal (Griffiths and Matthewson 2018: 312). For instance, if diseases of affluence, such as diabetes or obesity, continue to spread, they may become statistically typical in some parts of the world, or even globally. In such a case, the BST account would counterintuitively force us to label them as non-diseases. Boorse (2002) has retorted to the argument

by biting the bullet: statistical typicality with the relevant sex and age group requires the condition to be counted as a disease. However, his answer seems counterintuitive as a condition's status as illness would vary according to its changing prevalence. In the second case, our folk intuitions contradict, or at least are not clear about, some statistically rare conditions being disorders. The most common criticism is that increasingly people globally would not consider homosexuality to be a mental disorder, even if it were to lower one's reproduction function statistically.[13]

One could try to address these objections by refining reference classes so that the subgroups that clearly are not diseased would form their own reference classes, while maintaining that some other groups where diseases have become prevalent would not constitute their own classes. As an example, homosexuals could be understood to form their own reference class (Cooper 2005; Kingma 2013). Since homosexuals would not be statistically unreproductive within their own reference group, they would not be considered to have a disease or a psychiatric disorder. Similarly, the idea of a reference class could be employed to try to distinguish accidental functions from genuine ones (Cooper 2002). As an example, although sunscreen protects one from sunburn, just like dark pigmentation, only the latter has the genuine function within the reference class made of people with dark skin. This is because protection from sunburn is *typical* within their reference class, while using sunscreen in the human population in general is merely accidental. On the other hand, since obesity and diabetes have become predominant in some populations, these groups of people should not be understood to form their own reference group, so these conditions would retain their disease status.

[13] Some studies seem to suggest homosexuality may carry reproductive evolutionary benefits for their population (e.g. Pillard and Bailey 1998)

However, this approach of refining reference classes seems ad hoc and raises a question about the BST account's ability to determine reference classes value neutrally. One problem with making reference classes more fine-grained is that this procedure becomes subject to slippery slope counter-arguments. Cooper (2002) argues that because there is always bound to be individual variation, partly due to the fact that individuals may belong to many minorities, we could eventually end up with one-person reference classes. Moreover, some groups of people that would be held as forming a reference group of their own on account of some rare but normal functions, could be held as not forming a reference group on account of some endemic diseases that are not prevalent among the general population. The only way to rule this out seems to be by making normative judgements as to which possible groupings are reference classes in their own right, and which are malfunctioning subgroups within those reference classes (cf. Cooper 2005, cf. Kingma 2013). In addition, there is an underlying epistemic problem inherent within the referent-class idea. In practice, it may be impossible to calculate whether a given individual falls within or outside a reference class (Cooper 2005). For example, it seems impossible to calculate the reproduction and survival statistics of sexual minorities.

The problem with reference classes is due to the causal-role account of function. It does not seem to offer a natural means to demarcate functional categories, which in turn undermines the BST account's ability to provide a value-neutral account of disease. Apparently, Boorse's primary argument over his choice of reference classes is based on "differences in normal physiology between males and females, young and old" (Boorse 1997: 8). However, as Griffiths and Matthewson (2018: 313) point out, the reliance on "normal physiology" to demarcate reference groups is unhelpfully circular, because it does not offer an independent way to decide between

alternative, but seemingly equally natural ways to construct reference classes.

The causal-role account of the BST view is also subject to a more principled objection. Arguably, because membership in the causal-role account is determined by functional ability, it cannot account for why organs that cannot perform their defining functions should be lumped together with organs that can function as defined (Millikan 1989: 295; Amundson and Lauder 1994). For example, if hearts are defined based on their function to pump blood, a putative heart that lacks this function should not be categorized together with ones that are able to perform the function. Such a scenario is evident in congenital disease when a heart does not have the necessary parts to pump blood. Hence, the causal-role account of function would counterintuitively and against medical practice exclude the organ from being a heart.

In sum, as these objections demonstrate, the statistical method of deciding disease in the BST account may force us to be revisionary in cases where we would prefer not to be. As mentioned, the BST account could try to address this problem by refining reference classes so that some cases would become *typical* or *normal* within their own reference classes, while abstaining from doing so with minority groups where a disease has become predominant. However, as has been argued, this runs into difficulties with the account's purported value neutrality of demarcating reference classes. I will next argue that whether something is considered typical or untypical would also need to be culturally qualified.

The second problem with the BST account is its strict distinction between culture and biology. The distinction is problematic because human behaviour is not merely determined by biological needs, it is also regulated by culture-specific norms and is supported by individual choices. More specifically, the ability of an individual to survive and reproduce is regulated by a combination of biological, individual, and

sociocultural factors (Pörn 1995). This means that cultural norms may hinder reproduction or survival for the otherwise most biologically capable individual, while enabling biologically less capable individuals to survive or reproduce, creating a situation where social norms trump biological functions. This can be exemplified with the relatively benign skin disease pinta, which is an infectious tropical disease causing skin lesions and discoloration. According to Honko (1960), pinta was until the 1960s so common among some tribes in the Amazon basin that these tribes had formed cultural norms that prohibited marriage to individuals that lacked the disease. This means that in these cultural contexts having the disease would have been a reproductive advantage. Boorse might try to answer this objection by arguing that the BST account is not based on how humans actually behave, but rather on how they could have behaved had their sociocultural circumstances been different (i.e. in counterfactual circumstances). The underlying idea would be that although cultural and individual variation plays a part in explaining human behaviour, unlike biology, cultural and individual factors do not restrict and determine behavior to the same extent. However, this argument faces the problem that human evolution is intertwined with cultural evolution in feedback loops (see Bolton 2013). That is, human evolution is in fact coevolution with cultural factors, so that culturally upheld habits have caused evolutionary changes. This is shown, for example, in the way that the invention of cooked food led, through natural selection, to smaller jaws in our evolutionary development.[14] Moreover, in human evolution, the predominance of some diseases may be due to cultural selection. Hence, for instance, some diseases similar to pinta may have become widespread because individuals without that genetic predisposition have not had as many

[14] This means that our ancestors' jaws, which were adapted for eating raw meat, would nowadays lower one's ability to survive and have offspring, On cultural scaffolding, see Sterelny (2010).

offspring. (I am not arguing that pinta is in fact such a disease, only that there may be such diseases.) Therefore, according to Boorse's answer to epidemics, conditions such as pinta should not be considered diseases, at least if those people form their own reference class. As mentioned, this would also go against our intuition that the same conditions should not be considered a disease at one point and a non-disease at another only due to its statistical standing.

In conclusion, the biostatistical account does not seem to have the resources to demarcate between functional categories, and hence between healthy and unhealthy organs or minds. Alternative accounts of disease have addressed this problem either by arguing for (i) a more robust evolutionary account of biological dysfunction, or (ii) by admitting that an explicit normative judgement does the logical work of determining disease. Next, I will look at Wakefield's hybrid account, which combines (i) and (ii), and thereafter normativist theories that take the latter approach (ii).

The Harmful Dysfunction Account of Disorder

Jerome Wakefield's (1992a, 1992b) hybrid harmful dysfunction account grants a role for a normative element in determining mental disorder and employs a selected effect theory to determine dysfunction. According to his account, mental disorder consists of two individually necessary and together sufficient components.

(i) Social value judgement that the condition is harmful.
(ii) Objective identification of a malfunctioning internal mechanism

(Wakefield 1992a: 374-375.)

The idea is that inner psychological mechanisms (i.e. mental modules) have been selected by evolution to execute various functions. This means that dysfunctions should be understood as psychological or cognitive mechanisms that are not performing their naturally selected functions. Moreover, although dysfunctions cause harms, not all the harms they cause qualify as mental disorders. Wakefield (1992a: 384) provides two reasons for this. First, selective pressures may have changed so that a malfunctioning mechanism does not produce the harmful consequences it initially produced. Second, some consequences of breakdowns in internal mechanisms may not have clinically significant consequences in the current social environment. According to Wakefield, his account implies that psychiatry should first work out the dysfunctional mechanism that is causing the problem, and only thereafter a judgement based on social values can be made whether the dysfunction is harmful. (Based on this, Murphy formulated his own two-stage picture of psychiatry.) What is crucial for my analysis is that Wakefield argues that the dysfunction condition can be naturalized and separated from the norm component. This means that the harmful dysfunction account, like the BST account, can be understood as providing a naturalistic and realistic account of psychiatric disorders.

The harmful dysfunction account relies on the selected effect account of biological function (i.e. the etiological account), which is based on the evolutionary history of natural selection. According to the selected effect account, an organism's function is not determined by the present goal it serves, but by whatever the organism was selected to perform in its evolutionary history. In other words, natural selection explains why certain parts and processes have had the function of contributing to survival and reproduction in the species' evolutionary history (Griffiths and Matthewson 2018: 304).

At first sight, the harmful dysfunction theory's advantage over the BST account is its robust theory of dysfunction and the necessary role

it grants for harm. The evolutionary function analysis comes with the advantage that it seems to facilitate a more robust distinction between genuine functions and accidental or culture-adapted ones than the causal-role account of function. For example, we saw that the BST account had problems in distinguishing between accidental and genuine functions. This does not seem to be a problem for the evolutionary function analysis. For example, although sunscreen serves to protect us from sunburn, unlike dark pigmentation it does not have the genuine function of doing so because it was not naturally selected to do so (cf. Griffiths and Matthewson 2018: 307). Moreover, because of the dual requirement of value judgement and biological function, Wakefield's account is more socially sensitive than the BST thesis. His account can explain why some (putative) dysfunctional conditions are not considered disorders since they are not socially harmful, while some patterns of behaviour or conditions are not labelled as disorders since there is no underlying dysfunction. For instance, if the social withdrawal disorder found in the DSM (See APA 2013) really were caused by an internal malfunctioning mechanism, it would not count as a disorder according to the harmful dysfunction account in cultures where social withdrawing is not considered harmful (presumably in East Asian cultures, such as Japan). On the other hand, even though Confederate slave owners considered the escape tendency of slaves to be socially harmful, it was not a psychiatric disease because the underlying reason for escaping was not a dysfunctional subsystem but slavery itself.[15]

However, the harmful dysfunction account has also been rightly criticized for conflicting with our intuitions due to its account of dysfunction and its strict biology-culture dichotomy. The central problem with the etiological account of dysfunction is that evolution is a

[15] It seems natural for any slave to attempt escaping, but perhaps the term was evoked to explain the difference between individuals who escaped and those who did not.

complex and changing process. Rachel Cooper (2005), for instance, has argued that evolution does not offer the means to fix the function of an organism because it may have originated from a different selection pressure than that which has determined its current use. For example, apparently wings initially evolved as heating regulators in insects before they started to serve as a means for flying (Cooper 2002: 268). In such a situation, Cooper (2005) asks us to consider which selection pressure is the one we should consult to determine the function and dysfunction of an organism. That is, the harmful dysfunction account seems to leave us in the dark as to why we should not consider the original selection pressure as more important, so that an insect incapable of flying should not be considered as having a disease.[16]

Perhaps even more fundamental are the arguments according to which a mental disorder may be due to a condition that has never had any evolutionary function or no longer has one (see Kingma 2013). Dominic Murphy and Robert Woolfolk (2000), for instance, point out that mental problems could be understood as spandrels or vestigial traits. Spandrels are the side effects of selected traits, whereas vestiges are traits that are not selected anymore. As an example of the former, apparently sickle cell illness became widespread in sub-Saharan Africa as a side product of a mutated haemoglobin structure that has the benefit of creating immunity against malaria (Hales 2006: 84). The reason for the co-evolution of sickle cell illness and immunity to malaria is that natural selection works in genetic bundles. An example of the latter, the appendix and wisdom teeth no longer have selected effect functions. Finally, having some mental disorders may have carried evolutionary advantages, and thereby may have been naturally selected. Kingma (2013), for instance, argues that psychopaths may carry evolutionary

[16] Recently, Griffiths and Matthewson (2018: 304) have defended the harmful dysfunction account by arguing that we should concentrate on recent evolutionary history.

advantages for the individual. This can be supplemented by arguing that in evolution, individual variation is the norm because it provides the best means for the population to survive by enabling adaption to changing environmental circumstances. In sum, if mental mechanisms are akin to spandrels or vestiges, or are selected, they may cause problems and yet not work against their natural functions. The underlying rationale in these cases is that the harmful dysfunction theory, just like the BST account, needs to rely on interests or values to support decisions over dysfunctions. In the case of the harmful dysfunction theory, this means deciding which selection process, evolutionary side effects, or selected effects are the ones that do the logical work of distinguishing between normal and abnormal groups (cf. Cooper 2005: 16).

The second problem with the harmful dysfunction account, as with the BST view, is its implausibly strict distinction between culture and biology that does not correspond with scientific discoveries (Bolton 2013). According to the harmful dysfunction account, mental disorders are dysfunctions in the mental modules that have developed to perform specific evolutionarily beneficial tasks. Murphy and Woolfork (2000) point out that this is an unlikely hypothesis concerning the way in which the structure of the human mind and brain will turn out because the latest findings do not support strong modularity. The co-evolution of cognition and culture is a case in point. Although Wakefield acknowledges that the mental modules have been formed in interaction with culture, he nonetheless maintains that malfunctioning mental mechanisms or modules are grounded in natural selection and can be cleanly separated from socialization processes. This is problematic because cognitive science and evolutionary psychology suggest that complexity interaction has taken place between culture and biology in the evolutionary development of the mind/brain. More specifically, in some

cases, the content of our minds may be determined by universal features, whereas the structure may be shaped by culture (see Nisbett and Norenzayan 2002). At any rate, the issue has not been settled and there is an ongoing debate over the cultural influence of cognition (see Chapter 6). The same situation is apparent in gene-behaviour interactions. Bolton (2013: 442) points out that some evolved functions are social (such as mating and child-rearing), while psychological and behavioural phenotypes are typically the product of interaction between gene-environment interactions.

In conclusion, both the harmful dysfunction and the BST accounts clash with our intuitions of what the concept of mental disorders should pick out. Some conditions that we classify nowadays as disorders would not count as such according to the accounts, while other conditions or behavioural traits that we would prefer not to label as disorders would be classified so according to the accounts.[17] Moreover, both the BST and the harmful dysfunction account contradict scientific studies on culture-biology interaction. The harmful dysfunction account, in addition, does not seem to correspond with recent scientific discoveries due to it placing greater weight on the nature of dysfunction. The problems of both accounts are manifest especially when we consider borderline cases where we need to refine and decide what the concept should pick out. In fact, Schwartz (2007b) argues that both accounts have a "line-drawing" problem because they have trouble in distinguishing low-normal functions from dysfunctions. The problem is that the normal range of biological functions fall on a scale in which abnormality is only at the extremes of that scale. Hence, neither Boorse's nor Wakefield's account seem to able to provide a natural means to decide where the line should be drawn.

[17] However, as to what extent this a problem for the HD account, is debatable. Griffiths and Matthewson (2018), for instance, endorse this revisionist outcome.

However, in spite of these problems, both the BST view and the harmful dysfunction account are compatible with our intuition that mental disorders are not merely determined by value judgements. Therefore, some philosophers conscious of the mentioned problems, but nonetheless with realist inclinations, have argued that a weaker form of naturalism can be saved (see Kingma 2014, Murphy 2006, Rashed and Bingham 2014). For instance, Kingma (2013) argues that although concepts of health and disorder are not value-free, this does not lead to wholesale normativism. Rather, it is indicative of naturalistic pluralism. In this she relies on Boorse's account, and argues that although the choice between alternative reference classes is value-laden, these alternative classes can nonetheless be distinguished in a value-free fashion.

Nevertheless, the harmful dysfunction account has been more influential, in part due to the similar definition found in the DSM and the fact that it grants an explicit role to values. For instance, types of hybrid accounts have been offered by Cooper (2005) and Murphy (2006). The underlying idea in both is that although the superordinate concept of mental disorder may be irreversibly value-laden, some of the conditions that fall within its extension can in principle be scientifically relevant natural conditions. This approach resembles Wakefield's hybrid account, except that it strips away the evolutionary dysfunction account from the natural component. Cooper and Murphy exemplify their accounts by comparing the concept of psychiatric disorder with the concept of "weed". Although it is interest- or value-laden which plants are classified as weeds, they are nonetheless natural kinds of plant species. Similarly, while the concept of psychiatric disorder may not be value-free, the world still consists of interest- and value-free conditions, which, according to our interest and values, we have labelled psychiatric disorders (see also Beebee and Sabbarto-Leary 2010b). However, Cooper and Murphy employ the weed metaphor to

describe opposite views. While Cooper argues for a normative view of mental disorder, Murphy argues for a full-blown brain-based account. More precisely, whereas Cooper asserts that there are genuine psychological or neurological joints or conditions that we label psychopathological, Murphy argues that psychiatric disorders represent different ways our brains can malfunction or deviate from typical rationality. In general, I agree with such hybrid approaches, although I believe that value-laden and natural factors are intertwined due to the social nature of psychiatric disorders. Next, I will explore alternative approaches to the concept and subsequently present my own account in relation to them.

2.2.2 Normativism

Normative or constructivist approaches to the concept of mental disorder hold that evaluative considerations play an intrinsic role in defining psychiatric disorder. The means that the concept of disease or psychiatric disorder is inalienably value-laden so that purported biological or psychological facts about abnormality, such as facts about dysfunctions, are not the determinative factors for a condition to count as a psychiatric disorder.[18] In contrast, judgements over harmfulness, suffering, and disvalue determine when something is a psychiatric disorder.

Normative approaches to mental disorder should be distinguished from anti-psychiatric views, such as Thomas Szasz's (1999) account, according to which psychiatric disorders are mere social deviances and myths. Unlike Szasz, normativists usually argue that only conditions that are medically treatable, at least in theory, count as diseases. On the

[18] See e.g. Cooper (2002, 2005), Fulford (1989), Goosens (1980), Nordenfelt (1986, 1995, 1997), Sedgwick (1973).

other hand, normativists disagree with naturalists by denying that psychiatric disorders can be analysed based on biological or psychological facts alone (or primarily) because they involve the whole wellbeing of human beings as well as judgements over undesirability. Unlike naturalist approaches, normative approaches position human goals, such as healing and wellbeing, as taking precedence in the definitions (Salmela 2004). Although normativist accounts make different commitments, the main point is that disease or psychiatric disorders can only be understood in relation to the holistic wellbeing of humans.

However, since there is no space to analyse all normativist approaches, I will concentrate on Cooper's account. Cooper (2002: 271; 2005) defends a normative account of diseases in general, and psychiatric disorder in particular, according to which the concept is anthropocentric and depends on our interests. Her view is contemporary and well argued for, and thereby suffices as a representative of constructivism in general. Cooper relies on descriptive conceptual analysis to uncover the distinctions that are implicit in how the concept is used. Therefore, unlike Boorse and Wakefield, her primary intention is not to uncover what psychiatric disorders are based on conceptual analysis, but rather to uncover the intuitions we purportedly share about the concept. While Cooper (2020) argues that the concept of mental disorder has changed over time, she maintains that there is considerable consensus mainly due to the predominant role of the DSM classification manual and its definition.

According to Cooper, the use of the concept "mental disorder" consists of three criteria that are individually necessary and together sufficient for a condition to qualify as a disease. These criteria are (i) it is a bad thing to have; (ii) the afflicted person is unlucky; and (iii) the condition can potentially be medically treatable. According to condition (i), the inflicted individual needs to feel harmed by the condition. In

practice, this means that the same condition can be harmful for one individual and benign for another (Cooper 2002: 274). Based on this, Cooper argues that concept of "psychiatric (or mental) disorder" resembles the concept "weed". Just as a plant can be unwanted and considered a weed by some gardeners while not by others, a condition's status as a mental disorder is interest- and value-dependent. In practice, this renders the label "mental disorder" a subjective judgement as "people have different aims, different abilities, and different preferences" (Cooper 2002: 274). The second condition (ii) states that the condition is "unlucky as judged by the uninformed layman" (p. 276). This condition arguably avoids the problems inherent in statistical infrequency since some individuals in the majority could feel unlucky, while others in the minority could feel lucky. The last condition (iii) is naturalistic, in which case the conditions that medicine find treatable are mental disorders. This condition is meant to differentiate disease from other unlucky and bad things in life, such as poverty and social problems. Whether something is potentially medically treatable is a question of functional goals, not about the nature of the objects themselves. As the function of medicine is to treat diseases, those conditions that are potentially treatable should be considered diseases. In other words, the objects are not united by their common similarity of being diseases, but by medicine deeming them to be treatable despite their different natures.

Cooper's account of psychiatric disorder seems to lead to relativism (Ereshefsky 2009: 224, Stegenga 2018: 30) The reason is that Cooper hangs the ontological status of psychiatric disorder on the folk-psychological concept(s) of psychiatric disorder. This seems to put, as Ereshefsky (2009) argues, medical conditions on a par with drapetomania. As an example, had there been medical means to intervene on homosexuality, the account seems to lead us to believe that homosexuality would not only have been considered a psychiatric

disorder previously, but would have actually been a disorder. The problem is, however, that although folk and expert intuitions may change about the status of particular conditions, the general intuition is that it is a matter of fact whether a specific condition is a mental disorder. Moreover, an increasingly shared intuition is that experts and lay people were previously wrong to hold homosexuality as a mental disorder. Consequently, the relativistic conclusion stands against our shared intuition that the concept of psychiatric disorder is not relative. This is a problem for Cooper, since according to her account, our intuitions underlie the correct application of the concept of psychiatric disorder.

Another related challenge for Cooper's account is that it does not seem to be able to settle existing individual or group disagreements over whether a condition is a disorder. The account seems to portray disease as a thin concept to the extent that although there would be relatively stable necessary and sufficient conditions for its employment, they cannot settle disagreements over the concept's specific application. The reason is that the inflicted individuals themselves, according to Cooper's account, are the experts over the harmfulness and unluckiness of their condition. This leads to the unintuitive outcome that the same individuals may hold their condition to be a disorder at one point while not at another time or context. Cooper could try to restrain relativism by maintaining that there is an objective matter of fact as to how conditions are experienced by different individuals at various times and places. But this would be tantamount to merely stretching the point further because we could ask whether the *same experience* can be held harmful and unlucky at one point and benign at another. A similar problem is manifest in group disagreements. For example, although pinta was not considered unlucky by the tribesmen in the Amazon basin, it was a disease according to medical knowledge. Therefore, it is reasonable to presume that since the 1960s many tribesmen have

reassessed their position based on the availability of medical knowledge and the readily available penicillin treatment. In the light of this, Cooper's account would lead to the unintuitive conclusion that for some people pinta was a non-disease at one point, and became a disease later. But surely a more plausible explanation is that the tribesmen were initially wrong, while correctly acknowledging later, in the light of new evidence, that pinta is a disease. The unintuitive outcome of relativism becomes even more evident when considering cases where the inflicted individual's rational faculties over the harmfulness and unluckiness over his or her condition are influenced by the condition itself, as is the case with severe delusions and addictions.

Cooper could try to counter relativism by falling back on the condition that something must be medically treatable to qualify as a disease. This would not only exclude cases like poverty or oppression as psychiatric disorders, but also drapetomania and homosexuality, since it was later determined that they are not medically treatable. Indeed, ideally it would restrict relativism and disagreements only to borderline cases where intuitions are not clear. However, as Stegenga (2018: 32) argues, many conditions are medically treatable and yet we do not want to label them as diseases. The problem is, as Stegenga points out, that something is medically treatable because it is a disorder, not that something is a disorder because it is medically treatable. As an example, we would like to determine whether we to administer drugs to people diagnosed with ADHD, autism, and social phobias. Moreover, medicine administers enhancement treatments, such as plastic surgery (Schwartz 2014). The problem for constructivism is that arguing that medicine decides whether something is a disease requires us to answer how medicine makes those decisions. In other words, we should then determine whether the research and treatment goals of psychiatry are naturalistic (treating real psychiatric disorders) or value-laden (treating unwanted conditions). The former would mean that

normativism collapses into naturalism, whereas the latter would be tantamount to turning a blind eye to the problem of relativism. In sum, the fact that a condition is in principle treatable by psychiatry does not mean that it should be treated, and thereby that it is a psychiatric disorder.

A novel approach to addressing the relativism of harm-based accounts has been to explore whether harmfulness can be understood objectively. Rashed and Bingham (2014), for example, point out that since dysfunction cannot be used as a demarcation criterion, the distinction between genuine mental illnesses and social deviance could instead be approached by analysing harmfulness (see also Aftab and Rashed 2021). Through their examination they came to the conclusion that harm or distress caused by a genuine psychiatric disorder was intrinsic rather than relational. However, as they willingly admit, it is not easy to determine what such a distinction amounts to. While I doubt that this approach can be applied to all disorders, I nonetheless believe that in some cases we could locate objective harmfulness in, for instance, social interactions. For instance, Schilbach (2016) argues that impairments in social interactions may not only be symptoms, they may also contribute to the development and constitution of psychiatric disorders.

Moreover, recently Cooper (2020) has accommodated conceptual change into her account and shifted toward a more naturalistic account. She argues for a "belt and braces" approach on account of the fact that the consensus over the concept of mental disorder is shifting and therefore her original conceptual account requires adjusting. In particular, she points out that the DSM-5 has given up the notion that a disorder has to be harmful. Cooper succinctly argues that there are three possible ways for conceptual analysis to answer conceptual change: (i) to give up on conceptual analysis, (ii) to be revisionary, (iii) or to embrace her "belt and braces" approach. She points out that these

should not in fact be considered separate approaches, but instead they complement each other. However, the "belt and braces" approach is meant to capture the multiple points that would need to change for the concept of mental disorder to become value-free. Cooper maintains, based on two observations, that it is not foreseeable that the concept would be revised to the extent that value judgements would be abandoned altogether. First, she points out that political and ethical norms are currently employed to set thresholds between the normal and the pathological. That is, statistical means alone cannot be used, and are not used, to determine the boundaries of disorders. As one example, ADHD arises at the extreme end of normal distribution (Cooper 2020: 154). Second, values dictate decisions over whether to ameliorate one's condition by altering the individual or the environment. In the case of homosexuality, intolerant societies would like such individuals to alter their behaviour, whereas societies are increasingly tolerant in considering that the suffering of the individuals is due to social intolerance.

2.2.3 Problems with Top-Down Approaches

Although I am sympathetic towards the mentioned weaker forms of normativism, there is a question over what role conceptual analysis can have in determining the nature of psychiatric disorder. As Murphy (2008) argues, the debate over normativism and constructivism is problematic because it relies extensively on conceptual analysis. This can be seen in how the debate is based on employing conditions to demonstrate that they intuitively or unintuitively either fall in or outside the extension of the proposed definition (Lemoine 2013). Murphy maintains that conceptual analysis does not only rely on there being a shared folk-psychological concept to be had, but also maintain

implicitly that it ought to constrain research by having the ultimate say on the identification of psychiatric disorders.[19]

Conceptual analysis has been criticized for relying on armchair intuitions to uncover the meanings of concepts (Hintikka 1999, Stich 1992, Schwartz 2014, Griffiths 1997). In psychiatry, empirical studies have demonstrated that intuitions over the concept of psychiatric disorder and particular disorders diverge on multiple dimensions (Aftab 2021). And even if traditional conceptual analysis can settle some inconsistences between alternative definitions, it cannot provide reasons to decide between two coherent and consistent accounts of psychiatric concepts (Lemoine 2013). These problems have led to an impasse and controversies over how to define the concept of psychiatric disorder, whether particular conditions are disorders, and in the case of some purported conditions, where to draw the line between normal and pathological (Schwartz 2014: 574).

In this light, the accounts of psychiatric disorder that I have covered this far can be called top-down approaches in contrast to bottom-up approaches. The reason is that they largely take for granted that we can discover what disorders are by analysing how the concept of psychiatric disorder is currently applied. In other words, the presupposition is that we have already identified psychiatric disorders, and now only need to define them. The method largely relies on analysing stereotypical cases of psychiatric disorder, and thereby try and reach a general definition of psychiatric disorder. Once such a definition is reached, it is in turn employed to distinguish between borderline cases and evaluate empirical research. For this reason, such research emphasizes an understanding of what psychiatric disorders are in their totality. An alternative, bottom-up approach would be to concentrate on the

[19] Hyman (2010) has labelled the extant diagnostic categories as "epistemic prisons" that impede research. Currently, the problem is that over 90 percent of research is based on the DSM categories (Tabb 2019).

empirical research of specific disorders, or even smaller units of analysis, and then move on either to refine the general concept based on those discoveries, or alternatively, delegate it an explicitly pragmatic role in psychiatry. This approach would not only mean refining the intension of the concept of psychiatric disorder, but also its extension. The distinction between these approaches is meant as an analytic tool. In reality, the difference is more in the emphasis because psychiatry is largely in a "pre-paradigmatic" stage and therefore needs to "get on" without clear foundations.

Both naturalist and normativist approaches can be understood to be committed to a top-down approach to defining psychiatric disorders. While the naturalistic accounts of Boorse and Wakefield largely rely on conceptual analysis to pinpoint the putative underlying natural phenomenon of our folk concept of psychiatric disorder, normativists seem to employ conceptual analysis to argue that there is no such natural phenomenon.[20] Therefore, although their employment of conceptual analysis is more justified because it concentrates on the nature of the concept, and the kind of role it plays, instead of the putative underlying disorders, constructivists nonetheless seem to rely on armchair conceptual analysis to make assumptions about the ontology of psychopathology (or rather its putative biological or psychological non-existence). Moreover, even this more modest employment of conceptual analysis is problematic for at least three reasons. First, naturalistic conceptual analyses suggest that lay and professional intuitions over the concept vary (Haslam et al. 2007, Haslam 2013). Indeed, according to studies, people hold contrastive explanatory models that are relative to their profession and culture (Colombo et al. 2003, Harland et al. 2009, Tikkinen et al. 2019). This means that analysing, for instance, the DSM definition may not capture

[20] Although Boorse's account is typically seen as a naturalist account, it can alternatively also be understood as an account that explicates how we employ the term.

how the term is used in general. Second, it is not clear to what extent we desire lay intuitions, or even professional presumptions, to play a significant role in constraining and guiding psychiatric research and informing clinical judgements about discoveries. Third, although judgements over harm and unluckiness may be value-laden and relativistic, some judgements may be more objective and morally warranted than others. But to determine which judgements are warranted, we would need to analyse their reasons, not merely our own intuitions. In sum, the conclusions that can be drawn from normativist accounts, such as Cooper's view, about psychiatric disorders is restricted. In sum, armchair conceptual analysis may not be the correct method to discover putatively shared intuitions about psychiatric disorder, to reveal assumptions that should constrain research, or to determine whether they are warranted.

A normativist could try to address these problems by discarding armchair conceptual analysis in favour of adapting a form of realism about values. For example, Wrigley (2007) has pointed out that a realist account of mental disorder needs to rely on moral realism (see also Stegenga 2018). Assuming value realism, adapting relativism about disorder would be to commit oneself to a type of naturalistic fallacy. That is, one would not be entitled to infer from the fact that there are different and contradictory judgements over psychiatric disorders so that they are all as warranted. Instead, there could be warranted ways to use the concept so that some people would have values that they are morally unentitled to. This account could be coupled with philosophical analysis over warranted reasons for those values. For example, the fact that the conception of homosexuality as a disorder was common among specialists and lay people does not mean that their assessment was justified, as shown by later analyses. This historical shift could be backed up, for instance, by arguing that we have later discovered ethical principles that exclude homosexuality from being considered a

disorder. That is, we could claim to be in a better position to recognize that the view that homosexuality is a mental disorder was all the while unwarranted. In sum, according to this approach, values concerning mental disorders would not be considered to depend on subjective or social values but would instead be universal. However, to distinguish psychiatric disorders from social problems, such as poverty and discrimination, there would still have to be an underlying condition. Hence, if this approach is applied to drapetomania, one could argue that we have discovered that it does not have an underlying condition, and therefore was not a disorder. On the other hand, even if there had been an underlying condition that produced the "the desire to be free", the judgement that it was a disorder would not have been warranted. Consequently, this type of value objective view would be Wakefield's hybrid theory turned upside down, so that value judgement over harmfulness would be the objective determining feature, and although a psychological or biological condition would be necessary for a disorder, it would not do the logical work of distinguishing abnormality from normality. That said, I will not pursue this argument any further here, but instead argue later that we should engineer the concept by being value-sensitive to its consequences (see Chapter 7).

Naturalists, in turn, could try to address the objections I have presented by supplementing conceptual analysis with externalist semantics, such as the causal-historical theory of reference. According to this theory, the beliefs and descriptions we associate with our theoretical concepts and terms do not necessarily determine their reference (Kripke 1980, Putnam 1995). Rather, their reference can be set by ostension or definite descriptions when the object or kind is baptised, and from there it is passed on to other users in a causal-historical chain. Consequently, individuals can refer to the same objects or kinds although they may associate conflicting and false beliefs with them. If this approach is applied to psychiatric disorders, it would

require that we have initially identified psychiatric disorders correctly, although the varying folk-psychological conceptions we have associated with them may have been contradictory and incorrect. On the face of it, such a view seems to require that psychiatric disorders are natural kinds that have underlying essences or causal mechanisms that are the same in all possible worlds, and which generate sufficiently stable and distinct surface properties.[21] The central motivation would be empirical: although we may not be sure what the properties of psychiatric disorders are, we can talk about them and conduct commensurable research related to them, and thereby trust to find their true nature in the future. Murphy (2014: 199) calls this view the "vindication project", because it relies on our current concepts or terms of disorder to be vindicated by future discoveries over their referents.[22]

However, a simple causal-historical reference view of psychiatric disorders is implausible in the light of historical alterations in the conditions that have been considered mental disorders. That is, not only our conceptions of what constitute psychiatric disorders have varied historically and cross-culturally, but also the conditions, their outcomes, and their representations. This implies that the concept of psychiatric disorder does not resemble such traditional natural kind terms as "water" or "gold". Part of the reason is that psychiatry is not only trying the uncover the underlying structure of an already fixed (under controlled circumstances) stable surface property, as with water and gold, but is instead trying to discover whether the conditions we have labelled pathological are sufficiently stable and distinct, while simultaneously trying to determine whether they should be labelled pathological. In practice, this has been manifested by the challenge to

[21] Wakefield's (1999) "black-box essentialism" is an example of this. See section 6.2.5

[22] Another alternative would be to couple a dysfunction theory with the meaningfulness-based view of psychiatric disorders, so that mental disorders are the conditions that patients bring to the clinic.

pick out homogenous groups of patients that share a given psychiatric disorder.[23] As Aucouturier and Demazeux (2013) argue, the difference is that unlike prototypical natural kinds, such as water and gold, there is no agreement in psychiatry over the nature of the research object, or even whether it is a unified kind.[24] In consequence, new discoveries may not only require changes in the intensions of our concepts, but also in their extensions.

Moreover, if normativists are right, even if there is a unified object of research in psychiatry, there is the further question whether it ought to be labelled a disorder. That is, even if an externalist semantic theory of the term "psychiatric disorder" or particular terms for disorders were correct, they are not by themselves enough to support naturalism. The reason is that social constructivist accounts could also rely on externalist semantics to support their case. This can be seen in the light of how it has been suggested that some social kind concepts, such as race and gender, may be real kinds. Mallon (2016) argues that the term "gender" or "sex" may refer to a (relatively) stable cluster of properties, because most of the clustering is generated by the very actions caused by the stereotypical beliefs we associate with genders, rather than their purported underlying biological mechanisms or essences. That is, by being considered "the natural way to be and act", socially determined patterns of behaviour have become entrenched in institutional, material, and social structures, and socially determined patterns of behaviour associated with genders have become stable and are thereby considered natural. Similarly, constructivists could argue that although (some)

[23] Lakoff (2005) writes about a clinic in Buenos Aires that provided blood samples to a multinational pharmaceutical company who were researching bipolar disorder, a condition that their psychodynamic approach did not previously acknowledge.

[24] When Kripke and Putnam formulated the causal-historical theory of reference, they were careful to choose folk terms that have stable and unified meanings shared by everyone.

psychiatric categories pick out stable patterns of behaviour, they are just not the type of kinds that naturalist theories would have us believe. In sum, whether a condition supports inductions by being a real kind should be separated from the reason it supports those inductions. Although some of the conditions we currently label psychiatric disorders may not support inductions for the reasons we believe, they can nonetheless be real in the sense that they actually do support those inductions.

In addition to these challenges, the traditional naturalist and normativist accounts have a problem with scientific advancement. In general, the problem with settling on a strict concept of mental disorder is that at best it may not have much pragmatic value, and at worst, it could constrain research and clinical practice. The reason is that science progresses through self-improvement, which includes its ability to adjust our scientific concepts with the world (Kincaid 2013: 148). This means that if we discover that our current psychiatric concepts do not have the referents we thought that they had, we may either adjust our beliefs and theories associated with the concepts, adjusting also their extension (the objects that fall under the concept), or we can abandon the concepts. In psychiatry, this does not merely mean adjusting our extant concepts with the world. Instead, a further complication is that because the concepts may be value-laden, they cannot be merely fitted with the causal structure of world. Rather, we also need to decide whether the referents of our psychiatric concepts are those we would like them to be. That is, while the world may offer limited objects for psychiatric labelling, if normativists are to be believed, it is partly a normative choice how we choose between those objects.

My idea is that the same problem that has plagued attempts to provide necessary and sufficient criteria for demarcating science from pseudoscience also concerns the demarcation between psychiatric

disorders and normality.[25] A strict definition of a psychiatric disorder is bound to be either too narrow or wide to capture the conditions that psychiatry studies and will study. That is, the criteria cannot be simultaneously sufficiently universal and precise to encompass all the desired conditions while excluding those we do not want. In fact, we may lack sufficient knowledge about our brains and minds to settle currently even on a more open-ended definition of psychiatric disorder, even if such conditions were in theory attainable and warranted. Rather, it seems plausible that once the sciences of the mind evolve, we will also become better at making judgements about psychiatric disorders. Such an approach is taken by Murphy (2006) and Zachar (2014a), who argue that the concept of psychiatric disorder needs to be open-ended to be able to accommodate empirical discoveries. A further problem is that advances in pharmacology, genetic engineering, and robotics are likely to provide ways to enhance or augment our biological and psychological makeup to such a degree that the current conceptions of health and disease will not meet future challenges. Similarly, scientific, cultural, and technological evolution may change our social and physical environment to the extent that we will develop novel disorders and the old ones will disappear.

To summarize, I have argued against a strict definition of psychiatric disorder. This seems to pose a problem for psychiatry as a legitimate science since the object of research is not a unified natural kind. For this reason, Scott Lilienfeld and Lori Marino (1995), for example, maintain that the concept should be abandoned. However, contrary to Lilienfeld and Marino's assumption, scientific concepts do not need have strict boundaries to be useful. One would in addition need to argue that the concept is trivial if it picks out merely the properties it defines, and ethically and scientifically questionable if it inadvertently

[25] See Hansson (2013) about an open-ended approach to science and pseudoscience demarcation.

forces distinctions where none exist, and thereby inflicts stigma. But this conclusion seems too radical and empirically unfounded. In fact, psychiatry may resemble comparative religion studies, archaeology, and parts of biology, so that the concept of "psychiatric disorder" may resemble concepts such as "religion" and "gene", which are both vague, yet scientifically useful.[26]

2.2.4 Bottom-up Approaches

I have argued this far against naturalistic and constructivist views of mental disorder, contending that they are at least partly based on traditional conceptual analysis. While naturalistic accounts based on conceptual analysis fail because they cannot provide a purely factual account of psychiatric problems as dysfunctions, constructivist views are problematic because they run counter to our intuition that psychiatric disorders are non-relative. Moreover, the relativism of the normativist accounts does not seem to match the undeniable clinical reality of the seriousness and universality of many psychiatric disorders. In this section, I analyse alternative bottom-up empirical approaches to determine the nature of disorders, and in the light of these approaches, formulate and defend my own bottom-up constructivist approach. I will first introduce Murphy's strong biomedical model (and RDoC in Chapter 6), followed by contextual approaches.

As I argued, top-down naturalist approaches consider it to be a given which conditions are prototypical mental disorders, and therefore, through conceptual analysis, try to figure out a definition that would apply to them. This means that traditional naturalist approaches rely on

[26] Kincaid (2014) compares "mental disorder" to the concept of a "gene", which is scientifically useful, although it does not have necessary and scientific conditions.

the conviction that psychiatric disorders resemble, for example, biological species or somatic diseases so that the subject matter is already given. However, in psychiatry, the ontological question about the nature of disorders and the epistemic question about how to determine their nature, cannot be cleanly separated. Unlike with biological species, we are not in agreement whether ADHD is a disorder, or how many disorders schizophrenia is. Instead, the hope is that uncovering the correct psychiatric explanation will bring clarity not only to what disorders are, but also how to identify them better. In other words, the starting point of identifying the objects of research and classification in psychiatry is messier or more complex than in many other sciences.

Biomedically oriented bottom-up approaches, such as Murphy's (2006) account and the RDoC (see Chapter 6), are more pessimistic about our current concepts and their referents. Such approaches consider that the problems of current disorder categories, for instance, do not resemble questions about what the nature of a particular species is, such as whether whales are mammals or fish, but rather whether we have even identified the research objects correctly. Hence, they consider that both the intension and the extension of the superordinate concept of psychiatric disorder and particular concepts may need to be revised based on further discoveries. Hence, according to these accounts, the concept of psychiatric disorder would resemble Griffiths' (1997) account of the folk concept of emotion, which according to him, picks out three different kinds of states or kinds: affect programmes, higher cognitive states, and social constructions. Similarly, bottom-up naturalistic approaches to psychiatric disorder discard conceptual analysis as a guide to their nature in favour of empirical research. This does not rule out the possibility that psychiatry may eventually be able to discover the true causes of mental disorders, or even common features that all disorders share. Rather, the bottom-up approaches, such

as Murphy's (2006) account and the RDoC program, remain open to the possibility that these discoveries may lead to substantial revisions concerning our current beliefs about the nature of disorders.

Murphy's Disease Model

Dominic Murphy (2006, 2013) defends a bottom-up approach according to which cognitive neuroscience provides the means to explain *mental dysfunctions* as breakdowns in information processing neurological systems in our brain. Based on this, he believes that psychiatry will eventually merge with cognitive neuroscience. This means Murphy's view is naturalistic and revisionist. According to him, conceptual analysis cannot provide the means to find out what psychiatric disorders are. In the light of this, the empirical lack of success of DSM categories is explained by their saturation with folk conceptions. Consequently, according to Murphy, the task of defining and determining the nature of psychiatric disorders should be left to cognitive neuroscience, which may lead to a thorough overhaul of our current classifications and conceptions.

In defence of this view, Murphy (2006, 2013) has coined the term "two-stage picture" to describe the biologically oriented strong view of the medical model of psychiatry. The two-stage picture resembles Wakefield's harmful dysfunction theory, with the fundamental difference that Murphy's account is revisionist and does not rely on an evolutionary analysis of dysfunction. According to his account, psychiatric research is and should be divided into descriptive and normative stages. In the first stage, psychiatric research aims to discover and describe the breakdowns that cause and maintain mental disorders. The second stage is evaluative and refers to the value-laden judgements of which breakdowns are mental illnesses and how to treat them. Therefore, whereas the first stage is based on the scientific work

of researchers, the second stage relies on the judgements of lawyers, bureaucrats, social workers, and ethicists (Murphy 2006). The second stage is where the use of the concept psychiatric disorder comes into play. Murphy (2006; 25, 98) exemplifies this with the gourmet lesion, a brain lesion that leads the inflicted individual to seek fine dining. Although the reason is clearly a brain trauma, it nonetheless raises the question of its harmfulness, and therefore whether it should be called a psychiatric disorder. Therefore, like Cooper, Murphy (2006: 97-98) holds that the superordinate concept of "mental disorder" resembles "weed", "pest" and "vermin". Although they are not natural kinds themselves, they nonetheless pick out different types of natural kinds. But unlike Cooper's account, Murphy's view does not imply that these various types of conditions could not be genuine natural dysfunctions or malfunctions in our brains.

The central argument that Murphy (2006: 85, 199) makes is that psychiatry can, in some areas of research, be free of values by being based on neurocognitive mechanistic explanations.[27] This means that we can come up with mechanistic explanations about neurocognitive systems without resorting to an evolutionary account of function, and thereby provide ahistorical explanation of an underlying malfunction in the brain. The main point is that a (brain) system malfunctions when it fails to contribute according to its customary role in maintaining the overall system. This requires understanding the role that the system, i.e. parts of our brain, has in contributing to its overall functioning (Murphy 2006: 81). Crucially, such malfunctions can come about, for instance, because of pathogenic information or an advertent environment. For these reasons, Murphy maintains that psychiatry should ideally merge with cognitive neuroscience by investigating how neurological

[27] Murphy (2006) does not believe that this approach is applicable to all disorders. Although in certain cases it is not, there can still nonetheless be objective standards of rationality.

functions, understood mechanistically, do not normally contribute to the overall wellbeing of the system. Consequently, according to Murphy (2006: 349), the boundaries of psychiatric disorders can be identified by combining statistical and causal reasoning.

As an example of this approach, Murphy argues that the human visual system can be explained mechanistically so that blindness can be understood as a malfunction that deviates from the normal functioning of the eye (apparently, deafness can be explained similarly, as it deviates from normal functioning). However, relying on the two-stage picture, Murphy maintains that this does not settle the question over whether blindness should be considered a psychiatric disorder or part of human variation.

The difficulties that underlie this eliminativist approach to psychopathology make it more of a promissory idea than a robust empirical hypothesis. I will return to these in Chapter 6, though I will mention a few here that are related to the concept of psychiatric disorder in general. Basically, the problem with Murphy's brain-based view is that it seems to consider psychiatric disorders as *malfunctioning brains in a vat*.[28] This implies that abnormal mechanisms can be studied and identified independently from their larger environment. I will argue instead that psychiatric disorders can only be understood in relation to their interaction and role in the larger environment, so that in many cases the explanation may need to abstract away from the purported neurological basis. In addition, Murphy's account is not a fully bottom-up account of psychiatric disorders, because it takes one explanatory approach, and argues that it applies to all disorders. This suggests that Murphy argues by philosophical means that the pluralism of disciplinary approaches to psychopathology found in psychiatry is a problem that needs to be overcome.

[28] I am indebted to Tomi Kokkonen for drawing my attention to the appropriacy of this philosophical scenario.

Contextual Accounts of Psychiatric Disorder

Contextual approaches to the concept of psychiatric disorder, or to the disorders themselves, do not concentrate so much on what disorders are, but instead examine the social or natural processes that have lead us to label some conditions disorders. Hence, they are bottom-up accounts because they are not based on trying to discover the meaning of the concept, but instead investigate the reasons and consequences for holding such a concept. A contextual approach to the concept of mental disorder can be coupled with both naturalism or normativism. In the following, I present Broadbent's and Canguilheim's naturalistic approaches and Hacking's more construct-oriented account. Thereafter, in relation to these approaches, I present my own contextual account.

Alex Broadbent (2020) has recently defended non-objective naturalism by arguing that health is a second-order property to the extent that relativizing health based on an age group of a sex of a species (according to the BST account) has contributed to the survival and reproduction of our species. Broadbent follows Peter Menzies and Huw Price (1993), who have argued that causality is a secondary quality, and suggests that health is also a subjective yet value-free concept. Menzies and Price endorse an agency-based approach to causation, according to which it is a secondary property in the sense that although it is not physically perceivable, it is nonetheless grounded in our interaction with the world. That is, causation should be seen as an extrinsic quality based on our ability to manipulate things in the world. And yet, it is also a subjective experience similar to colours, because it is projected by humans onto the world. Based on this, Broadbent suggests that the concept of health is like causation in that it is simultaneously a subjective and a natural feature of the world. Furthermore, Broadbent argues that although qualifying reference groups based on sex and age may not be more natural than having some other reference groups, we

make that qualification because it has been beneficial in the history of our natural selection (Broadbent 2020: 623). Hence, Broadbent agrees with Kingma (2014) that the age of a sex in a species is not the only possible natural or explanatory qualification for reference classes. However, he argues that holding such a view has been evolutionarily beneficial. In other words, the idea is that we do not share a concept of health primarily for cultural reasons. Rather, individuals or groups holding different conceptions of health would not have survived and reproduced to the same extent. In effect, Broadbent flips Boorse's account on its head. Instead of arguing that Boorse offers the most natural concept of health and disease, the concept may have been the healthiest to hold in the light of Boorse's account. That is, it can be considered the most natural because it has played a crucial evolutionary role in our survival and reproduction. That is, Boorse's concept of health can be called "natural" in the sense that it has provided an advantage in natural selection.[29]

However, Broadbent's account of health as a secondary property is suggestive at best. The account requires that we have had species-wide shared beliefs about health that have helped us to survive and reproduce. In other words, although there may have been other culture-dependent ways to qualify reference groups, only the qualification over an age group in a sex of a species has been shared to the extent that it has increased our evolutionary fitness. Consequently, Broadbent's account would have us believe that other conceptions of reference groups than that of sex and age have not been as evolutionarily beneficial at any point in our history, or have been weened out later due to an evolutionary mismatch. To establish the former case, however, we would need an evolutionary account of why alternative conceptions of

[29] Although Broadbent does not mention it, causation cognition may have played an evolutionary role for mammals. Apparently, for instance, reptiles are not capable of causal reasoning to the same extent.

alternative reference classes have not benefitted us in our natural selection (e.g. reference classes based on biological ethnicity or homosexuality). To establish the latter point, the view would have to be supplemented with an evolutionary mismatch theory of beliefs and conceptions, and based on that, an empirical account of how alternative conceptions have been weened out in natural selection due to such a mismatch. The basic idea would be that although some conceptions of health may have evolved to fit certain environments, unlike the qualification based on sex and gender, they have later become adaptively unfit in a changed environment to the extent that they have not facilitated the overall reproduction and survival of the population. One place to look for such an account could be mismatch theories of health (although they have been challenged, see Morris 2018). Nevertheless, a plausible problem with such an account could be that because the evolution of species happens on a population level, the concept of health may have developed to reflect more the wellbeing of individuals than whole populations. That is, just as it is evolutionarily beneficial for individuals in a population not to share all their traits, it could have hindered evolutionary fitness for everyone to share a similar concept of health. Furthermore, another challenge is that cultural and technological evolution may have played a role in the evolutionary mismatch of the concept of health. As one example, the human jaw volved to fit with the cultural evolution of eating cooked food (Sterelny 2010). Nowadays a human jaw dapted to non-cooked food would be considered an evolutionary mismatch. Similarly, some aspects of the health concept of our forefathers would probably not contribute to our reproduction and survival in society today. This raises the question to what extent the qualification of sex and age would have maintained its purported evolutionary role throughout our natural selection.

Another challenge with the account is that it seems to rule out culture as an endogenous explanation. The challenge concerns the

empirical question whether natural selection can trump cultural beliefs concerning health and disease. Although it seems plausible that we qualify capabilities according to sex and age, it is not self-evident that we attribute these capabilities in a robust way to health and disease. As many anthropologists have argued, ecological circumstances may limit the possibilities for values and cultural categories, but they do not determine them. There can be many, quite possibly endless, different but equally ecologically and evolutionarily fit cultural categories and values (see e.g. Sahlins 1985). And even if the reference class of sex and age were determined by nature, culture may have influenced how we have emphasized its importance, just as culture determines which nutritional plants and animals are considered food. At any rate, we cannot simply assume that ecological and material circumstances form the substructure that uniformly realizes cultures, their categories, and their values. As an example, as I previously argued, the skin disease pinta may have in fact contributed to reproduction for cultural reasons.

There is also a further question raised by moral realism. Even if we had shared the qualifications of sex and age with our concept of health, it does not imply, by itself, that we ought to have. Moral realists could argue about disease that the most evolutionarily fit concept may not be the most morally justified. For example, I doubt that we would be willing to admit that we are worse off in comparison to our foraging forefathers and foremothers, even if their environment may have been a better match with our biological nature. Similarly, even though alternative qualifications of reference classes of health may not have been as healthy, in the sense of contributing to our survival and reproduction, they cannot be ruled out as being less warranted.

A contextual approach to the concept of health can also be defended from a more individualistic perspective. Such an approach is offered by Canguilhem (1991), who argues that health is the individual's dynamic ability to adapt to new environmental situations.

This includes the social/cultural and material environment as well as values (see Tiles 1993: 738). This means that Canguilheim's account turns the BST account of disease on its head, so that instead of the average determining what is normal health, what is considered normal causally influences the average health (when it is understood as the ability to reproduce and survive). The reason is that values structure our social and material environment, which in turn influence average life expectancy and fertility. Consequently, an individual's health can be determined quite precisely based on his or her situation and environment, whereas species-wide generalizations cannot because of individual and environmental variations (Tiles 1993: 737). As an example, the poor survival rate of post-natal girls in China was partly causally determined by the cultural disvalue associated with girls (McGuire 2020). Canguilhem also argues that environments change to the extent that the same condition can be beneficial in one environment and unhealthy in another. In other words, culture may trump or enhance one's chances to reproduce and survive. This matches with the previously mentioned example according to which having the skin disease pinta was considered normal. The individuals that did not have pinta were considered abnormal and were forbidden to get married and reproduce offspring. In addition, it is fair to presume that they did not enjoy the same respect and caretaking, which plausibly lowered their average life-expectancy. Therefore, the culture-dependent conception of normality and health may profoundly influence an individual's health understood in Boorse's statistical sense.

Ian Hacking's view of a *sociocultural or ecological niche* can also be understood as a contextual approach to some psychiatric conditions (see also Chapter 5). According to Hacking (1998), some transitional psychiatric syndromes (e.g. fugue, hysteria, and multiple personality disorder) are enabled and maintained by sociocultural vectors, including the classificatory practices that target them, to the extent that

they form an "ecological niche" for the given syndrome to flourish. Hacking's point about the ecological niche is not to identify what mental disorders are but rather to clarify how some symptoms and patterns of behaviour become thought of as disorders, and how that influences the conditions or syndromes so considered. Hence, Hacking's view can be considered neutral as to what makes a condition a real psychiatric disorder.

However, as such, Hacking's view is too exclusive of other causes to be applicable to all psychiatric disorders. Rather, it is more plausible that most conditions that we have labelled disorders are not brought about by niches, but rather those niches may create "conceptual spaces" for those conditions and symptoms to be recognized, and in this way they influence how they are experienced and expressed. This is indicated by the fact that some severe disorders do not vary cross-culturally very much, and are held to be problems of the mind almost everywhere, whereas other disorders seem to be either culture bound, or if universal, substantially culturally shaped.

2.2.5 A Value-Sensitive Pluralistic Account

Based on what I have argued so far, I will now present my value-sensitive account of the concept of psychiatric disorder. It is a normativist account because it acknowledges the irreversible role that non-epistemic values play both in how the general concept of psychiatric disorder is defined, and also, as I will argue in the following chapters, in how particular disorders are specified. However, it is also a bottom-up account, because it suggests empirical research not only on the putative disorders themselves, but also on the non-epistemic values underlying their conceptions and classifications, as well as on the societal conse-

quences of those classifications. In addition, it is a realist account because I argue that conditions can in principle be genuine disorders in different ways without being merely problems of living or social forms of suffering. More specifically, I doubt that a single explanatory approach can be employed to understand the complexity of disorder. Hence, there may be different legitimate ways to conceptualize and classify those conditions.

In the light of what I have argued thus far, I am doubtful that the superordinate concept of psychiatric disorder, at least as it is now conceived, would pick out a group of phenomena that are naturally united by sharing one feature or a conjunction of features. This account is motivated by the fact that the concept of psychiatric disorder cannot be compared with, for instance, the concept of biological species, because there is no comprehensive theory, like evolution in biology, that could be employed to explain all psychiatric disorders, as evolution can explain species. Rather, there is a need for different disciplinary approaches that may also classify the disorders in different ways. Hence, the question is not so much which suggested definition is correct, but rather which definition is best suited to achieve relevant epistemic and non-epistemic goals. My account can be called a fully bottom-up one, because it embraces the plurality of explanations of psychiatric disorders, and does not find a full integration presently plausible (see Chapter 4).

However, although psychiatric disorder as a unified kind in a naturalist sense may not exist, it may exist in a weaker, yet non-conventional social sense. The concept of psychiatric disorder, and the psychiatric practices it informs, may contribute to the "sociocultural niche" which (weakly) glues otherwise diverse conditions together. The idea is that part of what makes conditions conspicuous is the fact that when they are considered as disorders, they are not only experienced as such, but also come under the influence of similar causal consequences (see

Chapter 5). In other words, the consequence of being diagnosed with a disorder leads to being subject to unified institutional practices and social roles. This idea reflects the fact that social norms and values play a dual role in psychiatry. While social norms and values play a conceptual role in partly determining which particular disorders we lump together under the general heading of psychiatric disorder, they also play a causal role in influencing and contributing to the social context where disorders originate, develop and are experienced.[30]

This naturally prompts a question as to how individual psychiatric disorders differ from social deviance if not by sharing the common feature or features of being dysfunctions. As I pointed out, antirealist accounts cannot account for the genuine and universal suffering that many psychiatric disorders cause. Normativism, on the other hand, runs into trouble with relativism. Hence, some other approach is called for. It is plausible that part of the challenge of trying to provide an account of genuine disorder lies with the ambiguity of the concept of realness. One way to address this is by refining the question of what is a real disorder with the help of contrasts (Zachar 2014b, Cooper 2010, Hacking 1995a: 11; see Austin 1962, see also Mäki 2008). One can ask, for instance, whether a condition is a real disorder in contrast to social deviance or a real contrast to the non-brain disease. Understanding the nature of different disorders may lead to answering these questions differently. While many disorders are not real brain diseases, they may nonetheless be real disorders in virtue of not being social deviances.

Moreover, some real disorders may be shaped by sociocultural factors to the extent that they are culture-bound. This implies that instead of there being one universal feature or features that all disorders share, it is plausible that many of the conditions currently considered disorders

[30] As an example, Kendler (2012) mentions social norms as causal factors as something that can potentially be manipulated to influence pathological drinking behaviour.

share a family resemblance (see Lilienfeld and Marino 1995). As Wittgenstein (1953) pointed out, some concepts resemble families or games. Although all family members may not share a single feature or features, they may nonetheless share overlapping features. Similarly, particular psychiatric disorders may not share the same properties, but instead share overlapping properties, while nonetheless being genuine disorders.

Moreover, in the light of the multifaceted nature of psychiatric disorders, explanations based on social factors are bound to play an important role in their definitions and explanations. Social factors are not only triggers or learned reactions, but also form the contexts where disorders originate, as well as where they are experienced and interpreted. This does not contradict the view that biological accounts can be better fitted to explain some disorders, while psychological and social approaches are more appropriate to explain others. Rather, it implies that some disorders may be real as biomedical entities, others as psychological conditions and some others may be real as impairments in social interaction. The outcome is that one can be a realist about individual kinds of psychiatric conditions, such as schizophrenia and ADHD, while being suspicious about psychiatric disorder as natural kind (Beebee and Sabbarton-Leary 10b). Moreover, the complex nature of those conditions may license re-classifications based on their contrastive epistemic and non-epistemic aims.

Scientific progress and social change also seems to imply that we should have an "open ended" stand towards revising the superordinate concept of psychiatric disorder in order to encompass all the conditions that we may want to hold as pathological in the future. That is, even if a consensus over the definition of a concept were to be reached at some point in the future, revisions could still be needed from time to time due to our cultural, technological, and biological evolution. The reason is that our sociocultural evolution is bound to alter our social and

physical environment, on the one hand, and our pharmacological and technological development can enable better interventions on our cognitive and physical makeup, on the other. These factors together will most likely generate new forms of mental suffering and ways to modify or treat old ones.

Therefore, I suggest a bottom-up, value-sensitive approach to the concept of psychiatric disorder or mental disorder, and to the specification and classification of particular psychiatric disorders.[31] It is based on the acknowledgement that psychiatric classification and concept reflect, but also reciprocally affect, the social context where disorders originate, are experienced, and are shaped. [32] Value-sensitivity is based on making explicit and studying the norms and the social contexts where psychiatric disorders originate, but also on the fact that how they are studied, classified, and treated shapes that context. In other words, my approach is bottom-up because it is naturalistic: I investigate the reasons – values and facts – why we hold disorders as mental disorders, not just whether they are based on values. This approach is intended to be an analysis of scientific research and classification rather than clinical practices, which are clearly value-laden (see Cooper 2007: 127). For instance, Brigandt (2020) mentions that psychiatry has non-epistemic aims that should play an explicit role in classificatory decisions. These include the aim to ensure the right to treatment, which may require us to draw the boundaries of disorders somewhere, even if they are fuzzy in nature. On the other hand, Brigandt also mentions the aim of avoiding stigmatization. This means that classifications may need to take into account how psychiatric disorders are characterized in addition to how their boundaries are

[31] I am borrowing the notion "value-sensitivity" from technological design (see van Wynsberghe 2015).
[32] Friedman and Hendry (2019: 1) suggest a similar approach to value-sensitive technological design.

specified. Moreover, there may be a trade-off between the right to treatment and avoiding stigmatization: the former may suggest wider definitions, while the latter may suggest more restricted ones. I agree with these considerations and suggest more empirical research not only on the psychiatric disorders themselves, but also on the consequences of their classifications (I will return to these issues in the final chapter).

In conclusion, my account can be called *value-sensitive explanatory pluralism* because it is based on the idea that conceptions and classifications of psychiatric disorder have wide-ranging social consequences that require investigation and value-sensitive weighing (see Table 1). Unlike the mentioned normativist approaches, which concentrate on defining the concept of psychiatric disorder, my account is a naturalist one because it concentrates on the *causal significance* of the concept and classification of psychiatric disorder. Nevertheless, knowledge of the consequences of psychiatric classificatory practices may bring them (partly) under our control. It is also a realist account, because there is no reason to suppose that those conceptions and their consequences are all there is to psychiatric disorders. Rather, the complexity of disorders requires both empirical studies by different disciplines and value-sensitive weighing of how to weigh the results of those studies for different explanatory, pragmatic, and ameliorative purposes.

Table 1. Philosophical approaches to the concept of mental disorder

	Top-down	**Bottom-up**
Naturalism	Weak medical model (DSM) Biostatistical theory Wakefield (dysfunction)	Strong medical model (RDoC) Murphy's disease model
Normativism	Harmful accounts (e.g. Cooper, Fulford, Nordefelt)	Value-sensitive pluralism (Vesterinen)

3 Psychiatric Disorders as Scientific Kinds

I have argued that the superordinate concept of psychiatric disorder is value-laden, and thereby does not pick out a single natural kind of psychopathology. In addition, I have suggested that non-epistemic values may play a role in how particular psychiatric disorders are specified. To address these challenges, I suggested an explicit value-sensitive approach to the process of classifying and specifying psychiatric disorders. The idea is that we should balance the value-laden aims of classifications and conceptions with our empirical discoveries to determine whether a given condition is a disorder and how to draw its boundaries. This is a co-fitting process that balances between what we want with the classification and what nature has to offer. Nevertheless, although I have argued that the superordinate concept is value-laden, this does not, by itself, mean that particular psychiatric conditions could not be genuine conditions in nature that we, partly based on our non-epistemic values, lump together. Hence, there is a need to investigate the nature of particular disorders. Particular disorders may not only differ substantially from other classificatory kinds, such as chemical elements and biological species, but also from each other and somatic diseases.

In this chapter, I ask whether particular psychiatric disorders can be scientifically relevant objects of research – real or natural kinds – and if so, what type of theory can be formulated to explain them. I explore the idea, defended especially by Cooper (2005) and Beebee and Sabbarto-Leary (2010b), that although the concept of psychiatric disorder may not pick out a real or natural kind, particular disorder concepts nevertheless in theory can. However, even this more modest realism about psychiatric disorders needs to address the relatively poor inductive success of psychiatric classifications. The reason is that natural kinds are usually invoked to explain the robust inductive success of a

scientific classification. But since psychiatric classification has not been particularly successful, a thesis of psychiatric disorders as scientifically relevant kinds needs both to account for psychiatry's modest inductive success this far, and if possible, suggest improvements for classifications based on the formulated account of psychiatric kinds.

I begin by providing an overview of the philosophical debate over the notion of natural kind and its relevance to psychiatry. I argue against essentialist views of natural kinds on account of the fact that they would exclude the kinds studied by the special sciences in spite of the relative inductive success of their classifications. Instead of essences (or ironclad laws of nature), I argue that kinds studied by the special sciences are better explained by exception-ridden causal and mechanical explanations. In the light of this, I present Boyd's (1999) homeostatic property cluster (HPC) view of natural kinds and its application to psychiatric kinds by Kendler et al.'s (2011). Thereafter, I provide an overview of Craver's (2009) criticism, which concerns our ability to individuate or specify a property cluster by finding a non-conventional way to delineate mechanisms from each other and the environment. To address this challenge, I explore alternative ecumenical theories of natural kinds. However, I argue that the HPC view is better suited to explain psychiatric kinds than its alternatives, and the criticism it has received relies too much on an ambiguous and technical understanding of the mechanisms that are responsible for the property clusters. Nevertheless, I point out that a better account of the connection between inductive explanation and homeostatic mechanism is required to understand how property clusters can reliably support inductive inferences.

3.1 Natural Kinds of Psychopathology

Scientific realists generally argue that successful scientific classifications and concepts pick out natural kinds. This view of scientific realism can be called classificatory realism or natural kind realism. The idea is that kinds are natural when they are not arbitrary, or when their classifications are not imposed on nature by the classifier but are instead found there through empirical research. This means that classifications that manage to divide the world according to its natural kind structure, to paraphrase Plato, "carve the nature at its joints". A common motivation for promoting kinds in the philosophy of science comes in the form of inference to the best explanation. Hilary Kornblith (1993), for instance, argues that the best explanation for the inductive success of science is the existence of natural kinds (cf. Boyd 1989: 7). If there were no underlying natural kinds that bind properties together, it would be a miracle how we are able to infer reliably from one property to another. As an example, the targets of the periodic table in chemistry should be held as natural kinds because the taxonomic table has been more inductively successful than any proposed alternative. Consequently, discovering natural kinds and describing them better grounds the advancement of science.

Although the term "natural kind" has not been commonly used in psychology and psychiatry as such, the idea has nonetheless been implicit (Zachar 2000). The hope has been that successful psychiatric classifications would pick out real psychiatric disorders, as well as their properties and relations, rather than imposing arbitrary or value-laden constructs on the world. The idea has been that such a classification would be able to organize knowledge, enable accurate definitions, enable communication between researchers and clinicians, and facilitate further research as well as diagnoses and treatments (see Thakker and

Ward 1998). Consequently, realism about psychiatric kinds can be understood as a central tenet of psychiatric realism.

A realistic account of kinds can be contrasted with conventionalism. According to a constructivist or nominalist view of conventionalism, nature is not divided into pre-packaged kinds of things, only into individuals that we may classify into conflicting categories based on our interests, values, and conventions. As I pointed out in the previous chapter, the anti-psychiatric movement held a nominalist view about psychopathology along these lines. Ian Hacking's dynamic nominalism is a more moderate type of conventionalism (see Chapter 5). It holds that although classificatory practices causally influence the purported members of the classifications, the classificatory concept and its causal consequences need not be the only thing that those members have in common. Another type of conventionalism in psychopathology is instrumentalism according to which successful scientific categorizations are merely pragmatic instruments for predictions and interventions, but do not pick out natural kinds. Zachar (2000, 2014a, 2014b) argues for a pragmatic approach to classifying psychiatric disorders. As I argued in the previous chapter, although the DSM classification manual (i.e. the weak medical model) is realist about the definition of psychopathology in general, in practice it is instrumentalist about particular disorders because of its a-theoretical approach to their definitions.

The discussion over natural kinds in philosophy has mainly concentrated on the classificatory objects of the natural sciences, such as chemical elements and even biological species. The idea of natural kinds has traditionally been invoked to explain why the natural sciences have been successful. The naturalistic answer is that our scientific inferences are reliable because they are supported by natural kinds: inferences between properties are reliable because those properties are

bound together in natural kind structures. However, psychiatry has a notorious track history of unreliable predictions, failed explanations, and inefficient treatment interventions. Moreover, there is no agreement in psychiatry on the nature of mental disorders, and consequently, how to classify and explain them. This means that there is no obvious scientific progress that would invoke natural kind explanations. Instead, what requires explaining is the lack of scientific progress and sound inductive success. Therefore, inference to the best explanation does not straightforwardly apply to psychiatric kinds (or many other kinds in the special sciences). Therefore, the same approach to kinds cannot be used for studying and arguing for kinds in psychiatry as, for instance, for chemical elements or biological species.

However, this does not mean that establishing what kinds of scientific objects psychiatric disorders are would not be relevant. If anything, the lack of apparent success in psychiatry means that understanding the nature of the objects studied by psychiatry is important not only for scientific, but also for ethical, social, and pragmatic reasons. Part of the reason is that the idea that psychiatric disorders are natural kinds has been employed to defend psychiatry as a medical science, and thereby to justify medical interventions as well as policy choices. Another motivation has been the common intuition and hope that there is a generic psychopathology to be found.[33] Such an account would not only be able to demarcate between genuine psychopathology and normality, but also answer questions about responsibility and inform about human nature in general. However, as I tried to demonstrate in the previous chapter, a general non-normativist account of psychiatric disorders that could accommodate all the conditions we currently want to label as pathological, and will most likely want to continue to label in the future, is not achievable.

[33] For example, a motivation has been to find one latent reason for psychopathology – the p-factor (see Caspi et al. 2014).

Nonetheless, I now want to explore whether there may be particular psychiatric kinds that support robust inductive inferences. As mentioned earlier, although the general concept of psychiatric disorder may be value-laden, and thereby does not pick out a natural kind, concepts of particular disorders may nonetheless do. According to Brigandt (2011), a philosophical theory of a scientific kind should aim to answer two questions. First, what inferential and explanatory goals does the discipline have when it studies and classifies kinds? Second, how well do the classifications and grouping into kinds satisfy those discipline-relative explanatory and inferential goals? I will examine the goals set especially by the strong and pragmatic views of the medical model and argue that the former make too weak and the latter too strong demands for most psychiatric kinds. In addition to these questions, an account of psychiatry kinds also needs to consider the fact that psychiatric classification does not only have epistemic aims but also pragmatic and moral ones. Therefore, a purely theoretical realist view of psychiatric disorders according to which they would not be merely subjective distinctions or social constructs imposed on nature, but nonetheless would not support sound predictions and therapeutic interventions, would not satisfy the pragmatic aims of psychiatric classification.[34] Rather, a realist account of psychiatric disorders, one that can support theoretical and pragmatic progress in psychiatry, would require an account of psychiatric disorders as relatively stable objects that support explanations, predictions, and therapeutic interventions.

A central challenge in the philosophical debate over psychiatric kinds has been to come up with a scientifically relevant theory of natural kinds, and to maintain, in the light of empirical evidence, that such an account is applicable to the research and diagnostic objects in

[34] This could be the case, for instance, if social, psychological, or biological interaction makes psychiatric kinds so complex and unstable that their nature can only be explained in retrospect.

psychiatry. Cooper (2013) argues that the challenges facing realist accounts of psychiatric disorders are that they are interactive, their boundaries are fuzzy, and their identification is at least partly value-dependent. I agree with these qualifications, elaborating them with empirical examples in the coming chapters and presenting my own answers in relation to them. I will begin by providing a general analysis of natural kinds, and subsequently, in later chapters, provide my own account of psychiatric kinds.

3.2 Essentialism and Naturalism in Natural Kind Realism

Realism about natural kinds can be approached from an essentialist or naturalist perspective (or strong realism and weak realism) (Kornblith 1993, Reydon 2009, Bird and Tobin 2018). Whereas naturalism holds the inductive success of science as the defining feature of natural kinds, essentialists argue that the definition of natural kinds cannot be based on the varying inductive successes of the sciences but needs a robust metaphysical foundation instead (Beebee and Sabbarton-Leary 2010a: 2). In the following, I will first provide an overview of general conditions set by essentialism and then compare it to the naturalist view by using biological species as an example.

Essentialist accounts generally provide natural kinds with two conditions. First, the members of a natural kind should share properties that are necessary for the membership of the kind. Usually this is supplemented with a sufficiency condition, so that essential properties are independently necessary and jointly sufficient for membership of the kind. The idea is that when properties are an essential part of a natural kind, they are a posteriori discoverable, instead of being based on a priori definitions. Traditionally, philosophers have connected the necessary nature of essential properties with the laws of nature (e.g. Kripke

1980, Putnam 1975, Ellis 2001). Peirce, for example, argued that what renders natural kinds interesting for research is that they are bound together by the exceptionless laws of nature (see Hacking 1991: 119).[35] However, in the 1960s, the necessity condition was modally formulated so that each member of the natural kind has the essential properties necessarily in every possible world that the member exists. Initially, the idea was developed by Ruth Marcus, and later picked up by Saul Kripke and Hilary Putnam, when they argued for external semantics of natural kinds. According to Putnam's famous twin-earth argument, two individuals on different planets can associate the same descriptions with the natural kind term "water", and still refer to different natural kinds, H_2O and XZY because the term's references were locked externally by different external surroundings instead of the term users' equal mindsets. Conversely, people can associate different and conflicting descriptions with a natural kind term, and yet refer to, and disagree over, the properties of the same natural kind.[36]

Second, perhaps the central condition of essentialism is that the properties of natural kinds should be intrinsic (e.g. Locke). Here the idea is that the existence of natural kinds does not depend on flimsy external and relational circumstances, but that they are based on fundamental non-relational, distinct, and stable properties. The idea is that intrinsic properties determine the natural kind's contingent and observable properties, as these superficial properties are prone to varying external forces. Different requirements as to the specific nature of these intrinsic properties have been proposed. Influentially, Saul Kripke and

[35] Ellis (2001), on the other hand, has argued conversely that essence explains why the laws of nature work.

[36] This last point was central in arguments against incommensurability and anti-realist views of the advancement of science. See Chapter 7.

Hilary Putnam defended the view that intrinsic properties are microstructural.[37] Thus gold, for instance, has the atomic microstructure of 79 protons and 118 neutrons in its nucleus (atomic number 79), which cannot be observed with the naked eye. This intrinsic microstructure, in turn, determines its superficial properties, such as its malleability, shininess, yellow colour, and ductility (Kornblith 1993: 30).

On the other hand, the naturalistic view of natural kinds is a weaker form of realism that defends natural kinds from an epistemic point of view without committing itself to essentialism. It stresses how natural kinds ground robust inductive inferences, explanations, and predictions (Mill 1843, Boyd 1989, Millikan 1999, Dupré 1993). In general terms, the naturalistic approach to kinds is based on two basic convictions. First, as stated before, the central feature of natural kinds is that they support inductions. This line of goes back to J. S. Mill, who argued that natural kinds share conceptually and logically unrelated properties (Hacking 1991).[38] This means that if we encounter a member of the natural kind, we are warranted to infer that its kind-typical properties are no doubt present as well. This renders natural kinds scientifically and empirically interesting because they and their associated properties are susceptible to a posteriori empirical research instead of merely being imposed on nature by the classifier. For instance, all instances of gold have the same melting point of 1337 K, a density of 19.3 cm-3, and are under similar external conditions malleable, shiny, and yellow in colour (Khalidi 2013). Therefore, if we learn that a certain lump is in fact a piece of gold, we can infer that it has all the kind-typical properties that we know of, and others that we can learn about.

---

[37] LaPorte (2004) argues that had Putnam and Kirpke been empirically better informed, they would not have supported microstructuralism. Kripke, however, does not seem to be as adamant about microstructuralism as Putnam.

[38] Mill (1843) argued that projections should be open-ended, but as we will see, this is not the case with kinds in the special sciences.

The naturalistic tradition of natural kinds holds that scientific classifications and categories provide the best examples of what the natural kinds are and where to look for them (Khalidi 2013: 65). This implies that natural kinds are needed to explain why some scientific classifications ground epistemic projects more than others. Richard Boyd (1999b: 69) has emphasized this point with the accommodation thesis: natural kind concepts are projectable because they accommodate with the causal structure of the world. This means that epistemically useful classifications and concepts are more natural because they latch onto the metaphysical causal structure of the world. However, it should be emphasized that argumentative precedence is given here to empirical research, and the epistemic usefulness of the natural kind concepts, over any metaphysical justification and a priori definition of natural kinds.

A naturalistic approach to natural kinds can take either a positive or a negative stand on essentialism. A positive stand would claim that naturalism contrasts with an essentialist view of natural kinds, or at least, there is nothing an essentialist or strong metaphysical view adds to our understanding of natural kinds.[39] A negative stand, on the other hand, would hold that there is no contradiction between the two approaches, and the correct type of metaphysical theory could supplement the naturalistic approach. Although I argue for a naturalistic view of natural kinds that is as much as possible metaphysically non-committal, I see no a priori reason why it could not be supplemented with a robust metaphysical theory of natural kinds. Therefore, although I do not provide an explicit argument for it, I tend to side with the negative argument. However, central to my approach to kinds is whether the inductive success of science requires a robust metaphysical theory of them to back it up. My argument is that it does not.

[39] This is perhaps the view that Ladyman and Ross's (2007) structural realist would take. I thank Alexander Bird for pointing this out to me.

Nevertheless, a strong essentialist view of natural kinds does not concur with a naturalistic approach to how kinds are employed in the special sciences. This is because the necessity and intrinsic conditions do not match how the special sciences explain and describe their classificatory targets, and in many cases, these conditions are not necessary for grounding their inductive success. At any rate, as I will argue, they are not necessary for holding kinds to be real or scientific, although some kinds may possess them. Nevertheless, the essentialist account of natural kinds may be the case in some sciences, for instance in chemistry, where necessary and sufficient conditions seem to apply more straightforwardly.

It is clear that the conditions that essentialism sets for natural kinds do not straightforwardly match the nature of biological species. This is because evolution is a dynamic and partly external process that makes intra-species properties varied and inter-species boundaries fuzzy. This means that biological species do not adhere to the necessary and intrinsic condition of natural kind essentialism. Devitt (2010) seems to argue that genes are the best candidates to fulfil the conditions of essentialism (see also Ereshefsky and Reydon 2015). This is not believable. Because the evolution of the species takes place on the population level, intra-species genetic variation is beneficial for a population as it renders the population better equipped to adapt to environmental changes. Consequently, individual property variation within species is the norm rather than the exception (Dupré 1981, Ylikoski and Kokkonen 2010). By the same token, there are no intrinsic properties that would determine the manifest and stereotypical properties of species, such as their morphology and physiology (Griffiths 1999: 203, Dupré 1981: 84). Instead, environmental factors, such as genealogical and ecological pressures, together with intrinsic determinants, such as genetic exchange within a population, are the reasons why members of the same species tend to share most of their properties (Griffiths 1999:

210). The dynamic nature of evolution also means that species have fuzzy rather than discrete boundaries.[40]

Nevertheless, although biological species do not adhere to the conditions set by essentialism for natural kinds, they still ground inductive generalizations and explanations. A central reason for this is that although biological phenomena are not governed by exceptionless essences or the laws of nature, they are nonetheless based on tractable causes and causal mechanisms. The importance of this difference for the conception of scientific kinds becomes explicit by delving into the historical origins of why realism has been conflated with essentialism. Historically, logical empiricists founded scientific explanation and the advancement of science on the laws of nature, and the exceptionless generalizations that they arguably support. But when the positivist view of science by the logical empiricists ran into numerous problems, including the fact that it cannot ground the advancement of science on a cumulative approach of combining explanatory laws, most philosophers with a realist inclination towards the advancement of science turned to metaphysical realism and essentialism for support.

During the past few decades, however, there has been a shift from the normative view that all legitimate scientific explanations should be based on the laws of nature, to the realization that scientific explanations in the special sciences rely on causes and mechanisms. Moreover, since causal and mechanistic explanations admit exceptions, the reductive distinction between the sciences has given way to granting that the special sciences, such as biology, are held to offer as good explanations as chemistry or physics. Importantly for my analysis, this realization undermines the reasons for holding onto the essential view of natural kinds as the model for all legitimate scientific kinds. That is, if there is no principled difference between explanations between the sciences,

[40] These reasons have led some philosophers to argue that species are not kinds but individuals with a common history (Hull 1980).

there is no a priori reason why the objects of research in the special sciences could not ground inductive inferences as well. In other words, in this light, legitimate scientific kinds can be understood more on an epistemic and explanatory basis, since the essential view of natural kinds has lost its historically grounded explanatory purpose. In sum, as I will argue, the advancement of science and scientific explanation does not require a robust metaphysical picture of kinds, but realism can be based on an explanatory footing.

Consequently, since the essentialist view of kinds is in contradiction with the naturalist view of kinds, especially in the special sciences, we are faced with two options. Either we grant that the kinds invoked by the special sciences are not kinds in the same sense as kinds found in the basic physical sciences, i.e. physics and chemistry, or we relax the conditions for kindhood to match the objects studied by all the sciences. The former choice would have us rely on a strong metaphysical account of kinds, whereas naturalism and an explanatory view leads us, as I will argue, to conjecture in favour of the latter choice. This has liberated the use of talk of kindhood to cover groupings that support scientific generalizations (e.g. Griffiths 1997, Kornblith 1993, Griffiths 1999: 216, Mallon 2016, Khalidi 2013, Cooper 2013, Godman 2020).

Nevertheless, because of the term "natural kind" connotes the research objects of natural sciences (albeit the distinction is fuzzy and partly conventional), a better term for all the kinds that the sciences study is "real kind" or "scientific kind" (I will use these interchangeably). In this way, the term "natural kind" can be preserved for kinds of natural phenomena, while the term "human kind" can be employed for human and social phenomena. In sum, thus far I have argued that although essentialism does not fit the kinds studied by the special sciences, this does not rule out the fact that they can in principle be scientific kinds that play a robust inductive and explanatory role.

3.3 Naturalistic Approaches to Kinds

Thus far I have argued that naturalism about natural kinds leads to non-essentialism. This opens the door for an understanding of special science kinds as scientific kinds. It is increasingly common to argue that divisions and groupings studied by the special sciences, including human sciences and psychiatry, are real kinds to the extent that they are projectable. However, an ecumenical and naturalistic approach to kinds needs to address several challenges. First, it has to account for the heterogeneity (including fuzziness and gapless) nature of the purported kinds studied by biology and the social sciences. Second, it needs to provide a scientifically relevant account of how kinds can be mind-dependent and yet support inductions. Third, it needs to account for the non-intrinsic and interactive nature of the kinds. In the following two sections, I analyse the question of heterogeneity, while the question over mind-dependency and interactivity will be analysed in the fifth chapter.

3.3.1 The Homeostatic Property Cluster View

Thus far I have argued that the essentialist account of kinds does not fit with the way kinds are invoked in the special sciences, especially biology and psychiatry. However, in the light of naturalism, I have also argued that there is no principled reason to preclude special science kinds from being scientifically useful. In the following, I introduce the dominant view of special science kinds and the criticism is has received. Subsequently, I examine the challenges it faces in the human sciences generally, and in psychiatry in particular.

Richard Boyd's (e.g. 1999a) homeostatic property cluster (HPC) view of natural kinds has become the standard view in the philosophy of the special sciences (e.g. Griffiths 1997, Murphy 2006, Kuorikoski and Pöyhönen 2012, Hauswald 2016). The HPC view of natural kinds is based on four criteria (Boyd 1989 16-17, 1991, Craver 2007: 574):

(i) Property clusters occur regularly together.
(ii) Homeostatic mechanism(s) explain why the clusters co-occur and hold together.
(iii) Successful classifications are accommodated with inductively important causal structures.
(iv) Property clusters figure in causally important generalizations.

The HPC view holds that natural kinds consist of (i) co-occurring property clusters and (ii) homeostatic mechanisms or mechanisms that are responsible for and realize or cause them (Boyd 1989 16-17; 1991). The thinking here is that although the properties of a natural kind may vary from member to member, the mechanisms nonetheless hold the properties in a state of homeostasis under environmental pressures. The idea of homeostatic mechanism is meant to accommodate the fact that although property clusters in the special sciences are imperfectly captured by their stereotypical definitions, they nonetheless ground inductive inferences. That is, if we find an underlying homeostatic mechanism or mechanisms that serve as the common cause for the co-occurrence of certain properties, we can reliably infer that the property cluster kind grounds inductive inferences. In other words, according to the HPC view, natural kinds serve both as explananda and explanandum: causal mechanisms explain the co-occurrence of property clusters, and the clusters together with the mechanism or mechanisms ground other explanations and inductive inferences (Reydon 2009, cf. Kuorikoski and Pöyhönen 2012).

The HPC view builds on the realist inference to the best explanation: the best explanation for our inductive success is the existence of natural kinds. To this end, Boyd provides two other criteria (iii) and (iv) that describe how kinds are invoked universally and specifically by the sciences (Boyd 1989: 16-17, 1991, Craver 2007: 574). The accommodation condition (iii) provides a general realist justification for natural kinds. It requires that the naturalness of a classification is based on the "accommodation between the relevant conceptual and classificatory practices and independently existing causal structures" (Boyd 1999a: 55).[41] This means that natural kinds do not have a posteriori definitions based on their manifest properties, but instead require research on the causal structures that enable the properties to cluster. In other words, successful accommodation between our concepts and natural kinds renders epistemic projects, such as inductive inferences, reliable and possible.

The causal importance condition (iv), in turn, holds that homeostatic property clusters should be scientifically relevant. The criterion is meant to exclude scientifically irrelevant homeostatic property clusters as natural kinds. One might argue, for instance, that although the behaviour of an air pressure barometer is caused by a common homeostatic mechanism, it should not be considered a natural kind since it does not ground a posteriori interesting generalizations. Moreover, the causal importance criterion also implies that the naturalness of property clusters depends, in part, on the goals and methods of the specific discipline, since different disciplines have different criteria for the properties or success of their classifications and conceptualizations. A good example is that while hysteria is not considered an interesting or real homeostatic grouping in psychiatry anymore, it is nonetheless a

[41] Boyd (1991: 144) claims that classifications may have causal influence, but cannot make logical or conceptual changes in the causal powers of the targeted kinds. I return to this idea in the following chapter about looping effects.

scientifically relevant social kind in history and sociology, since it was produced by psychological and cultural mechanisms.

On the face of it, the HPC view avoids the pitfalls of essentialism by being able to accommodate a naturalistic approach to kinds. In other words, it can account for how kinds are varyingly used in the special sciences and still offers a viable account for their inductive success. First, members of a natural kind do not have to satisfy the necessary (and sufficient) condition. As an example, although biological species imperfectly share morphological, physiological, and behavioural features, they nevertheless ground epistemic projects (Boyd 1991: 142). Another example offered by Boyd is the human liver. Although it varies from individual to individual, it has an important explanatory role in medicine (Boyd 1999b: 91). Both cases show how the HPC view is able to capture how property clusters are multiply realized. Neither biological species nor human livers are realized by exactly the same mechanism in their every instantiation.

Second, the HPC thesis does not require that natural kinds are intrinsic and microstructural, but instead the clustering of properties can in part be caused by extrinsic mechanisms (Boyd 1999b: 84). For instance, external mechanisms, such as environmental pressures, together with the internal mechanisms of interbreeding and genetic descent, cause the properties of biological species to cluster together (Khalidi 2013: 73). Crucially, external mechanisms do not render definitions of HPC kinds a priori, since discovering them requires empirical research. Consequently, according to the HPC theory, scientifically relevant natural kinds can potentially also be found in the special sciences, not only in sciences that study kinds that more likely to fulfil the requirements of essentialism, such as chemistry or physics. This is also true about psychiatric kinds, where external mechanisms can include social norms as well other environmental pressures.

Based on these reflections, replacing natural laws and essences with the mechanistic view of kinds paves the way for an inclusive interpretation of the success of scientific classification. Thus, not all scientific kinds need be like gold, whose intrinsic atomic structure produces exactly the same properties in its every instantiation. On the contrary, multiple homeostatic mechanisms, some of them external, can produce and maintain imperfect property clusters that nonetheless ground robust epistemic projects.

In conclusion, I have pointed out that a difference needs to be made between two types of realism. Essentialism is based on ontological realism, whereas the HPC theory can be accommodated within an epistemic view of kinds. The latter approach defends realism against nominalism and is not concerned about the material basis of kinds. This matches with the naturalistic account of the human sciences (that their explanations are not different from the natural sciences). Since the homeostatic mechanistic account does not set restrictions to the types of properties and mechanisms involved, in the human sciences those forces can be external norms or other social factors. In other words, homeostatic mechanisms can be social and bring about and sustain social properties (Boyd 1991: 140). In short, on the face of it, the HPC view can accommodate social and human kinds, as well as more traditional kinds studied by biology and chemistry.

3.3.2 Psychiatric Disorders as Property Clusters Kinds

A common realist approach to psychiatric disorders is to interpret them as homeostatic property cluster kinds (HPC view) (e.g. Beebee and Sabbarto-Leary 2010b, Kendler et al. 2011; Kendler 2012, Murphy 2014, Tsou 2020). If psychiatric disorders are property cluster kinds they would not be mere pragmatic categorizations, but would instead

support sound inductive inferences while not falling prey to the unattainable demands of essentialism. The HPC view of psychiatric disorders can potentially accommodate the fuzzy boundaries of disorders, include multiple causal mechanisms as their explanations, and account for how different combinations of mechanisms in different individuals can produce the same psychiatric kind. In spite of these ecumenical features that differentiate HPC kinds from essential kinds, they can nonetheless support robust inductive inferences. Further, the HPC view is compatible with external social and environmental mechanisms participating in generating and sustaining those clusters (Boyd 1991, Mallon 2003). Moreover, the identification of the mechanisms that are responsible for the property clusters enables extrapolations to alternative situations. That is, knowledge about a shared homeostatic mechanism enables secure extrapolations between property clusters and reliable projections over kind-typical properties that the members are likely to share. Nevertheless, different psychiatric disorders probably fall on a scale, where some of the kinds share many properties while others only some of them.

Especially Kendler et al.'s (2011) version of the HPC view, i.e. mechanistic property clusters (MPC), has become the standard for arguments in favour of understanding psychiatric disorders as natural or real kinds. They argue that MPC kinds are closer to essentialism than mere property cluster kinds. Hence, according to their account, psychiatric kinds should be equated with the responsible causal mechanisms which may include "complex and multi-level causal mechanisms that produce, underlie and sustain our psychiatric syndromes" (p. 15). However, they argue that the MPC account differs from essentialism on three accounts. First, the MPC account can facilitate causes such as "evolutionary, developmental, genetic, physiological, psychological, behavioral, social" (p. 1148). More specifically, they "include underlying etiological pathways (genetic,

physiological and cognitive-affective) but also the overt properties themselves (symptoms for psychiatric illness)" (p. 1148). Second, instead of postulating deterministic relations between causes and properties, they argue that the relations should be understood as probabilistic. That is, identifying relevant causes does not determine whether certain symptoms are present, but instead determines their probability. Third, the MPC view holds that different combinations of etiological, maintaining, or realizing mechanisms can be responsible for the same set of properties. For example, depression can interact with cognition to the extent that negative expectations become self-fulfilling prophecies. This model, according to them, could also include biological causes.[42]

Despite these promising aspects, the idea that psychiatric disorders are property cluster kinds still remains contested. There are multiple challenges with understanding psychiatric disorders as HPC kinds. Zachar (2000; 2014a), for instance, has (previously) advocated a pragmatic kind view of psychiatric disorders, according to which value preferences, along with epistemic preferences, are weighed against each other to figure out a classification that works in practice. Haslam (2014), on the other hand, argues that only some psychiatric disorders meet the criteria for HPC kinds. Most of the psychiatric disorders should instead be categorized as pragmatic or dimensional constructions. One way to answer heterogeneity is by giving up the categorical approach. Tabb (2019), for example, argues against understanding psychiatric disorders as mechanistic kinds based on recent developments in precision-oriented psychiatric research (such as RDoC). According to her, these approaches are based on the idea that instead of the current diagnostic categories, we should have smaller

[42] Kendler et al. (2011) argue that the MPC account can accommodate Boorsboom's (2008) causal systems perspective.

units to intervene on. These worries, and their proposed answers, are related to a more theoretical problem, which I will consider next.

3.3.3 Individuation Problem of the HPC view

I have argued in favour of the property cluster view of natural kinds. However, the HPC view faces a thorny challenge with special science kinds, especially psychiatric disorders. The challenge concerns our ability to individuate or specify a property cluster by finding a non-conventional way to extract mechanisms from each other and the surrounding environment. When this challenge is answered pragmatically by appealing to explanatory relevance, we are faced with another problem of when to lump and split types of mechanisms. In this section, I provide an overview of the problems. In the next section, I present alternative approaches to kinds that could be employed to answer these problems. In the chapter 4 I formulate my own thesis as an answer.

The individuation problem

The dual nature of the HPC view, according to which natural kinds consist of both co-occurring property clusters and underlying homeostatic mechanisms, faces an individuation problem. The problem is whether it is the underlying mechanism(s) or the property clusters that determine the boundaries of natural kinds. Trying to answer this problem exposes an apparent circularity that underlies the reasoning behind the HPC view. That is, although homeostatic mechanisms are invoked to explain the similarity of property clusters, those properties may vary, and therefore we need to rely after all on homeostatic mechanisms to decide when two property clusters are similar enough to

count as the same natural kind (Ereshefsky 2010: 677, Tobin 2018). Consequently, the individuation problem seems to demonstrate the apparent inability of the HPC view to account for the heterogenic nature of natural kinds, the very feature that supposedly makes it superior to the essentialist accounts of natural kind.

The most straightforward answer to the individuation problem would be to rely on a strong interpretation of the accommodation thesis by arguing that "nature's joints are located at the boundaries of mechanism" (Craver 2009). However, as Craver argues, this would lead to problems in deciding when to split and lump a property cluster.

The strong mechanistic interpretation of the HPC view would mean that we should be able split and lump together property clusters (i.e. natural kinds) based on their underlying mechanisms. This means that property clusters that are maintained by more than one mechanism should be split, while two or more property clusters that are maintained by a single mechanism should in turn be lumped together. Crucially, Craver argues that the split and lump strategy demonstrates that the strong mechanistic interpretation of the HPC views leads to conventionalism because there are no interest-free methods to subtract mechanisms from the causal structure of the world. In particular, according to him, pragmatic considerations determine when two mechanisms are the same, and where one mechanism ends and another begins (Craver 2009: 575). The split strategy has been used by, for instance, Paul Griffiths (1997) when he argues that our folk concept of emotion in fact picks out three different kinds of mechanisms (affect programmes, central emotional processes, and socially constructed patterns of behaviour). The lump strategy has been employed, in turn, by Pascal Boyer (2011) when he maintains that religions and delusions are in fact maintained by the same psychological mechanism.

Craver (2009) illustrates the split and lump problem with three arguments. First, when different mechanisms underlie, produce, or

maintain a property cluster, the kind can be split or lumped together differently depending on which mechanism(s) one consults. This can be exemplified by major depression. Arguably, the same co-occurring symptoms of clinical depression can be caused by different etiological mechanisms, although the symptoms themselves are sustained by the same constitutive brain or psychological mechanism (cf. Craver 2009: 584). Therefore, as a putative psychiatric kind, major depression should be split if we consult the etiological mechanisms that lead to the manifestly similar symptoms, or lumped together if we address the psychological or brain mechanisms that maintain the symptoms. Or alternatively, the same etiological mechanism can give rise to a different combination of symptoms of clinical depression under different circumstances. This, in turn, would lead one to lump the depression kind based on its etiology, or split it based on its constitution. In sum, the point is that different combinations of mechanisms, and the emphasis put on them, lead to different ways of lumping together or splitting the kind.

The second challenge is how to sort token mechanisms under type mechanisms. The central benefit of the HPC view in contrast to essentialist accounts of natural kind is its apparent ability to account for property cluster variation and multiple realization. This means that different components and organizations of mechanisms should be able to realize the same type of mechanism. However, in cases where the property cluster varies, the mechanism(s) that realize it should also be different. This raises the question of which differences in property clusters and their underlying mechanisms are relevant enough to warrant splitting the kind. For instance, the properties of ADHD differ from individual to individual, which means that the underlying neurological mechanisms must differ as well. Hence, according to current research, ADHD could be split based on two different

neurophysiological sub-mechanisms (Campaner 2016: 125).[43] It could either be split into sub-kinds based on these constitutional differences, or the kind could be lumped by glossing over them. At any rate, according to Craver's argument, this is a conventional decision.

The third problem concerns the boundaries of mechanisms. According to Craver (2009: 589), the entities, activities, and organizations of causal mechanisms do not have clear-cut boundaries. Instead, their delineation depends on the goals and interests of the explanation. For example, the mechanism of action potential transgresses the physical boundaries of membranes: some components of the mechanisms are within and some outside the physical boundaries of the membrane (Craver 2009: 590). The boundary problem is especially problematic for psychiatric kinds, whose etiological and constitutive mechanisms intertwine and participate in feedback loops in complicated ways to produce and underlie symptom clusters. Raffaella Campaner (2016), for example, argues that the explanation of ADHD may require an extended model that situates low-lever mechanisms within a larger mechanism that encompasses causal factors from social and psychological levels.

The first two problems can be addressed by means of explanatory pluralism. Although different combinations of mechanisms give rise to different classifications of kinds, they may nonetheless match different classificatory goals. In the first case, if several etiological mechanisms generate the same property cluster that is realized by the same constitutional mechanism (or mechanisms), the kind can be equated with the property cluster and its constitution, or it can be split according to the etiological mechanisms. These would presumably support different generalizations, so that the classification would satisfy different explanatory goals. Nevertheless, unlike in biology where

[43] Currently, this is a conventional decision in Italy and other Western countries.

evolutionary history is important for classification, the etiology of a psychiatric disorder does not play a significant role. Perhaps an exception could be if the goal is to prevent disorders. In such a case, preventive interventions on the etiology of disorders might play an important role.

In the second case, when property clusters are realized by different sub-mechanisms, we simply need to decide the level of description that we are most interested in. For example, if the property cluster of depression can be realized by a slightly different mechanism, the umbrella property cluster grounds generalizations and need not be eliminated. In this case, we can argue that the superordinate term of depression picks out a kind cluster that supports generalizations as do the underlying mechanisms separately (i.e. sub-depressions).

The third, the boundary problem, is trickier to answer. Craver's arguments are ultimately based on his own causal account of mechanisms (which is partly based on Woodward's interventionalism, see Chapter 4). He argues that for mechanistic explanations the splitting and lumping mechanism is not a problem because modelling the components, activities, and organization of mechanisms is guided by causal and explanatory relevance set by the explanandum phenomenon (Craver 2007: 144). In Craver's (ibid.: 123) words "The boundaries of mechanisms – what is in the mechanism and what is not – are fixed by reference to the phenomenon that the mechanism explains." However, delineation is a problem for the HPC view because depending on which properties of a putative kind we consider relevant may lead us to split or lump together the kind differently (Craver 2009: 584). This implies that there is no interest- or value-free means to carve the kinds, because how the explanandum phenomenon is specified depends on our explanatory interests. On the other hand, this is not necessarily a problem if it is granted that the delineation is partly an interest- and value-laden question. Indeed, I will argue in the following chapter that

this enables an ecumenical, non-metaphysical, and pluralistic account of property cluster kinds.

Further, the boundary problem does not seem to be a universal question because many kinds have clearly identifiable property clusters that can fix the explanandum, and thereby the explanatory mechanism. Hence, one plausible answer to the individuation problem is offered by naturalism. Looking at how sciences classify things, the problem is clearly more relevant to some kinds than others. The inductive power of some special science kind concepts is a testament to the success of their delineation. In this light, the individuation question does not so much undermine realism about special science kinds than it concerns our ability to fine-tune the HPC view to account for the undeniable inductive success of these kinds. Moreover, I wish to emphasize that explanatory choices are governed by classificatory aims. In psychiatry, there are cases where explanatory interests, such as whether the explanations are to serve preventive or therapeutic interventions, may lead to different kind delineations. Moreover, although the way the explanatory questions are framed in psychiatry may be partly governed by value-laden aims, the answers to these explanatory questions are nonetheless provided by nature.

The strong mechanistic interpretation of the HPC view leans toward the essentialist (or the laws of nature) account of natural kinds. This means that the old problems that previously caused difficulties appear again in novel disguises. However, the whole point of the HPC view is to provide an ecumenical, family-resemblance approach account of kinds, not to provide the old approach to natural kinds in new guises. Moreover, the discussion over mechanisms is technical and at times ambiguous, and when it is applied in order to understanding the HPC view, it carries over those ambiguities.

Nevertheless, it has to be admitted that the questions mentioned are not merely academic problems in understanding psychiatric kinds,

whose delineations do not have a successful track record due to their heterogeneity. The DSM and ICD classification manuals, for instance, are commonly accused of comorbidity and blurring the distinction between normal health and disorders. In consequence, it would seem that the individuation problem either requires us to provide a better defence of the strong mechanistic approach to natural kinds or to bite the bullet and opt for a more pluralistic and promiscuous view of kinds. In this light, I will next provide an overview of different theories that are purportedly better equipped to answer the individuation problem. After going through some of the challenges they face, I will provide my own answer by arguing that a larger role should be given to property clusters rather than concentrating solely on responsible mechanisms. I do this by arguing for a non-metaphysical interpretation of mechanistic explanation and pointing out how non-epistemic values play a role in how property clusters and mechanisms are specified.

3.3.4 Alternative Ecumenical Approaches to Kinds

Alternative ecumenical theories of natural kinds may not have the same problem with kind specification. In the following section, I examine other ecumenical theories of kinds based on their potential ability to avoid the individuation problem. I distinguish between theories that do not rely on mechanisms and those that do not rely on property clusters. The former approach is notably taken by Slater, Griffiths, and Dupré, whereas Khalidi's simple causal view and Millikan's etiological views can be understood as examples of the latter. In the light of the naturalistic approach to kinds, the central question they should answer is whether mechanisms and causes are needed to explain and ground inductive inferences.

An obvious first strike at answering the individuation problem would be to manage without mechanisms altogether. According to the HPC account, homeostatic mechanism(s) do the ontological work of binding properties together into a co-occurring cluster, and hence also explain why our inductive inferences work. But since we run into difficulties in trying to individuate the kinds based on mechanisms, the explanation for inductive success could be structured on the property cluster instead. Such approaches have been provided by Dupré (1993) with his promiscuous realism, and lately Matthew Slater (2015), who replaces mechanisms with stability in his theory of stable property clusters (SPC).

John Dupré (1981, 1993: 83) argues for a promiscuous realism that supports a pluralistic view of property clusters. According to his thesis, natural kinds should be understood as property clusters that can be individuated in cross-cutting ways. In other words, the world consists of clusters of properties "mapped onto a multidimensional quality space" (see Cooper 2005). (In later work he has restricted promiscuity, which is the version I am considering here.) These property clusters come in different levels of fine-grainedness on the one hand, and different disciplines may focus on different similarity relations between the properties on the other. For instance, whereas fish form a similarity bundle on the genus level, so do different sharks on a more fine-grained level. Moreover, mental states may not be considered physical kinds, but they may nonetheless be scientific kinds for psychology.[44]

[44] A more restricted theory has been offered by Griffiths (1999), who nonetheless privileges the property cluster over its mechanistic explanation (see also Murphy 2006). An even more promiscuous view is Zachar's (2014) pragmatic approach to psychiatric kinds. His view is maximally ecumenical, since according to it, any

Promiscuous realism has been criticized for not being able to distinguish between folk-classification and scientific classification on the one hand, and between accidental and natural property clusters on the other (Khalidi 2013). First, in his earlier writings, Dupré (1981) maintained that folk classifications are legitimate candidates for natural kinds. He argued, for example, that classifying whales as fish is no less natural than classifying them as mammals. However, this view has a problem with explaining why we do not settle with folk classifications but instead seek scientific ones. The reason, according to Khalidi (2013), is that although folk classification and scientific classification share some of their epistemic aims, ultimately scientific taxonomies serve our epistemic aims better. Consequently, Khalidi (2013) argues that it does not seem reasonable to place folk taxonomies on the same footing as scientific ones. I agree with Khalidi that this is a challenge for promiscuous realism, but I also believe that it only demonstrates that some classifications are more natural than others. For example, if the concept of a whale was only important in biological explanations, classifying them as fish would not be legitimate. However, in line with Dupré's account, since classifying kinds is partly goal-relative, classifying whales as fish may play an epistemic role, for instance, in the fishing and cooking industries. This means that a classification that lumps whales together with fish due to their similar phenotype, can ground some inductive inferences that are important for fishing or cooking purposes, although admittedly very few that a classification of fish that excludes the mammal would not. Importantly, these are different epistemic projects that are not in conflict with each other, and therefore the choice between them is not based on values (cf. Griffiths 1997: 199). Moreover, the similarities between whales and fish are due to the

non-arbitrary classification is taken to correspond with the natural kinds structure of the world.

fact that during their evolution they have shared some environmental pressures. These can be understood as extrinsic mechanisms responsible for their shared properties. And since I have already pointed out that extrinsic mechanisms can be as responsible for property clusters as intrinsic ones, there is no a priori reason to rule out the whale-fish group as a weak property cluster kind. As a result, this line of reasoning implies that although some kind groupings ground more inductive inferences than others, those weak kind groupings may nonetheless be legitimate for some epistemic goals.[45]

A greater problem facing promiscuous realism is that properties may co-occur accidentally (Cooper 2005). At first sight, for example, it may seem that all the small and salty cucumbers placed in the same basket at my local grocery store comprise a natural kind apart from the regular non-salty and relatively big cucumbers. However, upon further research, the clustering of these properties may turn out to be merely accidental or they may comprise a conventional class of fermented pickled cucumbers. If they are an accidental collection of cucumbers, their classification would have no counterfactual power. On the other hand, if they are fermented pickled cucumbers, their properties have clustered together due to socioculturally-driven culinary needs, and thereby could support some inductive inferences. In the case that they truly were a biological species of their own, they would ground robust inductions. At any rate, whether the clustering of properties (smallness, saltiness, etc.) ground inductions due to biological or social reasons, or are the result of an accident, merits further investigation. Importantly, based on the forgoing example, it is the potential property binding cause or causes that lends inductive power to kinds, such as to a biological or

[45] However, biological explanation is more robust in two ways: it is more stable under various counterfactual situations, and it explains the clustering of more properties.

even a pickled cucumber, in contrast to the accidental grouping of "small cucumbers".

One way to try to avoid the problem of accidental groupings is by concentrating on the nature of the stability that grounds epistemic projects. This is the line that Matthew Slater's (2015) account of natural kinds as Stable Property Clusters (SPC) takes by separating accidental property clusters from genuine homeostatic ones based on the specific nature of the latter (Slater 2015: 96). The idea is that by giving up on the idea of the causal explanation for the stability, we can also avoid the explanatory circularity that it brings to the delineation of the property cluster. In the place of the mechanism, Slater offers cliquish stability of the property cluster that plays the epistemic role of uniting property clusters. By cliquish stability, Slater means that once a subgroup of the cluster is encountered, it is probable that the other properties of the cluster are present as well. Probability entailment of properties in the cluster, in turn, is understood counterfactually, so that the properties in a cluster cling stably to each other over a relevant set of counterfactuals. Furthermore, the relevant range of counterfactual stability is determined by the discipline relevant goals of the delineation. Consequently, natural kind is merely a domain-relevant status given to property clusters rather than a robust ontological category.

Therefore, the apparent advantage of concentrating on the nature of stability over the HPC view is that it retains the crucial distinction between property clusters that ground inductions and merely accidental or subjective groupings or delineations. At the same time, this concentration on the nature of stability does not succumb to the circularity of the delineation problem affecting the HPC view due to its dual nature of including both mechanisms and property cluster. Slater does not reject causal grounds completely, but states instead that cliquish stability can be multiply realized by, for example, homeostatic mechanisms, strict essence, or historical essence. Thus, his view is that

"uncovering certain homeostatic mechanisms underlying the stability of some properties can contribute to the construction of epistemically fruitful classification (…) without supposing that such identification is necessary" (Slater 2015: 403). In other words, as he admits, stability should be taken merely as a "brute fact" (Slater 2015: 395).

However, my argument against the SPC view is that property clusters' causal grounds are a necessary part of natural kinds' epistemic and scientific value. Slater's account falls short in explaining what separates the epistemic value of scientific classifications and folk classifications and therefore he is unable to explain the inductive success of scientific classifications (cf. Martinez 2017). As the history of science demonstrates, knowledge about the cause for the explanandum phenomenon has enabled generalizations over different contexts.[46]

A historical example of the discovery of syphilis's causal ground serves as a case in point. Syphilis was one of the most devastating diseases encountered in mental hospitals until the discovery of penicillin at the beginning of the 20th century. Syphilis was called the "great imitator" because of its variable clinical course and diversity of manifestations (Carneiro et al. 2013). Syphilis is an infectious disease caused by the *Spirochete* bacterium that influences the central nervous system and is transmitted primarily by sexual contact. It was only when the bacterium, and the causal mechanism by which it produces the symptoms, were discovered, that the various courses and manifestations of the disease were lumped together under the same treatable kind heading. As a response, Slater could argue that it was not the disease mechanism, but the spirochete bacteria that was discovered and facilitated the development of treatment. The idea would be that the bacteria in fact constitutes the core properties, not the symptoms it

[46] Griffiths (1999) argues that kinds based on mechanisms are maximally predictive (see also Pober 2013 and Murphy 2017).

causally brings about. However, what are considered the explanantia properties and what are considered the explanans, depend on our explanatory interests. And since here the aim was to explain the symptoms and course of the disease, they should therefore be considered to be the explanandum property cluster. Moreover, it was the discovery of the specifics of the causal mechanism by which spirochete bacterium produced syphilis that enabled us to truly understand and predict the dynamic nature of the symptom cluster. So, although it may be that the discovery of the bacterium already enabled the development of penicillin treatment, it was the discovery of the disease mechanism that enabled precise predictions and an understanding of the disease.[47]

Consequently, the causal ground of a kind sets its applicability domain for epistemic projects. In other words, the counterfactual stability of a property cluster is made explicit by knowing its causal ground. By knowing what causes a property to cluster can we explain and predict the kind's behaviour under different counterfactual circumstances. That is, while the SPC view only presupposed the counterfactual power of kind classifications, scientific research advances by discovering the causes for that power. Hence, although the individual problem is an interesting philosophical and theoretical challenge, it does not cast doubt on the empirical fact that mechanistic explanations in the history of science have endowed scientific kinds with epistemic value.

In sum, I have argued that promiscuous realism and the SPC view are problematic because they do not provide reasons why kinds support inductive inferences. This point is especially problematic because knowledge about the causes of a property cluster enables us to extrap-

[47] Eduardo Martinez (2017) offers a similar critique, although it is based more on predictability.

olate the kind to different circumstances. In addition, promiscuous realism is problematic because it cannot explain why classifications ground inductive powers to different degrees. In the light of these objections, there needs to be something that explains why, how, and to what extent property clusters warrant inductions. This is precisely the role that mechanisms in the HPC view play.

Alternative Approaches to Causal Ground

Another way of addressing the individuation problem could be offered by providing a different analysis of the reasons why property clusters support inductive inferences. Muhammed Khalidi's (2013) simple causation theory of kinds does this by offering an ecumenical causal analysis of natural kinds, whereas Ruth Millikan's (1999) view of historical kinds provides an account based on etiological mechanisms.

Khalidi (2013: 78, 2015) claims to follow Craver's criticism to its logical conclusion by giving up the idea of mechanisms in favour of a more permissive simple causal view. He argues that the idea of a homeostatic mechanism is too restrictive to be applicable to all kinds that support inductive inferences, because inductions are possible in some cases where there is no homeostasis or a single mechanism that would explain it. Rather, Khalidi argues that the co-occurrence of property clusters is better explained by causal threads that end and start in concentrated nodes or hubs. In support of this idea, he points out that kinds enable projections not only from causes to effects, but also from effect to their causes, and to common causes. For example, once we know that something is a piece of gold, we can project that is malleable, has a certain melting point, and has a crystalline structure. According to Khalidi, these projections are based on causal relations between properties rather than relying on a single underlying mechanism. The upshot is that natural kinds ground inductive generalizations because

they are "highly connected nodes in a causal network" instead of being produced and maintained by causal mechanisms. These nodes, in turn, explain why some kinds enable more inferences than others. The more causal threads leave and end in a node, the more robust inductions it enables.

It seems to be that Khalidi's account is ambiguous whether the causal nodes should be understood as being similar to a mechanism, or as having given up the role that mechanisms (understood as networks of causal dependencies) play in explaining the co-occurrence of property clusters. If the former is the case, I fail to see their advantage over mechanisms. If the latter is the case, Khalidi's account seems to be troubled by the same problem as the SPC view. That is, in exchanging mechanisms with causal threads, the view seems to lose its ability to explain why properties cluster together in the first place. The central idea behind a mechanism is to "open up" the black box between the cause variable and the effect variable. Hence, if we abandon the idea of a mechanism, we also give up the ability to explain why causes and effects are connected, and thereby also the knowledge of when an inference between them is warranted. For Kahlidi's account, this would mean that we would not be able to explain why certain "causal nodes" tend to concentrate together. In general, we would not have a sufficient explanation as to why the kind concept grounds inductive generalizations.

Millikan offers an account of kinds studied by the special sciences by arguing that only a certain type of mechanism grounds inductive generalizations (e.g. Millikan 2005, see also Khalidi 2013: 75). She argues that "historical kinds" or "copying kinds" should be distinguished from "eternal kinds" typically found in the natural sciences. Both kinds have the ability to ground generalizations, but for different reasons. Whereas "eternal kinds" of the natural sciences share properties because of the laws of nature, the kinds found in the special

sciences share properties due to common etiological copying processes or mechanisms. According to Millikan, these include the kinds targeted by the classifications in the social sciences and biological sciences, as well as some artificial kinds. For example, individual members of a biological species are similar because they have descended from the same ancestor (Khalidi 2013: 136) through a copying mechanism. In particular, Millikan (1999: 55) argues that the historical copying process has two central features. (1) All the kind members are reproduced or copied from the same model. (2) The copying process is usually sustained by a historical environment and function, which is preserved by the copying process (Millikan 2005: 307-308, Khalidi 2013: 136). Hence, the historical-copying process is not simply a form of etiological causation, but instead historical kinds should be individuated based on their similar causal powers, which are due to both historically shared copying processes and environmental processes (Khalidi 2013: 137).

The historical-copying account of kinds, nevertheless, clearly does not fit all the kinds classified in the special sciences. This is especially the case with psychiatric kinds. The reason is that historical-copying mechanisms and constitutive mechanisms may pull towards different ways of delineating kinds, just as Craver (2007: 584) argues that etiological and constitutive mechanisms do. Moreover, although Millikan would have us delineate the kind based on the historical process that brought about the kind (e.g. some dysfunctional biological process) and environmental processes together, many psychiatric kinds are in fact grouped together based on their synchronic (constitutive) causal structures instead.[48] For example, clinical depression can be split or lumped together depending on which mechanism one consults.

[48] It may be a stretch to explain the etiology of psychiatric disorders with Millikan's copying mechanisms. Although, perhaps in cases where unconscious copying plays a role this can be justified (see Cooper 2013, Hacking 2010).

Although the etiology of major depression may include physical and meaning-based causes (such as light deprivation or meaningful trauma, see Arpaly 2005), the symptom cluster of the disorder may still be the same. In other words, Millikan's copying mechanism idea would lead us to split depression into two different kinds. However, perhaps it is clinically more useful to lump the kind together based on the shared synchronic mechanism that constitutively explains the symptom cluster. Crucially, it should be left to empirical research, and its particular epistemic aims, to decide between these groupings (see Chapter 7).

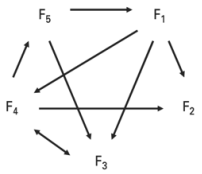

Figure 1. (Kendler et al. 2011).

This point can be further stressed with the network model of mental disorders. Denny Borsboom (2017: 82), who endorses the theory, argues that there need not be an explanatory etiological or constitutive mechanism underlying the symptom clusters of mental disorders, but instead the causal interaction between the symptoms themselves can sustain the syndrome.[49] Similarly, Kendler et al. (2011) argue that the symptom cluster can be understood as a mechanism (see Figure 1). For example, according to the network account, schizophrenia could be the sum of its reinforcing symptoms, so that, for instance, hallucinations produce delusions (Kendler et al. 2011: 1147). It is plausible that the

[49] Boorsboom has said that there cannot be symptomless mental disorders.

network model explains at least some psychiatric disorders. As a consequence, since according to this account of mental disorders it is the combination of symptoms and their causal relations that constitutes the causal mechanism, a purported preceding copying mechanism would not explain why the kind supports inductive inferences.

In sum, these alternative ecumenical approaches to kinds fall short of explaining how property clusters are glued together so that they can ground epistemic projects. In particular, I have argued that they have a problem in distinguishing between an accidental and a scientifically legitimate property cluster, and between degrees of epistemic usability of kinds. However, as Cooper (2005) and Murphy (2006) have pointed out, there is no a priori reason to believe that one theory of natural kinds fits to explain all psychiatric kinds (see also Hacking 1991). Instead, it is more plausible that the kinds vary, and therefore some accounts may be better fitted to explain some kinds than others to the extent that some of those may fit with the mentioned theories. Nevertheless, although I have argued in favour of the HPC view as a general account of scientific kinds, a better account of the connection between inductive explanation and homeostatic mechanism is needed to understand how property clusters reliably ground inductive inferences.

4 The Applicability Domain Account of Psychiatric Kinds

I have argued that the homeostatic property cluster (HPC) view matches the way kinds are employed for inductive inferences and generalizations in the special sciences, including psychiatry. Nevertheless, I also described how the view has an individuation problem over how to draw the boundaries of mechanisms and property clusters. I also pointed out the value-ladenness of the concept of psychiatric disorder, the heterogeneous and fuzzy nature of the particular disorders, as well as the modest inductive success of their classifications. I will now formulate and defend a domain-relative explanatory solution to those challenges. It is based on the dual role that homeostatic property cluster kinds can play in psychiatry. While property clusters support inductions, they also call for mechanistic explanations of how, and to what extent, they support those inductions. My aim is to provide an account of mechanistic explanation that is fitted to specify the range over which kinds and their property variations are explained.

I begin by examining the philosophical discussion about mechanistic explanation and argue that the contrastive counterfactual theory of explanation can provide a conceptually refined account that fits my purposes. Although mechanistic explanation is common in all the special sciences, such as biology and the social sciences, it lacks an agreed general philosophical account. Therefore, I will first provide an overview of mechanistic approaches, then provide a permissive philosophical account of mechanistic explanation based on the contrastive counterfactual theory, and supplement it with the heuristic idea of causal relevance based on invariance under interventions. In particular, I point out that the benefits of applying the contrastive counterfactual theory in order to understand mechanistic explanation in psychiatry are that it provides a non-metaphysical account of

explanation, an account of explanatory relevance with contrasts and counterfactual dependency, and a non-absolute account of explanatory stability, as well as severing explanations from predictions. This makes it capable of being employed to understand how various factors, such as genetic predisposition, psychiatric humiliation, and economic situations, can contribute to causing and maintaining psychiatric disorders. I then apply the contrastive counterfactual theory to disentangle causal and constitutive explanations in psychiatry and introduce a heuristic approach to modularity requirement to mechanistic explanation in psychiatry.

In the second section, I formulate my own account of psychiatric kinds. It is based on the idea that mechanistic explanations of property cluster kinds have specific domains of applicabilities that represent the range over which mechanistic explanations can account for kinds' aspects in alternative situations. The idea is that the domain can be supplemented by knowing more about the various causal mechanisms that are responsible for the property cluster, thereby increasing the identifiable range of reliable projections the kind-category supports. To explicate these domains, I employ the contrastive counterfactual theory of explanation. I argue that for a property cluster kind, such as a psychiatric kind, to be useful for classifications and inductive inferences, its mechanistic explanation should support counterfactual information so that it represents the causal structure that causes or realizes that kind. That is, a mechanistic explanation of a property cluster kind should be able to answer what would happen to specified aspects of the property cluster within specified domains if the responsible mechanism(s) were to change in specified ways. The central motivation for this account is to provide a non-reductive and non-metaphysical heuristic account of psychiatric kinds so that different discipline-relative explanations can be understood to supplement those explanatory domains. Moreover, since according to

the contrastive counterfactual theory the explanandum phenomenon determines the appropriate mechanistic explanation, and the identification of psychiatric phenomena is partly value-relative, so too is the specification of responsible mechanistic causal structures.

Finally, I explore discussions about the possibility of integrative explanatory pluralism, or partial reductionism, that have been suggested in relation to the medical model. This is an important question because there are multiple competing and partly incompatible explanatory approaches to psychiatric disorders. However, in contrast to these approaches, I suggest that the applicability domain approach offers a superior non-metaphysical account of psychiatric kinds according to which they can function heuristically to facilitate interaction between various disciplinary-related explanations.

4.1 Mechanistic and Causal Explanation in Psychiatry

4.1.1 Mechanisms in Science and Psychiatry

Mechanistic explanation plays a central role in psychiatry, as it does in the special sciences in general, to the extent that philosophers and psychiatrists alike believe it can facilitate a multifactorial and realistic approach to psychiatric disorders and their classification (see Campaner 2016: 115). However, there is no unified philosophical account of mechanistic explanation, a fact which also reflects, as I argued in the last chapter, on the HPC view of natural kinds, including Kendler et al.'s (2011) mechanistic account of psychiatric kinds. In view of this, I believe that a more fine-grained explanatory approach of the contribution of different causal factors to psychiatric kinds within a mechanistic framework can be formulated. In this section, I will first provide an overview of mechanistic explanation in philosophy and

science, then an account of explanation based on the contrastive counterfactual theory, and finally, I will supplement this account with an interventionist approach to explanatory relevance.

In the philosophy of science, mechanistic approaches and analyses of explanation have become increasingly dominant since the 1990s. Causal mechanistic approaches to explanation in philosophy grew partly out of the need to account for the shortcomings of the deductive-nomological model (DN model or covering law model) of explanation by logical empiricism (Hempel 1965). In the DN model, an explanation is a deductive or statistical argument that has a particular event as a conclusion and a general law and initial conditions as its premises (Hedström and Ylikoski 2010: 54-55). This means that explanation takes place by subsuming a particular event under a general law. The model is widely rejected primarily because it is too wide according to some accounts while being too narrow to others (Halina 2017: 216). The model is too wide by making none-explanatory things explanatory. For instance, according to the DN model, a forthcoming storm can be explained by the barometer dropping. It is too narrow because non-law-based explanations are excluded, such as those employed in biology and the social sciences. Also, it fails to account for the asymmetry of explanatory relations, so that the length of the flagpole can be explained by its shadow (Salmon 1989: 103). Moreover, the DN model allows irrelevant premises, such as that the consumption of contraceptive pills by a man can explain why he does not become pregnant.

Causal mechanistic explanations can overcome these problems because they are based on revealing networks of causal dependences in the world. The barometer does not explain the storm, or the shadow the length of the flagpole, because these purported explanations do not track the direction of causation. Moreover, causal mechanistic approaches to explanation match how explanation is conducted in the special sciences, such as biology and psychiatry, which have exceptions

and are not based on the laws of nature (Halina 2017: 216). That is, although there are no exceptionless natural laws in psychiatry that would explain psychiatric disorders, we have been able to formulate (partial) mechanistic explanations to account for some of their aspects. In sum, unlike the DN model, philosophical approaches to mechanistic explanation follow closely how explanation is empirically done in the sciences.

The central tenets of the approaches to mechanisms and mechanistic explanation are captured by the minimalist account according to which "a mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon" (Ilari and Williamson 2012: 120). Thus, according to this account, a mechanism consists of causally interacting parts organized so that they explain either the production, maintenance, or constitution of the targeted phenomenon.[50] The minimalist view encompasses four central features of mechanistic explanation (see Hedström and Ylikoski 2010). First, mechanisms are identified by their effects, so that a mechanism is always a mechanism *for* something (Darden 2006, see also Craver 2007: 123). Second, mechanisms have a structure that produces the phenomenon of interest. The idea is that mechanisms explain by providing how answers to why-questions by opening the "black box" between the causal variable and the effect variable, the cogs and wheels of the causal dependency. Third, the elements in a mechanism are explained, in turn, by other mechanisms, so that mechanisms form a hierarchy. The idea is that elements at one level can be explained by mechanisms at a lower level of description or organization (Craver 2007: Ch. 5). This means that mechanistic explanation or modelling requires a "bottoming out" of the relevant

[50] Slightly different formulations have been offered by the new mechanists. For example, see Machamer et al. (2000), Glennan (2002), Bechtel and Abrahamsen (2005).

entities and their mutual organizations and interactions for the mechanism (cf. Woodward 2015, Campaner 2016). A mechanistic representation, therefore, necessarily picks a level of description for the explanation. And finally, mechanism is inherently a causal notion. It explains by providing information about the activities of the elements in a causal process that produces the phenomenon of interest.

The mechanistic approach to explanation has been influential in psychiatry.[51] As an example, Kendler (2008) argues that explanatory problems caused by the heterogeneity of psychiatric disorders can be overcome by applying a mechanist approach. He points out that a mechanistic approach provides a middle way in between reductionism and emergence in explanation (Kendler 2008: 696). Similarly, Kendler et al. (2011) argue that psychiatric kinds can be identified with shared physiological mechanisms that are generated and maintained by their symptoms as well as the "mechanisms investigated by the molecular, physiological, computational, psychological and social sciences" (p. 1148). Part of this appeal for mechanistic explanation in psychiatry is that there is no common theory that would unite all explanatory approaches. Moreover, mechanisms are often proposed the end goal of an explanation of a psychiatric disorder, rather than relying on those mechanisms to support explanations and inductions themselves. For example, currently the reference to a mechanism for ADHD or schizophrenia should be understood more as a heuristic call to explain those proposed mechanisms, rather than knowledge of what those mechanisms really are. Moreover, mechanisms can facilitate interaction of different explanatory approaches whose interests are in different factors on different levels of description. In sum, mechanisms can facilitate research on heterogeneous factors in different disciplinary approaches, they can function as heuristic suggestions for further

[51] See Campaner (2011) for mechanistic explanation in medicine.

research, inform therapies, and – my primary interest here – support classifications.

Mechanistic explanations can be divided into constitutive explanations and causal explanations. While etiological causal mechanistic explanation is meant to capture the causal process that brings about the effect phenomenon of interest, constitutive causal mechanistic explanation describes the elements, activities, and organizational features that give rise to the phenomenon of interest (Hedström and Ylikoski 2010). This means that an etiological explanation is required when we are interested in how an effect phenomenon is brought about by the interaction and activities of the intermediating mechanism's parts. Recent philosophical discussions over mechanisms (i.e. new mechanists), in contrast, have typically concentrated on constitutive mechanistic explanation, so that the phenomenon of interest is taken to be constituted by the activities, organization, and interactions of parts of a mechanism as a whole (see Glennan and Illari 2018 and Figure 2).

Although my account is meant to accommodate the general idea underlying various types of mechanistic explanations, it requires simplifications since different disciplines employ mechanistic explanations that have different commitments. For instance, while cell biology (Bechtel 2006) and neurosciences (Craver 2007) study integrated systems by decomposing and localizing them into concrete components, mechanistic explanations in biology and the social sciences study types of interaction through abstraction (see also Kokkonen 2021: 67; Sarkia et al. 2020). According to Kuorikoski (2009) the difference lies in whether the causal properties of the mechanisms are taken to be monadic or relational. Social mechanisms, for instance, are composed of agents (i.e. individuals or groups) and their interactions, but unlike the components of the neurosciences, they do not have intrinsic causal powers that would allow them to be studied in isolation. Rather, the

causal power that these agents and their interactions exhibit are relational, so that they depend on other individuals and social contexts (Kokkonen 2021). Hence, according to Kuorikoski, the explanatory aim in social mechanistic explanation is not decomposition, but rather abstraction by intentionally leaving out some non-relevant aspects of the system. This also means that while components of the system – the causal properties of agents and their interactions – cannot be replaced and studied individually, those individuals can nonetheless be replaced with other individuals as long as their context remains the same.

In contrast, constitutive explanations of decomposable monadic systems are common in medicine and psychiatry (i.e. the new mechanist approach) (see Levy 2013). They are based on decomposing the system into constituent parts, with the aim of trying to understand each part separately, and thereafter integrating them to see how they jointly produce the complex mechanism of interest (Kendler 2008, see also Bechtel 2019). Although this means reducing the mechanisms into parts and organizations and their interactions, the approach is not reductionist in a hard sense because a mechanism is always part of a larger context, and its organization requires an understanding of the phenomenon as a whole (Kendler 2008: 696). Rather, the idea is that smaller units can be studied in isolation and reorganized to see how they function together. The following figure (2) illustrates constitutive relation between a phenomenon and the mechanism that realizes it.

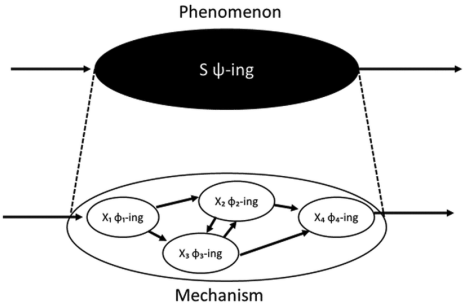

Figure 2. (Craver 2007)

At the top of the figure is the phenomenon that illustrates the explanandum, which is explained by the underlying organization of the entities (circles) and activities (arrows). That is, "S's φ-ing is explained by the organization of entities $\{X_1, X_2, ..., X_m\}$ and activities $\{\phi_1, \phi_2, ... \phi_n\}$ (Craver 2007: 7). The interlevel relation between the phenomenon of interest and the underlying mechanism is constitutive, while the intralevel relations between the entities of the mechanism are causal (Craver and Betchel 2007).

Despite these differences between mechanistic explanations, my aim is to provide a mechanistic account that can help to understand different explanatory approaches to psychiatric disorders. That is, because of the plurality of explanatory approaches, it is important to find ways to understand how these approaches can supplement the explanation of different aspects of the same psychiatric disorders. Therefore, for the most part, I rely on a minimalist approach that can facilitate various explanations with mechanisms. As Kuorikoski (2009: 151-152) points out, relational and monadic mechanistic approaches are compatible with the minimalist view of mechanism: "they are structures performing functions by virtue of their component parts, component operations, and their organization, and the orchestrated functioning of

these mechanisms is responsible for one or more macro phenomena". For now, I will not delve into the details of different approaches to mechanistic explanation further, but instead analyse mechanistic explanation based on explanation in general. Nevertheless, I will return to the distinction between constitutive and causal explanation when I have provided a better understanding of explanation in general.

4.1.2 Contrastive Counterfactual Theory of Explanation

A conceptually refined account of mechanistic explanation requires a general view of explanation. This kind of view is missing in some philosophical approaches to mechanisms (see Craver 2007: 105). The contrastive counterfactual theory (CC theory) of explanation (Ylikoski 2001, Woodward 2003) is naturally suited to making understandable what makes mechanistic descriptions explanatory (Kuorikoski and Ylikoski 2010, Ylikoski 2012, Pöyhönen 2013b, Kokkonen 2021). The reason is that it provides an analysis of causal and explanatory relevance of mechanistic representations in terms of their ability to answer counterfactual questions. In my argument below, I follow especially Ylikoski's (2001) and Ylikoski and Kuorikoski (2010) account of the CC theory. It is largely similar to Woodward's (2003) version, namely the interventionist account of causation and causal explanation, with the exception that its emphasis is more on explanation rather than causation.

As mentioned, the DN model has been largely rejected in favour of causal and mechanistic explanation.[52] A counterfactual interpretation of causation can be traced back to Hume (1748 Section VII), who writes: "Or in other words, where, if the first object had not been, the second

[52] The DN model is partly based on Hume's regularity theory of causation, which is his alternative to the counterfactual account.

never had existed." This idea can be represented with the lighting of a match: if the match had not been struck, it would not have been lit. The striking of the match is a *difference-maker* between it being lit and not being lit, and thereby is causally and explanatorily relevant for the explanandum phenomenon that the match was lit. While counterfactual theories of causation (e.g. Lewis 1986) have met with problems (see e.g. Psillos 2002), a nonreductive counterfactual theory of causal explanation can largely overcome them.

According to the contrastive counterfactual theory (CC theory) (Woodward 2000, Ylikoski 2001), explanations are answers to why-questions that specify real and objective causal (and constitutional) dependencies in the world (Woodward concentrates on causal explanation). The counterfactual idea of the theory is that explanations provide descriptions of modal dependencies that answer *what-if-things-had-been-different* questions (*what if*-questions). Moreover, according to the theory, explaining is an epistemic activity that requires conceptualizing the targeted phenomena or event in a certain way. This means that explanations hardly ever explain the whole phenomenon, only some aspects of it. For this reason, explanations are ideally spelled out in a contrastive structure: they answer questions of why fact rather than foil (e.g. Garfinker 1981, van Fraassen 1980, Lipton 1991, Ylikoski 2007, Kokkonen 2021). The idea is that such explanations answer why-questions that identify the topic explanandum within a contrast class (Kokkonen 2021). The contrastive structure makes explicit the aspects to be explained and thereby determines whether a putative explanation is relevant. Ideally, the explanans is also specified contrastively, so that it identifies in contrastive terms the cause of the explanandum topic in the contrastive class (Schaffner 2005, Kokkonen 2021). In sum, according to the contrastive counterfactual theory, explanations describe how a change in the value of the explanans variable x_1 (instead of x_2, x_3, $x_{4...}$) would bring about a change in the

value of the explanandum variable y_1 (instead of y_2, y_3, y_4…) (Woodward, 1984, 2003, Ylikoski and Kuorikoski 2010, Ylikoski 2013).

So, for instance, if we want to explain some symptoms of a disorder, we would like to know why it has these specific symptoms in contrast to other or no symptoms. The explanation then needs to answer what makes the difference between the explanandum and its contrastive alternatives. Answering this question singles out the causal factors that are relevant for explaining the actual explanandum rather than the presented alternatives. This means that it makes sense to ask different contrastively specified explanatory-seeking questions of the same phenomenon to gain a better overall understanding of the different factors responsible for it.

In sum, the non-reductive (or non-metaphysical) approach of the contrastive counterfactual theory to explanation means that different factors, for example psychological and genetic ones, can function as legitimate causal factors in explanations. Moreover, the contrastive approach helps to make explicit whether a causal factor is explanatorily relevant based on whether it can contribute to answering contrastive questions about the explanandum. In addition, contrasts can help to specify the correct level of description for explaining specific aspects of a given psychiatric disorder. As Kuorikoski and Ylikoski (2010: 222) point out, lower-level factors are irrelevant for the given explanatory task if they cannot make a difference to the higher level variables. The reason is that a putative explanation that cannot answer more what if-questions, is explanatorily irrelevant. Based on this, Kuorikoski and Ylikoski say that "*explanations find their own level*" (p. 222). Finally, uncovering counterfactual dependencies need not provide robust predictions, but can instead progress piecemeal in describing the causal system (Hedström and Ylikoski 2010, Pöyhönen 2013b).

The Interventionist Approach to Causal Explanation

When the contrastive counterfactual theory of explanation is applied to causal explanation, it can be supplemented with Woodward's idea of causation as *invariance under interventions*.[53] According to Woodward, causal and explanatory relevance can be revealed by a hypothetical intervention that exposes how by manipulating cause variables we can systematically change effect variables. Hence, the account specifies counterfactual dependencies through manipulation. It is through hypothetical manipulation that we can determine whether a putative cause variable can *make a difference* to the effect variable of interest. As Hausman and Woodward (1999: 533) put it, "causes are levers that can be used to manipulate their effects". And crucially, the account makes understandable how explanations in the special sciences can be robust even though they are not based on laws of nature.

> A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by interventions. (Woodward 2003: 55)

This passage highlights three aspects of the interventionist theory of causation and causal explanation (see Risjord 2014). Since intervention is irreversibly a causal notion, the definition is meant to describe the central concept that is fundamental to many other causal notions, rather than providing an ontological account of the nature of causation in non-

[53] Woodward (2002) has also provided an interventionist account of mechanisms. The interventionist approach is endorsed in psychiatry by Kendler (2012), and Kendler and Campbell (2009), Woodward (2015), Campaner (2018).

causal terms (Woodward 2000: 201-202). The account insightfully captures the idea of how scientists themselves think of causal explanation as manipulation (Woodward 2003: 9). That is, the difference between mere historical description and explanation is that unlike historical descriptions, causal information allows objects to be brought about or changed by manipulating other objects. The second thing to bear in mind is that the interventionist theory ideally requires that other variables, i.e. background conditions (or "causal field"), can be held put while the direct cause is manipulated. However, this modularity supposition is problematic for complex causal structures, such as mechanistic explanations, because the complexity of the structure of the mechanism may rule out surgical interventions that would not alter other causes. Finally, since there are events that cannot be manipulated, e.g. historical or cosmological events, interventions should be understood counterfactually. These ideal interventions reveal how the effect variable would change if we were to intervene on the cause variable. This caveat renders the interventionist theory immune to the anthropocentric criticism that many variables in science, such as cosmological events, cannot be manipulated. Earlier manipulationist theories (e.g. von Wright 1971) were unable to account for this

According to Woodward, the notion of intervention in causal generalization can be further supplemented with the broader idea of *invariance*. The idea of invariance is that a generalization between two or more variables is invariant if it explains how the effect variable(s) would change under hypothetical interventions (Woodward 2000: 205, 2003: 15).

> A generalization G (relating, say, changes in the value of X to changes in the value of Y) is invariant if G would continue to hold under some intervention that changes the values of X in such a way that, according to G, the value of Y would change – "continue

> to hold" in the sense that G correctly describes how the value of Y would change under this intervention. (Woodward 2003: 15)

Invariance relations exclude spurious generalization, such as the correlation between a barometer reading and the rising of a storm. The connection between a barometer reading and the weather is not invariant, since an intervention on the barometer reading would have no effect on the weather (Woodward 2000: 210). Moreover, in the interventionist account, *explanations* appeal to interventions that reveal invariant generalizations. Descriptions of invariance generalizations are explanatory because they can answer a range of counterfactual or 'what-if-things-had-been-different' questions about the explanandum (Woodward 2000: 209).

The interventionist theory of causation accommodates explanation especially well in psychiatry and other special sciences for at least three reasons (see Kendler and Campbell 2009). First, the idea of invariance is a metaphysically non-committal view that can support various generalizations. That is, causal variables can stand as a shorthand for any entity or activity that can be theoretically intervened and manipulated. Hence, for instance, if hypothetical interventions can be targeted at genes, propositional attitudes, as well as social institutions, they can potentially serve as causal variables in explanations (see Woodward 2015). Second, unlike explanations based on the laws of nature, the interventionist theory can for the fact that although generalizations in the special sciences are fragile and historically contingent, they nonetheless are explanatory (Craver 2007: 94). The reason is that although many generalizations in the special sciences depend on background conditions, they are nonetheless invariant under certain interventions. Third, what is specifically important for a mechanistic account of scientific kinds, unlike explanations based on laws, is that invariance permits degrees of generalization. The more a

generalization is invariant under important interventions, the deeper the explanation it provides. Whereas explanations based on the laws of nature are putatively exceptionless, invariance is not universal, but instead holds under certain range of interventions and background conditions.

There is no straightforward answer as to how stable the invariance relationship should be to be scientifically relevant. For the invariance generalization to be causal and explanatory, it should be above a certain threshold by answering relevant discipline and domain-relative counterfactual questions. Consequently, there is no general answer as to how stable a mechanism should be to qualify as an explanatory mechanism for a psychiatric syndrome. Instead, the interventionist account leads to the conclusion that there are different degrees of mechanistic and kind stability. In addition, the idea of intervention is especially fit to describe medicine, where explanations are subservient to diagnosis and healing interventions, rather than satisfying purely theoretical curiosity. Murphy (2011), for instance, points out that the interventionist approach naturally fits the study of the difference-making factors of medicine. Moreover, according to Woodward (2015: 156), mental or psychological states can be held as causes for behaviour and other mental states in psychiatric disorders because they can be invariant under interventions. Further, medical interventions, like other interventions, require that the system of interest is stable enough to ground generalizations so that system's behaviour can be changed predictably (Craver 2007: 93).

Contrastive Counterfactual Approach of Mechanistic Explanation

The contrastive counterfactual theory of explanation, together with the intuitive idea of explanatory relevance as invariance under

interventions, provides a permissive account of mechanistic explanation that can match with most accounts of mechanisms and mechanistic explanation. According to the contrastive counterfactual theory of explanation, mechanistic explanation can be understood as describing networks of counterfactual dependence (Ylikoski 2013: 291). When the heuristic idea of intervention or manipulation is added to this view, we can provide a more detailed picture of mechanistic explanation. Explaining with etiological causal mechanisms is then tantamount to asking counterfactual questions about how changes in the mediating causal mechanism would make a difference to the output of interest. Constitutive explanation can also be understood counterfactually (Ylikoski and Kuorkisko 2020), and according to Craver (2007), as simultaneous multiple manipulation of the parts of the system and the system as a whole in scientific experiments.[54] In this case, a mechanistic explanation can answer counterfactual questions of how changes in the mechanism's components, or their interactions and organization, would make a difference to the system as a whole, and vice versa (see Kuorikoski and Ylikoski 2013, Pöyhönen 2013b, cf. Craver 2007). A plausible way to understand the difference between these mechanistic explanations is that causal explanations track events, while constitutive explanations track causal capacities (Ylikoski 2013). Moreover, if mechanistic explanations are understood as tracking ontic structures in the world (see Craver 2007), so that it is the actual parts, actions, and organizations' features that do the explaining, mechanistic descriptions carve out networks of causal relations in the world (see Pöyhönen 2013b).[55]

[54] There is debate over the credibility of mutual manipulativity (Craver et al. 2021).
[55] See Illari (2013) concerning the difference between ontic and epistemic explanation.

This permissive account of mechanistic explanation is compatible, for instance, with Kendler and Campbell's (2009) approach to psychiatric explanation (see also Kendler 2008; Kendler, Zachar, and Craver 2011). These scholars distinguish between an interventionist approach to causal explanation and more profound mechanistic explanation, although, according to them, the former can supplement the latter. Similarly, in Kendler, Zachar, and Craver (2011), it is argued that causal mechanisms can be employed to identify psychiatric kinds to support psychiatric classifications. The view of mechanistic explanation in both of these papers seems to be related to Bechtel's mechanistic explanations of the physical structures in biology and cognitive science (see Campaner 2016, 2018). However, they acknowledge that what complicates matters in psychiatry is the need for a multilevel approach to mechanisms that can explain the interaction of different factors on different levels of description, such as genetic, neurological, psychological, and social. Hence, they argue that in many cases psychiatric explanation can rely on the rigorous but less demanding interventionist account of explanation.

However, unlike Kendler and Campbell (2009) and Kendler, Zachar, and Craver (2011), I do not want to commit myself to a specific interpretation of mechanistic explanation or to the specifics of Woodward's interventionist account of causation or his account of mechanisms (2002). The central motivation for this permissive or "thin" view of mechanistic explanation, and explanation in general, is to provide an explanatory account that can accommodate mechanistic explanations in psychology and the social sciences. In short, my account is motivated by the complex and dynamic nature of psychiatric disorders, as well as the plurality of their explanations. A successful mechanistic explanation of the complex and dynamic nature of psychiatric disorder, one that can facilitate explanations of all or most of the relevant aspects, may need to abstract away from the disorders'

purported "material" bases. A non-metaphysical explanatory account of psychiatric disorders can help us understand how different disciplinary approaches, such as those that concentrate on social, psychological, and biological factors, manage to explain different aspects of psychiatric disorders. That is, a permissive mechanistic approach has the advantage of being able to facilitate different explanations, while at the same time not succumbing to (currently) unrealistic demands of providing an ontological mechanistic account that could integrate all the factors that cause and maintain psychiatric disorders. The challenge with robust mechanistic approaches, such as Betchel's, is to be able to separate levels of description cleanly in order to rule out interlevel causation (see Woodward 2020). Moreover, the presented permissive mechanistic account offers a seamless non-reductive account of mind-body interactions in explanatory terms. Simply put, there is a counterfactual dependency if by wiggling psychological factors we can manipulate brain states, or vice versa. Moreover, according to this approach, psychiatric disorders can be understood as heuristic or investigative kinds that invite further examination yet nonetheless support specified inductions. This is especially important in psychiatry, where the primary aim is to discover reliable therapeutic interventions on psychiatric disorders, instead of being able to answer metaphysical questions about their nature.

In sum, contrastive counterfactual explanation, supplemented with the heuristic notion of the interventionist idea of causal relevance, provides a conceptually intuitive interpretation of mechanistic explanation. With this in mind, I will next examine how the contrastive counterfactual theory of explanation can be employed to explicate causal and constitutive explanations in psychiatry.

4.1.3 Disentangling Causal and Constitutive Explanations

The contrastive structure in the contrastive counterfactual theory of explanations can be employed to explicate whether explanations do in fact target the same explananda by making explicit and refining underlying explanatory-seeking questions. For example, when explaining some aspect of a psychiatric disorder, it may not always be clear whether we are interested in constitutive or etiological mechanisms. This has caused considerable confusion in philosophical debates over psychiatric explanation. Sometimes descriptively lower-level variables are irrelevant for explaining a descriptively higher-level explanandum (Kuorikoski and Ylikoski 2013). On the other hand, sometimes focusing on social factors can be relevant to understand their causal influence on the neuro-cognitive explanandum downstream, or neuro-cognitive factors can explain higher-level factors downstream.

The contrastive counterfactual theory can be employed to specify whether we are interested in causal or constitutive information (Ylikoski 2013: 289). Rather than asking whether a particular disorder is a brain or psychological problem, it is better to specify the question by posing smaller questions with contrasts. This means that contrasts can make explicit and refine what we are really asking. As mentioned earlier, explanations are ideally spelled out in a contrastive structure: they answer questions of why fact rather than foil (e.g. Ylikoski 2007). Based on this, simply asking whether some disorder is real or what it is made of, are too large questions to be examined fruitfully. Instead, a better explanatory approach is to split the question into smaller ones. Answering these smaller questions can also shed light on larger ones (Ylikoski 2018).

According to Ylikoski (2013), constitutive explanations address causal capacities, while causal explanations follow events and behaviours. That is, while causal explanations rely on etiological

counterfactuals, constitutive explanations rely on organization and its parts. Consequently, the doubly contrastive structure of explanation can be represented in the following way:

a (a*) explains b (b*) in the context of U

(Ylikoski 2013).

When this approach is applied to explanations of psychiatric disorders, the causal explanation b and b* can be about the generation, sustenance, and changes in the disorder, while a and a* would be about antecedent events. In constitutive explanation, b and b* can be about the disorder's causal capacities, and a and a* are about the capacities of the disorder's parts or their organization. In both cases, the causal field U can contain facts about the physical and social environment as well as other facts about the disorder. Moreover, in both cases the interest is the contrast between b and b*, while their similarities are assumed to belong to the causal field U.

According to this account, when trying to explain, for instance, some aspect of a disorder and its symptoms, such as major depression, we come up with the following questions (following Ylikoski 2013). (The exemplary questions I use are very general, and in specific cases they would concern more specific aspects of a disorder.)

(i) How did this individual become depressed?
(ii) What makes this individual depressed?

While question (i) is causal, question (ii) is constitutive. The answer to the first question (i) tells about the causal history of the patient's psychiatric disorder and its symptoms. It aims to explain facts about the individual's depression by referring to earlier events in the development of the disorder, based on their contribution to the difference between

the fact and the foil (cf. Ylikoski 2007). The facts can include psychological, social, or biological factors. The foil can be, for example, why this individual has these specific symptoms rather than other or no symptoms. Or why did the symptoms appear at a certain time or location rather than at another time or location.[56] Similarly, it can concern social, cognitive or genetic factors in contrast to other factors on the same level of description. The answer to the second question (ii) is constitutive because it tells about the things that the patient's disorder is made of as well as their capacity to realize the mental disorder. These factors can include components, actions, and organizational features from different levels of description or organization such as psychological, genetic, and neural factors, and possibly also external social factors. The contrasts can include, for example, differences between social, psychological, or neurological organization. Ideally, a robust explanation of a major depression would require both an etiological account of the causal processes that led to the disorder and a constitutive explanation of the pathological capacities that realized the disorder at various stages during that process.

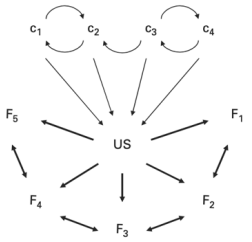

Figure 3. (Kendler et al. 2011).

[56] See Ylikoski (2007) about possible contrasts.

In practice, many psychiatric explanations combine both causal and constitutive explanations. That is, many explanations of instances of a psychiatric disorder may require constitutive and etiological mechanistic explanations, some of them intrinsic and others extrinsic (see Figure 3). For example, such a combination could address questions concerning an individual's biological or psychological disposition to become mentally ill. This can be exemplified by alcohol addiction. Although Indigenous peoples in North America were more (genetically) predisposed to becoming addicted to alcohol than most native Europeans, the ones that realized that potential, only did it after being exposed to alcohol and social injustice. On the other hand, had there been established social mechanisms to counter alcohol abuse, people predisposed to alcoholism may not have developed the addiction. Hence, the property cluster of interest can be explained by supplementing various explanations, so there may be external causal mechanisms that trigger the mechanisms that exhibit the phenomenon, as well as other external environmental mechanisms that contribute to the properties maintaining their homeostasis, including factors that explain the addiction experience. The explanatory interests determine which of these mechanisms we should concentrate on.

Moreover, according to Ylikoski (2013), some developmental explanations are hybrids of causal and constitutive explanations in a stronger sense. He describes the gene knock-out experiment as an example. In such an experiment, the target is to study how making a gene non-functional affects the causal capacities of the adult organism. In this example, the explanans and the explanandum are both constitutive, while their connection is causal. Quite plausibly, some psychiatric explanations are similar hybrids of constitutive and causal explanation. One example might be a study on how genetic differences at birth would bring about differences in the cognitive capacities of schizophrenic adults.

4.1.4 Modularity as a Heuristic Aim in Psychiatric Explanation

The interventionist view of explanation can be employed to reflect on the possible theoretical requirements, and practical possibilities, for successful mechanistic explanation of psychiatric disorders. In particular, I want to investigate whether such explanations should be modular, so that they do not alter the causal structure that they represent.

An interventionist account of mechanistic explanation may require that the representations of its components are modular under interventions (Hausman and Woodward 1999: 550, Woodward 2002: 374, 2003: 48, 2013: 51).[57] The idea is that if mechanisms are understood as causal structures, they can be characterized by an interventionist view of causation, which in turn may require modularity of their representations (Kuorikoski 2012).[58] According to Woodward (2002), a model is modular if an intervention on a putative cause does not alter the subsequent causal relations in the underlying structure of the represented causal mechanism. The unaltered causal structure ensures that the model can predict the outcome of an intervention. This means that a mechanistic representation is a modular system if its generalizations or equations are invariant under interventions (Raerinne 2011). In other words, a modular representation of a system enables interventions on the variables of its generalizations or equations without altering its other generalizations or equations. This modularity ensures that a complex mechanism can be carved into independent components whose causal relations can be examined while holding other components of the mechanism put.

[57] In some places Woodward (2010, footnote 11) does not require this. See Steel (2008) about modularity of social mechanisms.
[58] A mechanistic account based on the CC theory does not make such a commitment.

If attainable, modularity could be an important aim for the mechanistic approach to psychiatric disorders. For example, Campaner (2018) argues that in theory modularity could be employed to disentangle different causal components underlying the comorbidity of psychiatric disorders. That is, if co-morbid disorders could be decomposed into smaller modular components, so that these parts could be intervened upon separately, a better picture of their relations could be built. Consequently, Campaner advocates a *heuristic ideal* of isolating the modular causal relations of psychopathological systems to the extent that they could provide a better in-depth analysis of particular disorders and their comorbidity with other disorders. Hence, one could argue that while this may not be possible in practice, the goal could nonetheless be to investigate in theory what would happen if such an intervention were possible.

Modularity requirements could also have a bearing on the individuation problem of the HPC view of kinds. As argued by Craver (2009), kinds that are brought about by more than one mechanism should be split, whereas two or more kinds that are realized by one mechanism should be lumped together. Hence, if the modularity requirement holds for homeostatic mechanistic explanations, we should be able to decompose the mechanisms into modular components, and thereby gain a better understanding of how psychiatric kinds can be demarcated.[59]

Modularity, however, is a problematic requirement for a mechanistic explanation of psychiatric disorders and special science kinds in general.[60] The reason for this is that many mechanistic representations, or the represented mechanisms themselves, are so

[59] Illary and Williamson (2012) argue that modularity is a problem for mechanistic modelling and it does not necessarily describe how actual mechanisms function.
[60] It is not clear that everyone is talking about the same thing (see Kuorikoski 2012). There is confusion, for instance, over whether modularity is a feature of the ontic mechanism or its representational model.

fragile that they may change under interventions. Nancy Cartwright (2002, 2004) and Sandra Mitchell (2008), for example, claim that modularity should not be considered relevant for mechanistic modelling because in many actual cases it does not hold. Mitchell (2008) in particular argues that modularity often fails in biology, where the organisms are "constructed" to maintain their causal structure under changes to their general functions. Similarly, modularity is a problematic requirement in the social sciences, where interventions can be structure altering (see Steel 2008). Illari and Russo (2014: 105) exemplify this point with Lucas's critique of macroeconomic policy making (see also Woodward 2000: 200-221). Lucas argued that the very models that underlie policy decisions become engines of change when economic agents base their economic decisions on them. Hence, this would make the models invalid in explaining the initial economic behaviour.

One answer to these critiques would be argue that many models initially represent the targeted causal structure correctly. For instance, in the Lucas critique, while knowledge about the explanatory models alter the causal structure downstream, it does not necessary change it immediately, so that the models may enable predictions and causal interventions for a while. Moreover, when new interventions change the causal structure, it may be enough to incorporate the new information into the model, i.e. render some background information as endogenous variables, so that this information correctly represents the additional dependencies (Kuorikoski 2012: 366).[61] Lucas (1976) himself argued that a more fine-grained mechanism based on rational choice theory

[61] This can also be done so that the first mechanism switches to another mechanism under certain interventions. In other words, the new causal structure can be understood as a new mechanism.

could be used to explain why the interventions alter the original mechanism.

Another possible way to address these criticisms is by arguing that the modularity requirement should not be overemphasized (see Raerinne 2011). As Steel (2010) argues, although it is almost always possible to intervene on a causal system so that the consequences are not tractable, the point of modularity is to model the causal structure of the system. Such modelling involves one of the many possible mechanistic representations of the system that describe relevant dependencies that succumb under interventions or scrutiny, while other parts of the system are taken as given. Hence, as Steel puts it, modularity does not describe what causal systems are in their totality, but rather how some parts of them can be idealized for desired explanations. This means that it is enough that a model is modular for the explanatory purpose it was built for.

Constitutive and causal explanations may also come with their own modularity requirements. Kuorikoski (2012) argues that constitutive explanation requires a strong parameter form of modularity and more background information than causal explanations that can rely on a weaker variable modularity. This is because the targets of explanation are different: while parameter modularity requires that we are able to situate a particular mechanism in a wider causal nexus by uncovering its constitutive parts, variable modularity requires that interventions on the variables within the mechanistic model do not change its causal structure downstream. When this approach is applied to psychiatric kinds, a constitutive explanation of the kind would locate it within the counterfactual space of different ways it could have been built, while causal explanation situates some of its aspects (represented by variables) in dependency relations with their antecedent causes.

In the light of this discussion, it may sometimes be helpful to rely on the heuristic aim of building a modular representation of causal

structures in psychiatric explanations in order to disentangle their complex causal relations. However, many psychiatric disorders may be too complex and dynamic to be explained by means of decomposable monadic mechanisms. For instance, according to Campaner (2016: 115), "the causal networks investigated present multiple nonlinear interaction between biological, psychological and socioeconomic processes, and often causal loops". This lack of modularity, or weak modularity, weakens the mechanistic explanatory power in psychiatry. However, it should be borne in mind that not all explanations of psychiatric disorders need to be constitutive (i.e. mechanistic representations need not be parameter modular). In fact, it seems that at the moment we do not have enough knowledge about the inner structures of most disorders to attain physiological constitutive explanations, while for some disorders constitutive biologically oriented explanations may never be attainable, due to, for example, feedback loops between our brains and social determinants. However, this does not mean that we could not have some understanding of how their manifest symptoms are brought about by antecedent causes. These causal representations may submit to different strengths of modularity so that we may be able to trace the outcomes of interventions downstream by therapy or by pharmacological interventions. The implication is that although it is extremely difficult to come up with a complete picture of a given psychiatric disorder, we may nonetheless come up with partial causal explanations of its various aspects targeted by different disciplinary approaches.

4.2 The Applicability Domain Account of Kind Explanation

The account of mechanistic explanation based on the contrastive counterfactual theory can be employed to understand property cluster kinds in general, and psychiatric kinds in particular. Pöyhönen (2014), for example, has employed the contrastive counterfactual account to analyse the mechanisms responsible for property cluster kinds. In this section, partly relying on Pöyhönen's account, I will argue that mechanistic explanations of property clusters have domains of applicabilities, which should be ideally identified, especially when the aim is to explain as dynamic and complex a phenomenon as a psychiatric disorder. The central motivation for this approach is to put forward an account that can be employed to identify how, and to what extent, mechanisms explain property clusters. This approach enables a more rigorous understanding of how different factors can contribute to explaining the same kind, and also how these kinds can be delineated for different epistemic and value-laden purposes. In other words, it offers an epistemological account of psychiatric kinds.

4.2.1 Explanatory Depth and Scope

As pointed out, the HPC view grants scientific kinds a dual role: property clusters support inductive inferences, while responsible mechanisms explain why and how these properties co-occur in a cluster. In other words, while causal mechanisms explain the clustering of properties, the clusters together with their mechanistic explanations enable secure extrapolations and projections (Kuorikoski and Pöyhönen 2012; see also Reydon 2009). Therefore, the more we learn about the mechanisms that generate property clusters, the more securely we can extrapolate that token clusters are of the same type.

Moreover, when a specimen is identified as a member of a well explained kind, we can reliably generalize and predict its behaviour based on that membership. Simply put, the better explanation we have for a property cluster, the more secure the inductive inferences and extrapolations it supports.

My view is that the non-reductive epistemological mechanistic approach to HPC kinds can be supplemented with the range over which these mechanistic explanations are reliably applicable. A model of the mechanism(s) that is responsible for a property cluster kind ideally specifies a *domain of applicability* over which it stably explains the kind and its property variations. The domain of applicability has a *scope* and a *depth* dimension (cf. Griffiths 1999, p. 217).[62] The *scope* dimension describes the actual properties, places, and times where the explanation is applicable. For instance, an explanation can cover specific aspects of a kind for a determined duration and location. The *depth* dimension describes the counterfactual stability of the explanatory relation, that is, how dependent the inferences that the explanation enables are on non-included situations. Hence, explanatory depth is a modal notion. The idea is that explanations of a homeostatic property cluster fall on dimensions of explanatory power or goodness described by the contrastive-counterfactual theory. That is, a mechanistic explanation has an applicability domain over which it can answer what would happen to aspects of psychiatric kind if the components, activities and organizational features of the responsible mechanism were to change in specified ways. Crucially, applicability domains are limited in the special sciences, including psychiatry, because explananda phenomena are contingent on complex

[62] Griffiths (1999, p. 217) distinguishes between the force and the scope of projections that kind categories support, whereas I am interested in identifying the stability of the mechanistic explanations that warrant these projections.

environmental circumstances, feedback loops, and dynamic processes. In sum, an explanation can exceed the limits of its spatial and temporal applicability by disregarding the dynamic and context-relative nature of the targeted kind. An explanation should therefore make explicit and explain why the given applicability domain is optimal for a set of phenomena. This is tantamount to identifying better the limits of the underlying causal mechanistic structure of the phenomena. That is, identifying the applicability domain of an explanation spells out the conditions under which the explanation is expected to be reliable and when it can breakdown.

The applicability domain approach can be exemplified by Ian Hacking's (1995) account of multiple personality disorder (now called dissociative identity disorder). Hacking argues that a prototypical conception of multiple personality disorder, and diagnostic practices built on it, shaped the way people behaved and perceived themselves (see Chapters 5 and 6). This possible explanatory account could potentially supplement an explanation of the multiple personality syndrome by being able to account in detail for the counterfactual dependence between the diagnostic category and the birth of a multiple personality kind of person. In accordance with Hacking's account, the connection may be at least partly explained by prototypical expectations and practices associated with the social role of being this kind of person (see also Mallon 2016). Moreover, such an explanation can potentially contribute to making an explanation deeper by being able to explain why the syndrome mainly existed in three different cultural locations (France, USA, the Netherlands) as opposed to other locations (cf. Kihlström 2005). Instead, the dominant trauma memory explanatory model, according to which the syndrome has its origins in trauma (e.g. childhood sexual abuse), is not very convincing (Kihlström 2005). It cannot explain why, for instance, when childhood abuse was added as an explanation for the multiple personality disorder, the diagnosed individuals started to

amend their memories accordingly (Hacking 1995a, Kihlström 2005). Hacking argues that the reason was a feedback loop between the general prototype associated with the syndrome and the classified people themselves. However, Hacking's proposed explanation is not very stable because it does not include psychological or cognitive factors. In other words, the explanation's scope only covers certain aspects of the syndrome, while other psychological or cognitive factors should be held as stable background conditions.

This domain-relative approach can be employed to answer the previously mentioned challenge to the HPC view of psychiatric kinds. I have argued that the crucial challenge to the property cluster view of kinds, and thereby to a HPC view of psychiatric kinds, is to account for their limited inductive success, facilitate the interaction of different factors and discipline-related explanations, as well as to determine the boundaries of these property clusters in spite of their fuzziness. Identifying the boundaries of psychiatric disorders also needs to account for the role of non-epistemic values (see Chapters 3 and 7).

Since explanatory stability according to the contrastive counterfactual theory comes in degrees, which is intuitively captured by the idea of invariance under intervention, the theory can explain the empirical fact that inductive inferences that kind concepts in the special sciences, including psychiatry, support are not open-ended. Rather, according to my account, knowledge of the causal and constitutive mechanisms responsible for a property cluster kind sets a domain over which inductive inferences are warranted. In the case of causal explanation, the idea of causal generalization as describing invariance relationships under interventions can make understandable the domain-relative nature of inductive inferences of property cluster kinds, which

are not as robust as law-based or essential kinds.[63] In sum, a mechanistic explanation of a property cluster kind identifies a domain over which it can answer counterfactual or "what-if-things-had-been-different" questions about how some of the aspects of the property cluster would behave in a systematic way under targeted (hypothetic) interventions.

The applicability domain approach can be applied to specify the explanatory contribution of alternative explanations to the same scientific kind, such as a psychiatric disorder. The contrastive form of explanation facilitates comparisons between alternative explanations to the extent that it may turn out that they are not necessarily competing (Pöyhönen 2013a, Van Bouwel 2014,). The reason is, as already mentioned, that explanations always answer questions concerning some aspects of the phenomenon of interest, not the whole phenomenon (Ylikoski 2018). For this reason, after making explicit the alternative explanation-seeking questions with contrasts by including only the factors that describe the difference between the given explanandum and its foil, it may turn out that the explanatory answers in fact complement each other. In this case, they would provide information about different aspects of the same kind (Pöyhönen 2013b: 39). In other words, these explanations could thereby contribute to the applicability domain of a psychiatric kind explanation. This can be put in counterfactual terms so that a better explanation of a property cluster kind answers more counterfactual questions because it locates the kind within a larger space of alternative possibilities.

The idea of explanatory depth and scope can be employed to compare the explanatory import of alternative explanations to the same psychiatric kind. Moreover, explanatory aims may also lead to different ways of demarcating the explanandum property cluster kind, so that the

[63] I do not want to claim that this is the only way to understand stability. Craver (2007), for instance, argues for mutual manipulativity in constitutional explanation.

corresponding explanation would carve the responsible causal structure differently. In this section, I will analyse comparative explanations, and in the following section, I will analyse kind delineation.

According to the contrastive counterfactual theory, explanatory power or *depth* is not an all or nothing matter. Rather, the depth of an explanation depends on how many relevant what if-questions it can answer. According to the CC theory, explanatory depth can be explicated with explanatory insensitivity and precision. The insensitivity of an explanation is the measure of the invariance of the explanatory dependency under different background conditions (Ylikoski and Kuorikoski 2012, Woodward 2006, 2003). The power of an explanation increases the less the counterfactual connection it depicts is relative to background conditions. In other words, an increase in sensitivity renders an explanation more fragile, whereas a decrease in sensitivity renders it more powerful. This is because explanations that hold only under certain background conditions cannot answer as many what-if questions as explanations that are not sensitive to those conditions. As Pöyhönen (2010) has pointed out, insensitivity is an especially important feature of the mechanistic explanation of property cluster kinds: an explanatory mechanism should be resistant to changes in at least some background conditions for the kind concept to support epistemic projects such as inductive inferences.[64] The precision of an explanation is an aspect of the explanandum (ibid). Deeper or more powerful explanations provide more details of the contrast space relative to the explanandum (Ylikoski and Kuorikoski 2010). That is, a powerful explanation is able to characterize in a fine-grained way why something is the case in contrast to something else.[65]

---

[64] It is important to distinguish insensitivity from scope: an explanation of a kind can be very insensitive, although there are not many actual tokens of the kind that it applies to. That is, an explanation may apply only to very few actual cases, but still provide a very robust explanation (Woodward 2003).

[65] Ylikoski (2007) has provided a comprehensive analysis of possible contrasts.

According to Ylikoski and Kuorikoski (2010), mechanistic detail is not an independent virtue, but can rather be analysed with sensitivity and precision. By including mechanistic details or providing more mechanisms to explain a phenomenon, the explanation can answer more what-if questions. More details make the explanation less sensitive to background conditions because the explanatory dependency applies to more circumstances. Moreover, there is an explanatory trade-off between sensitivity and precision (Ylikoski and Kuorikoski 2010). In general, the more precise an explanation is, the more sensitive it is to background conditions. This is natural because the counterfactual dependency between an explanans and a fine-grained explanandum is more prone to external disruptions than a dependency between explanans and an abstract explanandum (ibid.: 211).

The *scope* of a mechanistic explanation's applicability domain of a psychiatric kind, on the other hand, can be further explicated by both horizontal and vertical levels of description. The scope of the explanation is set by the contrastive explication of the aims of the explanation, that is, which aspect of the property cluster we are interested in explaining. The scope can be distinguished between a horizontal and a vertical level of description. The horizontal level includes the aspects of properties on the same level of description, whereas the vertical level includes aspects of properties from different levels of description. For instance, a robust complex mechanistic explanation of major depression may require one to consider psychological, neuro-cognitive, and genetic factors and their interactions, and it may include various etiological environmental causal factors.

The domain-relative approach can be employed to analyse how explanations can contribute to explaining the same psychiatric kind. This can be demonstrated with a potential hybrid mechanistic explanation that would describe the connection between migration and

the onset of schizophrenia. According to statistics, there is a nine-fold higher risk of schizophrenia among the black Caribbean community in the UK (Pinto, Ashworth, and Jones 2008; see also Howes and Murray 2014). The mediating social determinants between immigration and the onset of schizophrenia could include, for instance, social-economic status, cultural meanings, cannabis use, racism, and an urban environment. Changes in any one of these background conditions could have an impact on the course-grained explanatory connection between immigration and schizophrenia, and probably also alter the symptom profile of the disorder. Hence, an explanation that would refer to social factors merely as triggering influences on genetic predispositions would be sensitive to alterations in these social determinants. Instead, a social mechanism of discrimination and racism experienced by the Caribbean community, could contribute to explaining the increased incidence of schizophrenia, and thereby make the general explanation deeper. The scope of this explanation could cover, for instance, Caribbean immigrants in the UK in contrast to immigrants who have experienced less discrimination. The example demonstrates how the non-metaphysical account of mechanistic explanation matches the view that the HPC account can accommodate external mechanisms that are responsible for property clusters (Boyd 1991; see also Mallon 2003). It also fits nicely with the idea that uncovering social mechanisms enables extrapolations (see Elster 2015; cf. Steel 2008).[66]

However, it should be borne in mind that deeper explanations of psychiatric kinds require hybrid mechanistic explanations. As an example, Kirmayer and Gold (2012) suggest that there are at least three ways social factors may interact with genetic vulnerabilities to produce

[66] Roughly, a social mechanism can be said to consist of individual agents whose relations and actions are responsible for a social phenomenon. See Hedström and Ylikoski (2010).

psychiatric disorders. First, genes can induce behavioural patterns that predispose the individual to behave in certain ways that in turn have psychological effects. These in turn can contribute to the individual becoming sick. As an example, they speculate that a psychological disposition to be bullied as a child could influence the changes of developing psychosis later. Hence, Kirmayer and Gold (2012: 324) point out that the "social environment may thus be part of a loop that affects mental life". It is plausible that similar environmental loops can also influence our brains more directly. I think it is possible, for example, that the prevalence of schizophrenia among some members of the Caribbean community the UK could be partly due to a genetic predisposition to seek pleasures, such as cannabis smoking, that stimulate specific areas of the brains. Second, Kirmayer and Gold argue that some people may be genetically more susceptible to stress, which in turn may predispose them to mental disorders in urban environments. Third, they also point out that social life, such as abuse, may influence genetic expression (i.e. epigenetics processes) and thereby contribute to the development of psychiatric disorders. Therefore, in the light of these points, a hybrid mechanistic explanation, which would incorporate heterogeneous (etiological) causal factors, including feedback effects between social environments and genetic predispositions, would substantially widen and deepen the explanation's domain of applicability. As Kirmayer and Gold (2012) point out, the social world has its own structure, and therefore an adequate social explanation of a mental disorder cannot be reduced to mere "triggering" causes. This knowledge, in turn, could contribute to more secure projections based on the classificatory concept of schizophrenia in general.

Although it may seem that some disorders require only a simple explanation based on one cause, their better understanding may still require a mechanistic explanation that involves causal factors from several levels of description or organization. For example, the genetic

explanation of Huntington's disease offers a deep explanation within a relatively narrow scope (cf. Murphy 2015: 95). Huntington's symptom cluster consists of behavioural, cognitive, and psychological dysfunctions, such as slow eye movement, difficulty in concentrating, and feelings of irritability, respectively. These symptoms are caused almost completely by a single gene, IT15 on chromosome 4, to the extent that the onset of the symptoms is not dependent on specific social or psychological factors (ibid.: 95-96). In other words, the genetic explanation of Huntington's disease seems to be vertically narrow as it requires only one level of description. Nevertheless, such an explanation grounds generalizations that are highly invariant under many environmental, social, and psychological variations. Changes in these background conditions would not make a difference to the disease. Nevertheless, there are multiple mechanisms that contribute to how the gene produces the disease. Understanding these would require descriptions of causal factors from multiple levels, such as genetic and neurological (Gil and Rego 2008). The hope is that revealing these mechanisms would ultimately lead to effective cures. Hence, if the scope of the explanation is widened to cover the mechanisms that produce the disease, other facts than the gene are needed. Moreover, if the scope is to explain the experience and larger consequences of the disease, psychological and social explanations would be needed. For instance, if the aim is to explain the experience of a certain psychiatric disorder, while knowledge of its possible neurological constitution would help, it would not be enough to explain the disorder in its totality. This is also the case when they are directly aligned in a constitutive relation, in which case there is counterfactual dependency between the whole and its parts.[67] Rather, phenomenological explanations, for

[67] Although I not claiming that there is top-down causation, I nonetheless think that the question is pragmatically problematic in psychiatry. See Woodward 2020, Eronen 2021.

instance, would be crucial to understand why a disorder, such as Huntington's disease, is experienced in one way rather than in another. This is based on the fact that factors on different levels of description or scales succumb to genuine causal relations.

With most other syndromes, however, a stable explanation even of their etiological cause requires us to understand the interaction of many factors. Culture-bound syndrome hysteria serves as a good example. The causal processes leading to hysteria were historically and culturally bound since the syndrome manifested itself only at a certain point and location in history and generally only with upper middle class women. This is also exemplified by the disappearance of the syndrome when the cultural setting changed. For this reason, the existence of hysteria seems to have been contingent upon a specific historical and cultural context, so that a deep explanation of the onset of the syndrome would not only require psychological or cognitive explanations, but also an understanding of the role that sociohistorical factors played in that temporal and spatial period. These social and psychological mechanisms probably interacted to produce the syndrome as well as its symptoms and signs.

4.2.2 Value-Sensitive Specification

The contrastive counterfactual theory of explanation can be employed to understand how explanatory power and non-epistemic values can play a role in carving the boundaries of psychiatric kinds. The basic idea is that epistemic interests govern how explanatory-seeking questions are explicated with contrasts. As mentioned above, once the explanandum is set with the help of contrasts, the causal factors that are relevant for its mechanistic explanations can be obtained. Hence, if different epistemic interests are stressed, such as some explanatory

virtues in contrast to others, they may support specifying the kind phenomenon differently (Pöyhönen 2014). However, in psychiatry, epistemic interests are often also guided by value-laden classificatory and therapeutic interests. That is, explanatory relevance is partly fixed by holding aspects of a given phenomenon as unwanted and thereafter conducting research to find means to intervene on it.

Complex causal structures, such as those responsible for psychiatric disorders, may not support clear specifications of property cluster kinds. As Bechtel (2009) argues, mechanisms need to be situated in a larger environment to understand how they produce the phenomenon of interest. Similarly, Craver (2007: 144) makes this point by arguing that the boundaries of action potential are fixed by explanatory relevance, not by physical structures. Therefore, as Campaner (2006) argues, consideration over social "external" factors may lead to a more comprehensive extended mechanistic picture of a psychiatric disorder, which would require redefining its boundaries. In other words, if mechanistic explanations are understood as representations of the causal structure of the world, alternative explanations could carve that structure differently. Based on this, Pöyhönen (2014) has argued that the boundaries of complex mechanisms, such as cognitive systems, and the property clusters they are responsible for, can be legitimately delineated differently.

For instance, schizophrenia can be specified differently depending on the purpose of the classification, for example whether it is clinical treatment or epidemiological prevention. This becomes clear by comparing schizophrenia and schizotypy (Zachar 2013: 90). Schizotypy has fewer severe symptoms than full-blown schizophrenia, although it may lead to schizophrenia, and apparently both have the same underlying genetic structure. Based on their genetic uniformity, the two types of illness should hence be lumped together. This would make the classification insensitive to different environmental and

cultural variations. On the other hand, based on their different symptom profiles, schizotypy and schizophrenia can be split as different kinds. This would make their classifications more precise for treatment purposes. However, these different approaches are not epistemically in conflict with each other: they can both be retained because they support different sets of explanation and induction (see Griffiths 1999: 199).

However, the concept of psychiatric disorder and psychiatric classifications also serves non-epistemic purposes. This means that value-laden considerations, rather than purely epistemic ones, can either influence how the explanandum phenomenon is delineated for research or how the results are interpreted for classification. There are several reasons why specification interests in psychiatry may be partly value-laden (see Chapter 7). As mentioned before, at least currently a naturalistic definition of the general concept is not available, and this can reflect on how phenomena are chosen for research. This in turn may influence the way in which the causal structure is parsed for explanation. Moreover, apparently the boundaries of many psychiatric disorders are not clear-cut. This in turn leaves room for value-laden and pragmatic considerations. For example, ADHD can be classified with different sets of criteria. In the USA, hyperactivity is enough, but in Italy a severe case of attention deficit is also needed.

Another challenge is that even if there are clear-cut ways to delineate some disorders, pragmatic or ethical considerations may pull in opposite ways of classifying them (these in turn can influence the kinds through looping effects, see Chapter 5). This can be exemplified with Griffiths's (1999) distinction between investigative and normative kinds (see also Brigandt 2003). Investigative kinds can be understood as natural kind concepts that are open-ended to the extent that their intension and extension are modified to match empirical discoveries. The intension and extension of normative kinds, instead, can in addition be altered for normative reasons, such as social reform. Griffiths

suggests that diagnostic categories are normative kinds, because they are held to be unwanted.

This view can be exemplified with Brigandt's example of a trade-off between ensuring the right for treatment and avoiding stigma. For example, we may deem that auditory hallucinations should not be considered core symptoms of schizophrenia because of their stigmatizing influence (see Chapter 5). There is a possibility that this interest-related explanandum choice would carve the complex mechanistic structure as legitimately as if they were considered an important feature of kind. Or even if the underlying causal factors were an eliminable part of the kind, ethical considerations could be employed to exclude them from the classification. On the other hand, they could be deemed a relevant part of the disorder, even if the causal structure by itself would not suggest either way. The reason could be to ensure better treatment of schizophrenia. In the light of this, in addition to explanatory power, non-epistemic values may govern how the explananda are fixed for specification and classification purposes in psychiatry.

This account supports my value-sensitive pluralistic account of psychiatric kinds. That is, alternative specifications of mechanisms and property clusters can facilitate different epistemic and value-sensitive classificatory aims (see also Chapter 7). Therefore, questions over specification and classification may require explicit value-sensitive considerations over the causal significance of these classificatory choices.

4.2.3 The Multiple Mechanisms Approach

Much of what I have pointed out matches with Ylikoski and Pöyhönen's (2015) idea of a hybrid mechanistic account of addiction. They maintain that a psychiatric kind, such as addiction, can be

understood permissively as consisting of two elements of the HPC view:

i. A cluster of typical properties (etiology, symptoms, response patterns to treatments, etc.) that identify the disorder (we call this the syndrome),
ii. A matrix of causal mechanisms that are responsible for the co-occurrence of the properties in the cluster.

(Ylikoski and Pöyhönen 2015).

According to the analysis, the matrix of causal mechanisms (i) are responsible for the properties (ii) that our classifications try to fix on to. Ylikoski and Pöyhönen argue that although behaviour patterns are employed to identify the disorder, causal mechanisms are nonetheless required to maintain or bring about those clusters. Crucially, the mechanisms need not be singular biological mechanisms, which means that the account can retain the causal idea of the medical model, while not giving in to its implausible essentialism. The account can explain the heterogeneity of the causal processes responsible for the property cluster, and in the property cluster itself. This implies, in addition, that comorbidity is the result of similar mechanisms underlying various disorders. Moreover, they argue that all addictions neither need be located within an individual, nor do individual cases of addiction need to be exactly the same. Rather, it is enough that the symptoms and the mechanisms are considered to be sufficiently alike. A central benefit of this approach, according to Ylikoski and Pöyhönen, is that it can facilitate an empirically testable general theory of addiction. The reason is that information of a certain addiction facilitates empirically testable hypotheses about other forms of addiction (2015: 24). Moreover, according to their account, the general concept of addition is meant as

a heuristic tool, whose ultimate scientific usefulness needs to be tested empirically.

The idea that a general account of a psychiatric kind can function as a heuristic or investigative kind to facilitate the interaction of various mechanistic explanations fits my domain of applicability approach. That is, the more contrastively-fixed mechanistic explanations can supplement an explanation of a psychiatric kind, the wider and deeper the explanation's domain of applicability becomes. This mechanistic knowledge warrants more secure inductive inferences based on the kind concept. Consequently, a particular psychiatric disorder can be considered a scientific kind in the domain-relative sense when its property cluster profile is sustained and brought about by relevantly similar mechanisms. Moreover, relevance can be understood to depend on the particular classificatory, pragmatic, and explanatory aim.

The above notions can be fitted with a value-sensitive approach to classification and specification. In theory, psychiatric kinds could be decomposed to their different responsible mechanisms, and thereafter analysed how they interact to produce the kind (Pöyhönen 2013a). If, for instance, some aspects of a psychiatric kind were deemed to be caused primarily by social determinants, these factors could be considered the primary target for preventions. In the next chapter, I analyse whether one of these mechanisms could be a feedback mechanism which describes an interaction between classificatory projects and the targeted psychiatric kinds. Knowledge of this process, in some cases, may enable a better understanding of the causal significance of the values and norms commonly associated psychiatric kinds and thereby provide a naturalistic approach to the concept of psychiatric disorder. That is, it would enable moving from conceptual or historical (retrospective) analysis of the concept's significance to studying its causal significance. Ideally, this could enable the exploitation and control of those consequences. This account would

reinforce my bottom-up approach of the causal influence of the concept of psychiatric disorder.

In conclusion, the applicability domain account of explanation facilitates explanatory comparison and kind delineations. Moreover, the non-metaphysical interpretation of causal explanation provides an understanding of how biological, psychological, and social factors can be connected in a mixed mechanistic explanation. These different causal factors, which are usually studied by different disciplines, can be understood in the light of the contrastive counterfactual theory as explaining different aspects of the same psychiatric kinds.

4.3 Explanatory Pluralism in Psychiatry

A successful explanation of a psychiatric disorder needs to account for great heterogeneity in explanations and causal factors:

> The field of psychiatry incorporates more viable explanatory approaches than almost any other discipline in a modern university. Serious scholars have attempted to understand causes of psychiatric illness from the perspective of molecular neurobiology, molecular genetics, cellular neurophysiology, systems neuroscience, neuropsychology, clinical psychology (including a wide diversity of theories incorporating an array of mental constructs such as personality, cognition, and unconscious processes), epidemiology, genetic-epidemiology, sociology, and anthropology. (Zachar, Kendler, and Parnas 2020: 1)

This heterogeneity raises questions over how to combine different causal factors in explanations. For instance, what is the relation between genetic and social factors, on the hand, and neurocognitive and psychological factors, on the other? In addition, there are various

discipline-related explanatory approaches that concentrate on different aspects of psychopathology. How, then, should one relate to biological, psychological, and social explanations? I will first present common approaches to accommodate explanations from different areas of science in a pluralistic framework. Thereafter, I will put forward my own explanatory pluralistic approach based on the applicability domain account.

A common strategy to address this heterogeneity has been to endorse explanatory pluralism rather than (epistemological) reductionism (Eronen 2021: 930). According to explanatory pluralism, "(i) the best form (and level) of explanation depends on the kind of question one is trying to answer by the explanation and (ii) that in order to answer all explanation-seeking questions in the best way possible, we will need more than one form (and level) of explanation" (Van Bouwel et al. 2011). Pluralistic approaches that aim to integrate various levels of description or types of explanation have been offered, for example, by Murphy (2013), Kendler (2005), and Kendler and Campbell (2012). As Campaner points out, they all aim to avoid simple reductionism to biology, try to include multiple explanatory levels of description, and include higher-level explanations (Campaner 2014: 95; see also Van Bouwel 2014).[68]

Murphy (2013) has suggested that a way to overcome heterogeneity is by starting with an idealized exemplar of a psychiatric disorder in a patient, that is, an ideal version of the symptoms and how they typically evolve. Rather than concentrating on complex

[68] Schaffner (2006, 2008, 2013), on the other hand, believes that a partial reductionism is compatible with explanatory pluralism. He distinguishes between "patchy or creepy reductionism" by contrasting it with "sweeping reductionism". The gist is that "psychiatry should strive for 'patchy reductionism' with the goal of 'piecemeal integration' in trying to explain complex etiological pathways to illness bit by bit".

interactions in actual cases, Murphy maintains that an exemplar can render the individual variations of the given disorder under scientific scrutiny, and ultimately enable mechanistic explanations (Murphy 2013: 977). That is, the exemplars, in turn, are explained by models of pathogenic processes.

However, Tabb (2019: 13) argues that Murphy is not clear about the role of exemplars because he allocates exemplars the dual role of grounding classifications and being the targets of explanation. This seems to imply that the starting explananda may be influenced by folk psychology and the DSM. In the light of what I have argued, ths consequence may be even more problematic for Murphy's account. If, as Murphy argues, brain mechanisms are the only place to look for disorders, and they cannot be found there, one is led to admit that folk psychology and the DSM diagnostic categories play an irreversible role in how disorders are determined. One would be led to this conclusion, if as I argued, explanatory interests play an inalineable role in determining how mechanisms are delineated. This conclusion, on the other hand, can be avoided by a more ecumenical approach to explanatory pluralism, that grants the role of seeking disorders to other disciplinary-approaches as well.

As already mentioned, Kendler (2012) and Kendler and Campbell (2009) have argued for a causal interventionist position, which can, but does not necessarily need to be, supplemented with a stronger mechanistic explanation.[69] Part of the motivation is that the mechanistic approach to psychiatry is currently too demanding.[70] Similarly, Kendler

[69] Campbell (2008) argues that causation between "levels" should not to be ruled out on a priori grounds.

[70] Campaner (2016: 115) mentions a central challenge that nonlinear interaction and loops occur between biological, social, and psychological levels of explanation. Also, according to Woodward (2015: 156), mental or psychological factors submit to causal explanations of behaviour and other mental states in psychiatric disorders because they can be invariant under interventions.

(2005) has previously advocated local and patchy integration through inter-level integrative pluralism. He argues that although thoughts and feelings, as well as other subjective experiences, are causally explanatory, they do not exist independently of the brain. Nevertheless, Kendler (2005: 436) argues that although humiliation is ultimately realized in the brain, it may not be the best "level" to explain humiliation, or the depression it may contribute to. Instead, if humiliation is realized by different brain processes in different individuals, the preferred site of intervention can be the mental "level" (Kendler and Campbell 2009). On the other hand, if there are different psychological forms of humiliation that can be explained by one brain state, the correct level of intervention would be the brain state instead. Kendler and Campbell (2009) add, nonetheless, that because the interventionist approach can be employed to specify causal dependencies, it can identify the need for the deeper underlying mechanisms of those connections. For instance, according to them, the acknowledgement that "stressful life events cause the risk of major depression" offers a deeper explanation than "divorce increases the risk of major depression", may invite a search for an even deeper mechanistic explanation that underlies stress. This would not only explain the role of stressful life events such as divorce in depression, but also causal factors such as chronic difficulties and hassles. Hence, according to them, although there is a causal connection between divorce and major depression, identifying the causal mechanism (e.g. a type of stress mechanism) would provide a deeper explanation. This approach seems correct to me, and my account is not in direct contradiction with it. The only reservation I have is that their mechanistic approach may not be able to accommodate sufficiently well social and psychological explanations.

Another account of explanatory pluralism has been offered by Mitchell (2003, 2008). According to her (2008: 35), integrative

pluralism is explanatorily useful because of "the unlikelihood that there will be a simple, single causal account at the end of a reductionist investigation into the causes of behavior of complex systems". She provides a framework of integrative pluralism, which is compatible at least with Murphy's exemplars (see Murphy 2013). While exemplars are formed to determine how different causes can bring about the typical effects of a disorder, integrative pluralism studies how each level or cause contributes on its own, so that causes can be thereafter integrated into models in different ways to explain a singular concrete case (ibid.: 130). Murphy (2008) adds that to understand the multilevel complexity of psychiatric disorders can only be done by picking relevant variables from multiple levels of description that do not align to correspond with the same phenomena. Thus, variables at one level are not sufficient to explain a disorder because they may depend on variables at another level (see Maung 2016). Consequently, these arguments seem to suggest that a single comprehensive theory of psychiatric disorders that would support useful generalizations is hard to reach (Zachar 2008: 44). However, this does not rule out the possibility of more local domain-relative explanations for particular kinds of psychopathology.

Sullivan (2017) has also supported an integrative approach to psychopathology, although she is sceptical about its possibilities in the short term. She argues that only the integration of multiple causes provides a satisfactory explanation. She (p. 461) points out that "Although we currently have a lot of piecemeal explanations for different kinds of mental illness and such explanations have shed light on avenues for therapeutic interventions – some of which have been successful – we still have no cures. Such piecemeal explanations are not considered to be ultimately satisfactory." For example, social and environmental contexts together with genetic susceptibility are important in explaining the onset of major depression. However,

Sullivan (2014) has rightly noted the pragmatic problems facing integrative explanations across different levels of description (see also Tabb and Schaffner 2017). These problems include the fact that different disciplines have different ways to operationalize where to search for mechanisms, how to do it, and where to intervene (Sullivan 2014: 261). The gist is that scientific research into psychiatric disorders necessarily needs to concentrate on some aspect of a disorder, to the extent that different explanatory approaches employ different experiments, and that these various methods may even cure the world in different ways (Sullivan 2017: 460). In addition, the dynamic nature of psychopathology contributes to the instability involved in these differences. That is, even if we can offer an account of how the interaction in principle would work, there is still an epistemic challenge concerning the influence of folk concepts, economic aspects, and so on, in real life (I will return to this the next chapter).[71]

For these reasons, I believe that a strictly understood explanatory integration may be unattainable as a general goal in psychiatry, on the one hand, and may sometimes be empirically unfruitful when attainable, on the other. Integration requires that different explanatory approaches must be directly compatible to the extent that they explain the same phenomenon (Mitchell 2003), or that some explanations will eventually be eliminated. However, it does not seem reasonable to require that explanatory success is only attained when various explanatory and disciplinary-related approaches to psychiatric disorders can be integrated or eliminated to provide one true theory of psychiatric disorder(s). One reason is that it is highly unlikely that alternative discipline-related explanatory programmes will voluntarily relinquish their own approaches. Moreover, as I argued, sometimes various explanatory approaches can be understood as emphasizing and

[71] Bental (2003) argues that due to the heterogeneity of causes, the biomedical model should be abandoned.

conceptualizing different causal aspects of psychopathology (i.e. parsing the causal reality differently).

My positive argument, however, is that the applicability domain approach to psychiatric explanation can facilitate a less demanding approach to explanatory pluralism. It seems plausible that when full integration is unattainable, as I believe it may not be in many parts of psychiatry, more limited success can be attained through interaction between different explanatory approaches. For this reason, the applicability domain approach can be understood as an attempt to understand psychiatric disorders epistemically as heuristic kinds whose explanations do not require an understanding of "material" interactions between the social, the mind, and the brain. This approach can nonetheless facilitate different methods, different research perspectives, and diagnostic and professional interaction between various scientific programmes to contribute to an understanding of different aspects of the same psychiatric phenomena.

The non-metaphysical explanatory pluralism that my account supports can be exemplified with schizophrenic voice hallucinations. A mere neurological explanation of voice hearing in schizophrenia is unable to explain why such voices are experienced in a certain way. To establish that kind of understanding, explanations that refer to social and cultural factors are needed. Moreover, the nature of voice-hearing experiences may interact with other social determinants in one's life, such as the ability to maintain a job or a relationship, which in turn may have consequences for how the disorder develops. As Lende (2014) puts it, "The brain is not like the heart – subjective 'symptoms' is part of how it functions". Hence, a fuller picture of voice hallucinations requires a hybrid explanation that consists of causal factors from different levels of organization, such as neurological and psychological organization, and their feedback interactions with external factors, such as social and cultural contexts. This means that even in cases where we can

explain aspects of some higher-level factors with lower-level ones, a more robust explanatory picture may require explanations that refer to higher-level factors and their interactions with external determinants. That is, even if neurological and psychological factors align in a constitutive relation, explanations on a higher level of organization may supplement the general applicability domain of the kind explanation. While this approach does not provide a biological mechanistic account of how those heterogeneous factors are linked to produce schizophrenia, it does refer to their difference-making dependency. Hence, this account can overcome the challenge of how to connect psychological states with biological factors in a mechanistic account.[72] On the other hand, it does not rule out "deeper" biological explanations when they can be provided.

I believe that Longino's (2013, 2015) moderate pluralism provides in science provides realistic account of interaction between various explanatory approaches and matches my non-metaphysical approach to the explanation of psychiatric disorders. The central idea can be summarized in the following way:

> [T]he plurality in contemporary science provides evidence that there are kinds of situations produced by the interaction of factors each of which may be representable in a model or theory, but not all of which are representable in the same model or theory. Each factor is necessary for the phenomenon to have the various characters it has, but a complete account is not possible in the same representational idiom and is not forthcoming from any single investigative approach (as far as we know). A more complete representation of some phenomena requires multiple

[72] See Eronen (2021) on the problem of linking psychological and biological factors to explain psychopathology.

> accounts, which cannot be integrated with one another without loss of content.
> (Kellert et al. 2006: xiv).

Hence, Longino (2013: 2) believes that rather than trying to determine which explanatory approach is best, it may sometimes be more fruitful to analyse what each explanatory approach can offer to an overall understanding of the phenomenon. This means that the success of explaining some phenomenon is not always to integrate various explanatory approaches, but instead to try to understand the explanations in their context. This is because attempts at integration can result in explanations that are too general, vague, and cumbersome (Van Bouwel 2014). As with Longino's study on aggression, the differences in some parts of psychiatry may run so deep that it is unlikely that one unified explanatory approach can be achieved either through integration or by elimination. Nevertheless, according to Longino, the benefit of interaction between different explanatory approaches is that they can identify strengths, limitations, and biases in each other's explanations. That is, in the light of this, psychiatric phenomena may not be comprehensible by adopting a single theoretical approach, but rather requires multiple approaches that can answer different questions concerning different aspects of the same phenomenon of interest (Longino 16: issues II). Van Bouwel (2014: 113) has provided a similar account that he calls *interactive pluralism*. He argues that interaction and engagement between explanatory approaches in psychiatry may sometimes be more successful than trying to reach integration and consensus (see also Van Bouwel et al. 2011). Faucher and Goyer (2015: 212) point out that according to Van Bouwel "theories at different levels are required to answer different types of questions, they don't necessarily need to be integrated in a coherent whole". Likewise, the idea behind the applicability domain approach to psychiatric kinds is

that understanding psychiatric kinds heuristically can facilitate interaction in the various approaches leading to an explanation. This in turn is able to provide (some) consensus over how to classify and conceptualize psychopathology (see also Dupré 2015).

Longino also points out that competing explanatory approaches may not carve the causal structure of the world in the same way. That is, different explanatory approaches pose different explanatory-seeking questions and consequently focus on the causal structure of the world differently. However, this should not be considered a bad thing. As mentioned earlier, rather than aiming to integrate competing explanatory approaches, there may be trade-offs between explanatory virtues in what is considered a successful explanation, and on how to carve the causal structure of the world (see also Pöyhönen 2014, Van Bouwel 2014: 112).

Longino's approach fits my applicability domain account of scientific kinds. Alternative explanations may either interact or supplement when explaining the kind, or different explanatory approaches may successfully carve psychiatric kinds differently to the extent that they match their respective explanatory aims. Moreover, it is plausible that an internalist approach to explaining psychiatric problems is not always the most suitable one when giving explanations. For example, in contrast to Kendler and Campbell's (2009) example, perhaps the devastating effects of divorce are realized in dynamic social interactions, so the problem is neither in the mind nor in the brain of the individual.

In conclusion, I have argued in favor of moderate or interactive pluralism in explaining psychiatric kinds. Based on it, the applicability domain approach can be seen as facilitating the interaction of different disciplinary approaches of psychiatric kinds. In the next chapter, I will explore how this can be done in practice.

5 Identifying the Explanatory Domain of the Looping Effect

I have argued that in principle psychiatric disorders can be scientifically relevant kinds because they can support domains of explanatory applicability. By tracking these kinds, psychiatric research can in principle explain and predict mental problems, as well as facilitate treatment interventions. In particular, I argued that knowledge of the applicability domain of a psychiatric kind explanation identifies the conditions under which the explanation is expected to be reliable and when it can break down. However, this realist approach to psychiatric kinds faces a challenge from social constructivism, according to which psychiatric classificatory concepts do not refer to independently existing patterns in the world. While I have argued against the anti-psychiatric view that psychiatric disorders are nothing but myths, another more naturalistically oriented constructivist argument can be presented based on Hacking's (1999) dynamic nominalism. According to Hacking, many kinds studied by the humanities and social sciences are not stable natural kinds, but are instead unstable interactive human kinds whose existence depends on our knowledge and our conceptions of them. Hacking has coined the notion of looping effect to describe the interaction that gives rise to these human kinds. The looping effect describes how classificatory practices in the human sciences interact with the members of targeted classificatory kinds, arguably rendering the kinds unstable and ruling out their realist interpretation.

Hacking argues that at least some psychiatric syndromes are interactive human kinds. In essence, psychiatric classifications can contribute to creating "conceptual space" for a suffering kind of person, a commonly recognized conception of the normal way to be abnormal. This enables relevant individuals to recognize and interpret their experiences and feelings as kind-typical, and thereby learn the proper kind-typical

intentional actions to take. As a consequence, (some) psychiatric classifications would not only sort out bundles of symptoms, but also kinds of persons, whose symptoms are partly glued and shaped together in virtue of their looping interactions with classificatory practices. On the other hand, Hacking argues that this interaction also induces unpredictable reactions and behavioural changes, which in turn may require classificatory modifications. Hence, the possibility that psychiatric kinds are subject to looping effects raises the question whether they are fundamentally different from the natural kinds studied by the natural sciences to the extent that they would require different types of classifications and explanations.

Hacking's description of the looping effect has instigated a debate over its scope and the nature of the instability it generates. Critics (e.g. Cooper 2004, Khalidi 2013, Mallon 2016) argue for a realist interpretation of interactive human kinds on the basis that some prototypical natural kinds are also subject to looping effects, such as domesticated animals and disease entities, and that looping effects can be stabilizing in addition to destabilizing. These critics argue that the looping effect cannot be used as a general demarcation criterion between the kinds of objects that human and natural science classify. However, recently, Laimann (2018) and Allen (2018) have defended Hacking's position. They argue that although some biological kinds are subject to the looping effect, human kinds differ from non-human kinds because their classificatory-induced reactions are generally unpredictable. In sum, the looping debate is based on the dichotomy of whether interactive human kinds are real kinds that support robust epistemic projects. However, both sides of the debate mostly approach looping effects, and the realness of affected kinds, dichotomously. My contribution is to argue that looping effects influence kinds in various ways, and this makes a difference to interpreting the realness of interactive kinds, such as psychiatric disorders. In particular, I wish to explore how the idea of the

looping effect can be employed to explain classificatory influence on psychiatric disorder in the context of a realist account.

In this chapter, I defend a realist account of interactive psychiatric kinds by arguing for an explanatory domain account of the looping effect. I begin by arguing that previous accounts of the looping effect have been based on a needlessly coarse-grained dichotomy. They have largely taken for granted that either kinds are subject to the looping effect or they are not. By relying on the contrastive-counterfactual theory of causation (Ylikoski 2007, Woodward 2003), as well as on the homeostatic property cluster theory of natural kinds (Boyd 1989), I assert instead that looping effects have dimensions of causal relevance and explanation. Based on this, I argue that knowledge of the feedback mechanisms that mediate the looping effect can supplement the applicability domain over which a kind and its variations are reliably explainable. In other words, if we can account for the possible variation of an interactive kind under alternative circumstances, it is stable enough for reliable scientific generalizations.

In the last section, I apply my explanatory domain account to analyse case studies. Influential research programmes in social psychology, sociology, and psychological anthropology build on realist interpretations of human categories and rely on feedback explanations. By applying the explanatory domain account especially to cross-cultural case studies of psychiatric disorders, I argue that there are two types of feedback mechanisms that mediate the looping effect. Congruent feedback mechanisms describe matches between classifications and kinds. They can supplement for instance, explanations of the cross-cultural variation of schizophrenia. On the other hand, incongruent feedback mechanisms describe mismatches between classifications and kinds. They can help to identify unintentional misclassifications and their effects, or more problematically for epistemic projects, value-driven classificatory and

behaviour adjustments. For example, classificatory practices that disregard culture-dependent manifestations of disorders exceed their applicability and hence cause incongruence. Finally, I suggest that knowledge of feedback mechanisms facilitates classificatory adjustments and interventions on interactive kinds such as psychiatric disorders.

5.1 The Looping Effect, Interactive Kinds and Realism

Ian Hacking (e.g. 1986, 1995b) uses the looping effect to characterize a phenomenon that underlies many social constructivist arguments. The looping effect describes the interaction between classifications and the targeted "kinds of people" or human kinds that purportedly share behaviour and traits.[73] The idea is that classificatory practices induce reactions in the members of the human kind by enabling new intentional ways of being and acting. Tracking these changes requires revisions in the original classification, which may in turn lead to further changes in the members of the kind. Consequently, the interaction between the classification and the affected members of the human kind creates a feedback loop that renders the kind a moving target. According to Hacking, this classificatory instability generated by the looping effect distinguishes the human sciences from the natural sciences. In particular, the interactive human kinds studied by the human sciences do not support the robust explanations, predictions and interventions (i.e. epistemic projects) that the natural kinds picked out by the natural sciences do.

[73] I use the term 'real kind' to cover all the kinds that ground robust epistemic projects.

The looping effect's plausibility as a demarcation criterion between the natural and human sciences depends on the nature of the generated instability and whether it prevents interactive kinds from being real kinds. As examples of the kinds subject to the looping effect, Hacking (1995b, p. 351-352) has concentrated on "kinds of people" defined by their behaviour, condition, actions, tendencies, emotion and experience. For example, he has provided case studies of psychiatric disorders, such as fugue, multiple personality disorder, schizophrenia and autism spectrum, in addition to cases like teenage pregnancy, child abuse and homosexuality (see Hacking 1986, 1995a, 1998, 1999). According to the analyses, classificatory activities have affected the targeted kinds to such a degree, that the classifications have had to be amended.

Hacking has provided a particularly detailed analysis of the looping effect of multiple personality disorder (i.e. dissociative identity disorder) (Hacking 1995a).[74] Until the end of the 1970s, the diagnosis was based on individuals having two or three alternating personalities. However, once knowledge about the syndrome spread among specialists, media and lay people, the number of diagnoses as well as the diversity and number of alters associated with the syndrome started to increase. Hacking argues that popular knowledge about the syndrome created a prototype, which in turn induced more people to conform with it or otherwise react to it in varied ways. These reactions needed to be explained and to be integrated into the classification, which again affected the syndrome. Consequently, by the 1990s people were diagnosed with having hundreds of fragmented alters.

[74] Hacking (2007a) has later amended his analysis to include five axes instead of two. The axes include the experts who classify, the institutions in which the classification and classified interact, and the knowledge gathered in this manner. As these axes interact, they change the self-concepts and behaviour of the targeted people.

The reactions generated by the looping effect should be distinguished from at least two other ways in which kinds can change. First, the kind may change within the range set by the classification for non-classificatory reasons, and therefore there is no need to amend the classification. Second, the kind may change outside the range set by the classification for non-classificatory reasons. In such a case, the classification may need to be amended to match the changes. However, I am solely concerned with the third type of case, where the targeted kind is affected by the classification and may change because of it.

Although Hacking's examples mostly concern scientific classificatory practices and their objects, folk categories and their objects are subject to the looping effect as well. In general, scientific classifications reinforce and sharpen the boundaries of pre-existing folk categories or create new categories that become folk categories (see Root 2000, p. 631). An example of the former is psychiatric disorder terms. Madness was historically an unspecified folk category that has become increasingly sharpened and recategorized due to scientific research, thereby affecting the behaviour of the classified people. Hacking (1986, 2002) describes homosexuality as an example of the latter. Following Foucault (1978), Hacking argues that the homosexual as a kind of person only came into being through legal and medical thinking. Subsequently, the kind of person was transformed when individuals began to adjust their self-descriptions and actions in reaction to the classification and emergent prototypes. Finally, the 'gay movement' changed the values and beliefs associated with the classification.

A classification may either generate merely reactions from the classified or also genuinely shape the attributes targeted by the classification.[75] In the first case, a classification may prompt reactions

[75] Tsou (2007) argues that Hacking concentrates on classificatory (status) looping effect.

due to the status or stigma associated with it, while the targeted kind-typical behaviour is not affected. According to Hacking (1999, p. 114), for instance, once auditory hallucinations became an integral part of schizophrenia diagnosis, people ceased to report them. This in turn led to their diminished importance in the diagnosis. In the more substantive feedback effects, classifications and associated stereotypes influence the behaviour and content of the kinds targeted by the classification. Along these lines, Hacking mentions that the content of the hallucinations also changed due to becoming diagnostically important. In such cases, the classification does not merely successfully or unsuccessfully pick out a pre-existing kind, but is part of the casual structure of the social world, and may prompt reactions that lead to changes in the kind as well as the classification itself. If the classification is perceived as a match and meaningful to the classified and the larger social audience, it may generate congruence (i.e. conformity with the classification), while a perceived mismatch may lead to capricious behaviour and increased incongruence (i.e. nonconformity) (see Table 2. and section 5.2.2).

Table 2. Looping effects

	Status reaction	**Kind looping**
Perceived match	No modification	Congruence
Perceived mismatch	Classificatory modification	Incongruence

Classifications of human kinds induce reactions primarily because they are value-laden (Hacking 1995b: 370).[76] Although Hacking concentrates mostly on negatively value-laden classifications, positively viewed human groupings, especially ones that have political

[76] Hacking (1995a: 370) writes that "the greater the moral connotations of a human kind, the greater the potential for the looping effect".

importance, are also subject to the looping effect.[77] This is exemplified by how the idea of an aristocrat as a kind of person was connected with high expectations concerning character and behaviour. The expression *noblesse oblige*, for example, describes the conferred social obligations that were associated with aristocratic titles. These perceived obligations interacted with the behavioural patterns and traits of the aristocrats.

The looping effect seems to challenge kind realism in the human sciences, including psychiatry. Hacking (1983; 1991; 2007a) maintains that there are no prepackaged kinds united by their common naturalness, but instead there are different ways of classifying that correspond with the varying nature of the targeted kinds. Some natural kinds classified by the natural sciences are better described according to the essentialist approach, while others according to the naturalist tradition (Hacking 1991, p. 123).[78] But more importantly, whereas the natural kinds studied in the natural sciences are indifferent to our classifications and manipulations, classificatory activities in the human sciences lead to the looping effect that renders the targeted human kinds moving targets or interactive kinds (Hacking 1999). The reason is that social phenomena do not constrain classificatory possibilities to the same extent as natural phenomena (Hacking 1986: 166; cf. Martinez 2009, p. 214). Because of the dynamic nature of the kinds studied by the human sciences, Hacking has labelled his view as dynamic nominalism.[79] In short, he uses the looping effect to draw a principal distinction between

[77] Hacking (1997, 2007a) mentions the "genius" applied to the Romantics as an example. I believe that the categories of normal and abnormal are subject to looping connections, so that when category is altered, the other changes too. For instance, the idea of normal weight preceded the diagnosis of obesity.
[78] Lately Hacking (2007b) has argued that the concept of natural kind has become obsolete because of the confusion it creates.
[79] Alternatively Hacking (2002: 2) calls the view dialectic realism because of the interaction between our concepts and the world.

the instability of interactive kinds and the stability of natural or indifferent kinds.

The argument from the looping effect to non-realness (or nonnaturalness) of interactive kinds is not straightforward. Clearly, because the looping effect may lead to property changes, interactive kinds cannot be essential kinds. However, it is not clear why the looping effect would preclude all interactive kinds from grounding robust epistemic projects. Hence, there is an ontological and an epistemic question that require clarification: what renders interactive kinds susceptible to the looping effect, and why would their instability be epistemically problematic for classifications? The questions are co-dependent, but it is analytically helpful to keep them separate.

5.1.1 The Looping Debate

Hacking's account of the looping effect has stirred a debate over whether the affected kinds can be given a realistic interpretation (see Tekin 2016). Critics argue that the looping effect does not preclude human kinds from being real kinds because some prototypical natural kinds are interactive and some looping effects of human kinds are mediated by environmental changes.[80] For example, Douglas (1986, p. 101) has pointed out that microbes adjust themselves to our classificatory and medical interventions. Especially Cooper (2004) and Khalidi (2010, 2013) have developed this idea by arguing that the looping effect is not restricted to the examples involving human behaviour that Hacking offers, but affects biological kinds as well. For instance, labelling some species as domestic animals has led to selective breeding, which in turn has eventually created new breeds of dogs and

[80] By prototypical natural kinds I mean biological species and chemical elements that were considered natural by Kripke (1980) and Putnam (1975).

cats, for example. Consequently, we have had to adopt new labels and taxonomies to match these changes. Similarly, labelling bacteria and viruses as diseases leads to medical interventions that change them through natural selection. This in turn requires adjustments in treatments and labels. The gist of Khalidi's and Cooper's argument is that since some biological kinds clearly ground inductive inferences and explanations, there is no reason why other interactive kinds cannot do the same.

Hacking (1999, p. 106), however, has pointed out that genuine looping effects are mediated by the awareness of being classified, thereby distinguishing interactive kinds "ontologically" from indifferent kinds. This interpretation is supported by Hacking's (1986, 1995a, ch. 7) relying on Ascombe's theory of intentional action under a description. Once a new description, associated with a classification, becomes socially available, it prompts reactions by enabling new conceptual possibilities for being and acting. These new actions, in turn, render the kind a moving target. Cooper (2004) and Khalidi (2013), nevertheless, argue that even according to Hacking himself humans need not always be aware of the classification for it to influence their kind-typical behaviour. For instance, a child labelled as having attention deficit hyperactive disorder (ADHD) may be placed in a school where "stimulant-free schoolrooms" influence her behaviour (Hacking 1999, p. 103). Similarly, an individual can acquire refugee characteristics by being part of a refugee group (Hacking 1999, p. 32). Furthermore, realists argue that most of the time the instigated changes in human kinds are not fast enough to rule out epistemic projects (Mallon 2016), and looping effects can be stabilizing as well as destabilizing (Griffiths 1997; Murphy 2006; Kuorikoski and Pöyhönen 2012).

However, lately Laimann (2018) and Allen (2018) have defended a modified version of Hacking's original position. They argued that interactive human kinds differ from non-human kinds because they are

prone to wayward behaviour. Laimann (2018) argues that looping effects render human kinds epistemically capricious because their behaviour often invalidates existing classifications and knowledge about them. She seems to associate the problem with the complex nature of human interactions that undermine our ability to discover mechanisms that underlie patterns of change and stability. Allen (2018) comes to the same conclusion by arguing that intentional action is ontologically different because it enables humans to fake or be mistaken over their kind membership. Arguably, these creative reactions are harder to predict and rectify than mistakes over non-human (or non-aware) classifications.

These problems are not restricted to feedback effects that are mediated by subjective awareness on the part of the classified individuals. First, not all intentional beliefs and attitudes are propositional (Allen 2018). As Allen (2018) argues, intentional action does not require a linguistic description. She argues that it is enough that the agent can differentiate between different outcomes of action to intentionally choose between them. As an example, when my sister's pet dog Coco rolls over onto her back, she does it intentionally to get patted, not to get slapped. So, due to being brought up as a lapdog, some of Coco's kind-typical behavior is a collection of individual intentional acts that she has learned. And the small intentional choices Coco and other lapdogs make daily, are learned through training to become lapdogs, although none of them are presumably aware of their conceptualization as lap dogs. Consequently, this type of structural feedback is different from the non-intentional behavior brought about by breeding. Crucially, breeding does not bring about behavioral changes through the intentional acts of breeds, but instead by tinkering with their natural selection.

Second, not all members of a kind need be aware of being members of a kind, for the membership to influence their intentional choices. Rather, the members may imitate and learn the behavior of the other members of their in-group without being explicitly aware of their group-membership. For instance, according to the self-categorization theory in social psychology, individual opinions that guide our behaviour are reinforced when conformed by people in our in-group (Reicher, Spears and Haslam 2010: 53). Thus, for instance, refugees may behave in a refugee typical way through a learning process, in which at the other extreme are fully aware individuals, and at the other extreme, for example, children guided by their parents. A refugee child may follow fine-grained guidance on how to behave without knowing that the individual acts build up to a refugee-type action. Hence, although the action is not done directly under the description of being a refugee, it is nonetheless done because intentional choices between different courses of action have become possible for all the member of the in-group through the awareness of the classification by some of the members[81]. Finally, Mallon (2003: 343-345) argues that labels may render some intentional choices strategically salient so that none of the classified need be aware of the classification. That is, a label may change the social and physical environment so that certain courses of action become socially available or prohibited to the classified people. In sum, classifications may render some intentional actions possible without all the members, or any of the them, being aware that the actions are kind-typical.

These different grades of intentional action are manifest in classification of psychiatric disorders. For example, apparently one of the reasons for the ADHD epidemic in the USA is that some schools and caretakers pressure clinicians for diagnoses to obtain medicine and improved learning facilities (e.g. tutoring, smaller classes) for children

[81] Sometimes individuals in decisive positions may become the driving force in instigating a kind or its change (Sahlins 1985, Robbins 2004).

(First 2017). This makes misdiagnoses probable, leading to general diagnostic distortions. These problematic feedback loops can also occur when children act under the description of ADHD, without being explicitly aware of the classificatory description. It is enough that they learn the characteristic action pattern from the people in their "in-group". In fact, it is probable that a person's symptom profile depends on the specifics of the mediating feedback mechanisms and their complex interactions with the underlying psychological and neurocognitive processes, as well as on the larger social and cultural forces.

In conclusion, the creative and complex nature of human reactions to being classified complicate our ability to explain and predict kind-typical behaviour. However, in the next section, I argue that interactive human kinds, and the intentional actions that generate them, do not require a different type of explanation from non-intentional explanations. Therefore, the question about the ontological difference of interactive human kinds can be set aside. In the third section, I defend an explanatory account of feedback effects according to which their complexity does not preclude interactive human kinds from sustaining scientifically relevant epistemic projects.

5.1.2 Explanation and Epistemic Instability

The looping effect as a demarcation thesis can either be interpreted as a strong claim in favour of anti-naturalism about explanation, or as a weaker argument about causal construction which is compatible with naturalism. The former would mean that the ontological nature of human kinds in principle precludes them from being real kinds, whereas according to the latter approach the problem is not principal, but instead an epistemic problem due to causal complexity.

As an anti-naturalistic argument, the looping effect would be a thesis in favour of non-naturalist interpretivism, and would support the separation between the kinds studied by the human and natural sciences. Interpretivists generally argue that humans are self-interpreting, and therefore understanding intentional action requires interpreting its meaning to the agent, instead of explaining it by causes or laws (e.g. Winch 1958; Taylor 1971; Geertz 1973). Moreover, arguably interpretations need to account for the culturally situated and holistically determined beliefs, concepts, categories and the like. These make actions meaningful to agents and understandable, in the light of their context, to observers. Hence, one could hold that the efforts to explain human groups with causal generalizations induce new self-interpretations whereas the aim should be to understand the humans according to their own meanings and concepts.[82] The problem is that this muddles the distinction between classifications and kinds. [83] One way to understand the problem is that interpretivism seems to set the produced action in a conceptual (or quasi-logical) connection with its reasons (von Wright 1971). That is, while actions are identified based on the agent's own reasons (i.e. beliefs and desires), those reasons can only be established based on the actions they are reasons for (see Rosenberg 2016, ch. 3). Consequently, behaviour could be interpreted

[82] Hacking (1999: 31) seems to have something like this in mind when he writes about awareness being necessary for the looping effect. Perhaps the overall thought was that understanding human behaviour requires interpreting the concepts and meanings that guide our behaviour, instead of explaining behaviour with causes or laws (cf. Hacking 1999: 123).

[83] This would mean that the two axes of a looping description (from classification to kinds and back) would be logically connected to the extent that a description of a classification necessarily refers to its effect. However, the problem with arguing that there is a conceptual connection between classification and the actions they describe, is that it conflates events in the world with the way we describe them (See Davidson 1963). This problem has led some interpretivists to demand a more sophisticated view, according that although reasons can function as causes, they nonetheless do not enable relevant causal explanations.

and conceptualized only in retrospect, making objective and generalizable classifications virtually impossible. Lastly, the conceptual connection could describe how classifications constitute human kinds, not only how they can be explained. Indeed, in his earlier work, Hacking (1986: 166) argues that kind-typical actions are logically impossible before the invention of their explicit category descriptions. This implies that kind actions are constituted by their conceptualizations, not only conceptually connected in their explanations. Nevertheless, in all these cases, generalizations of human kinds would be mostly exhausted by their classificatory descriptions, and hence would not be real kinds that ground robust projections.

In the following, I will first argue against the non-naturalist explanatory view of the looping effect, and thereafter against the stronger constitutive view. The general view in the philosophy of social sciences is that reasons can function as causes in explanations and that interpretation and causal explanation need not be mutually exclusive (Tuomela 1977; Henderson 1993; Kincaid 1996; Ylikoski 2001). As an example, when an anthropologist conducts fieldwork by interpreting local customs and behaviours, she relies on the causal efficacy of cultural structures and beliefs. Especially the already mentioned contrastive-counterfactual theory of explanation (Ylikoski 2001; Woodward 2000, 2015) matches the thesis that explanations in the social and natural sciences do not differ in key features. According to the theory, explanations provide descriptions of objective causal (and constitutional) relations by describing counterfactual dependencies that answer *what-if-things-had-been-different* questions (*what if*-question). Explanations, in addition, have a contrastive structure, so that they answer questions of why fact rather than foil. The contrastive structure makes explicit the aspects to be explained and thereby determines whether a putative explanation is relevant. Hence, according to the theory reasons are causally explanatory because they can provide

counterfactual information on why someone acted in one way rather than another. This means that a putative intentional explanation is explanatory if it can answer questions concerning how the explanandum action would have been different, had the relevant beliefs and desires been different (see Ylikoski 2001, p. 97).[84] Based on this, a putative feedback explanation is explanatory if it can describe how a difference in classificatory related conceptions and beliefs would have made a difference to the behavior of the classified people. But having said that, understanding the causal process may require resorting also to lower-level explanations as well as structural explanations. Moreover, in the next section, I argue that some feedback descriptions are not just singular causal explanations but support stable generalizations.

The constitutive view of human kinds can be understood as a form of conventionalism (cf. Kornblith 1993). A strong interpretation of conventionalism would mean that human kinds are merely subjective distinctions or groupings imposed by scientists qua scientists. However, this would trivialize human kinds to the extent that kind membership would be merely a matter of opinion (see Griffiths 1997, p. 198) or "in-the eye-of-the beholder" (Cooper 2004)[85]. A weaker conventionalism would mean that interactive human kinds are akin to institutional facts which, according to Searle (1996), are epistemically objective although constituted by collective acceptance. For example, because central banks have the final say on what constitutes money, we can be individually wrong about which tokens are money, but cannot be collectively wrong about the type money. Crucially, however, interactive human kinds are not constituted merely by collective

[84] Relying on the interventionist theory, Woodward (2015, p. 157) argues that mental states can be invariant under interventions and hence causally efficacious in explaining the interaction between mental and cognitive factors of psychiatric disorders.

[85] Cooper makes a strong case in favour of understanding the looping effect as a causal process.

representations, or by representations in the minds of scientists qua scientists, but also by the behavioural patterns and traits they bring about. As an example, a common belief that money is not safe in a bank does not constitute the banks insolvency, but may instead bring it about (Kukla 2000). Similarly, the existence of human kinds does not merely depend on beliefs and minds of scientists qua scientists, but in addition on the classificatory conceptions and practices that constrain, shape and enable kind-typical behaviours and traits. Thus, the conceptions and beliefs associated with the label multiple personality disorder do not constitute multiples as kinds of people, but those beliefs as part of classificatory practices (treatments, institutional infrastructures, etc.) and prototypical expectations, may constrain, shape and enable kind-typical behaviours and traits. In other words, individuals do not need to hold correct beliefs about their own behaviour or the kinds. This means that an individual merely perceived as a member of a human kind does not become a member in consequence. This is exemplified in how a posteriori research is needed to uncover the reasons that brought about the behavioural pattern or its reinforcement.[86] And since I already argued that reasons can function as causes in kind explanations, nothing in principle prevents looping effects from describing causal interactions between classificatory descriptions and human kinds.

Shared conceptions may nonetheless be necessary to enable the complex social interactive processes that bring about and sustain human kinds. However, it is not plausible that new human kinds and kind-typical actions are born from conceptual stipulations. Instead, it seems credible that there are different degrees of conceptually and socially induced kind-typical intentional actions. This means that novel kinds come about incrementally, so that the kind and its conception are egging

[86] Guala (2016: 180) argues, as an example, that understanding social kinds requires studying how people behave, not what they themselves believe about their behaviour.

one another on in various ways. As an example, the diagnostic category of ADHD has grown more specific through institutional and social interactions (Lakoff 2000). In addition, kind-typical intentional actions may become possible before their linguistic expressions. Raymond Williams, for instance, has coined the notion "structure of feeling" to describe how social experience can take place when social structure is still in process (Williams 1978: 132). That is, "structure of feeling" describes how social awareness can be understood as a "lived identity", as something that cannot be expressed with the extant categories. The notion can be extended to cover kind formation: classificatory intentional action may become possible already when its linguistic expression is still in the making. The idea is that feedback processes of social interactions can brings about human kinds by generating compelling emotional experiences (see Collins 2004, p. xii). For instance, Siegel (1997) argues that Indonesian national identity first guided behaviour rather as a structure of feeling than as an explicit category. Crucially, stories that were spreading in the Indonesian language seemed to suggest that although the Dutch colonizers considered the Indonesians different, they nonetheless held them to be as good or even better than themselves. This in turn raised a feeling of nationality that could not be expressed within the dominant ideology.

The antirealist view of interactive human kinds can also be defended naturalistically so that the looping effect is a form of causal construction that destabilizes the affected kinds (see Hacking 1995b: 362).[87] However, those who take a realist approach to human kinds argue that looping effects can be stabilizing and that in general our theoretical knowledge can keep up with their rate of change (Mallon 2016; cf. Murphy 2001). Laimann (2018) claims, instead, that interactive hu-

[87] Hacking's retrospective approach describes historical kinds or "kinds-in-hindsight".

man kinds are generally capricious and problematic for epistemic projects because of the difficulty in discovering the mechanisms that underlie their patterns of change and stability. The reason is that feedback mechanisms can interact in complex and unpredictable ways with each other, and with larger social circumstances, *and because individuals may associate different meanings with their classifications.*[88] Consequently, she argues that secure extrapolations based on kinds in the human sciences are difficult if not impossible. This interpretation would mean that human kinds are historical in the sense that we can explain their alterations and context dependent stability only in retrospect.

I agree with Laimann's general idea to the extent that our ability to explain human kinds and their property variations, rather than the superficial stability of the kinds by itself, is the key to measuring the epistemic stability and scientific relevance of human kinds. That is, superficial stability itself is not much use if we do not know the underlying causes for the stability. Many human kinds are culturally and historically contingent, and unless we know which properties are kind-typical, we may easily be mistaken over which tokens are of the same kind. Conversely, knowing how the kind would vary under different circumstances supports its realist interpretation.[89] However, since I have argued that there is nothing in principle preventing human kinds from supporting epistemic projects, whether they do support projects is

[88] Laimann (2018) defends Hacking's distinction by arguing for the dual nature of interactive kinds, which are constituted by a base kind and a status kind. The category schizophrenia, for instance, refers to stable biological properties, i.e. the base kind, and in addition, to the social status associated with the category. Importantly, only the status kind is subject to the looping effect, thereby making it an unpredictable "capricious kind". (There are challenges in distinguishing biological and social properties, see section 6.2)

[89] Human kinds can be considered to have different dimensions of "realness" based on their susceptibility to explanations.

an empirical question. Next, I argue that the looping effect's explanatory relevance, and the epistemic projects that the human kinds support, are not all or nothing matters.

5.2 An Explanatory Account of the Looping Effect

5.2.1 Explanatory Domains of Feedback Mechanisms

In this section, I argue for an explanatory domain approach to the looping effect. The idea is that classificatory concepts of interactive kinds are associated with explanations that set a domain of applicability for explainable property cluster variation of a kind. Hence, inclusion of the looping effect into explanation of an interactive kind may enable more secure generalizations, and in some cases, predictions based on that kind.

My account is that the looping effect is mediated in some cases by feedback mechanisms that can supplement explaining the clustering of properties of human kinds. I follow Kuorikoski and Pöyhönen (2012) example and argue that some looping effects can be understood as middle-range theories of social, psychological and cognitive feedback mechanisms. Moreover, a common naturalist and realist approach to interactive human kinds has been to interpret them as homeostatic property clusters (HPC view) (Griffiths 1997; Kuorikoski and Pöyhönen 2012; Pöyhönen 2013b; Kokkonen and Koskinen 2016; Hauswald 2016). My argument is that if feedback mechanisms are part of the causal structure of human kinds, knowledge about them can support more secure domain-relative extrapolations and projections. That is, knowledge about a shared feedback mechanism may enable secure extrapolations between property clusters, and reliable

projections over kind-typical properties that the members are likely to share.

I argued in the last chapter that mechanistic explanations of human kinds have *applicability domains* over which they stably account for the kinds' properties under counterfactual situations. The domain of applicability – the range over which the explanation is reliable and stable – consist of a scope and depth dimension. The scope specifies the set of phenomena that the explanation is applicable to, whereas the depth describes how the explanatory relation would hold under alternative counterfactual situations. Hence, the applicability domain is better when it holds over more aspects of the kind in a wider range of alternative situations. This means that explanations of a homeostatic property cluster kind fall on a continuum of goodness described by the applicability domain. The idea is that a better mechanistic explanation of a kind enables more secure domain-relative projections (i.e. generalizations and predictions) based on the kind. Moreover, a stable explanation spells out why its domain of applicability is optimal for explaining the targeted set of phenomena. The reason is that human kinds support limited epistemic projects. In that case, classificatory projects that apply the explanation do not exceed the limits of their own applicability. In sum, identifying the applicability domain of an explanation spells out the conditions under which the explanation is expected to be reliable and when it can breakdown.

Based on this account, I will now argue that knowledge of feedback mechanisms can supplement the domain over which a human kind is explainable by accounting for some of its dynamic properties and identifying the limits of the domain's applicability. Hence, if feedback mechanisms are part of the causal structure of human kinds, knowledge about them can support more secure domain-relative extrapolations and projection. In that case, the scope of a feedback explanation describes specified properties of the kind in contrast to foils, while the depth

describes the explanations counterfactual power in explaining the scope. Nevertheless, the actual explanatory relevance of a feedback mechanism is an empirical question because interactive human kinds are affected in different ways and to different degrees by looping effects.

The explanatory domain of applicability can be illustrated with Luhrmann et al.'s (2015) study on the nature of the hallucinatory voices schizophrenia patients hear in the USA, Ghana and India. The study indicates that the voice-hearing experiences are exceptionally harsh in the USA in comparison to Ghana and India. The study seems to suggest that one of the reasons for the harsh voices in the USA is a feedback mechanism between the prototypical expectations associated with the diagnostic category and the hallucinatory voices. In this light, the feedback explanation's scope is the nature of the hallucinatory voices schizophrenia patients hear in the USA during a specified time-scale, whereas the explanatory relation is counterfactually relatively stable given that the diagnostic expectations are institutionally and socially entrenched. This means that if we or some social process were to manipulate the diagnostic expectations during that time, the nature of the voice-hearing experiences would change. Finally, such an explanation should spell out why its scope is limited to the USA and whether it has exceptions. Moreover, it should be established that the explanatory depth is correct for the explanation's scope, so that, for example, lower-level details, such as schizophrenia's different genetic subtypes, are irrelevant to explaining the voices in the USA in contrast to India and Ghana.

The explanatory domain account offers a method for identifying how relevant a feedback effect is in explaining a human kind. As I mentioned, explanatory relevance can be better determined by making explicit the contrastive structure of the explanatory-seeking question. Pöyhönen (2010, 2014). has supplemented this idea by pointing out

that it enables comparing alternative delineations and contributions of explanations (see also Van Bouwel 2014). A stable feedback explanation can supplement the domain of applicability of a kind explanation. Alternatively, knowledge of a feedback effect may help to identify how a putative explanation is unstable. Furthermore, feedback explanations fall on horizontal and vertical axes that represent their ability to supplement the applicability domain. An explanation is enhanced horizontally by widening its scope, or vertically by providing a deeper explanation within the scope. Although it is commonly argued that wider explanations are less deep, I rely in my explanatory analysis mostly on strong complementarity (Marchionni 2008), so that relevant feedback explanations can enhance both dimensions. This strategy is important for explaining interactive kinds that have inseparable biological, psychological and social properties. In such cases, integrating higher-level explanations, such as feedback mechanisms, with lower-level ones, can strengthen the explanation's applicability domain. In the schizophrenia case, the feedback mechanism can both provide a deeper explanation of the disorder's core symptoms in the USA, as well as account for why the explanation's scope is limited to the USA.

The depth dimension of a feedback mechanism's explanatory domain can be further explicated with the contrastive-counterfactual theory. I mentioned that the depth of an explanation depends on how many relevant *what if*-questions it can answer. Therefore, a feedback explanation can contribute to the explanation by providing fine-grained information of the counterfactual dependence between classificatory descriptions and properties of the kind. In this case, the feedback generalization describes how the kind would change, if its classification were to change in diverse ways, and vice versa. This counterfactual dependence can be explicated with the mentioned explanatory virtues

of *insensitivity* and *precision*, and in addition with *factual accuracy*.[90] The insensitivity of an explanation describes the invariance of the explanation under different background conditions (Woodward 2000; Ylikoski and Kuorikoski 2010). In other words, if a feedback mechanism makes a difference to the targeted aspects of the human kind (i.e. the explanandum), including it into explaining the kind (instead of leaving it as a background condition) would enable the explanation to answer more *what if*-questions.[91] An explanation of schizophrenia that includes a feedback explanation of diagnostic expectations is more insensitive if the expectations make a difference to the psychiatric disorder. The explanation's *precision* describes its ability to characterize in a fine-grained way why something is the case in contrast to something else (Ylikoski and Kuorikoski 2010). In the schizophrenia case, the feedback explanation could make an explanation of the psychiatric disorder more precise by describing in (more) detail the nature of schizophrenia voices in the USA in contrast to their nature in India and Ghana, or better still, everywhere. The *factual accuracy* describes the intentional falsehood an explanation incorporates (Ylikoski and Kuorikoski 2010). The most detailed and true explanation may not be the most relevant. [92] In the mentioned case, both the prototypical diagnostic expectations and the voice hallucinations are idealizations to make their explanatory connection salient.

[90] Other explanatory virtues could be relevant as well, see Ylikoski and Kuorikoski (2010).
[91] It is important to distinguish insensitivity from scope. An explanation of a kind can be very insensitive, although there may not be many actual tokens of the kind that it applies to.
[92] Intentional falsehoods are better understood to relate to models but not explanations. That is, multiple models may supplement an explanation of a psychiatric kind.

The applicability domains of feedback explanations can also be roughly compared. For instance, since severe autism has a strong neurobiological basis, and the individual's ability to communicate is limited, looping effects are not mediated by intentional reactions. The looping effect may nonetheless explain some behavioural responses to the actions of caretakers and environmental changes (Kuorikoski and Pöyhönen 2012). Moreover, a classificatory feedback explanation of pathogenic bacteria can supplement evolutionary explanations to account for antibiotic resistance. Nevertheless, the feedback explanation alone is unable to explain how subtle differences in classification would have made a difference to the bacteria. In sum, these feedback explanations are not as stable as the feedback explanation of schizophrenia.

Classificatory feedback mechanisms can be compared with self-fulfilling prophecies. They are based on expectations becoming a key component of the causal mechanism that generates the expected outcome (see Biggs 2009). This means that the prophecy can be invalidated if relevant people learn to intervene on the mediating causal mechanism. Consequently, if classificatory feedback mechanisms are self-fulfilling prophecies, disseminating knowledge about them could diffuse their causal efficacy. However, a classificatory feedback explanation's causal power to diffuse itself is limited. As Mallon (2016) argues, many social categories are firmly entrenched in larger social and material environments, thereby restricting one's space for action even if one becomes aware of their social nature. This is one of the reasons why true social change requires a highly concerted effort. Conversely, not just any enforced expectation or arbitrary claim will initiate a kind shaping or enabling looping effect. The reason is that feedback mechanisms bring about or reinforce interactive properties of kinds by interacting with other factors. It is commonly agreed, for example, that psychiatric disorders need multifactorial explanations

that combine social, psychological and biological causal mechanisms and causes (Kendler, Zachar and Craver 2011).

The modularity requirement for mechanistic models provides a way to understand the limits of feedback explanations. According to Woodward (2002), a model is modular if an intervention on a putative cause does not alter the subsequent causal relations in the underlying structure of the represented causal mechanism. The unaltered causal structure ensures that the model can predict the outcome of an intervention. Building on this idea, Steel (2006) argues that some interventions in the human sciences violate modularity because they are structure altering.[93] This means that the interventions cause unpredictable changes in the causal relations between the parts comprising the modelled social mechanism. In this light, the looping effect describes how classifications as part of classificatory or bureaucratic practices (understood as interventions here) alter the structure of interactive kinds inadvertently. The reason can either be epistemic, as argued here, so that sometimes the changes can be explained and anticipated by unboxing the mediating feedback mechanisms and their interactions with other mechanisms.[94] On the other hand, antirealists seem to claim either that the interactions are too complex to be modular, or that human reactivity eschews modularity in principle. This would also mean that if a putative feedback explanation is incorporated in classificatory practices, it will become less

[93] Strictly speaking, this is not the classificatory looping effect that Hacking writes about, but a feedback effect brought about by theoretical beliefs. Such critiques have been influential in economics. See Lucas (1976) and MacKenzie (2008) (see chapter 4 about critique toward modularity).

[94] This matches Mallon's argument (2003: 332) that the social origins of an interactive human kind need to be covert for the classified people and other related people for the kind to support robust projections.

explanatory by breaching the original feedback structure.[95] However, whether a feedback structure can be modelled so that it enables some interventions, is ultimately an empirical question. Indeed, a model need not provide a complete description of a feedback system, and its interaction with other mechanisms, to enable domain-relative interventions and causal predictions. Next, I provide some empirical evidence that knowledge of the feedback mechanisms' applicability domains facilitates explanations and predictions of interactive kinds as well as interventions on them.

5.2.2 Congruent and Incongruent Feedback Mechanisms

In the following, I employ empirical case studies to argue that the looping effect is mediated by congruent and incongruent feedback mechanisms. They are abstract and rough models that need to be filled with empirical details to generate generalizations and predictions. The distinction between congruent and incongruent feedback mechanisms is meant as an analytical tool for examining the looping effect's explanatory usefulness. In practice, the looping effect does not neatly divide between the two feedback mechanisms, but different feedback mechanisms may explain different aspects of the same kind. Moreover, the proposed feedback mechanism are meant to be understood as mechanistic schemes that require modifications to fit specific cases. Nevertheless, understanding these feedback mechanisms is relevant for identifying and explaining psychiatric disorders. Ideally knowledge of

[95] Woodward (2020: 441) points out that causal cycles are common in psychiatry. He also argues, however, that this does not rule out a causal explanation if it is possible to intervene on each variable independently (Independent Fixability Condition), and that cycles cannot be represented by a single directed acyclic graph (DAG).

the mechanisms can supplement, and help to identify the limits of, the applicability domain over which a disorder is stably explainable. This knowledge can also facilitate mitigating negative feedback effects.

Congruent mechanisms explain how feedbacks generate, reinforce and maintain stabilizing loops between classifications and interactive kinds.[96] Intentionally mediated congruence ensues when the classification is found meaningful and natural by the classified so that, for example, it seems to explain and exonerate one's experience, condition and behaviour (see Hacking 1995b).[97] According to Mallon (2016, p. 73-93), classified behaviour as *a social role* can also be preferred for strategical reasons, reinforced culturally and amplified by non-intentional automatic processes.[98] Mallon maintains that these causal mechanisms may lead, under the right circumstances, to a social role becoming structurally entrenched in social, material and institutional environments.[99] The idea is that structurally entrenched social roles are stable enough to be homeostatic property cluster kinds. I interpret this so that the interaction between the mentioned psychological, *cognitive* and social mechanisms (and other factors) may form a feedback mechanism that explains congruence between classifications and kinds. In this case, the classificatory practices and conceptions do not only provide opportunities and constrain behaviours

[96] In most of the cases Hacking describes, the amount of congruence describes the number of people conforming to the classification and the classificatory properties they exhibit. For example, during the multiple personality disorder epidemy, both the number of people exhibiting the typical behaviour associated with the label and the properties associated with the classification increased.

[97] See also Appiah (2005) and Haslanger (2012).

[98] Mallon (2006: 6) considers social roles to be representations that can include, for example, attitudes, theories, and concepts. Griffiths (1997: 142) mentions that a behaviour pattern can become virtually reactionary when it is culturally reinforced.

[99] A distinction can be made between the primary entrenchment that one is born in (e.g. race, gender) and the secondary entrenchment that one may be placed in later (e.g. mental disorders).

from above but also render the behaviours and experiences meaningful and seemingly natural. However, the extent to which interactive human kinds support scientifically relevant epistemic projects does not only depend on the stability of the kinds, but also on our ability to explain their properties under domain-relative alternative circumstances. Consequently, it is crucial to identify how congruent mechanisms can supplement the domain over which kinds are explainable.

Congruent mechanisms described by the labelling theory can explain why classified members of a kind have a higher disposition to exhibit kind-typical properties in contrast to unclassified members. The labelling theory describes how labels cause the targeted people to adjust their behaviour and self-images to conform to the labels, although the theory does not describe how this can reinforce the theoretical beliefs associated with the labels. Becker's (1953, 1963) account of the labelling theory, for example, is based primarily on congruent mechanisms mediated by intentional pathways. He argues that labels stick when individuals learn and internalize the concepts and meanings associated with them. As an example, to become a marihuana user, one needs to learn to use the drug, learn to recognize the effects, and learn to enjoy them (Becker 1953). Scheff's (1966) account of the labelling theory, instead, concentrates more on the influence of societal reactions and material environments. However, a problem with Scheff's approach is that when the inflicted individuals do not find their labels meaningful and natural, they may oppose them. As an example, Mclorg and Taub (1987) demonstrate that while anorectics tend to vigorously and openly deny their label, bulimics find their label more meaningful and natural. The reason is that dieting is not as readily conceived as deviant as excessive eating and vomiting. Moreover, some critics (e.g. Gibbs 1971) of the labelling theory have argued that labels cannot be the initial or primary cause of deviant acts because also unlabelled individuals in similar situations perform them. The explanatory domain

of the labelling theory is limited especially in cases where there is an underlying psychiatric disorder (cf. Gove 1975).

Nevertheless, this does not rule out the fact that explanations based on labels and feedback effects can supplement other explanations. For example, Lemert (1967) divides labelling effects into primary and secondary deviance. The primary deviance describes the many causes for a person for performing acts of deviance. At this stage, the label may not be the primary cause for deviant action, and is not considered part of the identity or social role of the individual. Secondary deviance describes how the labelled person amends her behaviour due to being labelled and how she assumes the deviant role. According to Lemert, the labelling theory explains the transition from primary deviance into secondary deviance through the labelling process. Therefore, in accordance with the distinction between primary and secondary deviance, labelling cannot be the only causal explanation for a person's behaviour. However, although the pattern of behaviour may have existed before, and is the reason why it is labelled in the first place, the kind may still be reinforced by the interaction between the label and the behaviour. According to Link et al.'s (1989) modified labelling theory, for instance, limited social opportunities together with internalized expectations of being socially rejected, may reinforce patterns of behaviour and conditions that have resulted from other causes. In criminology, for example, labelling effects mediated by intentional and structural mechanisms are used to predict the development of and propensity for criminal behaviour.

Congruent feedback mechanisms can supplement the explanatory domain of psychiatric disorders in several ways. This is illustrated by the already mentioned anthropological study lead by Luhrmann (2015), which indicates that schizophrenia voice hallucinations in the USA are harsh in comparison to Ghana and India, where they can be guiding and playful. The study, and a book by Luhrmann and Marrow (2016), seem

to suggest that one of the reasons for the voice-hearing experiences in the USA is that the prototypical expectations associated with the diagnostic category bring about negative social consequences.[100] In the West, the prototype of schizophrenia holds that it is chronic and devastating. In India and Ghana, on the other hand, patients and their families rarely know or remember the diagnosis, but instead interpret the disorder's symptoms in culturally meaningful ways. This means that the individuals suffering from the symptoms do not expect, or are not expected by others, to become failures in social and professional life (p. 205). The stigmatizing conceptions associated with the diagnostic label may also indirectly influence the severity and outcome of schizophrenia. According to two major longitudinal studies of over 30 years conducted by the WHO, people who had received a schizophrenia diagnosis, for instance, in India, Nigeria and Columbia, suffered a milder form of the disorder than people with the diagnosis in the USA, Denmark and Taiwan (Hopper 2004). Approximately 50 per cent of the people diagnosed with schizophrenia are less impaired in the global south than in the developed world (Hopper et al. 2007). The dominant biomedical explanations of schizophrenia consider the Western feedback effects as stable background conditions and are therefore sensitive to the symptomatic, severity and outcome variations of the disorder.[101] If the feedback explanation is integrated to supplement the explanation's applicability domain, it becomes insensitive to these variations, while the explanatory scope is widened to enable more precise comparisons between different cultural, historical and individual manifestations of the disorder.

Incongruent mechanisms describe how feedbacks generate, reinforce and maintain destabilizing loops between classifications and

[100] See also Kent and Wahass (1996), Mawson et al. (2011), Woods et al. (2014).
[101] Pöyhönen (2010, 2013) makes a similar argument about norm dependency of bulimia.

interactive kinds.[102] Intentionally mediated incongruence describes how a classification is found meaningless, ambiguous, unpreferred or unnatural by the classified persons, and therefore causes them individually or concertedly to act against it (see Hacking 2007a). As an example, misclassifications can cause incongruence because they are found unfitting and meaningless. Similarly, incongruence can be due to a mismatch between a classificatory conception and the classified people's perceived self-images, experiences and behaviour. This maybe the case when the individuals perceived cultural or social standing is in conflict with the label, although the classification otherwise is accurrate. As an example, especially people with a higher social standing in the West may resist the stigma of obesity even when clearly overweight (Aronowitz 2008: 7). In such cases, social pressure may lead the diagnosed individuals to exercise and lose weight. In other words, whether a classification and its associated theoretical beliefs "stick" does not only depend on whether they accurately describe the kind, but also whether they are congruent with larger sociocultural factors and beliefs associated with the label. On the other hand, structurally mediated incongruence describes how a new classification undermines the practices and structures that brought about the behaviour in the first place, leading to alterations in the classified people's behaviours (Hacking 1995a, ch. 4; Mallon 2016: 72). That is, as classified individuals succumb under changed structural constraints and opportunities, they may prefer, find meaningful and natural novel behaviours. At the extreme, incongruent feedback mechanism is a self-defeating prophesy, when the classification is generally challenged.

The cross-cultural application of the diagnostic categories in the DSM-5 and the ICD-10 psychiatric classification manuals causes incongruence. The categories are primarily based on symptom

[102] I am not considering cases where the looping effect dismantles a kind (see Hacking 2007a, 305).

delineations drawn from Western patients and informed by Western folk psychology (e.g. individualistic view of the self) (cf. La Roche et al. 2015).[103] Consequently, individuals with different cultural backgrounds may be misdiagnosed, while correct diagnoses can induce opposition because they are found stigmatizing by some cultural minorities (cf. Kirmayer 2001). Both situations may lead to alterations in the behaviour of the diagnosed individuals and thereby further distort theoretical beliefs about the disorders. As part of "modernization" processes, the diagnostic categories can bring about destabilizing cultural and social tensions by transforming local structures and conceptions.[104] By this I mean that the categories may distort or alter culturally different ways of experiencing, manifesting and coping with psychiatric problems (see Kleinman 1988; see Kirmayer 2002). As an example, Kitanaka (2012) argues that the moods labelled as depression in the West were not commonly pathologized in Japan until the introduction of the diagnostic category at the end of the 1990s.[105] The diagnosis and the use of antidepressants have induced unpredictable alterations in patients, leading to further changes in the conception of depression (p. 184). A diagnostic category can also become part of a causal process that enables a novel condition. As an example, the conception of self-starvation as a sign of personal suffering (and the idolization of slimness) interacts with local cultures in Asia, creating dynamic and evolving forms of anorexia nervosa (Lee 1996; Pike and Dunne 2015). Finally, research in cognitive science and psychology indicates that even underlying cognitive mechanisms of disorders may be socially and culturally shaped (Murphy 2015; Washington 2016). In

[103] Folk-psychological conceptions shape psychiatric disorders, see Luhrmann (2011).
[104] Structural changes can also be instigated by endogenous cultural factors, see Sahlins (1985) and Robbins (2004).
[105] Kirmayer (2001) goes as far as to argue that the category pathologized moods that were previously revered and even aestheticized (see also Watters 2011).

the light of these examples, cross-cultural research can supplement explanations of psychiatric disorders by undermining some of the taken-for-grantedness of Western diagnostic feedback effects.

Incongruent feedback loops that describe the influence of non-epistemic classificatory adjustments arguably represent the biggest obstacle for human kinds to support epistemic projects (cf. Griffiths 1997; cf. Khalidi 2013). As an example, the value-laden shifts in the conceptions associated with ADHD have influenced those classified with the condition. The diagnostic category has strong political and ethical implications, and therefore motivates individuals as well as larger interest groups to react. These include pressure from ADHD groups, consisting of patients and their families, as well as lobbying from pharmaceutical companies. Moreover, careless use of standardized scales has enabled individuals who lack the symptom profile to pretend to have them, or to become diagnosed mistakenly by the experts or themselves. On the other hand, the ADHD prototype as predominantly a male condition causes some girls and their parents to conceal and play down symptoms (Mowlem et. al. 2019; Quinn and Madhoo 2014). The upshot is that the distorted picture of ADHD may prompt behaviour reactions from both genuine and misdiagnosed individuals (cf. Allen 2018). Although these changes are hard to explain and anticipate, incongruent feedback explanations can help to identify how the psychiatric disorder is unstable, and thereby contribute to its explanation and classification. While the kinds may not succumb under causal predictions, and thereby support proactive interventions, knowledge of the feedback may nonetheless enable reactive measures. In addition, a more robust feedback explanation could be achieved in virtue of a cognitive explanation.

It is important to point out that the same psychiatric kind can be explained by both congruent and incongruent mechanisms. As I mentioned in Chapter 4, explanations are always aspectual. Hence,

while the content of voice-hearing experiences can be partly explained by congruent feedback mechanisms, individual objections towards medicalizing voice hearing can be understood as *status reactions* (see Hacking 1999). That is, while many people disagree and resist their labels, if the larger social audience perceives them nonetheless to be a match, their indirect influence is hard to resist. Moreover, often a label may not be liked, but in virtue of culture and social structure, it may nonetheless feel natural and inevitable.[106] Moreover, individual liking or taste is partly culturally shaped (see Bourdieu 2012). On the other hand, if the label is not perceived as a match by the social audience and culture, it may not be perceived as natural and inevitable by the individual either. My point is that whether a feedback mechanim is destabilizing or stabilizing depends on its interaction with other sociocultural factors.

One example of an organization that aimed to change medico-social attitudes is the Hearing Voices Movement (HVM), which sought to alter the stigma and negative perceptions associated with voice hallucinations. Currently, the HVM is more of a status reaction than a genuine kind-changing feedback mechanism, but if it proves successful in the future, the Movement could alter the way Westerners in general would feel about hearing voices, and thereby could contribute to shaping the very nature of those voices (see Chapter 6 about cultural shaping of cognition). The opposite is the case with obesity. When people in the West with higher sociocultural standing are labelled obese, it may cause them to exercise to dispose of the potential stigma associated with the label. But in this case, although the label initially matches their condition, it does not match with their culturally perceived social standing, and thereby does not feel inevitable and natural to the individuals concerned. In contrast, people with lower social standing may not like to

[106] I am indebted to Rachel Cooper for pointing out that "liking" and accuracy should be distinguished.

be labelled obese, but may nonetheless come to feel that it is a natural and inevitable part of their self-identity. In essence, the process matches the idea of secondary deviance (see Lemert 1967).

The congruent and incongruent explanatory models can be represented with the help of Coleman's boat (or diagram) and structural individualism.[107] Coleman's boat has become a standard depiction of social mechanisms (Ylikoski 2017, Coleman 1990, Hedström and Ylikoski 2010). It provides a visualization of how to address micro and macro relations in sociological explanations (Ylikoski 2021). According to Hedström and Ylikoski (2010), Coleman's boat demonstrates how mechanistic explanation does not need to be coupled with methodological individualism. Rather, the idea is that macro-level relationships can be explained by understanding the underlying chain of situational, action formation, and transformation mechanisms. The following diagram is an adaptation of Ylikoski's (2021) description of Coleman's boat with a feedback loop.

[107] See Coleman (1990), Hedström and Ylikoski (2010), and Ylikoski (2021) about structural realism and the Coleman boat.

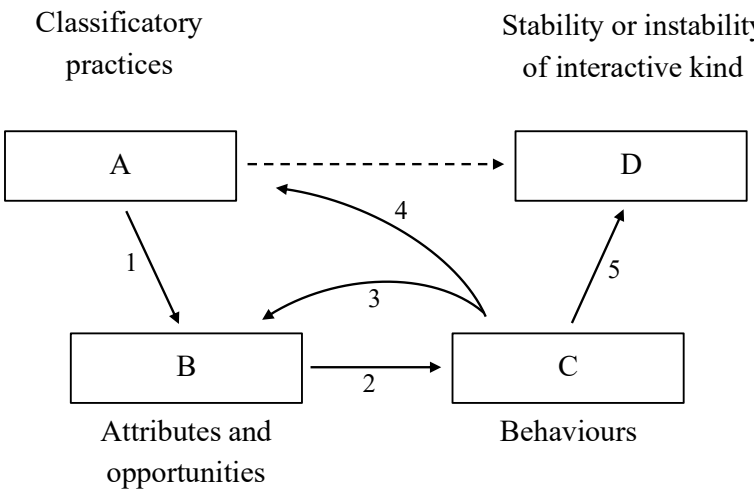

Figure 4. Congruent and incongruent feedback loop as a Coleman diagram (adapted from Ylikoski 2021 and Coleman 1990).

The diagram represents congruent and incongruent mechanisms. The first arrow (A-B) represents a situational mechanism that describes how conceptions and classificatory practices constrain and shape the behaviour and action (C) of classified agents by altering their beliefs, desires, and opportunities (B). The second arrow (B-C) represents a behaviour-formation mechanism that describes how a specific combination of these individual attitudes brings about kind-typical behaviours (C) (these may be further amplified by cognitive processes). The third arrow (C-B) describes how those behaviours may loop back to influence other people's beliefs, desires, and opportunities (B). The fourth arrow explains how the behaviours and their possible alterations may loop back to reinforce, sustain, or alter classificatory practices (C-A). Finally, the fifth mechanism (C-D) represents a transformation mechanism that explains how these behaviours can sustain, reinforce,

and produce structurally entrenched interactive kinds. On the other hand, it can also explain the instability of interactive kinds. That is, it can explain how behaviours undermine the classificatory practices and possible social structures that originally gave rise to them and shaped them. Hence, the rectangle (D) represents the net outcome of the feedback loops. The result can either be congruent or incongruent with the classificatory practices and conceptions. In other words, the feedback mechanisms can be understood as producing structurally entrenched kind-typical congruence or undermining such structures to produce incongruence.

The diagram can be illustrated with eating disorders (see Hacking 2006). Eating disorders like anorexia nervosa and bulimia were rare until the 1960s. The implementation of the body mass index (BMI) in clinical practice in the 1970s played a role in conceptualizing the importance of body weight. Increasingly, some people started to internalize the fact that they were overweight and obese, while others began to feel underweight. That is, the classificatory practices based on BMI influenced the beliefs and attributes of the people who fell under either extreme (arrow 1). It caused some individuals to pay more attention to eating habits and exercise, but it also facilitated a conceptual space for expressing mental suffering through eating habits (arrow 2). These altered behavioural patterns created a feedback loop that influenced other people's beliefs and attitudes (arrow 3), and thereby their behaviour (arrow 2). But it also influenced classificatory practices (arrow 4) and thereby structurally constrained and shaped the labelled peoples' beliefs and actions (arrows 1 & 2). Hence, through the iterations of these two pathways of feedback loops (3) and (4), kind-typical behaviours formed and became entrenched (5) in institutional and social structures. The net outcome (5) was the formation of psychiatric kinds such as anorexia nervosa, bulimia, obesity as well as the male-specific disorder, bigorexia (its symptom is to build massive

muscles to compensate for a constant feeling of "thingness"). Consequently, the net outcome has been the formation of relatively stable psychiatric kinds that congrue with their classifications.

The distinction between congruent and incongruent feedback mechanisms is meant as an analytical perspective on how the looping effect can succumb to explanatory applicability. In practice, the looping effect does not neatly divide between different feedback mechanisms. Rather, the feedback mechanisms of different pathways (3) and (4) may balance each other out, enforce each other, or trump each other, leading to a general outcome of congruence or incongruence. Consequently, congruence and incongruence should be understood as the net outcome of the different causal paths the looping effect can take. Moreover, I wish to emphasize that knowledge of the feedback mechanism, and making predictions and causal interventions based on them, depends on understanding the enabling sociocultural "niche" (see Chapter 5). The better those enabling sociocultural factors, which can be understood as the structural factors in the background causal field, are known, the more reactions to classificatory practices can be explained and anticipated. On the other hand, if relevant sociocultural factors that underlie interactive kinds are ignored, incongruent and congruent reactions to classificatory practices remain unexplainable and unpredictable.

Ideally knowledge of the feedback mechanisms can inform treatment and policy decisions by facilitating interventions that would weaken feedback and increase positive congruence (see Table 3). Luhrmann and Marrow (2016, p. 220), for instance, suggest that the diagnosis of schizophrenia should de-emphasize labelling and focus on behaviour rather than inner experience. The upshot is that individuals could find voice-experiences and treatments more positive and meaningful, which would increase positive congruence (see Thomas et

al. 2014). Feedback explanations could also be implemented proactively to mitigate or prevent incongruent effects of exported Western diagnostic categories (see Kirmayer 2001). Even value-driven incongruent loops could support some interventions based on models that represent individual motivations (but cf. Steel 2008, p. 158). In short, congruent explanations may facilitate domain-relative interventions because the feedback models are relatively modular. On the other hand, although models of incongruent mechanisms are non-modular or weakly modular, they may nonetheless enable preventive or reactive interventions. However, disseminating information about classifications, or impeding their dissemination, can by itself influence the classified people. Hence, modifying classifications based on feedback knowledge does not so much describe classifying kinds more accurately, as describing a co-fitting process that transforms both. This means that the conceptual engineering that a feedback explanation could enable would be tantamount to kind amending.

Table 3. Feedback knowledge and its applicability

	Congruent	**Incongruent**
Explains	Conformity with classification	Nonconformity with classification
Possible intervention	Proactive	Reactive

Lastly, I want to emphasize that I have only argued in favour of explanatory evaluation of interactive kinds and kind members on a case-by-case bases. This caveat in mind, however, I want to make a tentative suggestion. Although I have argued that there is no principled epistemic difference between interactive human and biological kinds,

there may be a quantitative explanatory difference. When the looping effect works through the direct or indirect intentions of the classified humans, their reactions are more fine-grained than is the case with environmental looping effects of, for instance, disease entities. The reason is that humans can rationalize, interpret and explain their own behaviour in a fine-grained manner. This means that knowledge of the looping effect is either relevant to enable stable explanations of human kinds, or as with value-driven incongruent loops, to understanding why and how such explanations are unstable and limited. This is exemplified by a self-diagnosing psychiatrist who adjusts his or her behaviour to be congruent or opposed to even the smallest diagnostic changes. The dependency could have systematic counterfactual power under the right motives and constraints, or alternatively, it could help to understand why such explanations are highly limited. On the other hand, counterfactual dependencies between classificatory practices and the behaviour of some disease entities are so course-grained, that they provide only weak explanations and projections.

6 The Medical Model and Cross-Cultural Variation

I have argued that explanations of psychiatric disorder are domain-relative, and that knowledge of looping mechanisms can supplement those domains. In this chapter, I will compare my applicability domain approach to the medical model based on the ability of both to accommodate discipline-related explanatory approaches and to account for cross-cultural variation of psychiatric disorders.

I begin by providing an overview of the explanatory and ontological commitments of the medical model. The medical model is the predominant explanatory approach in psychiatry. According to it, psychiatric disorders are real physiological entities that resemble somatic diseases. Following others, I distinguish between the DSM classificatory manual's weak medical approach and stronger versions of the medical model. It is commonly argued that the diagnostic categories in the DSM are partly constructed based on folk-psychological preconceptions and do not seem to pick out real disorders. To overcome this problem the US National Institute of Health (NIHM) has launched the Research Domain Criteria Program, which is a fully bottom-up "framework for collecting the data needed for a new nosology" (Insel 2013). The programme is based on the conviction that there is a need for a thorough overhaul in psychiatry to improve the present diagnostic categories that restrain and mislead empirical research. The RDoC considers psychiatric disorders to be real disorders on the level of neural circuits (Faucher and Goyer 2015: 200). I compare the RDoC and Murphy's (2006) biomedical approach, and argue that they seem to be too one-sided. In particular, while the NIMH acknowledges the importance of developmental and environmental aspects, the focus is nonetheless on the brain. Based on this, I argue that the RDoC's approach has problems in accommodating especially social and cultural explanations.

The second part of this chapter is dedicated to discussions about explaining cross-cultural variation and sociocultural causation. I argue that many aspects of psychiatric disorders cannot be understood without considering the complex ways in which cultures shape them. I begin by distinguishing between covert social construction and sociocultural explanation. I then investigate the strong medical model's problems of accounting for cross-cultural variation. Instead of the medical model, I propose a contrastive approach to explaining the sociocultural variation of particular psychiatric disorders. I argue that social causes and mechanisms can explain some aspects of psychiatric kinds, and that we should weigh their explanatory relevance in relation to how the explanandum phenomenon is specified. I then move to investigate arguments about cross-cultural cognitive variation. Since recent studies have demonstrated that it cannot be ruled out, I want to explore whether a bigger role should be given to culturally oriented explanations of psychopathology. The central point is that because culture shapes psychiatric kinds in complex ways, social and cultural explanations are indispensable. Moreover, I also investigate whether culture-bound syndromes should be understood as manifestations of universal disorders or as genuine cultural disorders. Based on this, I also investigate whether some disorders are best not explained by locating them entirely within the skin and skull of the sufferer but instead as being constituted at least partly by social dynamics. Finally, I argue that semantic arguments cannot settle questions about the nature of psychiatric disorders or psychiatric advancement.

6.1 The Medical Model's Ontological Commitments

The medical model is the predominant scientific programme in psychiatric research and clinical practice. Although the model

comprises various, mostly unspecified clinical and theoretical assumptions, the underlying shared commitment is to medical research based on observational and experimental methods. The programme is modelled on somatic medicine so that psychiatric disorders are taken to be categorical disease-like entities that can be explained, predicted, and treated through medical means (Kincaid and Sullivan 2010, Murphy 2017). Traditionally this is interpreted so that the medical programme makes an ontological commitment about psychiatric disorders as pathophysiological processes situated in an individual's brain, and that they should therefore be primarily explained by neurocognitive, biochemical, or genetic factors (cf. Kincaid and Sullivan 2014, Murphy 2017). For example, Emil Kraepelin advocated the strong approach by arguing that co-occurring symptoms of disorders are realized by abnormal nervous systems, making dementia praecox (i.e. schizophrenia) a neuropathological disease (Garrens 2014: 13).

Hence, when the medical model is strictly interpreted, it is committed to a biological view of psychopathology, where the symptoms and signs are manifestations of underlying discreet neuropathological processes situated within a person.[108] This seems to imply, as is commonly assumed, an essentialist account of psychiatric kinds so that they would have singular causes like most somatic diseases. However, as such, the view has been rightly criticized for not being empirically credible. Graham (2013: 53-63), for example, argues that the medical model is incorrect in excluding multiple causes, such as social and psychological factors, that contribute to psychiatric disorders. In this regard, for example, the biopsychosocial model's multi-causal approach is typically contrasted with the medical model.[109]

---

[108] Tsou (2021) also defends a similar biological kind account of disorders.

[109] In the model, biological, psychological, and social factors are given equal explanatory roles. The problem with the model is that it does not provide us with an

Murphy (2013: 927), however, points out that a distinction should be made between realizing processes and proximate causes. The idea is that although different factors, including social contexts and psychological factors, can causally contribute to bringing about pathological processes, brain processes nonetheless constitute and realize the observable symptoms of psychiatric disorders. In other words, according to Murphy, neurological processes are necessary but not sufficient to explain mental problems.

A central problem especially with the most reductionist versions of the medical model is that in spite of the huge funding and research input it has enjoyed, it has not been able to provide successful explanations and treatments (cf. Sullivan 2017: 456).[110] There are several reasons for this. As I argued in the first chapter, unlike the strong reductionist medical model implies, a strict distinction between social factors and biological factors about disorders is not easily made. Further, the strong model treats sociocultural factors merely as environmental triggers (e.g. in behavioural genetics) and personal experiences as epiphenomenal. This is a problem because many psychiatric disorders seem to be dynamic, dependent on physical and social environments, and perhaps even interpersonal. Further, the model seems to make an implausible commitment to a universalist stance on psychiatric disorders. As Washington (2016) points out, it is based on the implicit belief that human cognition – and by implication psychiatric problems – are the same everywhere. However, as Kincaid and Sullivan (2010: 368) argue, social factors can play a substantially stronger role than mere triggering factors in explaining psychiatric disorders.

---

understanding of how those factors interact to contribute to explanation (Ghaemi 2003: 82). Moreover, psychodynamic approaches are also opposed to the medical model.

[110] This includes the reductionist version based on genetics (Kandel 1998) and the biochemical version.

Because of the problems in, and criticism towards, the strong interpretation of the medical model, it has remained more a promissory research programme than a robust empirical resource for diagnoses and classifications. Consequently, less demanding interpretations of the medical model have been deployed in practice (see Murphy 2017, Cockerham 2017). Spitzer and Wilson (1975) in particular defend the pragmatic approach by arguing that the medical model should not be interpreted as making strict theoretical commitments, but rather that it should be evaluated based on how well it works in practice.[111] Further, they maintain that although an essentialist view is not applicable to disorders, this is not a problem for the medical programme in psychiatry since it does not apply to all somatic diseases either (e.g. vitamin deficiencies or endocrine disorders). Rather, the pragmatic approach holds that psychiatric disorders are syndromes that can be systematically characterized and described to enable predictions and treatments. This means that psychiatric disorder is taken to comprise clinically interesting behavioural symptoms and signs as well as responses to psychological and pharmacological treatments. This, for pragmatic purposes, covers all the clinical signs that can be reliably traced and classified. Since the DSM and the ICD classification manuals rely on this syndrome-based approach, it wields significant institutional power in research and clinical practice. The primary goal of the classification manuals is to facilitate predictions and treatments. Therefore, unlike the strong medical model natural kinds view of psychopathology, the descriptive view of the medical model implies a pragmatic view of psychiatric kinds.[112]

---

[111] In 1982, Spitzer resigned his strong interpretation of the medical model (Blashfield 1982).

[112] Zachar (2000; 2014: 91) defends an explicitly pragmatist capproach to psychiatric kinds. He maintains that rather than aiming to discover natural kinds, psychiatric classifications should balance different aims (e.g. treatments and standards of validity) so that they can be best met in practice.

However, as I argued, the pragmatic approach is simultaneously too strong and too weak to facilitate classificatory and explanatory progress in psychiatry. It is too weak because it defines the boundaries of psychiatric disorders operationally rather than causally or constitutively. This predisposes the diagnostic categories to problems of validity, which are manifest in high rates of comorbidity and the danger of medicalization.[113] Moreover, as I pointed out, it has raised criticism over the line-drawing problem: there is no interest- and value-free way to draw the line between what is pathological and what is normal (Schwartz 2007b). It is too strong, at least in the DSM version, because despite its operational definitions, it nonetheless advocates a biologically oriented dysfunctional approach to the superordinate concept of psychiatric disorder. This means that the DSM version is not explicit about the values and folk intuitions that underlie the decisions over its definitions. As I argued in Chapter 5, this implicit culture- and value-relative approache may lead to unintentional reinforcement of those ways of experiencing and expressing mental problems.

I will argue that the applicability domain approach to psychiatric kinds can answer these challenges. This approach offers an ecumenical account of psychiatric disorders, according to which different disciplinary approaches can supplement explanations of psychiatric disorders. In other words, it offers a heuristic account of how different explanatory approaches can explain different aspects of the same psychiatric kind, which neither implies the reductionism and essentialism of the strict medical model, nor inherits the problems of the pragmatic and descriptive interpretation of the medical model. Rather, the applicability domain approach supports a non-metaphysical pluralistic and value-sensitive explanatory approach to understanding

[113] However, An Loo et al. (2013) argue that if the discreet mechanistic approach is given up, comorbidity can instead contribute to an understanding of how causes can contribute to explaining psychiatric problems.

psychiatric kinds. Furthermore, I will argue that the non-metaphysical approach to psychiatric kinds means that the approach can accommodate the explanatory import of various causal factors, including intentional causation. That is, various disciplinary-relative explanatory approaches can be applied to supplement the domain of application of a psychiatric kind explanation.

6.1.1 The Medical Model and Psychiatric Kinds

The medical model has philosophical implications for the correct explanation and nature of psychiatric kinds. Although the strong version of the medical model is committed to a causal account of psychiatric kinds, and on building a classification based on those causes, it has not been able to integrate or facilitate multiple explanatory approaches sufficiently well. To address this, I will now compare my applicability domain approach to the explanatory approach of the medical model of psychiatric disorder. I argue that according to the applicability domain approach, psychiatric kinds can be considered heuristically to the extent that they can facilitate interaction between various discipline-related explanatory approaches.

The medical model's commitment to the nature of psychiatric disorders has implications for their explanation. The medical model endorses two types of explanation (Thagard 2000; cf. Maung 2016). The first type is conducted in medical research when the goal is to find mechanisms and pathoetiological reasons for co-occurring symptom clusters. The second type of explanation is diagnostic. It has the goal of using an explanation of a syndrome to diagnose symptoms based on diagnostic categories. Therefore, whereas the medical explanation starts from co-occurring symptoms (or a prototype) and looks for common causes, diagnostic explanation starts with the patient's

symptoms and tries to determine the syndrome cluster that explains them.[114] Hence, while the pragmatic minimal approach of the medical model emphasizes the role of diagnostic inferences, the strong interpretation implies that ideally medical explanation should dictate those inferences.

The diagnostic explanation resembles what Cooper (2007: 44) calls "natural-history style explanation", which relies on "natural kinds", and are as such typical in psychiatry. The idea of "natural-history explanation" is based on inductively inferring that if individuals have kind-typical symptoms, by identifying the membership, we can reliably infer how the symptoms and their condition in general will develop, and how best to treat them.[115] In other words, knowledge of the kind may provide us with information about the course and risk factors.

Murphy (2013; 2010: 606) points out that the natural-history explanation matches the minimal interpretation of the medical model and the DSM view of mental disorders, because it is based on causal differentiation. This means that we may know that two kinds differ, although we do not know the causal structure that realizes the difference. Hence, the minimal model of psychiatric disorder is content with pathoetiological causes. As Cooper and Murphy conclude, the historical-natural explanation can work up to a point. For instance, we know that smoking causes cancer before we knew the underling causal structure or mechanism that underlies that causal connection. However, such explanations are not necessarily very stable. Some people, for example, are genetically more disposed to cancer than others. Therefore, knowing the underling mechanism that realizes the causal connection between smoking and cancer means that we can extrapolate

[114] These explanations also facilitate other goals in medicine, for example, preventions and pharmacology (Illari 2017: 50).

[115] Cooper (2007) also distinguishes another type of explanation based on individual case histories, for instance, in Karl Jaspers's work (see also Bental 2003: 141).

more reliably from individual to individual. As Murphy argues, natural-historical explanations (or diagnostic explanations), should ideally be supplemented with a deeper mechanistic explanation. This matches the general belief that causal mechanisms should play a crucial role in psychiatric explanation and classification (Kendler 2008).

When the mechanistic approach of the medical explanation and the diagnostic inference are taken to complement each other, the view matches the dual nature of the HPC view of kinds. As I mentioned in Chapter 3, the HPC view is based on property clusters being the explanandum and the mechanism(s) as the explanans. Based on this, medical explanation can be understood as investigating the mechanism that underlie or produce the targeted co-occurring symptom cluster, whereas those mechanisms explain why and when we can infer reliably from one property or symptom to another. In other words, as I pointed out, to rule out accidentally lumping together unrelated property clusters, we need to uncover the responsible causal mechanisms.[116]

Although I agree with Murphy that mechanistic explanations are required to understand the nature of psychiatric disorders and to ground their classifications, I nonetheless believe that there is no a priori reason to commit oneself to a metaphysically binding material interpretation of the mechanisms. In contrast, Murphy (2006: 256) argues that in order to understand sociocultural factors, mechanisms should be interpreted as material.

> To understand how social and cultural forces shape behaviour, we need a mechanism, and the mechanism needs to be material, otherwise social and cultural causation will remain mysterious. The obvious place to turn for a mechanism of social causation is mental representation. The cognitive sciences suggest that aspects

[116] Murphy (2013: 973; 2017) argues that the HPC view can accommodate the medical model view of psychiatric disorders.

> of a social environment could be represented by the mind/brain. Yet because mental representations are physical entities they can stand in cause-effect relations within the organism. (Murphy 2006: 256)

According to this interpretation, the strong view of the medical model is committed to a biological interpretation of psychiatric realism. In other words, psychiatric disorders are biological entities that realize clinically observable symptom clusters. Moreover, Murphy (2017) himself defends an eliminative version of the strong medical model, according to which psychiatric disorders are pathological processes or mechanisms in the brain, and folk-psychological descriptions do not refer to real mental activities. In other words, higher level phenomena realized by the brain are not causally efficacious.[117] For this reason, psychological factors, as well as the other signs and symptoms that we use to identify psychiatric disorder, are merely epiphenomena.[118] Based on this, Murphy (2006) argues that pathological processes are best explained by cognitive neuroscience, which offers the appropriate framework to integrate different explanatory variables. The implications are that folk-psychological explanations will eventually be replaced by the precise methods of cognitive neuroscience[119]

Murphy's neuropsychiatric account should be distinguished from the older genetic and biochemical approaches of the medical model.[120] It is different, for instance, from Kandel's (1998) biological approach, which seems to rely exclusively on molecular and genetic explanations (Murphy 2006: 116). It also differs from the neurotransmitter (e.g.

[117] See Churchland (1981) and Stich (1992).
[118] Kraepelin argues similarly that psychological reports are useless because they are the manifestations of a broken brain.
[119] The RDoC project is tied with precision medicine. See Tabb (2019).
[120] Luhrmann (2012) calls this the bio-bio-bio model "for lesion, genetic cause, and pharmacological cure".

depression due to abnormalities in serotonin release) approach that has largely relied on the relative effectiveness of some psychopharmacological treatments. Unlike these approaches, Murphy's account gives external explanations, such as social and cultural explanations, a more substantial role. The idea is that psychiatric disorders can be described on many levels of explanation, such as "social pressures, computational processes, learning histories, or evolutionary mismatches" (Murphy 2006: 120). Crucially, according to Murphy, these explanations can be understood to be compatible only in the light of their representations in our brains.

However, in the light of what I have argued this far, it does not seem plausible that a single theoretical approach could be employed to understand all psychiatric disorders. In spite of the vast amounts of funding put into biological psychiatry, neurocognitive bases for psychiatric disorders have not been found, nor robust psychiatric cures developed. This implies that biomedical explanations may not be suited to explaining the dynamic and complex nature of many mental disorders. Moreover, as Sullivan (2017) points out, it is likely that folk-psychological explanations will continue to play a role in how clinicians and the patients themselves explain and manage disorders.

I will next analyze the RDoC project, which corresponds to Murphy's eliminativism. The project suggests that disorders can be comprehended within the context of normal neurological functioning, with the exception that it does not provide an account of how to demarcate psychiatric problems.

6.1.2 The Research Domain Criteria (RDoC)

The Research Domain Criteria (RDoC) project is committed to the view that psychiatric disorders are neurologically realized. RDoC was

instigated in 2010 by the National Institute of Mental Health (NIMH) with the aim of overcoming the problems of DSM. According to the RDoC, psychiatric disorders are brain diseases represented by malfunctioning brain circuits (Insel et al. 2010, Cuthbert & Insel 2013). The RDoC project is based on a bottom-up empirical approach to figuring out psychopathology as "complex combination(s) of disturbances in more fundamental processes, or dimensions of function, that do not necessarily align with currently identified categories of disorder" (Carter, Kerns, and Cohon 2009: 181, in Sullivan 2017: 462). The approach would lead to large revisions in the current classification of disorders as the supposition is that clinically identified behaviour cannot be mapped "on a one-to-one basis onto specific genes or neurobiological systems" (NIMH 2011). That is, neural circuits may link many current syndromes or their subgroups.

The RDoC project is arranged around a research matrix, which is a table for systematizing research on different explanatory approaches to psychopathology (see Sullivan 2017: 462). The purpose of the matrix is to study genetics and neural circuitry in the matrix's research domains that represent disorders as disruptions in neurocognitive functions. The domains are made up of five rows: positive and negative valence systems, cognitive and social processing, and arousal/modulation systems. In addition, the matrix contains columns that represent "levels of organization": genes, molecules, cells, circuits, physiology, behaviour, and self-reports. The idea behind the matrix is to encourage the integration of different disciplinary-related results under the same functional domains. Ultimately, the goal is to build a novel classificatory system based on neural circuits and genetics, which are explained by clinical neuroscience with the purpose of facilitating novel treatments (Poland 2014: 52-54). Hence, although the RDoC project does not really define psychiatric disorders, it does suggest that they are neurological problems.

There is no doubt that the RDoC is ontologically reductionist – psychiatric disorders are considered to be brain diseases. However, it not clear whether it is also epistemologically reductionist or eliminativist.[121] Although the project does not completely neglect environmental factors, such as social causes and phenomenological experiences in its matrix, it does seem to consider them to be placeholders for future neurological explanations. This is a problem because even if some psychiatric disorders are in fact brain problems, their explanations may nonetheless need to rely on higher "levels" of explanation (e.g. psychological or social). In the following, I will point out some of the criticisms that RDoC program has faced.

The central problem with RDoC is that not has enough focus been put on how different factors – biological, psychological, and social – are linked in realizing disorders (Lender 2014). Simply equating mental disorders as neurological disorders bypasses the role of other factors. The reason is that disorders may appear to be contrived and arbitrary on "lower" levels of explanation (i.e. neurological explanations).[122] Depression, for example, may be realized by our brains differently in different cultures or individuals.[123] If this is the case, a successful explanation may require concentrating on highly socially susceptible psychological processes, which are fragile and limited in scope, but may still ground some inductive generalizations that are clinically

[121] Kirmayer and Gold (2012) argue that RDoC is also epistemologically reductionist and hence unable to facilitate an understanding of human behaviour. See also Hutto and Kirchhoff (2015), Hutto (2016), and Sullivan (2017: 460). In contrast, Murphy (2006) argues for a multi-level approach.

122 Murphy's (2008: 109) brain plasticity implies that higher-level explanations are indispensable.

123 One could argue, for instance, that with depression the realizing brain mechanisms vary across cultures (Washington 2016).

significant.[124] In this case, phenomenology may provide the most suitable explanation and talk therapy the most functional treatment. Consequently, purely from an explanatory perspective, it is reasonable to consider intentional psychological explanations as causally efficacious (Woodward 2015).

Maung (2016) has also argued that a proximal neurological mechanism may be useless as explanation when the real interest is in the less proximal causal process. He points out, for instance, that although coughing is realized by a neurological mechanism, the real medical interest may lie in the tumour or infection that brings about coughing. Consequently, the primary medical interest need not be on constitutive underlying mechanisms, but can instead be etiological factors. This would render some psychiatric disorder explanations close to the bacterial model, where it is the antecedent causal factors that make a difference by bringing about diseases, rather than some abnormal bodily reactions.[125]

The brain-based views of some disorders may not only be explanatorily challenging, they may also be socially and therapeutically counterproductive. This can be exemplified by addiction, which according to the RDoC should be understood as a problem in the brain circuits (see Murphy 2017: 168). The problem is, as Kuorikoski and Uusitalo (2018) argue, that a biological picture of addiction is liable to harm the self-image of the labelled individuals. According to them, a

---

[124] Gerrans (2014), for his part, argues for an integrative model, according to which psychiatric disorders can be studied on both the brain and the psychological levels of organization, and that these levels of organization are complimentary. According to him, personal accounts provide just another level of explanation and intervention on the same disorders. Similarly, Bental (2003: 175) writes that "statements about brain biology and statements about mental processes represent different levels of descriptions of the same phenomena – in other words, that the mind and the brain are one".

[125] In some cases, a psychiatric disorder can be brought about by a normal reaction to an abnormal situation (as with PTSD).

more ethically sustainable approach would be to emphasize the multi-faceted interactive factors that underlie addictions. This is because studies have shown that a mere (biological) categorization may cause negative health outcomes (Mallon 2016: 87, Kaplan 2010). Haslam (2014: 25) argues that the essentialist view of psychopathology is stigmatizing because it represents inflicted individuals as "categorically abnormal, immutably afflicted, and essentially different". Currently, the problem is that the neurobiological medical model is baselessly applied, or conceived to be applied by lay people to all psychiatric disorders.

In contrast, I believe that there is no a priori reason to assume an ontological and internal "level" for all psychiatric disorders. Rather, sometimes a disorder may be best explained by the dynamic relationship between individuals and their social circumstances. For example, Kincaid and Sullivan (2014) argue that a better understanding of addiction requires considering the social role it bestows on the conflicted individuals (see the section 6.2.4 on culture-bound syndromes). Therefore, I believe that the non-metaphysical explanatory approach to psychiatric kind is both more ethically responsible and heuristically better in accommodating various discipline-related explanatory approaches.

Hence, a successful mechanistic explanation of the complex and dynamic nature of psychiatric disorder, one that links all the relevant factors, may need to abstract away from neurocognitive "material" bases. This is because whether a condition can be determined to be pathological requires contextualization. There are two reasons for this. First, even if it were deemed explanatorily necessary to hold that neurocognitive mechanisms realize psychiatric disorders, drawing their boundaries may still be interest dependent. According to the contrastive-counterfactual and interventionist explanatory theories, explanatory relevance determines in part how the causal structure is

delineated for explanations. Moreover, as Bechtel (2009) argues, understanding how a mechanism produces the phenomenon of interest requires "looking up" by situating the mechanism within a larger environment. Consequently, as I have argued, an emphasis put on pragmatic interests and social factors may license different mechanistic delineations. The reason we employ the concept of psychiatric disorder is to be able to intervene on the conditions so labelled. This value-laden aim in turn guides the explanatory-seeking quest to obtain causal factors that are deemed important for such interventions. Hence, for instance, clinical, neurological, and epidemiological interests may license different sets of causal structures as relevant explanations of the psychopathological phenomenon of interest. In sum, according to these objections, psychiatric disorders cannot be comprehended as isolated *malfunctioning brains in a vat*.

Moreover, psychiatric problems may not be deviances of normally functioning brains, but instead distinct mechanisms. As Kirmayer and Crafa (2014) point out, simply concentrating on normality to understand psychopathology neglects its specific nature. Nervi (2010), for instance, argues that pathological mechanisms cannot be understood merely as deviations from normally functioning ones, but rather as autonomous mechanisms. In a similar vein, according to Borsboom's (2017) network model, psychiatric disorders are novel mechanisms constituted by the interaction of symptoms. Finally, Washington (2016) has rightly pointed out that there may not be similar neurocognitive mechanisms that all humanity shares. This would put in doubt the RDoC program and Murphy's central idea that cognitive neuroscience can provide us with a picture of the normal functioning of our brains, which in turn could be used as a guide to delineate the putative malfunctions underlying psychiatric disorders.

The applicability domain account can overcome some of the problems related to the minimal or pragmatic interpretation of the medical

model when classifying psychiatric kinds, without making the reductive commitments of the strong interpretation. Whereas the medical model de-emphasizes social factors, my account holds them both as value determinants in definitions and as causal explanations. In addition, the mind and brain dichotomy can be better explained by my non-metaphysical approach. In contrast to the minimal interpretation, my account is realistic to the extent that psychiatric kinds can and should be classified based on various biological, psychological, and social causes. The reason is that my account is not based on mere descriptions of actual properties, but instead relies on identifying the stable counterfactual applicability domain of the kind explanation. Nevertheless, in contrast to the strong interpretation, I argue that the appropriate explanatory specification of a disorder may be a socially susceptible psychological process, or the social process itself, rather than an underlying neurocognitive or genetic factor.

6.2 Cross-Cultural Variation of Disorders and their Conceptions

I have argued that psychiatric disorders can in principle be scientifically relevant kinds because they can support domains of explanatory applicability. Ideally then, by tracking these kinds, psychiatric research can explain and predict mental problems as well as facilitate treatment and policy interventions. However, I have also argued against the medical model view that all psychiatric disorders are biological kinds, and against the view that they are merely concept-dependent human kinds. This raises a question: How do we account for the fact that psychiatric disorders have varied across times, cultures, and locations? If psychiatric kinds do not necessarily share an underlying biological structure, when are two property clusters the same psychiatric kind? Hence, cross-

cultural variation of psychiatric disorders presents a challenge when attempting to understand psychiatric disorders as inductive supporting scientific kinds. If some of the psychiatric conditions we classify as disorders are culture dependent, one may wonder whether they are merely social constructs rather than actual psychiatric disorders. Cultural variation also raises the further question of how to incorporate it into classifications and explanations. I will examine these questions in relation to the medical model, and argue that my non-metaphysical applicability approach is more suited to answer them.´

I investigate two interrelated questions related to cross-cultural variation of psychiatric disorders. First, to what extent are psychiatric disorders socially constructed? Second, what type of explanations are needed to explain sociocultural variation and social causation? While some anti-psychiatrics rely exclusively on sociocultural explanations based on the labelling theory to argue that mental illness is nothing but social deviance, some social scientists argue that psychiatric problems cannot be separated from their cultural interpretations, or that they should be understood as constructs created in patient-doctor interactions. I have earlier argued against the strong interpretation of the labelling effect and in favour of a realist account. I will now argue against a strong view of the interpretative account on the same grounds. But just as with the labelling theory, I nonetheless believe that the interpretative account has something to offer psychiatric kind explanations. I begin by first presenting a viable interpretation of social construction of purported syndromes, then argue that severe disorders are identified similarly almost everywhere. After that, I provide a causal account of the interpretive approaches to understanding mental illnes. I argue that a cultural approach to psychiatric disorders should not be understood to rule out other explanations. Rather, they can be understood as explaining some aspects of psychiatric disorders. Based

on this, I explore to what extent culture shapes our cognition and what it means for explaining culture-bound syndromes.

6.2.1 Social Construction and Cultural Conceptions of Disorders

The obvious culture-dependence of some purported disorders raises the question whether they are actually socially constructed kinds. The idea of socially constructed psychiatric kinds can be understood based on Mallon's (2003, 2016) idea that some human kinds are covertly socially constructed (see also Barnes 1983, Khalidi 2013, 2014, Appiah 1996, Griffiths 1999). According to Mallon, in cases of covertly constructed kinds, classified people are not in general aware that kind-typical behaviour is socially caused, but instead presume it to be natural. This means that the kind is not maintained by a priori distinctions, or groupings imposed by folk-categories, or scientific classifications, but instead by false beliefs and the kind-typical actions they generate. The central point is that the covert nature of social causes can unify behaviour and render it kind-typical because both agents and the general population believe that it is the natural way to be and act. When such beliefs become entrenched in practice, and thereby begin to constrain and shape the social and physical circumstances of the classified, their actions are bound together to become kind-typical. In such a case, knowing that certain individuals are members of a kind warrants inductive inferences over the kind-typical properties that they have.[126]

Clearly then some currently labelled syndromes could be such covertly constructed human kinds. In other words, the taken-for-grantedness of those categories makes them appear natural and inevitable.

[126] Although I do not argue for it here, I do not believe that the inductive powers of kinds are an all-or-nothing matter. Even definitional kinds ground their stipulated properties.

Griffiths (1997: 146) argues that, for example, multiple personality disorder (MPD) is a covertly constructed kind upheld by social pretences. Hacking (1995) writes that the status of having MPD was meaningful and provided social acceptance in the form of self-help groups, doctors, and society at large. Therefore, had it become commonly known that MPD behaviour is just pretence, the behavioural pattern would have lost its underlying social cause – the belief that it is naturally constituted (Griffiths 1997: 147). This is also the reason why it is common to doubt the legitimacy of someone's psychiatric problems by questioning the involuntary nature of that person's behaviour (e.g. Griffiths 1999: 150; Cooper 2010: 326).[127]

The fact that some conditions are socially constructed does not, however, mean that social causation could not play a substantial role in explaining genuine psychiatric disorders. Rather, what it means is that social construction cannot be all there is to a psychiatric disorder (Murphy 2006: 260, Schaffner and Tabb 2015). Consequently, the question about cultural variation and psychiatric disorders should be separated from the question whether they are exclusively socially constructed. Social and environmental context can obviously causally affect psychiatric disorders through their interaction with psychological and biological factors, but if we are able to explain at least partly the nature of this interaction, and thereby rule out covert construction, the question of exclusive social construction does not arise.

Although (most) psychiatric kinds are not covertly socially constructed, people nonetheless have associated social roles with them. This raises an important question about social causation. In the previous chapter, I argued that knowledge of the social roles associated with schizophrenia can supplement their explanations and thereby warrant more secure inductive inferences based on the kind. Based on this idea,

[127] *Latah* has been subject to accusations of fakery by being intentional behaviour, see sections 6.2.4 and 6.2.5 (see Winzeler 1999).

it is a separate question whether conditions that vary cross-culturally are psychiatric disorders or whether they are kinds that can support sound inductive inferences. In contrast, Tsou (2021) argues that only biological kinds support robust inductive inferences and this difference (partly) distinguishes real psychiatric disorder from covertly socially contrasted kinds. Likewise, Horwitz and Wakefield (2007: 199) argue that the emphasis should be on the underlying universal dysfunction, which can presumably be intervened upon irrespective of its purported unstable and cultural-dependent expressions. In contrast, I have argued for a non-metaphysical approach to psychiatric kinds, according to which their explanatory stability is an epistemic feature that falls on dimensions (with continuums). Based on this argument, cultural influence, and the relevance of the instability it creates, is a discipline-dependent epistemic question. Hence, rather than asking whether some kinds support inductions, we should ask why they support these inductions. As I have argued, many psychiatric disorders are socio-culturally and historically contingent, and unless we know which properties are kind-typical, we may easily be mistaken over which tokens are of the same type of kind. But if we can account for the possible variations of the kind in specific circumstances, the kind-concept can be used for reliable generalizations. In such cases, we know what the manifest properties of that kind are in those circumstances. Conversely, if we discover a common homeostatic mechanism for properties that tend to cluster together, we can infer that the property clusters are of the same type. On the other hand, I have also argued that the specification of psychiatric kinds is partly dependent on our classificatory goals. Hence, on the one hand, the naturalness or realness of a kind is indicative of our ability to explain and predict how the kind would vary under different circumstances, and on the other hand, it is relative to our classificatory goals.

I have argued for a partly value-laden view of psychiatric disorders, according to which values play an irrevocable role whether certain conditions are considered pathological and where to draw their boundaries. The implication of this approach is that we need to ask whether a wider consensus over which conditions are mental disorders can be reached. Anthropological and historical studies have suggested that almost everywhere people identify the most serious psychic disorders as mental problems. This is consistent with the fact that the rate of psychotic disorders does not vary across cultures (Horwitz 2002: 167). This implies that there is considerable cross-cultural conceptual and value-laden similarity concerning which conditions are labelled real mental problems. Sociologists Allan Horwitz (2002: 32) has concluded that "social groups universally distinguish norm-breaking behaviors that arise because offenders violate social standards from those that stem from some internal dysfunction". For instance, a study made in the 1960s showed that people who were not aware of Western medicine in Southeast Asia had folk concepts of mental disorders identical to those used in psychiatry. These included psychoses, organic brain syndrome, mental retardation, epilepsy, and childhood autism (Westermeyer 1979). A similar study carried out in East Africa showed that irrational psychotic behaviour (e.g. screaming, stripping naked, and running wildly) is considered an internal psychological (Edgerton 1966). This study also demonstrated that the identification of mental problems was conducted based on behaviour, not on "internal" experience, such as hallucinations (see also Luhrmann 2016: 8).[128] Further, a comprehensive study conducted among Yorubas and Eskimos proved that explicit labels for insanity exist in these cultures (Murphy 1976). The labels refer to beliefs, feelings, and actions that are thought to emanate from the mind or inner state of individuals and to be beyond their control to the

[128] Although the symptoms may be related to brain activity, and are thereby invariable produced, culture may dictate how they are managed.

extent that they are identified as "losing one's mind". Therefore, there is credible evidence that at least some mental problems are identified in many places in the same way.[129]

This is also my own discovery based on field work in Benin, West Africa.[130] In Benin, beliefs and healing practices related to mental illness are merged with vodun practices and beliefs. Vodun is a religion mainly practiced in Benin and Togo in West Africa, although it spread to the West Indies and Brazil during the slave trade. Vodun includes beliefs in witchcraft and magic, according to which the control of invisible forces enables predictions and interventions on our lives, but also facilitates spiritual healing. These beliefs shape conceptions of illnesses, how they are treated and, according to many studies, how they are experienced and even shaped (see Luhrmann 2016). In Benin, severe mental disorders are considered to be problems of the mind (*tadu* in the local *Fon* language). They are identified by out-of-place behaviour. As one of my informants told: "They can run around, do not recognize their own house anymore, take off their clothes, not recognize their parents." In such cases, if possible, the inflicted individuals are sent to a local hospital for analysis. Another informant mentioned that, "If the doctor cannot find anything, they know the sickness is from traditional vodou reasons and can be treated traditionally." This is an example of medical pluralism, namely the commonly shared conception that everything that helps should be used (Honkasalo 2016). Hence, where cultural conceptions in the West and Benin differ is over

[129] A plausible explanation for this similarity in cross-cultural identification of severe mental disorders could be related to our mind-reading abilities (see Hutto 2013, Gallagher 2013). Pascal Boyer (2011: 96) has argued that intuitive psychology constrains the detection of disorders because it provides people with intuitions that certain specific kinds of behaviour are signs of mental problems.

[130] I carried out my field work in 2020 in three villages in Benin by the coast near Togo. I conducted six interviews. One interview included five people, while the others were personal. The interviewees were bokono (diviners) and voduno (vodun priests).

what causes severe psychiatric problems and how to cure them, rather than whether such things exist.

6.2.2 Culturally Permeated Experience and Interpretative Explanation

In spite of the cross-cultural similarities in severe cases of psychiatric disorders, and in their identification, there is considerable cross-cultural variation in the prevalence, expression and conception of milder psychiatric disorders. For example, the rate of major depression is much higher in the West than in many other parts of the world, while some conditions, such as multiple personality disorder, seem to be particularly prevalent in the USA (Horwitz 2002:168). This implies that sociocultural factors play an important role in their explanation, and that their conceptualization requires explicit culture- and value-sensitivity weighing.

This is also my experience from Benin. Especially voice-hearing experiences are not held to be as bad to have as in the West. This is represented by what one bokono informant mentioned: “Some people are naturally like that, they can hear spirits, or they can be signs of god, they can also be signs of other things, or it can also be that they tempt you to do bad things. But they are easy to treat with herbs”. In the West, in contrast, the medicalized culture shapes these perceptions. They are held to be bad and signs of madness, which in turn influences how people are treated and even excluded from society (Luhrmann et al. 2016).

It has been common to draw a distinction between cultural expressions of mental illness and the purported underlying biological disorder. This distinction has been presented in different ways. Boorse (1974), for example, maintains that disease is a value-natural and objective condition, while illness represents self-identified and

negatively valued symptoms (Boorse 1974: Radden 2003: 43). A similar distinction has been made between explanatory models of pathogenicity and pathoplasticity (Yap 1952). According to Kleinman (1991), pathoplasticity describes expressions of unstable culturally shaped disease while pathogenicity represents the biological diseases that realize those expressions. Kleinman (1987: 107), for instance, has provided a well-known example of cross-cultural variation between depression in China and the West. According to his field work, in China depression is experienced or reported as physical aches and pains, especially as lower back pains, whereas in the West it is a distinctly mood-related problem. Based on these findings, Kleiman (1988) emphasizes the importance of pathoplasticity, arguing that rather than concentrating on the underlying biology of disorders, we should pay attention to the social and cultural circumstances that surround suffering and healing (see also Kirmeyer 2006: 10). Kleiman (1988: 107) writes: "Depression experienced entirely as low back pain and depression experienced entirely as guilt-ridden existential despair are such substantially different forms of illness behaviour with different symptoms, patterns of help-seeking, course and treatment responses that though the disease in each instance may be the same, the illness rather than the disease is the determinant factor."[131] Consequently, Kleinman argues that it would be a *categorical fallacy* to assume that Western psychiatric categories can be applied cross-culturally (Kleinman 1991: 14).

The interpretative approach to mental illness that Kleinman employs is based on the conviction that mental illnesses are not merely natural entities but also carry meanings and support human understanding (Good 1994: 53). This explanatory approach should be distinguished from the labelling theory of social interactive

[131] As Kleinman is writing about unknown underlying causes, Radden (2003) wonders how he can overcome the problem of incommensurability.

explanation, because the emphasis is not so much on labels being imposed from the outside, rather than on the two-way interpretation that mediates between the cultural categories and the individuals' own self-understanding. In psychological anthropology, pathoplasticity has been equated with "idioms of distress" (Nichter 1981), which describe culturally mediated illness experiences and behaviours (see Kleinman 1991, Good 1994, Luhrmann 2016). The idea is that cultural conceptions of illness enable patients to understand themselves, while "cultural interpretations interact with biology or psychophysiology and social relations to produce distinctive forms of illness" (Good 1994: 53).

A crucial consequence of this is that a larger explanatory role should be given to the meanings associated with mental symptoms. Aragona (2015: 599), for example, argues that "Mental symptoms should be considered as hermeneutic co-constructions occurring in a[n] intersubjective space created by the dialogue between sufferer and healer... To understand, handle and communicate these experiences, sufferers proceed to configure them by means of templates borrowed from their own culture." In an interview, Aragona told me that "I don't force Western categories but I try to find a way through the middle."[132] This approach in clinical practice is apparent in this vignette from an interview I conducted with Aragona at the Roman clinic where he works.

[132] I interviewed Aragona at the NIHMP (National Institute of Health, Migration, and Poverty) in Rome. In Italy, treatment is community based and there is a general attempt to avoid using psychiatric labels. Nevertheless, based on my interviews, diagnostic categories are also necessary there. Such categories are needed for bureaucratic purposes, to enable communication in international conferences, and these categories are also often looked up by the patients themselves.

> "One day [I met] a person maybe from Cameroon, [a] Muslim. He came and he had a homosexual history with his boyfriend. He was discovered, they risked being killed and they had to flee. And in crossing the desert they had nothing to eat, so they risked dying in the Sahara. Then his boyfriend said that he won't eat the food because it is not allowed in their religion… But the boyfriend died… he was forced to come to Italy. And he had a lot of problems. One of the problems was dreaming every night. So he could not sleep because in his dream there was a cow. He woke up when this cow was killing him… [and] said: 'What do you think it can be because I don't know what it can be?' He said that for Muslims the cow is the symbol for enemies… My enemies are the family of my boyfriend who are persecuting me because they think that I killed him. So I'm in pain and grief because my boyfriend died. And I feel guilty because I survived and he died. And also I feel persecuted by them… Step by step we found that marabout who is a… traditional healing figure in that country, [and he] could make something in order to let this cow disappear. So we arranged to make a phone call to Africa to ask some of the family if this is possible… the marabout gave his reply. And after [that] the dream changed. And in the end he was in the bus that was travelling in the desert, [as well as] above the bus, with the cow and the answer."
>
> (Aragona: Interview on October 17, 2017. Minor language corrections have been made.)

The individual had been labelled as having an adjustment disorder (at the clinic, a third of the patients were given this diagnosis). According to Aragona's clinical experience, the cultural meanings associated with

his trauma played a crucial role in the individual's behaviour, experience, and response to treatment. It is plausible that such experiences in turn shape the course of the disorder, evenwhen there is a biological universal to be found (see Kirmayer 1991: 26).

However, the interpretative approach and social interaction theory should not be understood to rule out other explanations, although this is the approach that some take. Rosenberg (2002: 253), for instance, has argued that "mechanism-oriented clinical medicine … assigns comparatively low priority to the multicausal, to the social, ecological, public policy, and quality of life perspectives". But as I argued in the preceding chapter, the interplay between psychical factors and sociocultural meanings (including values and folk conceptions) can be understood to supplement the applicability domains of psychiatric kind explanations. This is made possible by my non-metaphysical interpretation of the responsible mechanisms of property clusters. More specifically, because intentional actions can be given causal explanations, and thereby they do not require a different interpretative explanation, there is no a priori reason to make a sharp distinction between natural and meaning-oriented explanatory approaches. Rather, the applicability domain approach to psychiatric kinds can be understood heuristically to facilitate interaction between various explanatory approaches. In fact, I believe that previously a sharp distinction between law-based natural science explanations and interpretative humanistic explanations has hampered psychiatric explanation and understanding. Such a distinction underlies Jaspers's (1963) approach, while the strong medical model has opted for a fully naturalistic approach to explaining psychopathology. I have argued, on the other hand, that the contrastive counterfactual theory can accommodate meanings as causal explanations, and thereby meaning-based intentional explanations of sociocultural and narrative

approaches to psychiatric disorders can form a legitimate addition to an understanding of psychopathology.

This interactive pluralist approach can be illustrated with schizophrenic voice hallucinations. The more the sociocultural meanings associated with them are known, the more that knowledge enables extrapolations between various forms of schizophrenia. Alternatively, if classificatory goals and explanatory virtues so require, the schizophrenia kind may be split into sub-kinds. The feedback mechanisms that I examined in Chapter 5 should therefore be understood as part of a larger sociocultural "niche" that shapes psychiatric disorders and their expressions. Moreover, the better those external structural features are known and understood, the more securely we can make inductive inferences based on psychiatric disorders and their properties. Consequently, for instance, knowledge of cultural meanings that influence how psychiatric disorders are experienced and acted out, i.e. idioms of distress, should be incorporated into diagnostic practices and classifications not only to avoid blatant misidentifications (Kirmayer 2001), but also to avoid unwanted and unpredictable cultural distortions of disorders. This means that valid generalizations require an understanding of the culturally determined patient's own point of view (cf. Parnas and Urfer-Parnas 2017).

The importance of this pluralist approach to explanation is apparent in how most researchers working on psychiatric disorders agree that the influence of sociocultural factors comes in degrees. Arpaly (2005) has convincingly argued that even milder disorders or individual cases are more hard-wired than others. She explains this with the concept of content efficacy: unlike other bodily states, minds have content because they are about something. Furthermore, mental states may affect one another in such a way that mental content plays a crucial role. For example, I may get offended if someone calls me names in Finnish or

English. On the other hand, if someone called me names in Greek, it would not have the same influence because I would not understand what was said. In the former case the mental content is causally efficacious, whereas in the latter case the content has no impact on my mind. Arpaly also mentioned another more specific type of content efficacy: reason-responsiveness. In such cases, the person responds to reasons that the mental content raises. For instance, calling me names might lead me to infer that the name-caller just happens to be having a bad day, or it might lead to intense soul-searching of my own behaviour. Following this kind of reasoning, according to Arpaly, we can come up with three different possibilities for psychiatric causation.

(i) Laura is depressed because she has been through an Arctic winter and she is light-deprived.
(ii) Laura is depressed because media images of female beauty make her feel worthless.
(iii) Laura is depressed because she lives in a tyranny and witnesses a lot of injustice.

The first case is not content efficacious and resembles the etiology of somatic problems. The second case is content efficacious and the third reason-responsive. Arpaly stresses that most disorders and cases involve all three forms of causation to different degrees. Nevertheless, I would argue that both the second and especially the third case are perhaps best not explained solely by looking inside the individual's bodily situated mind or brain. For instance, the second case requires an explanation of the social practices that maintain norms of beauty, the individual and general internalization mechanisms of norms, and the factors that explain why only some people are susceptible to the negative impact of the norms (cf. Pöyhönen 2010, 2013a). These sociocultural factors and meanings may serve as a causal trigger for

depression. In the third case, in contrast, the dynamic interaction between the individual and his/her sociocultural context seems to be a sustaining or constitutive part of the problem. Therefore, sociocultural explanations are more relevant in explaining the latter two types than the first type of depression.

Hence, the interpretative approach, just as any other explanatory approach, should not be considered the only correct explanation. No cultural explanation applies equally to all disorders to the extent that it could rule out other explanations.[133] The contrastive counterfactual theory of explanation, however, is based on the idea that no explanation is basic, and instead, explanatory relevance is set by how the explanandum phenomena are characterized with contrasts. In line with this, my view is that a social explanation of psychopathology depends on the particular disorder, as well as on which aspects of the kind are considered explanatorily interesting. In this way, sociocultural explanations can supplement, to various degrees and ways, the psychiatric kind's explanatory domain.

In conclusion, while exclusive social constructivist explanations may apply to some disorders, these are exceptions. In general, people everywhere tend to see severe disorders as problems of the mind. However, I argued that cultural explanations can be understood causally and that they are crucial in understanding some aspects of psychiatric disorders.

6.2.3 Cultural Shaping of Cognition and Disorders

In the previous section I argued that an interpretative approach can be understood as a type of causal explanation of the cultural influence on

[133] Traditional ethnopsychiatry rooted in psychodynamics, such as Devereux's (1980) approach, can be considered to rule out other explanations.

psychiatric disorders. In this section, I explore another approach to cultural influence on disorders based on cognitive science. If the cultural influence on disorders can be understood in causal terms, there is a priori reason to think that it does not only shapes experiences of disorders, but also shapes the processes that give rise to those experiences. Based on recent studies, I argue that we cannot rule out that culture can penetrate cognition to the extent that it undermines the universalist account of human cognition and mind that the medical model seems to rely on.

As I argued in the last chapter, the challenge for the strong view of the medical model is that although it acknowledges social factors, it nonetheless considers them subservient to more profound biological explanations. Murphy (2006), for example, argues that cognitive neuroscience provides the means to explain dysfunctions as breakdowns in information processing systems in the nervous system. Similarly, RDoC is based on the conviction that mental disorders have a neurological ontology. However, research on cross-cultural variation of human cognition has changed over the past few decades. Joe Henrich et al. (2010), for instance, demonstrate that there is considerable cross-cultural variation in self-concepts, the visual system, and even in visual illusions usually considered to be cognitively impenetrable (e.g. the Müller-Lyer illusion). According to Segall et al. (1966), San foragers of the Kalahari Desert were not affected by the illusion possibly because they were not visually exposed to "carpentered corners" when growing up. These difference, nonetheless, do not prove that there is no cognitive unity rather than undermine the presupposition that WEIRD (Western, Educated, Industrialized, Rich and Democratic) people's similarities in normally considered cognitive processes are representative of human cognition in general.[134] That is, Henrich et al. (2010)

[134] For a discussion of cognitive penetration, see Firestone and Scholl (2016) and Arstila (2017).

demonstrate that there is no reason to presuppose that the cognitive processes usually regarded as universal would not be shaped by sociocultural processes.

Based on these possible cross-cultural cognitive variations, Washington (2016) presents a mechanistic variation problem for the brain-based approach to psychopathology: if cross-cultural cognitive variation is possible, so is deep mechanistic variation of psychiatric disorders. This implies that we cannot rely on finding universal neurological or cognitive mechanisms to make inferences about universal neuropsychological abnormalities. Empirical evidence seems to back up this argument. Seligman and Kermayer (2008), for example, argue for an integrative approach to the sociocultural influence on psychopathology. Although they believe that there may be cognitive-psychological universals, culture may nonetheless penetrate some parts of our cognitive psychology in a fine-grained manner (Sun 2012; See also Nisbett and Norenzayan, 2002, Carruthers et al. 2005, Semin and Echterhoff 2011). Similarly, Fessler and Machery (2012) argue that relevant cross-cultural differences in cognitive processes have been discovered. The moral is that there is no reason to presuppose that the human mind and cognition will suffer from disorders in similar ways, or even that psychiatric disorders are the same everywhere.

Fessler and Machery (2012) distinguishes between two basic types of cultural variations in the cognitive processes usually taken to be universal (see also Nisbett and Norenzayan 2002). First, the divided model is that people in different cultural and environmental circumstances learn to rely on different cognitive processes, although they share the potential for the same cognitive capacities. Second, the integrative model is that cultures vary in how people learn to rely on inferential procedures, strategies, and the environment to the extent that sociocultural factors may shape or are constitutive of those cognitive processes. Based on this distinction, it is possible that neurocognitive

processes and sociocultural factors in combination produce, and sometimes even constitute and shape, psychiatric problems (cf. Kirmayer and Seligman 2008). According to the stronger integrative model, sociocultural processes may shape cognitive or psychological processes, which in turn may instigate culture-dependent psychiatric problems. On the other hand, according to the divided model, some cognitive universals may constrain the possibilities for cultural manifestations of psychiatric disorders. As a consequence, even with universally occurring psychiatric disorders, when they interact with local environments, they may produce different outcomes. It is an empirical question whether the divided or integrative model, or some combination of the two, explains a particular psychiatric disorder.

One example of the divided model can be represented by schizophrenic hallucinations. Although there is no agreement whether schizophrenia hallucinations are realized by the same neurocognitive processes (Luhrmann 2011, see also Aleman and Laroi 2008), one suggestion is that they are due to disruptions in perceptual processes.[135] However, even if similar disruptions in cognitive processes underlie some or all types of schizophrenia, the content and experience of schizophrenia hallucinations vary across cultures (see Chapter 5). Luhrmann (2011) argues that hallucinations can be shaped by cultural learning in at least two ways. First, people learn and train to focus their minds in different ways. More specifically, people in different cultures learn to manage their symptoms in different ways. Second, people are influenced by culture-dependent representations of the mind. A good example of the latter possibility is how the Iban of Borneo apparently lack

[135] This is implied by how schizophrenia patients experience depth inversion illusions, such as the hollow-mask illusion, (Keane et al. 2013). Keane et al. (2013: 4) suggest that the effect may be "explained by disturbances in prediction-error monitoring, the statistical process by which the brain updates expectations about future states of the world based on how past predictions matched with past experiences".

thought insertion and thought withdrawal entirely in schizophrenia, probably because they do not think of the mind as a container (Barrett 2004; see also Lurhmann 2011: 10). I have already mentioned the study of schizophrenia hallucinations in India, Ghana, and the USA (cf. Luhrmann's study 2015). According to this data, Americans hear harsher voices while Indians have more visual hallucinations. Luhrmann suggests that the difference could be due to cultural variation in the emphasis on sensory importance, and that unlike Indians in general, Westerners are uncomfortable when the wall between the mind and the world is breached. I have also mentioned the feedback effect of medicalization of hallucination in America (see e.g. Mazza 2016). There are accounts of disturbed people's hallucinations being affected by descriptions of devils and other religious beliefs dating back to the Middle Ages (Pietikäinen 2013: 54). Sonya Pritzker (Danzinger et al. 2011) has studied the effects of learning to focus one's mind. According to her, learning yoga and Chinese medicine can transform the Western conception of emotional suffering as a mental feature to the idea that distress can be controlled by the body.

Second, sometimes trying to locate a psychiatric problem inside an individual's brain or mind is not explanatorily helpful.[136] Rather, a better explanation could be reached by situating the problem within larger social dynamics.[137] This matches the integrative model of cognition. Seligman and Kermayer (2008), for example, argue that sociocultural and cognitive-psychological processes may constitutively explain psychiatric behaviour. As an example, in different cultures people may learn to rely on different processes, strategies and environment, so that some psychiatric disorders are perhaps best understood as being

[136] See Miłkowski et al. (2018) for an overview of extended approaches to cognition.

[137] This is in line with recent attempts to understand psychiatric disorders through problems in social interaction (Schilbach et al. 2013, Schilbach 2016). See also Krueger (2020).

constituted or scaffolded by the dynamic interaction between cognitive-psychological processes and the social environment.[138] In their seminal article "The Extended Mind", Clark and Chalmers (1998) argue that notebooks and computers can form external constitutive parts of our minds and memories. If this thesis is applied to psychiatry, some psychiatric problems are perhaps not best explained by our brains but are instead formed and scaffolded by the surroundings.[139]

A good example of the extended approach is offered by Kincaid and Sullivan (2010: 368), who argue that social factors can play a substantially stronger role than mere triggering factors in explaining psychiatric disorders. They argue that understanding addiction, for example, requires one to explain the role of social causes at the onset, in remission, in relapse, and in relapses and treatments. Crucially, complex explanatory models of addictions indicate that addiction can be distributed throughout the brain rather than being realized in a particular part of the brain. Their conclusion is that it may not be possible to have a neurobiological model of all types of addictions and across all afflicted people (2010: 367). Rather, understanding addiction requires a better inclusion of social factors, which are of varying importance for the explanation of particular disorders. They also argue that sometimes a better understanding of psychiatric problems may require one to locate the problem outside of the brain and the individual (see also Broome and Bortolotti 2009, Fuchs 2012). This requires, according to Kincaid and Sullivan, considering the social role the problem places on the inflicted individual. The role of an addict is constituted by individuals' interactions with their families and friends. They point out that this role can also be rewarding. Although it comes

[138] See Sterelny (2010) for more about the scaffolded approach to the mind.
[139] Telakivi (2020) argues that the idea of psychiatric problems extend beyond our brains and bodies, and can be elaborated based on the extended mind thesis or the 4E programme – embedded, embodied, extended, and enacted. See also Huffman (2016)

with limited possibilities, it nonetheless enables some interactions within peers. Based on this, it seems to me that it is also possible that social roles sometimes interact with our cognition to the extent that it is their interaction that provides the most powerful explanation of addiction.

However, proponents of the extended cognition theory have struggled to demonstrate why the connection between internal and external factors should be considered constitutive rather than causal. To overcome this problem, the difference between constitutive and causal factors can be reformulated as a question of explanation and classification. Pöyhönen (2014) argues that an extended approach to cognition can sometimes be explanatorily more powerful. The idea is that different explanatory aims may lead to demarcating cognitive systems alternatively so that sometimes explanatory power is enhanced by specifying cognitive systems to extend beyond the skin-and-skull. That is, different explanatory programmes may emphasize different explanatory virtues to the extent that their mechanistic delineations are better at answering discipline-relevant sets of explananda (2003b: 51). According to Pöyhönen, if such property cluster demarcations nonetheless sustain robust inductive inferences and explanations, they can be called natural kinds.

In sum, the possibility of cross-cognitive variation undermines the seeming commitment of the universalist stance of the medical model. However, this does not rule out that there are real disorders that are constituted by various social, psychological, and biological interactions. Rather, what it does rule out is the view that those interactions could be understood only on the level of neurology. Next, I will explore how research into culture-bound syndromes supports this interpretation.

6.2.4 Culture-Bound Syndromes and Classification

Culture-bound syndromes offer an analytically fruitful example to analyse the causal influence of sociocultural factors, and the possibility of including the factors into classifying and explaining psychiatric kinds. Studies of culture-bound syndromes have concentrated on whether they are cultural constructs, culture-specific expressions of universal underlying biological disorders, or genuinely culture-specific disorders (Hugher 1985, Cooper 2010, Hacking 2010, Murphy 2015). In this section, I examine these possibilities based on a more contemporary approach to the dynamics of culture. My argument will be that in many cases deciding between the universalist and particularist interpretation of culture-bound disorders depends partly on the specific disorder, and partly on values and classificatory aims. The point of this analysis is to demonstrate how a brain-based approach to disorders does not seem to leave enough room for value-sensitive considerations and cultural explanations.

Culture-bound syndromes are defined by being seemingly culturally or geographically bounded. The term was coined by P.M. Yap (1952, 1966) during the time of British rule in much of Asia (see Winzeler1995). He employed it to describe *latah*, a unique syndrome or pattern of behaviour, which includes mimicking and shouting, and can still be found in parts of Malaysia and Indonesia (see the next section). However, the notion of "culture-bound" does not necessarily mean that it is somehow more culturally constructed, such as with multiple personality disorder or eating disorders. Rather, it can also mean that the syndrome is "culture-specific". A good example is kuru, which could be found in New Guinea (Hughes 1985: 9). The syndrome was caused by prion (as with mad-cow disease), which was transmitted through the local practice of eating the dead (see also Cooper 2010: 327). While DSM-VI-TR includes 25 culture-bound disorders, DSM-5

(832-837) lists nine cultural syndromes. DSM-5 (APA: 758) changed the term to cultural syndromes to reflect the fact that some syndromes may be cultural expressions of universal syndromes.[140] "Cultural syndromes" are defined as "clusters of symptoms and attributions that tend to co-occur among individuals in specific cultural groups, communities, or contexts and that are recognized locally as coherent patterns of experience".

The notion of culture-bound syndromes has an underlying connotation of ethnocentrism and an outdated view of culture. This is because its perspective is distinctly Western and is based on obsolete views of cultural boundaries. Kirmayer (2007), for example, mentions that the idea of a culture-bound syndrome originally had a racial and colonialist connotation (see e.g. Kirmayer 2007). Likewise, Hughes (1985) writes that culture-bound syndromes imply that they are "deviant deviances" because they are unfamiliar to Westerners. That is, culture-bound syndromes were considered to be "inferior" culturally mediated forms of disorders that diverted from the purely biological versions that Westerners experienced.

Hence, in order for the concept of a culture-bound syndrome to maintain an explanatory role, the concepts of "culture" and "bound" should be distinguished from their traditional understanding, which was based on the holistic, essentialist and "otherness" approaches to

[140] DSM-5 (2013) made revisions to DSM-IV by emphasizing culture's importance in the definition, identification, and expression of psychiatric problems. La Roche et al. (2015) point out that especially DSM-5's introduction of the novel Cultural Formulation Interview (CFI) is an improvement over DSM-IV's similar one. The aim of the CFI is to enable practitioners to "identify cultural and contextual factors relevant to the diagnosis and treatment of different problems; its main aim is to provide a cultural meaning to patients' symptoms by embedding them into a cultural context" (p. 185). However, La Roche et al.'s (2015) conclude that DSM-5 is still committed to a biological and universalist approach to disorders to the extent that cultural variation is taken to be superficial (p.187).

culture.[141] The idea of strict cultural boundaries upheld by traditional anthropological approaches has been criticized for not acknowledging the power relations that marginalize people and create conflicts within and between individuals, on the one hand, and for neglecting the dynamic and boundless nature of cultures, on the other (Ortner 2016). In line with this, psychiatric anthropology has lately concentrated on locally and historically contingent social controversies where "local actors come to articulate competing views on the nature of their distress" (Kitanaka 2012: 5).[142] This is related to a shift by psychiatric anthropology towards a more engaged and collaborative approach to clinicians and psychiatric research (Luhrmann 2016: 5). Further, the approach acknowledges that there are conflicts within cultural spheres about how to intervene and treat mental disorders, which in turn reflects the way in which disorders are experienced, recognized, and enabled (cf. Lakoff 2005: 17, Kitanaka 2012). Based on this, my approach is that the "bound" element of a culture-bound syndrome can be understood as dynamic family resemblances, whereas "culture" is something that permeates all illness experiences, including those of Westerners.[143] This approach retains the central idea of the culture-bound syndrome, which shifts the notion towards the dynamic and local ways in which people can experience, express, and interact with mental suffering. Hence, the basic idea of culture-bound syndromes is still that they only appear in some places but not in others.

A viable explanation for some culture-bound syndromes is offered by Hacking. He (1998: 13) has argued that transitional syndromes are brought about and maintained by dynamic sociocultural or ecological niches. This implies that dynamic interaction between local norms,

[141] In DSM-5, the term "culture-bound syndromes" has been replaced with "cultural syndrome".
[142] See also e.g. Young 1995, Kleinman 1988.
[143] Perhaps a better term to describe dynamic kinds of psychopathology could be culture-reactive syndromes (see Huger 1985).

power relations, and increasingly also global conceptions, may create their own culture-bound or reactive forms of psychopathology. Hacking (1998) has provided an analysis of the *fugue* or "mad traveller" syndrome (as well as hysteria and Multiple Personality Disorder) based on his idea of ecological niche. The syndrome appeared in France at the end of the 19th century, and its symptoms included a seemingly purposeless compulsion to travel, obscured consciousness, and loss of identity during the travel. Hacking argues that transitional syndromes, such as fugue, can come about due to "an ecological niche" created by four vectors. First, a *taxonomical* gap enables a specific kind of illness. Fugue could have fitted either hysteria or epilepsy, or both. This made the phenomenon interesting and salient for the physicians of the day.[144] Second, *cultural polarity* creates a tension. Because fugue was somewhere between romantic tourism and criminal vagrancy, it was appealing to inflicted individuals. The third vector is *observability*. For something to be a disorder, it needs to be strange, disturbing, and noticed. The fourth vector is *release*. For example, fugue behaviour enabled men of lower class status to escape their working lives. Murphy (2006) has suggested that norms could form a fifth stabilizing vector, whereas Kuorikoski and Pöyhönen (2012) have proposed that a biological vector could do the same.

This analysis of transient syndromes can be further supplemented with Hacking's (2010) "imitation and internalization" approach. He demonstrates the reasoning with the pathological withdrawal of refugee children in Sweden, whose families were trying to receive permanent residence between 2001 and 2007. This withdrawn behaviour comprised an unwillingness to communicate with others, a refusal to eat, and even mild cases of coma. This behaviour was voluntarily imitated by other refugee children to the extent that it caused an

[144] Murphy (2006) has argued that the looping effect is part of the taxonomical vector since it links the transient disorder with classification.

"epidemic". Hacking argues that although the behavioural pattern may have started voluntarily, in the end the inflicted children lost control over the symptoms and over their behaviour.[145] A somewhat similar example, where the niche plays more of a triggering role, could be eating disorders. It is likely that pathological fasting or compulsive eating starts by imitating culturally sanctioned patterns of behaviour, but once the behaviour has become constant, it may be almost impossible to give up.[146]

According to Murphy (2015), some disorders may only appear to be culture-dependent but are in fact universal disorders that are only recognized in some parts of the world. The reason can be that there is no conceptual space for a disorder and its symptoms in some parts of the world. For this reason, although the symptoms of the disorder are universal, they are only recognized in some cultural contexts. A plausible explanation could be that our minds unconsciously latch on to culturally sanctioned ways of expressing suffering (Waters 2010: 111). This could either mean that the symptoms are not recognized at all, or that they are not experienced as pathological or objectionable.[147] The latter seems to be the case, at least to a certain extent, with the mentioned cross-cultural variation in the experience of schizophrenia

[145] Recently, some of the children have come forth and admitted that they were faking due to their parents' pressure. Nonetheless, this does not prove that all of the children had control over their behaviour or that they remember their own experiences correctly.

[146] Murphy (2006: 262) argues that sociocultural factors cannot explain biological susceptibility to bulimia.

[147] E.g. in the fifties in the USA, when people claimed to have black and white dreams (Schwitzgebel 2002). It may either be that the colour of their dreams did change because of black and white television programmes, or that their experience of dreams changed. Antti Revonsuo (2000) has argued that dreams serve a biological function, namely that dreaming is a "dry run" to prepare for dangerous situations. Based on this idea, perhaps it is only the content or experience of dreams that is culturally amenable.

hallucinations (see also Vesterinen 2021). However, Murphy also succinctly points out the limits of explaining disorders with reference to cultural factors. Cultural explanations of behaviour may not be applicable, as such, to culturally permeated expressions of underlying neurocognitive conditions due to their distorted nature. On the other hand, cultural explanations may be applicable as such to some culture-bound syndromes that can in fact be covertly constructed human kinds, such as multiple personality disorder, or some of its versions.

Obsessive-compulsive disorder (OCD) and anorexia nervosa present a comparable example of a universal and niche-dependent psychiatric kind. Studies have shown that OCD has a strong neurocognitive and hereditary base. Hence, social factors and labels probably influence only the content of obsessive thoughts and the types of repetitive behaviours that inflicted individuals engage in.[148] This would match Murphy's example, where although there is an underlying universal disorder, whether its symptoms are recognized as pathological influences the illness experience. Anorexia, on the other hand, seems to exemplify a more substantive case of niche dependency, indicated by how we have been able to follow its emergence simultaneously with the implementation of the diagnostic category in China, Taiwan, Japan, Singapore, and Hong Kong (Lee 1996). In this case, there is no doubt over the realness of the disorder, although it may be generated and partly maintained by an "ecological niche". That is, the normative niche, together with the looping effect, may create conceptual space and sanction the correct manner for displaying one's suffering in eating disorders, i.e. a "normal" way to be abnormal. Once this "conceptual possibility" is internalized, it probably latches on to other psychological and cognitive factors, so that self-starving becomes

[148] Although it may be that these individuals are particularly susceptible to altering their behaviour according to the ideas associated with the disorder. (My thanks to those who pointed this out to me).

hard to give up for more profound reasons (i.e. this could be a case of integrated cognitive explanation, where cultural models fine-tune and interact with cognitive processes to produce instances of illnesses).

There also seem to be cases when culture (or the "ecological niche") does not play a causal but a constitutional explanatory role in culture-bound syndromes. This can be exemplified with the culture-bound syndrome of Hikikomori, which is a form of behaviour found among Japanese youth. It is expressed by social withdrawal from all face-to-face contact. Typical medical explanatory models concentrate on the isolated individual. In contrast, the mentioned interpretative approaches have argued that parents and society as a whole play an important role in shaping and maintaining the behaviour (Rubinstein 2016). In line with this, and based on the extended mind thesis, the interpretative approach enables us to explain this behavior as partly constituted by social interactions imbued with cultural meanings. This means that prevalent cultural narratives make the behaviour meaningful, and parents play their part in enabling it[149]. It is also possible that economic and social uncertainty in Japan are contributory factors (see Allison 2013)

In contrast, it does not seem to be possible to explain Hikikomori by referring to the brain or to the psychological processes of the individual or the causal impact of the surrounding culture. In fact, it is most likely that there is nothing internally detectable within the inflicted individuals' minds or brains that would allow us to individuate them from the rest of the population. Rather, what constitutes their kind-typical behaviour is better explained by the surrounding cultural norms and social relations. That is, although this behaviour is considered shameful, it is meaningful to the individual, to his or her parents, and to society in general.

[149] Narratives may play an important role for many aspects of cognition and psychophathology, see Kirmayer and Crafe (2014), Kirmayer (2005).

Some syndromes raise the question whether they are culturally local or variations of universal disorders. A good example is Tajin Kyofusho ("interpersonal fear disorder"), which is an extreme form of social phobia. Although its full-symptom profile can be found only in Japan, similar phobias are nonetheless manifest in the West (APA 2013: 837). For instance, DSM-5 (APA 2013: 837) suggests a resemblance with the fear found in the USA of having offensive body odour (olfactory reference syndrome). The symptoms of Taijin Kyofusho include extreme fear of being disliked or of not being worthy of respect by others. These feelings translate into major concerns about facial blushing (erythrophobia), having an offensive body odour (olfactory reference syndrome), inappropriate gazing (too much or too little eye contact), stiff or awkward facial expressions or bodily movements (e.g., stiffening, trembling), or body deformity (APA 2013: 837). As a consequence, the individuals suffering from Taijin kyofusho withdraw completely from social interaction. A common cultural explanation is based on the Japanese sense of shame, but due to the similarities with other phobias, there may also be a shared biological and psychological process. Kirmayer (1991: 24) nonetheless argues that the distinction between pathogenesis and pathoplasticity does not hold in explaining Taijin Kyufuru because "the same factors contribute to both the cause and content of the illness". He follows Kleinman and Good (1985), according to whom cultures can shape our thoughts, which in turn can affect cognition and distress. This, in turn, may alter the underlying physical constitution of the disorder.

Consequently, the explanatory role given to sociocultural factors is not only dependent on the particular condition, it is also related to the epistemic and value-laden purpose of the classification. Therefore, whether some syndromes should be considered constituted by sociocultural dynamics, or whether they should be considered mental problems at all, may be dependent on the classificatory aim (see

Chapter 4). Kirmayer (1991: 26) argues, for instance, that universal psychiatric classification is an impossibility. Rather, classifications should make explicit the goal and context of their application, which in turn determines to what extent cultural factors are relevant. This approach is consistent with my applicability domain approach. Especially in the case of purported culture-bound syndromes, the responsible mechanisms can be delineated differently for different explanatory and classificatory purposes. For instance, classificatory goals determine whether Taijin Kyofusho should be considered to be one underlying neurocognitive disorder, or split as a unique cultural disorder. This type of classificatory weighing is also important because of the increasingly unbounded nature of "culture-bound" syndromes.

Moreover, how culturally susceptible disorders are classified is not only an academic question, but also something that is relevant for ethical reasons. This is because the classification may causally interact with local beliefs and the syndrome itself. The relation between scientific classifications and cultural conceptions of syndromes can be understood with etic and emic knowledge. While emic knowledge describes local understanding of behaviour, etic knowledge describes the outsider observers' understanding. As I argued, etic knowledge may alter emic knowledge either by creating congruence or incongruence between the two conceptions or between the scientific conception and the kind-typical behaviour.[150] Moreover, clearly emic knowledge does not always pick out a real property cluster (Kokkonen and Koskinen 2016: 102). In psychiatry, this means that researchers may discover that local conceptions lump together or split psychiatric kinds that are not based on some underlying mechanism, but are instead based on value-laden folk beliefs. However, those local conceptions may nonetheless

[150] Incongruence can cause psychiatric problems. Franz Fanon (1952), for example, describes the detrimental influences of internalized racism (see also Beneduce 2016).

serve other cultural purposes, which may in turn contribute to understanding and managing local ways of suffering. This is something that may be important to keep in mind when exporting diagnostic categories. On the other hand, if Western diagnostic categories can support more robust inductive inferences, enable efficient treatments, and dispel local stigma, the moral question is not raised.

There are two ways in which scientific projects and folk conceptions can interact. In the first instance, larger cultural changes can bring about classificatory changes. This was the case with the removal of homosexuality from the DSM due to the gay movement and the general de-pathologization of homosexuality. A similar movement seems to be taking place now concerning transgender. Hence, it would be a mistake to think that classificatory practices are always the engines of change whenever something comes to be considered a mental disorder. In the second case, as I argued in Chapter 5, classificatory adjustments may give rise to alterations in symptoms, experience, and behaviour. For instance, arguably shyness has become pathologized (Cooper 2010), and the same fate may await sadness due to the ICD-11 introduction of prolonged grief disorder.

In conclusion, the notion of culture-bound syndrome raises questions about the effect of cultural influence on the experience of some disorder, how it shapes some disorder, and how it can also play a constitutive role in some disorders. Therefore, I have proposed that a value-sensitive and pluralist explanatory approach to sociocultural causation of psychiatric kinds provides an ecumenical understanding of the sociocultural factors that are responsible for psychiatric kinds. As I have argued earlier, the applicability approach can be supplemented with the idea that the explanandum is not only specified with epistemic interests, it can also include considerations about value-sensitive decisions or whether to pathologize something, and where to draw the line. This implies that especially in unclear cases, the consideration of

the realness, universality, and boundedness of a purported culture-bound syndrome could include value-sensitive consideration about the causal consequences of those classifications.

6.2.5 Problems with Semantic Approaches to Psychiatric Kinds

Semantic arguments have been a common strategy when defending a biologically oriented view of psychiatric disorders in the light of cross-cultural, historical, and theoretical changes. Wakefield, for instance, deploys external semantics to argue that psychiatric disorders are natural kinds. He argues that folk-psychological concepts of psychiatric disorders are "essence place holders" so that kind membership is not determined by observable properties but by a hitherto unknown essential property that explains the observed features (Wakefield 1999: 471).[151] Wakefield uses the causal-historical theory of reference, according to which the reference of a term can be fixed ostensively or by a definite description of the underlying cause of the observed phenomena. Consequently, such a term's reference is not determined by definite descriptions or by concepts, but instead by external relations represented by a historical chain from the initial baptizing to the current user. Hence, successful theoretical terms can retain their reference across conceptual changes. In Wakefield's example, this would mean that although we are not aware what causes mental problems, we can still use and refer to them meaningfully. The implication is that future scientific research will discover what the true referents of our lay concepts of psychopathology are.

[151] Murphy (2014: 119) calls this a vindication project.

Wakefield's account has challenges concerning the nature and identification of psychiatric disorders. First, according to his earlier account, it was the dysfunction that (in part) determined whether a pattern of behaviour was a symptom of an underlying psychiatric disorder or not.[152] But in this revised view, the order of determination is reversed so that it is the folk-psychological conception of abnormality that determines whether a condition is a dysfunction. The reason is that this novel view implies that folk conceptions cannot fail to refer to real psychiatric disorders. That is, this approach does not seem to provide any argument about how to establish whether a reference has failed or not. This seems to imply that it is ultimately our folk psychology that determines which conditions are considered dysfunctional (see Fulford and Thorton 2007; see also Rashed and Bingham 2014). Consequently, although Wakefield argues that scientific research on the dysfunctions underlying "black boxes" would take precedence in determining mental problems, in practice, it seems that value judgments would determine where to look for, and more importantly, what conditions can be labelled as, dysfunctions. Hence, the harmful dysfunction distinction would collapse into being an account of harmfulness (see Aragona 2009). In a similar vein, Murphy and Woolfork (2000: 248) argue that Wakefield's account reverses the order of understanding the normal function of the mind. According to them, Wakefield's account has the absurd outcome that by detecting abnormal behaviour we can deduce the complete normal function of the human mind.

Hacking has also suggested a semantic resolution to separate socially induced instability of psychiatric kinds from their underlying biological nature. More specifically, Hacking (1999: 122) suggests

[152] According to Fulford and Thornton (2007: 161), Wakefield presents "dysfunction fact-side up, while all the time it is the hidden value-side that is doing the (logical) work". This is because he employs value-laden terms, like "failure", to determine dysfunction (cf. Aragona 2009).)

"putting a theory of reference alongside social construction" to distinguish neurological and social properties, which are subject to the looping effect brought about by classificatory practices. The idea is that the causal-historical theory of reference can be employed to explain why some psychiatric kind terms refer to underlying neurocognitive indifferent kinds, while the kinds also seemingly have interactive social properties. As one example, he argues that "autism" as a psychiatric label could refer to the disorder's indifferent neurological pathology, while the folk-psychological conception of the disorder, the stereotype, could be subject to the looping effect. Hacking (1999: 123) writes that kind terms "exhibit a looping effect, that is, they have to be revised because the people classified in a certain way change in response to being classified. On the other hand, some of these interactive kinds pick out genuine causal properties, biological kinds, which, like all indifferent kinds, are unaffected as kinds, by what we know about them." While Hacking writes that social properties cannot be subject to causal explanations, the passage can be given a more favourable reading by understanding "genuine causal" properties as a reference to the lack of the looping effect. According to this analysis, some properties employed to individuate autism are causally efficacious to the extent that they are not susceptible to the looping effect, while interactive social properties that are susceptible to the looping effect are captured by the stereotype.

Hacking's suggestion, nonetheless, draws an unfounded distinction between the social and biological properties of disorders (see Murphy 2006). This becomes clear when comparing extrinsic mechanisms that contribute and maintain many psychiatric disorders to the intrinsic mechanisms that underlie water. For example, the boiling temperature and liquidity of water varies at different altitudes. However, if we are looking for the common denominator among all the causes for the superficial manifestations of water (e.g. boiling temperature and

liquidity) in different background conditions, we are led to fixing the reference to the common cause, H_2O (cf. Kitcher & Stanford 2000: 113). That is, the microstructure H_2O is enough to explain the shared properties of individual samples of water. The reason is that although the microstructure produces different superficial properties in different circumstances, the core molecular properties do not vary. In contrast, when we want to explain depression or autism, there is no unique microstructure or some other type of essence underlying all the variations. Even in prototypical somatic diseases, such as cancer, no two cases are identical. This is exemplified by how two patients with the same disease may respond differently to the same drug (cf. Radden 2003: 45). Moreover, Murphy (2001: 154) points out that Hacking's semantic demarcation cannot account for the social shaping of our cognition. As I mentioned, in the light of cross-cultural research on cognition, we cannot exclude the possibility that culture can penetrate even low-level psychological processes (see Murphy 2014: 107). In sum, as Murphy (2006: 274) points out, Hacking's semantic analysis can only describe when we revise the use of our terms, not how social and biological properties intermingle.

Another problem is that historical and contemporary employment of theoretical concepts are open to different semantic interpretations. This is exemplified by how semantic externalism has been employed to defend a realist account of socially constructed kinds. For example, Mallon (2016) argues that an external semantic account of human kind terms explains how they can pick out socially constructed kinds instead of nothing. The idea is that external semantics can answer the "mismatch" problem between false beliefs and social kinds. Because causal-historical theories allow reference to be maintained through external relations, they can explain how individuals can speak of the same social kinds, such as sex and race, although they may associate false essentialist beliefs with them. Moreover, as I pointed out before, Mallon argues

that the false belief that the kinds have an essence stabilizes the behaviour picked out by the covert social category, which in turn means that the kinds can support generalizations and predictions. In this way, externalist semantic arguments can be employed both to defend essential and social constructionist interpretations of kinds. The consequence is that a causal-historical account cannot be employed to settle ontological questions about psychiatric kinds.

These semantic approaches are also problematic for a more principled reason. Because our current explanations of psychiatric disorders are insufficient, semantic arguments about their ontology are tempting. However, it has become clear that semantic arguments cannot be employed to make ontological conclusions (Bishop and Stich 1998). The reason is that intuitions over semantic interpretations of case studies vary cross-culturally (see e.g. Stich et al. 2004).

Likewise, the even more pragmatic employment of semantics to analyse disagreements is problematic. This can be exemplified by a strategy that employs semantics to demonstrate that philosophical controversies over ontology are in fact due to resolvable semantic commitments. In line with this, following Stich and Mallon (2000), Murphy (2015: 106) suggests a pragmatic solution to the problem of comparing different approaches to cross-cultural variation in psychiatric disorders. Stich and Mallon (2000) point out that mental state terms are comparable to theoretical terms because they are embedded in culturally determined folk-psychological theories. To demonstrate this, they attribute a holistic descriptive theory to Lutz (1998: 10), who argues that conceptions of emotion are culture-dependent to the extent that a competent speaker needs to be familiar with the local folk-psychological theory of the self and general social interactions. However, according to Stich and Mallon (2000), such a culture-related interpretation of emotion terms need not contradict (or be incommensurable) with a universal view of emotions based on

evolutionary psychology. Indeed, a causal-historical semantic interpretation of emotion terms can support a universalist stance on emotions. The reason is that a biologically oriented medical researcher can use a different description to refer to a universal biologically grounded emotion, while at the same time local conceptions may refer to culturally determined sub-samples of those emotions. This means that while a descriptive theory would support a culture-bound view of emotion terms, an inter-theoretical and universalist stand on emotions can be understood in the light of causal-historical theory. Following this pragmatist solution, Murphy claims that if we are interested in culturally determined conceptions and manifestations of disorders, we can adapt a descriptive approach to culturally determined folk-psychological theories of disorder terms. This, however, does not rule out a universalist account of culture-bound syndromes if the theoretical terms of psychiatric kinds employed by scientists are interpreted according to the causal-historical account to refer to underlying biological kinds.

However, Murphy's pragmatic solution relies on the empirical supposition that there are breakdowns in universal neurological processes to be found. But, as I have argued, currently there is no empirical evidence that this applies to all or even any psychiatric disorders. This implies that we are not warranted to assume that our brains become diseased in similar ways everywhere. Further, local descriptively determined terms may not pick out neat subgroups of the purported universal underlying kinds, but could instead postulate cross-cutting accounts of the syndromes, which would complicate Murphy's semantic solution.

My explanatory domain approach to psychiatric kinds supports, instead, a classificatory pluralistic resolution to cross-cultural comparisons. A good example is the culture-bound syndrome *latah* found in South East Asia, especially among older women in rural areas.

The syndrome's symptoms include losing one's self-control, when startled, by mimicking, cursing, and making vulgar gestures (Clifford 1898: 198). A cultural explanation for the syndrome can be that it is a ritualized role that permits individuals to violate the normal social structure (Lee 1981; cf. Winzeler 1995). This implies that there is a feedback loop between the cultural expectations associated with the role and the pattern of behaviour. On the other hand, Simons (1996) argues that *latah* is a culture-specific variation of a neurocognitive startle-matching syndrome that may include Tourette's and some other culture-specific conditions. In this light, social feedback mechanisms could explain the local manifestations of the neurocognitive kind. Moreover, *latah* (or some of its forms) could simultaneously be held as a unique interactive kind upheld by the interaction of the neurocognitive mechanism and the social mechanism (cf. Murphy 2006, p. 276). This means that whether a kind concept is split or lumped together depends on the underling mechanisms and the explanatory relevance set by the discipline-dependent characterization of the explanandum.[153] Indeed, when classifications have different epistemic aims, they may benefit from carving the property cluster kind differently (i.e. emphasizing different mechanisms), because explanatory dimensions may have trade-offs (Pöyhönen 2014). For instance, clinical practice may benefit from psychiatric disorder concepts associated with deep proximal explanations, whereas epidemiological approaches may prefer disorder concepts associated with explanations with wide scopes (see Campaner 2014: 99). In some cases, there may also be a trade-off between shallower explanations with wider time scopes and deeper explanations with limited time scopes.

[153] Psychiatric disorders with neurocognitive subtypes can be split or lumped together based on the same argument.

I conclude that semantic approaches cannot by themselves settle the nature of psychiatric disorders or culture-bound disorders. Nevertheless, conceptual and cross-cultural differences do not cast doubt on the fact that some culture-bound disorders are real psychiatric kinds, only on a universalist interpretation. Next, I analyse conceptual comparison as a problem for the advancement of psychiatric research.

6.3 Conceptual Change and Psychiatric Progress

Contemporary and historical conceptual variation in scientific research challenges realist interpretations of scientific kinds, including realism over psychiatric kinds. As I have mentioned, realists typically argue that science progresses by describing better the natural kinds structure of the world. In other words, newer theories can offer a truer (or more truth-like) and more objective picture of the world by "cutting the world at its joints" (see e.g. Psillos 1999). On the other hand, if scientific concepts radically differ between historical contexts and competing contemporary theories, they may not be talking about the same kinds. This would hinder or rule out scientific progress because there would be no accumulation of knowledge about those things. In particular, Kuhn has argued that radical theoretical changes or paradigm shifts involve incommensurability. Arguably, incommensurability occurs when the meanings of theoretical terms employed by theories change in paradigm shifts to such a degree that there is no common measure to compare them. At the extreme, Kuhn interprets this to imply that competing taxonomic structures "create their own worlds" instead of describing independently existing kinds (Kuhn 1996:150).

Conceptual change in psychiatry raises the question of incommensurability between contemporary, historical, and cultural conceptions of psychiatric disorders. Horwitz (2002: 56-57), for example, argues

that the change from the psychodynamic to the bio-medically oriented medical model of psychiatry exemplifies a shift between incommensurable paradigms. Whereas the psychodynamic approach holds that mental problems are indeterminate manifestations of underlying unconscious mental mechanisms, the strong view of the medical model considers them to be distinct biological kinds with overt symptoms. This difference is demonstrated by Luhrmann (2001: 22), who argues that in the early 1990s, different paradigms in the USA led to concentrating together different diagnostic and treatment tasks (see also Cooper 2005). While psychoanalyst-oriented students learn to listen, and concentrate on individual life histories, biologically oriented students are taught to categorize patients based on diagnostic categories. Similarly, Lakoff (2005) documents how at a hospital in Buenos Aires, a shift from the psychoanalytical explanatory approach to the biomedically oriented one led to completely different perceptions about how to recognize and treat psychiatric disorders. Cooper (2007, Ch. 6) points out that incommensurability in psychiatry does not only concern historically succeeding theoretical approaches, but also that mutually incompatible explanatory theories of psychiatric disorders currently exist. This is further complicated by the fact that although the DSM and ICD diagnostic manuals are purportedly theory free, they are in practice filled with unspecified phenomenological and clinical language (cf. Murphy 2006). In a more specific historical case analysis, Radden (2003) argues that "melancholia" and "depression" are not co-extensive. This makes direct comparisons between their respective conceptions and explanations difficult.[154]

Problems of comparison or even incommensurability can similarly occur when moving from a descriptive to a causal conception of psy-

[154] Horwitz and Wakefield (2007) and Jackson (1986) maintain that melancholia and depression refer to the same disorder (see also Zachar 2014: 138)

chiatric kinds. Maung (2016) argues that a taxonomic change from syndromes to underlying causes may require lumping or splitting syndromes in cross-cutting way. Nevertheless, a realist could hold that the intention of a causal conception is not to build on former knowledge about psychiatric kinds. Rather, the purpose is to build a new nosology based on novel empirical knowledge of underlying causes. Therefore, incommensurability between old and new classifications is inconsequential to the advancement of psychiatry. This interpretation would mean that psychiatry is following in the footsteps of somatic medicine. Thagard (2000:15), for instance, points out that humoral and germ conceptions of disease are mutually untranslatable because Hippocratic classification divides up the world based on bodily locations, while the germ model is based on microbial causes. Nevertheless, partial comparison is possible because the symptom clusters of (some) diseases have stayed the same, and most importantly, a full comparison is unnecessary in the light of the undeniable pragmatic success of the germ theory.

These problems of comparison across historical and contemporary changes may seem to support general constructivist beliefs about the nature of psychiatric disorders.[155] That is, one could argue that the different ways disorders have been historically and culturally defined and classified implies that disorders themselves are relative to historical and cultural contexts. The problem with this argument is that it presupposes that theoretical or folk-psychological descriptions associated with the concept determine its referent. This argument implicitly underlies Kuhn's view of incommensurability. Kuhn implicitly employs descriptivism, according to which two theories cannot be about the same objects if they associate contradictory descriptions with their central theoretical terms, and consequently, there is no growth of

[155] Alternatively, this could be understood as a pessimistic induction (Kendler 2016).

knowledge about the referents.[156] The historical answer to the presupposed antirealist implications of incommensurability by scientific realists was to employ external semantics. However, conceptual change does not, by itself, suffice to draw strong ontological conclusions. This is because historical examples can be given alternative semantic interpretations. All that an interpretation based on semantic externalism can demonstrate is that referential continuity is possible despite conceptual and theoretical changes.

My argument is that the contrastive-counterfactual theory of explanations can help to address antirealist interpretations of the incommensurability thesis. The question whether psychiatric categories in alternative theories refer to the same kinds is too ambiguous to be given a commensurable answer (cf. Chang 2017, cf. Cooper 2020). Instead, by making the explanans and explanandum more explicit by means of contrasts, more constructivist views of incommensurability can be overcome. Hence, instead of asking whether "melancholia" was real, or whether "depression" refers to the same kind as "melancholia", we should ask whether certain causes, such as specific social factors (rather than others), brought about similar symptoms, such as sadness in individuals with certain social identities (rather than others). As Chang (2017) argues, conceptual change in psychiatry can be understood as an iterative process that is compatible with pluralistic realism. The idea of epistemic iteration is to "get on" in the absence of assured foundations (Chang 2017: 231). Chang (2017) questions the assumption that there is just one right answer to taxonomic questions, including psychiatric classification. According to Chang "successive stages of knowledge, *each building on the preceding one*, are created in order to enhance the achievement of certain epistemic goals." (Chang

[156] Kuhn's later taxonomic incommensurability does not threaten a pluralist interpretation of psychiatric kinds.

2004: 45, emphasis original).This pluralism encourages researching for as many classificatory answers as possible.[157]

Finally, a comprehensive incommensurabilty in the special sciences in general, and in psychiatry in particular, is not plausible because their language is not "total" and hence comparisons can be conducted in natural language (cf. Sankey 1994). For example, although the concepts of "melancholia" and "depression" may be theory-laden, we can nonetheless talk about and compare their symptoms in natural language. In addition, there is no reason why scientists could not be "bilingual" by being able to comprehend competing concepts and theories (Godfrey-Smith 2021). This means that these symptoms, perhaps excluding phenomenological self-reports, are empirically directly comparable. In sum, although there may be some incommensurability or problems of comparisons in psychiatry, by itself incommensurability does not support an antirealist interpretation of psychiatric kinds. Nevertheless, conceptual variation in the social sciences, including psychiatry, is more complicated than in (most) of the natural sciences because of social interaction.

[157] Poland (2014) argues that initial success in DSM has led to dogmatic outcomes.

7 Conclusions

I have argued that psychiatric disorders can, in principle, be scientific kinds that support sound inductive inferences, and thereby ground scientific classification, research, and clinical practice. I reached this conclusion by addressing two research questions. My first question was what kinds of scientific objects are psychiatric disorders, and is it possible to formulate a general theory of them? My second question was what implications does my theory of psychiatric disorders have for classificatory projects and explanations as well as for treatment interventions and policy decisions?

I addressed the first question by examining whether the concept of psychiatric disorder can be given a definition based on scientific facts, and whether particular psychiatric disorders can in principle be scientific or real kinds that support sound inductive inferences. I argued that while naturalist accounts based on conceptual analysis cannot naturalize psychiatric disorders as dysfunctions, normativist views contrast our intuition that psychiatric disorders are non-relative. In addition, I pointed out that such accounts largely rely on a top-down approach: by means of conceptual analysis they try to come up with ways to define psychiatric disorders in their totality. Moreover, I argued that these accounts have a problem with scientific advancement. Settling on a strictly defined concept could constrain research and clinical practice. Consequently, my answer to the question is that a general theory of psychiatric disorder based only on scientific facts is currently unlikely to be had and could even be scientifically counterproductive. Instead, I defended a bottom-up account of the concept of psychiatric disorder according to which we do not only need empirical research on the putative disorders themselves, but also on the non-epistemic values underlying their conceptions and classifications, as well as on the social consequences of those classifications. In sum, I

argued that an open-ended concept can facilitate scientific, social, technological, and moral development. The general idea then is that of a co-fitting process that considers both what we want with the concept of psychiatric disorder as well as what can be discovered.

I then examined whether an account of particular psychiatric disorders as scientific kinds can be formulated. I argued against essentialist views of natural kinds, and instead pointed out the benefits of the HPC view of natural kinds. More specifically, I argued that although the kinds studied by the special sciences do not have necessary and micro-structural conditions for their membership, these kinds can nonetheless be stable enough to ground inductions and that a posteriori research can be conducted to find their partly externally determined properties. I based my argument on the shift from considering explanations to be based on laws of nature and essences, to causal and mechanistic explanations. However, I discussed the HPC view's specification problem according to which it is unclear whether the property cluster or the underlying mechanism determines the kind's boundaries. I argued that this challenge cannot be overcome by concentrating solely on explicating the mechanisms that are responsible for the property clusters. Rather, I argued that the problems of the HPC view stem from a confusion over the division between explanatory labour of mechanisms and property clusters. As an example, I argued that Craver et al.'s (2011) mechanistic account of psychiatric disorders is problematic in this sense. Based on this problem, I discussed some ecumenical alternatives to scientific kinds, but argued that they neither can explain why inductions are warranted or why only certain kinds of inductions are warranted.

As an answer to the specification problem, I relied on the contrastive counterfactual theory's account of mechanistic explanation and applied it to the homeostatic mechanisms of property clusters. I argued that the contrastive counterfactual and interventionist theory offers a non-reductive and domain-relative approach to understanding

kind explanations. That is, the contrastive counterfactual theory is metaphysically non-committal and is thereby suitable to describe the various causal factors that are responsible for the properties of psychiatric kinds. Based on these notions, I formulated my own account according to which psychiatric kind explanations have applicability domains over which they can reliably explain aspects of explananda kinds. More specifically, I asserted that knowledge of various mechanisms and causes can supplement the domain of applicability over which a psychiatric kind and its variations are reliably explainable and warrant inductions. Simply put, specifying the applicability domain of an explanation spells out the conditions under which the explanation can be expected to be reliable and when it can fail. Moreover, following others, I argued that the contrasts can be employed to determine whether competing research approaches can supplement each other or whether they delineate kinds differently in the light of their explanatory and value-laden aims. Based on this, I showed how different research approaches, such as clinical and epidemiological, can stress contrastive reliability-domains that licence re-classifications of disorders. Nevertheless, I argued that my approach shifts the focus of the HPC view of natural kinds from the ontological question about their mechanistic individuation to the epistemic relevance of their specification. In addition, I argued that the contrastive counterfactual theory of explanation can help to disentangle causal and constitutive explanatory information in psychiatric explanation, while the interventionist theory suggests researching for modular causal structures. In then compared my account to various pluralistic explanatory approaches in psychiatry and argued that my approach is superior because of its non-metaphysical nature. The reason is that it can facilitate hybrid explanations that link heterogeneous causal factors to explain psychiatric kinds.

My second question was what implications does my theory of psychiatric disorders have for classificatory projects and explanations as well as treatment interventions and policy decisions? I addressed the question by investigating whether the effects of classificatory practices and sociocultural causes can be explained by my domain-relative theory of psychiatric kinds. More specifically, I employed my applicability domain account to explain classificatory instability, plurality in disciplinary approaches, and cross-cultural variation of psychiatric disorders. I began by arguing against Ian Hacking's account of interactive kinds, according to which human kinds require explicit conceptualizations and intentional reactions. According to Hacking, some classificatory projects in the human sciences are subject to the looping effect: classifications have to be constantly revised because of the intentional reactions of the classified people. Arguably, this instability renders the affected human kinds interactive and hence different from the natural kinds classified by the natural sciences. Against this view, I asserted that feedback explanations that mediate the looping effect can supplement the domains of applicability over which the kinds sustain epistemic projects. A feedback explanation's explanatory relevance depends on its ability to widen and deepen ways of explaining a human kind, in other words enhancing the domain of applicability. By applying this approach to empirical case studies, I demonstrated that congruent mechanisms can supplement, in various ways and degrees, the applicability domain of interactive human kinds, while incongruent mechanisms can help to identify why the domain is unstable. However, because human categorizations and their effects are constantly reproduced in social interactions, the epistemic projects that human kinds support are in general not as stable as the projects that prototypical natural kinds support. The underlying idea, nevertheless, is that whether an explanation of a human kind is sufficiently stable depends on the discipline-relative epistemic aims set for the explanatory domain.

Finally, I compared my approach with the medical model's ability to account for explanatory plurality and cross-cultural variation. Unlike the medical model, I argued that the applicability account does not necessarily require a demanding metaphysical integrative approach of various discipline-related explanations. Rather, it can facilitate heuristic interaction between various explanatory approaches to a psychiatric kind. I then demonstrated that the medical model, both the pragmatic approach and the stronger biological version, have problems in explaining sociocultural causation. In contrast, I asserted that my applicability domain approach has the advantage of being able to encompass heterogeneous causal factors to explain cross-cultural variation, and in some cases, to understand the social dynamics that constitute psychiatric problems. In particular, I argued that mental disorders cannot be identified without taking into account the sociocultural context in which they originate, develop, and are experienced. In short, I argued that psychiatric disorders cannot be identified and studied as "malfunctioning brains in a vat". Moreover, in some cases, explanatory power is increased by specifying the explanandum to cover larger social dynamics rather than the targeted individual's psychology or brain. Therefore, according to my account, cross-cultural variation should not be seen as an argument against realist interpretation of psychiatric kinds, but rather as a naturalist invitation to study social causation, knowledge of which can supplement kind explanations.

Consequently, an important methodological implication of my account is that classificatory and diagnostic practices should pay attention to the mechanisms that underlie the dynamics of interactive kinds such as psychiatric disorders. Feedback explanations may not only contribute to explaining some aspects of kinds and help anticipating when a classificatory adjustment is needed, they can also potentially facilitate conceptual engineering and thereby enable kind amendments. Just as criminology is used in assessing criminal policy measures,

feedback explanations could be implemented to anticipate and mitigate the negative effects of diagnostic practices.

Next, I want to reflect on what implications my applicability approach, and the conclusions I have reached this far, have in relation to conceptual engineering approach to scientific concepts. The application of conceptual engineering to psychiatric disorders is a novel approach and serves to emphasize and deepen my conclusions and provide a means to analyze what implications they may have for future research.

7.1 Towards Ameliorative Conceptual Engineering in Psychiatry

I have argued that while the superordinate concept of psychiatric disorder is value-laden, particular kinds of psychopathology can be identified and individuated based on their explanatory applicability. I want now to combine the value-sensitive approach to conceptualization with the domain-relative explanatory approach to psychiatric kinds, and explore what roles non-epistemic values play in how the classifications of particular kinds of psychopathology are determined. I argue that the choice made between alternative kind individuations is neither possible nor preferable based solely on epistemic values. Rather, I argue that the choice concerning which alternative explanatory model to employ, and thereby also how to individuate the boundaries of a psychiatric kind, and whether to pathologize certain borderline conditions at all, should be done in a value-sensitive fashion. In particular, since the conceptualization and classification of psychiatric kinds cannot only be based on empirical discoveries, they should also be informed by evaluating the possible social consequences of those classifications.

7.1.1 Conceptual Engineering in Psychiatry

I have argued that traditional descriptive conceptual analysis cannot settle the nature of psychopathology (see Chapter 2). For example, according to recent studies people have an implicit and conflicting conception of psychiatric disorder (e.g. Tikkinen et al. 2019, Harland et al. 2009: 967, Colombo et al. 2003). Moreover, even if there were a widely-shared concept, it would not necessarily correspond with empirical discoveries, on the one hand, and with how we may want to employ it in the future, on the other. However, an alternative approach to the concept of psychiatric disorders is offered by conceptual engineering (Schuur 2019). In general, proponents of conceptual engineering suggest revising concepts, or proposing novel ones, so that they can better accommodate either epistemic or non-epistemic goals (e.g. Eklund 2015, Cappelan 2018, Brigandt 2020). Conceptual engineering based on epistemic goals relies on Carnap's (1950) and Quine's (1960) work on conceptual explication, which aims to make concepts more exact and formal. An alternative approach which relies on the ameliorative engineering of concepts based on non-epistemic values has been developed especially by Sally Haslanger (2012). She argues that some social kind concepts, such as "race" and "gender", should be primarily engineered to accommodate morally and politically motivated goals.

Up until now, conceptual engineering has primarily been employed to explicate the superordinate concept of psychiatric disorder (e.g. Schuur 2019, Lemoine 2013, Aucouturier and Demazeux 2013, Schwartz 2014, Griffiths and Matthewson 2018). The explication project aims to make the concept more precise in order to serve scientific goals as well as lay and professional intuitions better. The idea is to identify and justify the epistemic goals set for the concept, and thereby to offer a more precise and scientific definition that serves those goals better. Along these lines, Schwartz (2014), for example, defends

Boorse's account, while Griffiths and Matthewson (2018) defend Wakefield's account. Their arguments are that the accounts (BST and hybrid account, respectively) are empirically grounded, and thereby should be employed to revise the concept of psychiatric disorder, instead of intuitions being employed to criticize the accounts. Schwartz (2014), furthermore, identifies "disease" and "health" as serving roles in biological generalizations as well as in the ethical, practical, and policy question of clinical medicine. According to Schwartz, the concept is presently too vague and unscientific to serve those goals and should therefore be refined. Crucially, he maintains that the concept of "disease" may be defined differently to serve different goals. In this case, as Schwartz (2014: 574) points out, the question is not which definition is correct, but rather which one ought to be adapted. However, the idea that psychiatric concepts need to be refined for different purposes has not been widely applied in psychiatry yet.

A value-driven approach to refining psychiatric concepts could rely on Haslanger's ameliorative approach. Haslanger has defended an ameliorative approach to refine concepts based on the value-laden goals they serve. According to Haslanger's (2012: 149) ameliorative approach, there are no unique social kinds, but instead multiple *objective types* of ways to classify and categorize the social world. Haslanger's (2012: 224) ameliorative strategy is at its core stipulative: "the world itself can't tell us what gender is, or what race; it is up to us to decide what in the world, if anything, they are." Consequently, scientists have a moral responsibility in their taxonomic choices, and objective types can and should be constructed to promote value-laden goals, such as feminism and anti-racism (see Godman and Marchionni 2022). The problem with this view is, as Bach (2016, 2019) argues, that settling on objective types instead of trying to construct classification based on real or natural kinds may not be empirically valid. This would be counter to the whole idea of being able to alter the type of kinds since value-laden

interventions would not be as effective. In the light of this, and my realistic account of psychiatric kinds, a full-blown ameliorative approach in psychiatry would not only be empirically invalid, but also morally unsustainable and socially detrimental.

Therefore, a crucial question in conceptual engineering of psychiatric concepts is the role of non-epistemic values. While I agree that the explication approach is and should be important in engineering concepts so that they pick out scientifically relevant kinds, I also believe that non-epistemic interests play an ineliminable role in the process, and therefore should be identified, refined, and exploited (see also Aftab and Rashed 2021, Gagné-Julien 2021). In the following, I suggest that although ameliorative approaches cannot stipulate or impose how to model psychiatric kinds, they can nonetheless play an explicit and necessary role in informing specification and taxonomic choices.

Thus far I have argued that a strict distinction between values and scientific facts in conceptualizing psychiatric disorder cannot be maintained. I now wish to elaborate the idea that the value-laden status of the superordinate concept of psychiatric disorder calls into question the possibility of demarcating particular psychiatric kinds merely based on epistemic interests. While I agree that the explication approach is crucial in engineering concepts so that they pick out scientifically relevant kinds, I also believe that non-epistemic interests play an ineliminable role in the process. This argument is based on the semantic and methodological criticism of the value-free ideal of science. The semantic criticism, particularly endorsed by Putnam (2002) and Dupré (2007, 2012), is that the distinction between facts and values in science is unattainable due to thick normative concepts. This means that concepts cannot be provided a value-free definition, and normative implications are built-in. Based on this, it has been argued that the concept of psychiatric disorder is a thick normative concept (Keil & Stoecker 2017,

Bueter 2019a, Gagné-Julien 2021, Dupre 2007). Methodological criticism builds on Rudner's (1953) work, according to whom methodological choices are partly non-epistemic, and on the Duhem-Quine thesis, according to which theories are underdetermined by evidence. Based on historical example, Cooper (2007: 132), for one, argues that when theoretical interpretation of psychiatric data is underdetermined, researchers tend to adjust their results based on value-laden prior beliefs and expectations.

Non-epistemic values that guide research and classificatory choices in psychiatry can be pragmatic and ethical. In general, values motivate decisions by invoking desirable goals (Schwartz 2014). Sadler (2013: 761) argues that psychiatry is guided by moral and ethical values in (1) the process of formulating classification manuals (2) definitions of diagnostic criteria, and (3) in the employment of these systems. Ideally, psychiatry and clinical practice should be governed by "good" goals, such as the desire to help, heal, care, and cure (Sadler 2015). Ideally, the methods of psychiatric classification and research should be unbiased to discover scientific kinds instead of imposing them. In the light of what I have argued, this is an unrealistic goal, at least for now. Rather, currently the problem is that non-epistemic values play an implicit and unspecified role in definitions, classifications, and explanations. As Sadler points out, the value-free goal of psychiatry causes "toxic denial of naïve to bad political processes that are nevertheless powerful, producing deviant knowledge, compromised categories, and poor practices".[158] Cooper (2007) and Gagne-Julien (2021), for example, point out that biased racial and gender values influence all stages of psychiatric research and classification.

The descriptive approach of the DSM and ICD classification manuals makes them especially susceptible to biased and non-transparent

[158] I have argued that causal knowledge and cross-cultural research are both needed to overcome some of these biases.

value-driven refinements. Haslam (2016) argues that the extension of the concept of psychiatric disorder in the DSM has expanded both horizontally and vertically from 1952 to the present. While horizontal expansion describes how the definitions have expanded to cover qualitatively new phenomena or conditions that were not included before, vertical expansion describes how the extant definitions have become more qualitatively inclusive. Horizontal expansion is represented by the fact that in 1952 DSM-I contained approximately 100 psychiatric disorders, while the latest version DSM-5 has close to 300. As one example, Cooper (2015) points out that the DSM-5 definition refinements have not been openly scrutinized. Unlike in the previous editions, the fifth edition does not require symptoms to cause distress or impairment before a disorder can be diagnosed. The harmful criterion has been dropped on the assumption of biological psychiatry. Vertical expansion is exemplified by Horowitz and Wakefield's critique. They argue that normal sadness has become labelled as depression because of the descriptive approach that ignores the context and causes of symptoms. For example, in DSM-5, bereavement is no longer excluded from depression diagnoses. In sum, the descriptive and biological approach of the DSM diagnostic manual has enabled its explanation.

The expansion of the concept of psychiatric disorder in the DSM has wide ranging consequences. The concept not only influences which conditions are classified as diseases, but also where to look for them, and how to demarcate them. For instance, according to the Epidemiological Catchment Area (ECA) study of the early 1980s, one-third of the USA population between the ages of 18 and fifty-four years has a diagnosable mental disorder. In the NCS-R study in 2013, that figure had risen to approximately half of the population. The rise is arguably in part due to the extension of the concept of the DSM. The definition of mental disorder also influences psychiatric research. Cuthbert (2012)

analysed that over 90% of articles in *The American Journal of Psychiatry*, *Biological Psychiatry*, and the *Archives of General Psychiatry* concentrated on a single DSM diagnostic category (see Tabb 2019). This is also a problem for the RDoC since the researchers involved still rely on the DSM diagnostic categories (Cockerham 2017: 33). Cooper (2015: 94) points out that not only does it influence epidemiological studies and clinical treatments, it also shapes lay perceptions of the normal and the pathological.

Some of these problems can be overcome by making explicit and striving for objectivity concerning the values that influence methodological choices and conceptual development (see Dupré 2015, Longino 1990, Cooper 2007). Bueter (2019a: 494) and Gagné-Julien (2020), for instance, argue that those social values can and should be objectively managed in psychiatry. A similar approach to conceptual engineering would mean that the values that guide the conceptual formation of psychiatric disorders should be made explicit and constrained. Hence, although ameliorative approaches cannot stipulate or impose psychiatric conceptions and classifications of psychiatric kinds, they can nonetheless strive to play an objective role in informing between labelling and taxonomic choices between epistemically equally (or relevantly equal) valid alternatives. The upshot is that instead of considering the lack of unified explanatory paradigm in psychiatry as a problem to overcome, explanatory and classificatory pluralism should be considered the correct means to account for the ontological pluralism of psychopathology. Just as there are multiple empirically legitimate ways to classify species, there are probably multiple empirically adequate and cross-cutting ways to classify and model psychiatric kinds.

Next, I will analyse in more detail the role that non-epistemic values should play in conceptual engineering and classificatory choices, rather than just how to avoid implicit biases and pernicious value choices. I wish to explore especially the role that non-epistemic values

can play in classifying and demarcating particular kinds. The argument is that because epistemic values are no sufficient in themselves to dictate choices over classification of biological kinds, the choice concerning how to demarcate particular psychiatric disorder is ultimately a value-laden question as well. I assert that the conception of psychiatric disorder serves an inextricable role in choosing between different ways of specifying particular conditions, and that conceptual engineering should also be considered an ameliorative process because it has social consequences for the people concerned.

7.1.2 Consequence-Sensitive Classification

I have argued that the applicability domain account can help to answer challenges related to the symptom-based approach to classifying psychiatric kinds. This is because the account is not merely based on identifying manifest properties, it is also based on domain-relative mechanistic explanations. Nevertheless, the identification of property clusters requires an explicitly interactive process between value-sensitive aims of the specification of the kind and empirical research. While empirical discovery of mechanisms may require us to revise existing concepts, value-sensitive choices should guide us on whether and how we want to consult these explanatory mechanisms. In particular, there are certain cross-cutting ways to specify psychiatric kinds that require value-sensitive choices.

According to Murphy and Stich (1999), for instance, it is likely that the current diagnostic categories will be either torn apart or eliminated. Following Griffiths's (1997) horizontal and vertical arguments over the vernacular concept of emotion, they suggest that the superordinate concept of psychiatric disorder should be eliminated or allocated a new openly value-laden role, while concepts about particular psychiatric

disorders may have to be taken apart in order to track natural kinds at different levels. The horizontal argument has it that a concept that does not track a natural kind will be eliminated, while according to the vertical argument, a concept may have to be taken to pieces according to different explanatory goals. Although Stich and Murphy (1999) argue that the superordinate concept of psychiatric disorder does not pick out a natural kind, they maintain that it nonetheless serves a normative role in unifying "all the ways that our mind can go wrong", which counteracts its non-scientific character. This leads them to suspect that the concept will not be eliminated. Their vertical argument is that many subordinate concepts track property clusters at different levels, and thereby the concepts may have to be torn apart. For example, the subordinate concept of "depression" tracks several property clusters at different levels of resolution. Depression is used in behavioural generalizations, psychological generalizations, computational generalizations, and chemical generalizations. Although the concept supports all these generalizations, their property clusters do not correspond to form a unified kind. Instead, according to the evolutionary approach, if we concentrate on the cognitive level of description, only humans can be depressed, while the other generalizations would lead us to argue that apes can be depressed as well. Consequently, Murphy and Stich (1999) predict that future science may pull the "depression" concept apart.

While I agree with Murphy and Stich that the superordinate concept of mental disorder does not seem to pick out a natural or real kind, I also believe that the status of psychopathology plays an inexplicable role in how the particular concepts of disorders are delineated. As I have argued, specifying a property cluster depends on what we want to do with the resulting model, not simply on the causal structure of the world (cf. Craver 2009, Zachar 2014a: 95). This raises the question of how and what interests influence the way in which kinds are split and

lumped. While I believe that individuation can be done in part by relying on explanatory norms represented by the applicability domain of explanation, drawing the boundaries of kinds is also guided by non-epistemic interests. When classifying psychopathology, we do not only have to decide which classification best matches our epistemic goals that are related to, for instance, epidemiology or proximal clinical practice. Rather, because the concept of psychiatric disorder is value-laden, we also need to weigh the ethical and practical goals of the definition and classification, especially in borderline cases where intuitions are unclear.

Recently there has been extensive discussion over the role of non-epistemic values in the classification of natural kinds (e.g. Reydon and Ereshefsky 2022, Ludwig 2016, Bach 2019, Godman and Marchionni 2022). These accounts share a conviction that classificatory choices cannot be made merely based on epistemic goals. Rather, moral and pragmatic considerations influence theories or models of kinds and thereby affect their classifications (Reydon and Ereshefsky 2022). The general motivation is that because there are so many things that could be investigated, science is guided by context-dependent practical interests (Kitcher 2001: 61). These interests have mainly been in biology as well as in the social sciences. Reydon and Ereshefsky (2022), for example, argue that the widely used phylogenetic species concept may not satisfy the ethical goals of preserving biological diversity in conservation biology. While according to the The Phylogenetic Species Concept (PSC) "A species is the smallest diagnosable cluster of individual organisms within which there is a parental pattern of ancestry and descent" (Cracraft, 1983: 170). The problem for the PSC is taxonomic inflation: it may yield up to 48% more species. Preserving so many species is pragmatically not possible, and may also demotivate the general population. Consequently, conservatory biologists prefer

the biological species concept, which is based on interbreeding populations reproductively isolated from other such groups, and thereby does not classify as many species. The gist of the argument is that the choice between classificatory systems is ultimately guided by non-epistemic aims, the ethical aim of trying to preserve biological diversity, and the pragmatic aim of what is in practice possible.

Likewise, the classificatory choices concerning psychiatric disorders are partly value-laden. We need, for example, to decide whether we want to impose our conceptions of pathology onto other cultures. That is, whether some cultural ways of managing grief, for instance, should be considered more valuable than the possible benefits. For example, the introduction of the concept of prolonged grief in the new ICD-11 modelled institutional help in Western countries. Classificatory considerations could also involve questions about how to refrain from stigmatization in a culture-sensitive way, and how to secure the right for treatment with limited funding available in much of the global south. This approach can be compared to a similar argument by Brigandt (2020). Following Potter (2016), Brigandt points out that oppositional defiant syndrome is disproportionately diagnosed in African-American boys. Potter (2016) points out that the concept picks out behaviour that is caused by social inequalities and institutional racism, and thereby should be re-examined and refined based on considerations of social justice. There are also pragmatic questions involved. Although the diagnostic categories in the DSM may not represent correctly all or most psychiatric disorders, they nonetheless enable (some) treatments as well as facilitate discussions between various researchers and research programmes across the globe. Moreover, they can channel help, albeit often by being inconsiderate of individual needs, within the abstract bureaucratic means available.

There are also value-sensitive decisions concerning the demarcation of kinds that need to be made in the research process itself. Ludwig

(2016) argues that data collection requires non-epistemic choices about the boundaries of kinds for both practical and logical reasons. He demonstrates this with research on domestic violence, which involves decisions about demarcating violence as well as choices about gender and race definitions. The point is that it is impossible to investigate all the logically possible variables related to domestic violence. Instead, research on domestic violence needs to make pragmatic and ethical choices, such as whether to make a distinction between minorities and how to define genders. Most likely similar data choices are also made in psychiatric research. Research on the prevalence of depression among minorities in the USA, for instance, would need to rely both on a definition of depression and on a definition of those minorities.

However, some of the value-laden options can be empirically disproved because of their false descriptive content. Risjord (2007: 43) has demonstrated how empirical investigation can disprove values because theories with thick concepts make values theory-laden. For example, the concept of "higher civilization" was descriptively entangled with social evolutionary theory, and once the theory was disproven, the values inherent in the concept lost their grounding. The same reason has contributed to homosexuality losing its status as a psychiatric disorder and the hypothetical condition drapetomania never gaining support. Drapetomania was imbued with racism and the theory of social evolution, while homosexuality was based on the mistaken conception that homosexuality is somehow "unnatural" (e.g. that it is non-existent among non-human species). Consequently, once these underlying empirical theories were disproven, the value-laden concepts lost their explanatory support as well.

In sum, normative judgements over taxonomic choices seem to be inevitable because the *ontological pluralism* of psychopathology may warrant multiple adequate explanatory models and classification. Consequently, it seems to be neither possible nor preferable to place the

choice between different conceptions of psychiatric kinds solely on epistemic values. In consequence, this implies that there is no one true unified classification that will emerge from psychiatric research, but instead there are different classifications for different partly value-driven purposes.

Thus far I have argued that the specification of particular psychiatric kinds should explicitly involve value-informative choices about their social consequences. Some of those consequences are direct. They include, for instance, statistical decisions concerning prevalence, policy decisions about preventive and treatment management, and most importantly, choices concerning who has the right for treatment (e.g. Cooper 2015, Haslam 2016). Brigandt (2020), moreover, argues that drawing the line with diagnostic categories should be evaluated based on balancing the right for treatment with the risks of stigmatization and reinforcing existing social inequalities.[159] Next, I analyse the connection between value-sensitivity and explanatory pluralism, and its implications in the light of the indirect consequences of classifying psychiatric kinds caused by the feedback effect.

7.1.3 Value-Sensitivity and Explanatory Pluralism

I have argued that a unified picture of psychopathology is unlikely to be gained in the light of current explanatory pluralism in psychiatry. I have also pointed out that this should not be seen as an epistemic problem to overcome, but rather that it matches the complex and dynamic nature of psychiatric disorders as social and natural phenomena. Based on this, I argued for value-sensitive classificatory pluralism. If we admit that there are different ways to classify psychiatric disorders, we also

[159] Questions have also been raised about epistemic injustice concerning different classifications. See Bueter (2019b).

need to admit the irreversible role that value judgements and ethical considerations play in those classificatory decisions. I believe that by acknowledging this, psychiatric research and classification could facilitate progress with greater attention to justice.

Such an explicit value-sensitive approach would need to consider all stages of psychiatric research and classification. It is important to understand that the classification is on ongoing process, and being clear about the value-laden aims in each stage of the process allows one to adjust the process according to the changing aims (see Chang 2017: 242). The process could consist of research and weighing of the non-epistemic reasons and causes for considering that some disorders are pathological (e.g. ensuring the right for treatment, severity of harmfulness, weighing evidence for the diagnostic category) and critically assessing them based on their consequences (do they inflict stigma, alter indigenous cultural conceptions of suffering, etc.). Moreover, the process could draw from research carried out in anthropology, psychology, philosophy, sociology, and other fields.

In particular, I have argued that knowledge of the feedback mechanisms that mediate the looping effect can supplement the domain of applicability over which a psychiatric kind and its variations are reliably explainable. Here my aim is to extend the argument so that the feedback mechanisms should not only be identified to supplement the explanations of psychiatric kinds, but also to make value-sensitive judgements on which specifications and classifications should be employed based on their anticipated social outcomes. My argument is that knowledge of the feedback mechanisms that mediate looping effects can facilitate judgements concerning value-sensitive engineering of concepts and classifications of psychiatric kinds.

As an example, the biomedical model, when applied uniformly too all disorders, may have negative health consequences. Currently, the problem is that the neurobiological medical model is baselessly applied,

or conceived to be applied by lay people, to all psychiatric disorders. As I pointed out, overemphasizing the biomedical model of schizophrenia may have negative consequences on patients because of the stigma associated with its immutability. In some cases, the question concerns which properties should be included as part of the pathology. I argued in Chapter 5 that in the conceptualization of schizophrenia it might be best not to emphasize phenomenological factors but instead concentrate on behaviour. Although schizophrenia can be construed to include auditory experiences, de-emphasizing the pathological nature of these experiences may contribute to the well-being of the patients (see Luhrmann and Marrow 2016). Finally, in cases where our intuitions are unclear, clearer value-weighing could be applied. This is the case with pathologization of prolonged grief in the latest ICD-11. Although the common argument is that the definition guarantees the right for treatment, studies have shown that a mere categorization of a condition as a disorder may also cause negative health outcomes (Mallon 2016: 87, Kaplan 2010).

Anthropologists have long argued that knowledge of cultural meanings associated with mental disorders, i.e. idioms of distress, should inform diagnostic practices and classifications to avoid misidentifications and inadvertent outcomes (cf. Kirmayer 2001, Kleinman 1988). This is because disorders cannot be uprooted from their cultural context and studied in isolation. Rather, they receive their form and causal powers in social contexts. This also means that classificatory practices may have inadvertent causal consequences for these contexts. For example, the introduction of depression in Japan pathologized previously revered feelings and behaviour (Kirmayer 2002, cf. Kitanaka 2012). Although normalizations can reduce stigma (Haslam 2016), biologization can also have negative consequences when patients become pessimistic about recovery and the ability to control their condition (see also Rosenberg 2002). In Japan, biologization also led

companies to argue that overwork was not the reason for depression, but it was instead caused by genetic vulnerability (Kitanaka 2012). Luhrmann et al. (2016: 25) argue that cultural conceptions of schizophrenia in the West produce social defeat "by repeatedly creating the conditions for demoralizations and despair". Moreover, the general concepts of psychiatric disorder, and especially its institutional version in the DSM, has spilled over from medicine into lay language and practices, to the extent that some disorders have become cultural categories (see Lindholm and Wickström 2020). Our individualistic society has a tendency to locate problems within individuals, whereas sometimes it might be better to investigate the wider social circumstances of suffering.

In the light of this, conceptual engineering of the concept of psychiatric disorder or mental disorder should not only be understood as refining the concept to match scientific discoveries, as it should also be seen as a causal ameliorative approach to the particular kinds that fall within its extension. Although these kinds are not constituted by sociocultural factors, nor by classificatory and diagnostic practices that seek to sort out and investigate them, they nonetheless play an important role in how disorders develop as well as how they are perceived and experienced. Although currently the concept of psychiatric disorder is not scientifically rigorous, it will no doubt become more technical and precise as psychiatry advances. This, however, would not make psychiatric disorders any less social or their conceptions value-free. Rather, it would provide us with better grounds for making value-sensitive assessments concerning classifications, diagnoses, therapeutic interventions, as well as policy decisions. In this sense, the concept of psychiatric disorder can be compared to poverty. Research on poverty has been able to provide a better understanding of the condition as well as stricter value-laden definitions, which are nonetheless morally justified and necessary (see Zachar 2014b: 154, Brigandt 2020). Similarly, the

concept of psychiatric disorder may require relatively strict boundaries to serve its bureaucratic and moral purposes (Brigandt 2020). But it should be borne in mind that just like the conception of poverty, the concept of psychiatric disorder serves the double role of defining conditions and causally interacting with them.

In conclusion, while research on the nature of psychiatric disorder is essential, it is also necessary to investigate and weigh the potential social consequences of its proposed conceptualization and classification in the society at large. Diagnostic categories, when spread and employed, may have inadvertent social significances on health when they become part of social practices and lay conceptions, thereby gaining causal efficacy irrespective of their original institutional roles.

References

Aftab, A. (2019). ”Social misuse of disorder designation, Part III: Harm and ethical validity”. *Psychiatric Times.* https://www.psychiatrictimes.com/view/social-misuse-disorder-designation-part-iii-harm-and-ethical-validity

Aftab, A. (2021). “Experimental Philosophy of Psychiatry”, *Bulletin: Association for the Advancement of Philosophy and Psychiatry* 28(1), 2–8.

Aftab, A. & Rashed, M. (2021). “Mental disorder and social deviance”, *International Review of Psychiatry (Review article)”* 33, 478–485.

Aleman, A. & Larøi, F. (2008). *Hallucinations: The science of idiosyncratic perception.* American Psychological Association. DOI: https://doi.org/10.1037/11751-000

Alexandrova, A. (2018). “Can the Science of Well-Being Be Objective”, *The British Journal of Philosophy of Science* 69, 421–445.

Allen, S. R. (2018). “Kinds Behaving Badly: Intentional Action and Interactive Kinds”, *Synthese*, DOI: https://doi.org/10.1007/s11229-018-1870-0

Allison, A. (2013). *Precarious Japan*. Durham, NC: Duke University Press.

Amoretti, M. C. & Lalumera, E. (2021). "Wherein is the concept of disease normative? From weak normativity to value-conscious naturalism", *Medicine, Health Care and Philosophy* 25, 47–60. DOI: https://doi.org/10.1007/s11019-021-10048-x

Amundson, R. & Lauder, G. V. (1994). "Function without purpose: The uses of causal role function in evolutionary biology", *Biology & philosophy* 9 (4), 443–469.

APA (American Psychiatric Association) (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM)*, fifth edition.

Appiah, K. A. (2005). *The ethics of identity*. Princeton, N.J.: Princeton University Press.

Appiah, K. A. (2015). "Race, Culture, Identity: Misunderstood Connections". In *Color Conscious*. Princeton University Press, 30–105. DOI: https://doi.org/10.1515/9781400822096-002

Aragona, M. (2009). "The concept of mental disorder and the DSM-V", *Dialogues in Philosophy, Mental and Neuro Sciences* 2, 1–4. DOI:10.1590/1415-4714.2015v18n4p599.2

Aragona, M. (2015). "The Hermeneutics of mental symptoms in the Cambridge School", *Revista Latinoamericana de Psicopatologia Fundamental* 18, 599–618.

Aragona, M. (2017): Psychiatrist at NIHMP institute (National Institute for Health, Migration, and Poverty). Interview at NIHMP on October 17, 2017. (Interviewer T. Vesterinen)

Aronowitz, R. (2008). "Framing disease: an underappreciated mechanism for the social patterning of health", *Social science & medicine* 67(1), 1–9. DOI: https://doi.org/10.1016/j.socscimed.2008.02.017

Arpaly, N. (2005). “How it is not “Just Like Diabetes”: Mental Disorders and the Moral Psychologist.” *Philosophical Issues* 15(1), 282–298. DOI: https://doi.org/10.1111/j.1533-6077.2005.00067.x

Arstila, V. (2017). “Cognitive penetration, hypnosis and imagination”, *Analysis* 77(1), 3–10. DOI: https://doi.org/10.1093/analys/anx048

Austin, J. L. (1962). *Sense and sensibilia*. London: Oxford University Press.

Aucouturier, V. & Demazeux, S. (2013). “The concept of “mental disorder”. In H. Carel & R. Cooper (ed.) *Health, Illness and Disease*. London: Routledge, 89–106.

Bach, T. (2016). “Social Categories are Natural Kinds, not Objective Types (and Why it Matters Politically)”. *Journal of Social Ontology 2*(2), 177–201. DOI: https://doi.org/10.1515/jso-2015-0039

Bach, T. (2019). “Real Kinds in Real Time: On Responsible Social Modeling”. *The Monist* 102(2), 236–258. DOI: https://doi.org/10.1093/monist/onz008.

Barnes, B. (1983). ”Social Life as Bootstrapped Induction”, *Sociology* 17(4), 524–545. DOI: https://doi.org/10.1177/0038038583017004004

Bechtel, W. (2005). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology* (Cambridge Studies in Philosophy and Biology). Cambridge: Cambridge University Press. DOI:10.1017/CBO9781139164962

Bechtel, W. (2009): "Looking Down, Around, and Up: Mechanistic Explanation in Psychology". *Philosophical Psychology* 22, 543–564.

Bechtel, W. (2019). "From parts to mechanisms: research heuristics for addressing heterogeneity in cancer genetics", *History and Philosophy of the Life Sciences* 41(3), 1–23. DOI: https://doi.org/10.1007/s40656-019-0266-x

Bechtel, W. & Abrahamsen, A. (2005). "Explanation: a mechanistic alternative", *Studies in History and Philosophy of Biological and Biomedical Science* 36, 421–441.

Becker, H. (1953). "Becoming a Marihuana User", *American Journal of Sociology* 59, 235–242.

Becker, H. (1963). *Outsiders: Studies in the Sociology of Deviance.* New York: The Free Press of Glencoe.

Beebee, H. & Sabbarton-Leary, N. (2010a). "Introduction". In H. Beebee & N. Sabbarton-Leary (eds.): *The semantics and metaphysics of natural kinds.* Routledge.

Beebee, H. & Sabbarton-Leary, N. (2010b). "Are psychiatric kinds real?", *European Journal of Analytic Philosophy* 6(1), 11–27.

Bell, M. (2014). *Melancholia: The Western Malady*. Cambridge: Cambridge University Press.

Beneduce, R. (2016). "Traumatic pasts and the historical imagination: Symptoms of loss, postcolonial suffering, and counter-memories among African migrants", *Transcultural Psychiatry* 53, 261–285.

Bental, R. P. (2003). *Madness Explained: Psychosis and Human nature with a Foreword by Aaron T. Beck.* London: The Penguin Press.

Betz, G. (2013). "In Defence of the Value Free Ideal", *European journal for philosophy of science* 3(2), 207–220.

Biggs, M. (2009). "Self-Fulfilling Prophecies". In P. Hedström & P. Bearman (eds.): *The Oxford Handbook of Analytical Sociology.* Oxford: Oxford University Press*,* 294–314.

Bird, A. (2014). "Human Kinds, Interactive Kinds, and Realism about Kinds." (unpublished)

Bird, A. & Tobin, E. (2018). "Natural Kinds". In Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy,* Spring 2018 Edition. https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/

Blashfield, R. K. (1982). "Feighner et al., Invisible Colleges, and the Matthew Effect", *Schizophrenia Bulletin* 8(1), 1–6. DOI: https://doi.org/10.1093/schbul/8.1.1

Bolton, D. (2001). "Problems in the Definition of 'Mental Disorder'". *The Philosophical Quarterly* 51, 182–199.

Bolton, D. (2008). *What is Mental Disorder: An essay in philosophy, science and values*. Oxford: Oxford University Press.

Bolton, D. (2013). "What is Mental Illness?" In K.W.M. Fulford et al. (ed.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press.

Boorse, C. (1975). “On the distinction between disease and illness”, *Philosophy and Public Affairs* 5, 49–68.

Boorse, C. (1976). “What a theory of mental health should be”, *Journal of Social Behaviour* 6, 61–84.

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

Boorse, C. (2002). “Rebuttal on Functions”. In R. Cummings, A. Ariew & M. Perlman (eds.): *Functions: New Essays in the Philosophy of Psychology and Biology*. Oxford: Oxford University Press, 63–112.

Borsboom, D. (2017), “A network theory of mental disorders”, *World Psychiatry* 16, 5–13. DOI: https://doi.org/10.1002/wps.20375

Bourdieu, P. (2012). *Outline of a Theory of Practice.* (Original: Esquisse d’une theorie de la pratique, precede de trois etudes d’ethnologie kabyle). Cambridge: Cambridge University Press.

Bowker, G. C. & Star, S. L. (2000). *Sorting Things Out: Classification and its Consequences.* Cambridge, MA: The MIT Press.

Boyd, R. (1989). ”What Realism Implies and What it Does Not”, *Dialectica* 43(1/2), 5–29. http://www.jstor.org/stable/42970608

Boyd, R. (1991). “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds”. *Philosophical Studies* 61, 127–148.

Boyd, R. (1999a). “Kinds as the ‘Workmanship of Men’, Realism, Constructivism, and Natural Kinds”. In J. Nida-Rümelin (ed.): *Rationalität,*

Realismus, Revision: *Proceedings of the Third International Congress, Gesellschaft für Analytische Philosophie*. Berlin: de Gruyter.

Boyd, R. (1999b). "Kinds, Complexity and Multiple Realization", *Philosophical Studies* 95: 67–98.

Boyd, R. (2010). "Realism, Natural Kinds, and Philosophical Methods". In H. Beebee & N. Sabbarton-Leary (eds): *The Semantics and Metaphysics of Natural Kinds*. New York: Routledge.

Boyer, P. (2011). "Intuitive expectations and the detection of mental disorder: A cognitive background to folk-psychiatries", *Philosophical Psychology* 24, 95–118.

Brigandt, I. (2003). "Species Pluralism Does Not Imply Species Eliminativism". *Philosophy of Science* 70(5), 1305–1316.

Brigandt, I. (2013). "Explanation in Biology: Reduction, Pluralism, and Explanatory Aims", *Science & Education* 22(1), 69–91. DOI: https://doi.org/10.1007/s11191-011-9350-7

Brigandt, I. (2020). "How to Philosophically Tackle Kinds without Talking about "Natural Kinds"". *Canadian Journal of Philosophy*, 1–24. https://doi.org/10.1017/can.2020.29

Brigandt, I. (2021). "Natural Kinds and Concepts: A Pragmatist and Methodologically Naturalistic Account". In J. Knowles & H. Rydenfelt (eds.): *Pragmatism, Science and Naturalism*. Frankfurt am Main: Peter Lang Publishing, 171–196.

Brinkmann, S. (2020). *Diagnostic Cultures. A Cultural Approach to the Pathologization of Modern Life*. New York: Routledge.

Broadbent, A. (2019). "Health as a Secondary Property", *The British Journal for the Philosophy of Science* 70(2), 609–627. DOI: https://doi.org/10.1093/bjps/axx014

Broome, M. & Bortolotti, L. (2009). "Mental illness as mental: In defence of psychological realism." *Humana Mente* 11, 25–43.

Bueter, A. (2019a). "Social Epistemology and Psychiatry". In Tekin and Bluhm (ed.) *The Bloomsbury Companion to Philosophy of Psychiatry.* London: Bloomsbury Academic, 485–504.

Bueter, A. (2019b). "Epistemic Injustice and Psychiatric Classification", *Philosophy of science*. 86(5), 1064–1074.

Campaner, R. (2011): "Understanding mechanisms in the health sciences", *Theoretical Medicine and Bioethics*, 32, 5–17.

Campaner, R. (2014). "Explanatory Pluralism in Psychiatry: What Are We Pluralists About, and Why?" In M. C. Galavotti, D. Dieks, W. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.): *New Directions in the Philosophy of Science, The Philosophy of Science in a European Perspective* 5, 87–103. Switzerland: Springer International Publishing.

Campaner, R. (2016). "Mechanistic Models and Modeling Disorders". In E. Ippoliti et al. (eds.): *Models and Inferences in Science*, Studies in Applied Philosophy, Epistemology and Rational Ethics, 25, 113–132.

Campaner, R. (2018). “The Interventionist Theory and Mental Disorders”. In Wenceslao J. Gonzalez et al. (eds.), *Philosophy of Psychology: Causality and Psychological Subject: New Reflections on James Woodward’s Contribution.* Berlin: De Gruyter.

Campbell, J. (2008). “Causation in psychiatry.” In K. Kendler and J. Parnas (Eds.): *Philosophical Issues in Psychiatry*. Baltimore, MD: Johns Hopkins University Press*,* 196–216.

Canguilhem, G. (1991). *The normal and the pathological.* New York: Zone Books.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago, IL: University of Chicago Press.

Carruthers, P., Laurence, S. & Stich, S. (2005). *The innate mind: Vol. 2. Culture and*
cognition. Oxford: Oxford University Press.

Cartwright, S. A. (2004/1851): “Diseases and physical peculiarities of the negro race.” In A. L. Caplan, J. J. McCartney, and D. A. Sisti (Eds.): *Health, disease, and illness.* Washington, DC: Georgetown University Press, 28–39.

Cartwrigth, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Cartwrigth, N. (2002). “Against modularity, the causal Markov condition, and any link between the two: comments on Hausman and Woodward”. *British Journal for the Philosophy of Science* 53, 411–453.

Cartwright, N. (2004). “Causation: One word, many things”. *Philosophy of Science* 71, 805–819.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S.J., Harrington, H., Israel, S., … & Moffitt, T.E. (2014). “The p factor: One general psychopathology factor in the structure of psychiatric disorders?”, *Clinical Psychological Science* 2, 119–137.

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. New York, US: OUP Usa.

Chang, H. (2017). “Epistemic iteration and natural kinds: Realism and pluralism in taxonomy”. In K. Kendler & J. Parnas (eds.): *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry.* Oxford: Oxford University Press*,* 229–245.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy 78*(2), 67–90. DOI: https://doi.org/10.2307/2025900

Clark, A. & Chalmers, D. (1998). ”The Extended Mind.” *Analysis* 58(1), 7–19. http://www.jstor.org/stable/3328150

Clifford, H. (1898). “Some Notes and Theories Concerning Latah”. In Hugh Clifford (ed.): *Studies in Brown Humanity*. London: Grant Richards, 186–201.

Cockerham, W. C. (2017). *Sociology of Mental disorder*. London: Routledge.

Cohen, L. (1995). "The Epistemological Carnival: Meditations on Disciplinary Intentionality and Ayurveda", In D. Bates (ed.): *Knowledge and the Scholarly Medical Traditions*. Cambridge: Cambridge University Press.

Coleman, J. (1990). *Foundations of Social Theory*. Cambridge, MA: Belknap Press.

Collins, R. (2004). *Interaction Ritual Chains*. Princeton, NJ: Princeton University Press.

Colombo, A., Bendelow, G., Fulford, B. & Williams, S. (2003). "Evaluating the influence of implicit models of mental disorder on processes of shared decision making within community-based multidisciplinary teams". *Social science & medicine* 56(7), 1557–1570. DOI: https://doi.org/10.1016/s0277-9536(02)00156-9

Cooper, R. (2002). "Disease", *Studies in History and Philosophy of Biological and Biomedical Sciences* 33, 263–282.

Cooper, R. (2004). "Why Hacking is Wrong about Human Kinds", *The British Journal for the Philosophy of Science* 55, 73–85.

Cooper, R. (2005). *Classifying Madness: A Philosophical Examination of the Diagnostic and Statistical Manual of Mental Disorders*. Dordrecht: Springer.

Cooper, R (2007). *Psychiatry and Philosophy of Science*. Stocksfield: Acumen.

Cooper, R. (2010). "Are culture-bound syndromes as real las universally-occurring disorders?", *Studies in History and Philosophy of Biological and Biomedical Sciences*, 41, 325–332.

Cooper, R. (2013). "Natural Kinds", In Fulford et al. (eds.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press.

Cooper, R. (2015). "Must Disorders Cause Harm? The Changing Stance of the DSM". In S. Demazeux & P. Singy (eds.): *The DSM-5 in Perspective: Philosophical Reflexctions on the Psychiatric Babel*. Heidelberg: Springer.

Cooper, R. (2020). "The Concept of Disorder Revisited: Robustly Value-Laden Despite Change." *Aristotelian Society Supplementary Volume* 94(1), 141–161. DOI: https://doi.org/10.1093/arisup/akaa010

Cracraft, J. (1983): "Species concepts and speciation analysis". In R. Johnston (Ed.): *Current Ornithology*, Volume 1, New York and London: Plenum Press, 159-187.

Craver, C. F. (2007). *Explaining the Brain*. Oxford, GB: Oxford University Press.

Craver, C. F. (2009). "Mechanisms and natural kinds", Philosophical psychology 22(5), 575–594.

Craver, C. F. & Bechtel, W. (2007). ”Top-down causation without top-down causes”. *Biology and Philosophy* 22(4), 547–563.

Craver, C. F., Glennan, S. & Povich, M. (2021). “Constitutive relevance & mutual manipulability revisited.” *Synthese* 199 (3-4):8807-8828.

Cuthbert, B. N. & Insel, T. R. (2013). “Toward the future of psychiatric diagnosis: the seven pillars of RDoC”. *BMC Med.* 11(126). DOI: 10.1186/1741-7015-11-126

Danzinger, E., Schieffelin, B. & Pritzker, S. (2011). “The learning of mind: how do you figure out what a mind is? Issues of language”, Suomen Antropologi 4, 51–56.

Darden, L. (2006). *Reasoning in Biological Discoveries: Essays on Mechanisms, Interfield Relations, and Anomaly Resolution*. Cambridge University Press.

Davidson, D. (1963). “Actions, Reasons, and Causes”, *The Journal of Philosophy*, 60, 685–700.

Demazeux, S. (2015). “The Function Debate and the Concept of Mental Disorder”. In P. Huneman, G. Lambert & M. Silberstein (eds): *Classification, Disease and Evidence: New Essays in the Philosophy of Medicine*. Dordrech: Springer, 63–92.

Devereux, G. (1980). *Basic Problems of Ethnopsychiatry.* Chicago, IL: University of Chicago Press.

Devitt, M. (2010). *Putting Metaphysics First: Essays on Metaphysics and Epistemology*. Oxford: Oxford University Press.

Douglas, M. (1986). *How Institutions Think*. Syracuse, NY: Syracuse University Press.

Dupré, J. (1981). "Natural Kinds and Biological Taxa", *The Philosophical Review* 90, 66–90.

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science.* Cambridge, MA: Harvard University Press.

Dupré, J. (2004). "Human Kinds and Biological Kinds: Some Similarities and Differences", *Philosophy of Science* 71, 892–900.

Dupré, J. (2007). "Fact and Value" In Kincaid, H. et al. (ed.) *Value-Free Science?: Ideals and Illusions.* New York: Oxford University Press.

Dupré, J. (2012). *Processes of Life: Essays in the Philosophy of Biology*. Oxford: Oxford University Press.

Dupré, J. (2015). "For objective, value-laden, contextualist pluralism". In K. Kendler and J. Parnas (eds.): Philosophical Issues in Psychiatry III: *The nature and sources of historical change*. Oxford: Oxford University Press, 20–23.

Echterhoff, G., & Semin, G. R. (2011). *Grounding sociality neurons, mind, and culture.* Psychology Press.

Edgerton, R.B. (1966). "Conceptions of Psychosis in Four East African Societies". *American Anthropologist*, 68: 408–425. DOI: https://doi.org/10.1525/aa.1966.68.2.02a00070

Eklund, M. (2015). “Intuitions, Conceptual Engineering, and Conceptual Fixed Points”. In Christopher Daly (ed.), *The Palgrave Handbook of Philosophical Methods.* Cham: Springer International Publishing AG.

Ellis, B. (2001). *The Philosophy of Nature: A Guide to the New Essentialism*. Chesham England: Acumen

Elster, J. (2015). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences.* Cambridge: Cambridge University Press.

Engel, G. (1977). “The need for a new medical model: A challenge for biomedicine”, *Science,* 196, 129–139.

Ereshefsky, M. (2010). ”Microbiology and the species problem”. *Biology and Philosophy* 25 (4):553–568.

Ereshefsky, M. & Reydon, T. A. C. (2015). “Scientific Kinds”, *Philosophical Studie,* 172, 969–986.

Eronen, M. (2013). “No level, no problems: Downward causation in neuroscience”, *Philosophy of Science* 80: 1042–1052

Eronen, M. (2015). “Levels of organization: A deflationary account”, *Biology and Philosophy* 30: 39–58.

Eronen, M. (2021). “The levels problem in psychopathology”, *Psychological*
Medicine 51, 927–933. DOI:
https://doi.org/10.1017/S0033291719002514

Fanon, F. (1952). *Black Skin, White Masks*. New York: Grove Press.

Faucher, L. & Goyer, S. (2015). "RDoC: Thinking Outside the DSM Box Without Falling into a Reductionist Trap". In S. Demazeux & P. Singy (eds.): *The DSM-5 in Perspective – Philosophical Reflections on the Psychiatric Babel*. Dordrecht: Springer.

Fessler, D. & Machery, E. (2012). "Culture and Cognition". In Margolis, E., Samuels, R. and Stich, S. (eds.), *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford: Oxford University Press.

Feyerabend, P. (1965). Problems of Empiricism. In Colodny, R. G. (ed.), *Beyond the Edge of Certainty – Essays in Contemporary Science and Philosophy Vol. 2.* Englewood Cliffs, N. J.: Prentice-Hall, Inc.

Firestone, C. & Scholl, B. J. (2016). "Cognition does not affect perception: Evaluating the evidence for "top-down" effects." *The Behavioral and brain sciences*, 39, e229. DOI: https://doi.org/10.1017/S0140525X15000965

First, M. B. (2017). "Factors in the Development of Psychiatric Epidemics". In Kendler, K. and Parnas, J. (eds.): *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry,* 130–142. Oxford: Oxford University Press.

Foucault, M. (1992 [1967]). *Madness and Civilization: A History of Insanity in the Age of Reason*. London: Routledge.

Foucault, M. (1978). *The History of Sexuality, Vol. I: An Introduction*. New York: Pantheon.

Foucault, M. (2006). *History of Madness*. London: Routledge.

Friedman, B. & Hendry, D. (2019). V*alue Sensitive Design: Shaping Technology with Moral Imagination*. DOI: 10.7551/mit-press/7585.001.0001.

Fuchs, T. (2012). "Are Mental Illness Diseases of the Brain", In S. Choudhury & J. Slaby (eds.), *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience,* 331-344. Chichester, West Sussex: Wiley-Blackwell

Fulford, K. W. M. (1989). *Moral theory and medical practice.* Cambridge: Cambridge University
Press.

Fulford, K. & Thornton, T. (2007). "Fanatical about 'harmful dysfunction'", *World Psychiatry* 6, 161-162.

Fulford et al. (ed.) (2013). "The Next Hundred Years: Watching Our Ps and Q". In Fulford et al. (eds). *The Oxford Handbook of Philosophy and Psychiatry,* Oxford: Oxford University Press, 1–11.

Gagné-Julien, A.-M. (2021). "Towards a socially constructed and objective concept of mental disorder", *Synthese* 198, 9401–9426.

Gallagher, S. (2013). "Intersubjectivity and Psychopathology". In K. W. M. Fulford et al. (eds), *The Oxford Handbook of Philosophy and Psychiatry*.
DOI: https://doi.org/10.1093/oxfordhb/9780199579563.013.0019

Geertz, C. (1973). *The Interpretation of Cultures*. New York: Basic Books.

Gerrans, P. (2014). *The measure of madness: Philosophy of mind, cognitive neuroscience, and delusional thought.* MIT Press. DOI: https://doi.org/10.7551/mitpress/9780262027557.001.0001

Ghaemi, S. N. (2003). *The Concept of Psychiatry.* Baltimore, MD: Johns Hopkins University Press.

Gibbs, J. P. (1971). "A Critique of the Labeling Perspective". In E. Rubington & M. Weinberg (eds): *The Study of the social problems*. Oxford: Oxford University Press, 193–205.

Gil, J. & Rego, A. (2008). "Mechanisms of degeneration of Huntington's disease, review article". *European Journal of Neuroscience*, 27, 2803–2820.

Glennan, S. S. (2002). "Rethinking Mechanistic Explanation", *Philosophy of*
Science 69: 342–353.

Glennan, S. & Illari, P. (2018). "Varieties of mechanisms". In S. Glennan & P. Illari (eds.): *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge, 91–13.

Godfrey-Smith, P. (2021). *Theory and Reality: An Introduction to the Philosophy of Science.* The University of Chicago Press.

Godman, M. (2013). "Psychiatric Disorders qua Natural Kinds: The case of the 'Apathetic Children', *Biological Theory* 7(2), 144–152.

Godman, M. (2016). “Cultural syndromes: Socially learned but real”, *Filosofia Unisinos/ Unisinos Journal of Philosophy*, 17(2), 185–191.

Godman, M. (2021). *The Epistemology and Morality of Human Kinds.* New York: Routledge

Godman, M. & Marchionni, C. (2022). “What should scientists do about (harmful) interactive effects”. *European Journal for Philosophy of Science,* 12(63). DOI: https://doi.org/10.1007/s13194-022-00493-7

Goffman, E. (1961). *Asylums: Essays on the Social Situation of Mental Patients and Other inmates.* Harmondsworth Penguin.

Good, B. J. (1994). *Medicine, rationality, and experience – An anthropological perspective.* Cambridge: Cambridge University Press.

Goosens, W. K. (1980). “Values, Health, and Medicine”, *Philosophy of Science* 47, 100–115.

Graham, G. (2013). *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*. New York: Routledge.

Griffiths, P. E. (1997). *What emotions really are: the problem of psychological categories.* Chicago: University of Chicago Press.

Griffiths, P. (1999). “Squaring the Circle: Natural Kinds with Historical Essences”. In R. A. Wilson (ed.): *Species. New Interdisciplinary Essays*. Cambridge, MA: MIT Press.

Griffiths, P. (2004). “Emotions as Natural and Normative Kinds.” *Philosophy of Science* 71, 901–911.

Griffiths, P. E. & Matthewson, J. (2018). “Evolution, Dysfunction, and Disease: A Reappraisal.” *The British Journal for the Philosophy of Science* 69(2), 301–327. DOI: https://doi.org/10.1093/bjps/axw021

Guala, F. (2016). *Understanding Institutions.* Princeton: Princeton University Press.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press.

Hacking, I. (1986). “Making Up People”. In T. Heller, M. Sosna & D. Wellbery (eds.): *Reconstructing Individualism*. Stanford, CA: Stanford University Press*,* 222–236.

Hacking, I. (1991). “A Tradition of Natural Kinds”, *Philosophical Studies* 61, 109–126.

Hacking, I. (1992). “World-Making by Kind-Making: Child Abuse for Example”. In M. Douglas and D. L. Hull (eds.): *How Classification Works: Nelson Goodman among the Social Sciences*. Edinburgh: Edinburgh University Press*,* 180–238.

Hacking, I. (1995a). *Rewriting the Soul: Multiple Personality and the Sciences of Memory*. Princeton: Princeton University Press.

Hacking, I. (1995b). “The Looping Effects of Human Kinds”. In D. Sperber, D. Premack, A. J. Premack (eds): *Symposia of the Fyssen*

Foundation. Causal cognition: A Multidisciplinary Debate. New York: Clarendon Press, 351–394.

Hacking, I. (1997). "Taking Bad Arguments Seriously: Ian Hacking on Psychopathology and Social Construction". *London Review of Books* 19, 14–16.

Hacking, I. (1998). *Mad Travelers*. Cambridge, MA: Harvard University Press.

Hacking, I. (1999). *The Social Construction of What?* Cambridge, MA: Harvard University Press.

Hacking, I. (2001). "Aristotelian Categories and Cognitive Domains", *Synthese* 126, 473–515.

Hacking, I. (2002). "How "Natural" are "Kinds" of Sexual Orientation?", *Law and Philosophy* 21, 335–347.

Hacking, I. (2007a). "Kinds of People: Moving Targets", *Proceedings of the British Academy* 151, 285–318.

Hacking, I. (2007b). "Natural Kinds: Rosy Dawn, Scholastic Twilight", *Royal Institute of Philosophy Supplement* 61, 203–239.

Hacking, I. (2010). "Pathological withdrawl of refugee children seeking asylum in Sweden", *Studies in History and Philosophy of Biological and Biomedical Sciences* 41, 309–317.

Hales, S. D. (2006). *Relativism and the Foundations of Philosophy*. Cambridge, MA: The MIT Press.

Halina, M. (2017). “Mechanistic explanation and its limits”. In S. Glennan and P. Illari (eds.): *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge, 213–224.

Hansson, S. O. (2013). “Defining Pseudoscience and Science”, In M. Pigliucci & M. Boundry (eds.): *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem.* The University of Chicago Press: Chicago*,* 61–78.

Harland, R., Antonova, E., Owen, G. S., Broome, M., Landau, S., Deeley, Q., & Murray, R. (2009). ”A study of psychiatrists' concepts of mental illness.” *Psychological medicine*, 39(6), 967–976. DOI: https://doi.org/10.1017/S0033291708004881

Haslam, N. (2013). “Reliability, Validity, and the Mixed Blessings of Operationalism”, In Fulford et al. (eds.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press.

Haslam, N. (2014). “Natural Kinds in Psychiatry: Conceptually Implausible, Empirically Questionable, and Stigmatizing.” In Kincaid and Sullivan (eds.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*. Cambridge, MA: The MIT Press.

Haslam, N. (2016). “Looping effects and the expanding concept of mental disorder”, *Journal of Psychopathology* 22, 4–9.

Haslam, N., Ban, L. & Kaufmann, L. (2007). ”Lay conceptions of mental disorder: The folk psychiatry model”, *Australian Psychologist* 42(2), 129–137, DOI: 10.1080/00050060701280615

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique.* New York: Oxford University Press.

Hausman, D. M. & Woodward, J. (1999). ”Independence, invariance and the causal Markov condition”, *British Journal for the Philosophy of Science* 50(4), 521–583.

Hauswald, R. (2016). ”The Ontology of Interactive Kinds”, *Journal of Social Ontology* 2(2), 203–221.

Hedström, P. & Ylikoski, P. (2010). “Causal Mechanisms in the Social Sciences”, *Annual Review of Sociology* 36, 49–67.

Hempel, C. G., 1965, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.

Henderson, D. K. (1993). *Interpretation and Explanation in the Human Sciences*. Albany: State University of New York.

Henrich, J., Heine, S. J. & Norenzayan, A. (2010). “The weirdest people in the world?” *The Behavioral and Brain Sciences*, 33(2–3), 61–83. DOI: https://doi.org/10.1017/S0140525X0999152X

Hintikka, J. (1999). “The Emperor’s New Intuitions”, *The Journal of philosophy* 96(3), 127–147.

Hoffman, G. *(2016). “*Out of our skulls: How the extended mind thesis can extend psychiatry*.” Philosophical Psychology* 29, 1–15. DOI: 10.1080/09515089.2016.1236369.

Honko, L. (1960). *Varhaiskantaiset taudinselitykset ja parannusnäytelmä*. Hautala, Jouko (ed.) Jumin keko. Tietolipas 17. Forssa.

Honkasalo, M.-L. (2016): "After Sacrifice: Deeds and Words for Protection and Cure", *Suomen Antropologi* 41, 29–45.

Hopper, K. (2004). "Interrogating the Meaning of 'Culture' in the WHO International Studies of Schizophrenia". In J. Jenkins & R. Barret (eds.): *Schizophrenia, Culture, and Subjectivity*. Cambridge, UK: Cambridge University Press, 62–87.

Hopper, K., Harrison, G., Janca, A. & Sartorius, N. (eds.) (2007). *Recovery from Schizophrenia: An International Perspective: A report from the WHO Collaborative Project, The International Study of Schizophrenia.* Oxford: Oxford University Press.

Horwitz, A. V. & Wakefield, J. C. (2007). *The loss of sadness: how psychiatry transformed normal sorrow into depressive disorder.* Oxford University Press.

Howes, O. D. & Murray, R. M. (2014). "Schizophrenia: an integrated sociodevelopmental-cognitive model." *The Lancet*, 383(9929), 1677–1687. DOI: https://doi.org/10.1016/S0140-6736(13)62036-X

Horwitz, A. V. (2002). *Creating Mental Illness*. Chicago. Ill: The University of Chicago Press.

Horwitz, A. V. & Wakefield, Jerome C. (2007). *The Loss of Sadness: How Psychiatry Transformed Normal Sorrow into Depressive Disorder*. Oxford: Oxford University Press.

Hugher, C. C. (1985). “Culture-Bound or Construct-Bound?”. In Simons, Ronald and Hugher, Charles. *The Culture-Bound Syndromes: Folk Illnesses of Psychiatric and Anthropological Interest*. Dordrecht, Holland: Reidel Publishing Company.

Hull, D. L. (1980). “Individuality and Selection”, *Annual Review of Ecology and Systematics*, 11, 311–332.

Hume, D. (1748). *An Enquiry concerning Human Understanding*.

Hutto, D. D. (2013). ”Interpersonal Relating”. In K.W.M. Fulford et al. (ed.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199579563.013.0018.

Hutto, D. H. (2016). “A Reconciliation for the Future of Psychiatry: Both Folk Psychology and Cognitive Science”, *Frontiers in Psychiatry*, 7. DOI: https://doi.org/10.3389/fpsyt.2016.00012

Hutto, D. H. & Kirchhoff, Michael D. (2015). “Looking beyond the brain: social neuroscience meets narrative practice”, *Cognitive Systems Research,* 34–35, 5–17.

Hyman, S. E. (2010). ”The diagnosis of mental disorders: the problem of reification”. *Annual review of clinical psychology* 6, 155–179. DOI: https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Illary, P. (2013). “Mechanistic Explanation: Integrating the Ontic and Epistemic”. *Erkenntnis*, 78, 237–255. DOI: 10.1007/s10670-013-9511-y

Illary, P. (2017). “Mechanisms in Medicine”, In Solomon, Simon & Kincaid (ed.): *The Routledge Companion to Philosophy of Medicine*. New York: Routledge, 48–57.

Illary, P. & Russo, F. (2014). *Causality: Philosophical Theory Meets Scientific Practice.* Oxford: Oxford University Press.

Illari, P. & Williamson, J. (2012). ”What is a mechanism? Thinking about mechanisms across the sciences”, *European Journal for Philosophy of Science* 2(1), 119–135.

Insel, T. (2010). “Faulty circuits”, *Scientific American* 302, 44–51.

Insel, T. (2013). *Transforming Diagnosis*. NIMH Director’s Blog. Available at: http://www.nimh.nih.gov/about/director/index.shtml

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C. & Wang, P. (2010). ”Research domain criteria (RDoC): toward a new classification framework for research on mental disorders.” *The American journal of psychiatry*, 167(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jackson, S. W. (1986). *Melancholia and depression*. New Haven, CT: Yale University Press.

Jaspers, K. (1963). *General Psychopathology*, translated by J. Hoenig and Marian W. Hamilton, Chicago: The University of Chicago Press.

Jefferson, A. (2018). “What does it take to be a brain disorder?”, *Synthese* 197, 249–262. DOI: https://doi.org/10.1007/s11229-018-1784-x

Kandel, E. R. (1998). ”A new intellectual framework for psychiatry”. *The American journal of psychiatry*, *155*(4), 457–469. DOI: https://doi.org/10.1176/ajp.155.4.457

Kaplan, J. (2010). “When socially determined categories make biological realities”, *Monist* 93, 283–299.

Keane, B., Silverstein, S., Wang, Y. & Papathomas, T. (2013). “Reduced Depth Inversion Illusions in Schizophrenia Are State-Specific and Occur for Multiple Object Types and Viewing Conditions”, *American Psychological Association*. DOI: 10.1037/a0032110

Keil, G. & Stoecker, R. (2017). “Disease as a vague and thick cluster concept”. In G. Keil, L. Keuck & R. Hauswald: *Vagueness in psychiatry.* Oxford: Oxford University Press.

Kellert, S., Longino, H. & Waters, K. (eds.) (2006). *Scientific Pluralism. Minnesota Studies in the Philosophy of Science.* Minneapolis: University of Minnesota Press.

Kendler, K. (2005). “Toward a philosophical structure for psychiatry”, *Am. J. Psychiatry* 162, 433–440.

Kendler, K. (2008). “Explanatory Models for Psychiatric Illness”, *The American journal of psychiatry 165*, 695–702. DOI: https://doi.org/10.1176/appi.ajp.2008.07071061

Kendler, K. (2012). “Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology”. *Molecular Psychiatry* 17, 1–18.

Kendler, K. (2016). “The Nature of Psychiatric Disorders”. *World Psychiatry* 15, 5–12.

Kendler K. (2021). ”Potential Lessons for DSM From Contemporary Philosophy of Science”. *JAMA Psychiatry* 79(2), 99–100. DOI:10.1001/jamapsychiatry.2021.3559

Kendler, K. S., & Campbell, J. (2009). “Interventionist causal models in psychiatry: repositioning the mind-body problem”. *Psychological medicine 39*(6), 881–887. DOI: https://doi.org/10.1017/S0033291708004467

Kendler, K. S., Zachar, P., & Craver, C. (2011). “What kinds of things are psychiatric disorders?” *Psychological Medicine* 41(6), 1143–1150. DOI: https://doi.org/10.1017/S0033291710001844

Kendler, K., Parnas, J. & Zachar, P. (eds.). (2020). *Levels of Analysis in Psychopathology: Cross-Disciplinary Perspectives*. Cambridge: Cambridge University Press. DOI:10.1017/9781108750349

Kent, G. & Wahass, S. (1996). “The Content and Characteristics of Auditory Hallucinations in Saudi-Arabia and the UK: A Cross-Cultural Comparison”, *Acta Psychiatrica Scandinavica* 94(6), 433–437.

Khalidi, M. A. (2010). “Interactive Kinds”, *British Journal of Philosophy of Science* 61, 335–360.

Khalidi, M. A. (2013). *Natural Categories and Human Kinds*: *Classification in the Natural and Social Sciences*. Cambridge, UK: Cambridge University Press.

Kihlström, J. (2005). “Dissociative Disorders”, *Annual Review of Clinical Psychology* 1, 227–253.

Kincaid, H. (1996). *Philosophical Foundations of the Social Sciences. Analyzing Controversies in Social Research.* Cambridge: Cambridge University Press.

Kincaid, H. (2014). “Defensible Natural Kinds in the Study of Psychopathology”. In Kincaid and Sullivan (ed): *Classifying Psychopathology: Mental Kinds and Natural Kinds*. Cambridge, MA: The MIT Press.

Kincaid, H. & Sullivan J. (2010). ”Medical Models of Addiction.” In Ross, Kincaid, Spurrett & Collins (ed.): *What is addiction?* Cambridge, Massachusetts: The MIT Press.

Kincaid, H. & Sullivan, J. (2014). ”Classifying Psychopathology: Mental Kinds and Natural Kinds.” In *Classifying Psychopathology: Mental Kinds and Natural Kinds.* Cambridge, MA: The MIT Press.

Kincaid, H., Zachar, P., Murphy, D. et. al. (2021). *Defining Mental Disorder: Jerome Wakefield and His Critics.* DOI:10.7551/mitpress/9949.001.0001.

Kingma, E. (2007). “What is it to be healthy?” *Analysis,* 67(294), 128–133.

Kingma, E. (2013). “Naturalist Accounts of Mental Disorder”, in K.W.M. Fulford et al. (ed.), *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press.

Kingma, E. (2016). “Situation-Specific Disease and Dispositional Function”, *The British Journal for the Philosophy of Science* 67, 391–404.

Kirmayer, L. (1991). “The Place of Culture in Psychiatric Nosology: Taijin Kyofusho and DSM-III-R”, *The Journal of Nervous and Mental Disease* 179, 19–28.

Kirmayer, L. J. (2001). “Cultural Variations in the Clinical Presentation of Depression and Anxiety: Implications for Diagnosis and Treatment”, *The Journal of Clinical Psychiatry* 62, 22–30.

Kirmayer, L. J. (2002). “Psychopharmacology in a Globalizing World: The Use of Antidepressants in Japan”*, Transcultural Psychiatry* 39, 295–322.

Kirmayer L. J. (2005). “Culture, context and experience in psychiatric diagnosis.” *Psychopathology*, *38*(4), 192–196. DOI: https://doi.org/10.1159/000086090

Kirmayer, L. (2007). “Cultural psychiatry in historical perspective”. In D. Bhugra & K. Bhui (Eds.): *Textbook of Cultural Psychiatry*. Cambridge: Cambridge University Press, 3–19. DOI:10.1017/CBO9780511543609.003

Kirmayer, L. J., & Crafa, D. (2014). What kind of science for psychiatry?. *Frontiers in human neuroscience*, *8*, 435. https://doi.org/10.3389/fnhum.2014.00435

Kirmayer, L. J. & Gold, I. (2012). “Re-Socializing Psychiatry: Critical Neuroscience and the Limits of Reductionism”, In S. Choudhury and J.

Slaby (eds.), *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience*. Chichester, West Sussex: Wiley-Blackwell, 307–330.

Kitanaka, J. (2012). *Depression in Japan. Psychiatric Cures for a Society in Distress*. Princeton: Princeton University Press.

Kitcher, P. (2001). *Science, Truth, and Democracy*. New York, US: Oxford University Press.

Kitcher, P. & Stanford, P. (2000). "Refining the Causal Theory of Reference for Natural Kind Terms", *Philosophical Studies* 97, 99–129.

Kleinman, A. (1988). *The illness narratives: Suffering, healing, and the human condition.* Basic Books.

Kleinman, A. (1991). *Rethinking Psychiatry: From Cultural Category to Personal Experience.* New York: The Free Press.

Kleinman, A. & Good, B. (1985). *Culture and depression*. Berkeley: University of California Press.

Kokkonen, T. (2021). *Evolving in Groups: Individualism and Holism in Evolutionary Explanation of Human Social Behaviour. Ph.D. thesis.* Theoretical Philosophy, Philosophy (Swedish), and Social and Moral Philosophy. Helsinki: Unigrafia.

Kokkonen, T. & Koskinen, I. (2016). "Genres as Real Kinds and Projections. Homeostatic Property Clusters in Folklore and Art". In Koski,

K., Frog and Savolainen, U. (eds), *Genre – Text – Interpretation. Multidisciplinary Perspectives on Folklore and Beyond*. *Studia Fennica Folkloristica* 22, 89–109. Helsinki: Finnish Literature Society, SKS.

Kornblith, H. (1993). *Inductive Inference and Its Natural Ground: An Essay in Naturalistic Epistemology*. Cambridge, MA: The MIT Press.

Kraepelin, E. (1896). *Lehrbuch der Psychiatrie.* Leipzig: Barth.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Krueger, J. (2020). "Schizophrenia and the Scaffolded-Self", *Topio*, 39, 597–609.

Kuhn, T. (1996 [1962]). *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Kuhn, T. (2002 [1987]). "What are Scientific Revolutions?" In J. Conant & J. Haugeland, (eds.), *The Road Since Structure*. Chicago: Chicago University Press.

Kukla, A. (2000). *Social constructivism and the philosophy of science*. Routledge. DOI: https://doi.org/10.4324/9780203130995

Kuorikoski, J. (2009). "Two Concepts of Mechanism: Componential Causal
System and Abstract Form of Interaction", *International Studies in the Philosophy of Science*, 23(2), 143–160. DOI: 10.1080/02698590903006875

Kuorikoski, J. (2012). “Mechanisms, Modularity and Constitutive Explanation”, *Erkenntnis* 77, 361–380. DOI: https://doi.org/10.1007/s10670-012-9389-0

Kuorikoski, J., & Pöyhönen, S. (2012). “Looping Kinds and Social Mechanisms”, *Sociological Theory* 30(3), 187–205. DOI: https://doi.org/10.1177/0735275112457911

Kuorikoski, J. & Ylikoski, P. (2010). “Explanatory relevance across disciplinary boundaries: the case of neuroeconomics”, *Journal of Economic Methodology* 17, 219–228.

Kuorikoski, J., & Ylikoski, P. K. (2013). ”How Organization Explains.” In V. Karakostas & D. Dieks (Eds.): *EPSA11 Perspectives and Foundational Problems in Philosophy of Science,* 69–80. (The European Philosophy of Science Association Proceedings, 2). Springer. DOI: https://doi.org/10.1007/978-3-319-01306-0__6

La Roche, M., Fuentes, M. & Hinton, D. (2015). “A Cultural Examination of the DSM-5: Research and Clinical Implications for Cultural Minorities”, *Professional Psychology: Research and Practice* 46(3), 183–189.

Ladyman, J. & Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Laimann, J. (2018). “Capricious Kinds”, *The British Journal for the Philosophy of Science*, DOI: https://doi.org/10.1093/bjps/axy024

Laing, R. D. (1987 [1965]). *The Divided Self: An Existential Study in Sanity and Madness.* Armondsworth: Penguin Books.

Lakoff, A. (2000). “Adaptive Will: The Evolution of Attention Deficit Disorder”, *Journal of the History of the Behavioral Sciences* 36, 149–169.

Lakoff, A. (2005). *Pharmaceutical Reason*: *Knowledge and Value in Global Psychiatry*. Cambridge: Cambridge University Press.

LaPorte, J. (2004). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Lee, R. L. M. (1981). “Structure, and Anti-Structure in the Culture-Bound Syndromes: The Malay Case”, *Culture, Medicine and Psychiatry* 5, 233–248.

Lee, S. (1996). “Reconsidering the Status of Anorexia Nervosa as a Western-Bound Syndrome”, *Social Science & Medicine* 42, 21–34.

Lemert, E. M. (1967). *Human deviance, social problems, and social control*. Prentice-Hall.

Lemoine, M. (2013). “Defining disease beyond conceptual analysis: an analysis of conceptual analysis in philosophy of medicine”. *Theoretical medicine and bioethics* 34(4), 309–325.

Lende, D. (2014). The Research Domain Criteria of the NIMH and the RDoC Vision for
Mental Health Research and Diagnosis. PLOS, 9.2.2014. Internet: http://blogs.
plos.org/neuroanthropology/2014/02/09/research-domain-criteria-nimh-vision-mental-health-research-diagnosis/

Levy, A. (2013). “Three Kinds of “New Mechanism”, *Biology & Philosophy* 28: 99–114.

Lewis-Fernández, R., Hinton, D.E., Laria, A.J., Patterson, E.H., Hofmann, S.G., Craske, M.G., Stein, D.J., Asnaani, A. & Liao, B. (2010). “Culture and the anxiety disorders: Recommendations for DSM-V”, *Depression and Anxiety* 27, 212–229.

Lilienfeld, S. O. & Marino, L. (1995). “Mental Disorder as a Roschian Concept: A Critique of Wakefield’s ‘Harmful Dysfunction’ Analysis”, *Journal of abnormal psychology*, 104(3), 411–420.

Lindholm, S. K. & Wickström, A. (2020). “‘Looping effects’ related to young people’s mental health: How young people transform the meaning of psychiatric concepts”, *Global Studies of Childhood*, 10(1), 26–38. DOI: 10.1177/2043610619890058.

Link, B. G., Cullen, F. T., Struening, E., Shrout, P. E. & Dohrenwend, B. P. (1989). “A Modified Labeling Theory Approach to Mental Disorders: An Empirical Assessment”, *American Sociological Review* 54, 400–423.

Lipton, P. (1991). *Inference to the Best Explanation*. London and New York: Routledge/Taylor and Francis Group.

Littlewood, R. & Dein, S. (2000). *Cultural Psychiatry and Medical Anthropology: An Introduction and Reader*. London: The Athlone Press.

Longino, H. E. (1990). *Science as social knowledge: values and objectivity in scientific inquiry*. Princeton, N.J: Princeton University Press.

Longino, H. E. (2013). *Studying Human Behavior: How Scientists Investigate Aggression & Sexuality*. Chicago: The University Press of Chicago Press.

Longino, H. E. (2015). "Pluralism, incommensurability, and scientific change". In K. Kendler & J. Parnas (eds.) Philosophical Issues in Psychiatry III: *The nature and sources of historical change*, 7–19. Oxford: Oxford University Press.

Lucas, R. (1976). "Econometric Policy Evaluation: A Critique". In K. Brunner & A. H. Meltzer (eds.), *The Phillips Curve and Labor Markets.* Carnegie-Rochester Conference Series on Public Policy, 1, 19–46. New York: American Elsevier.

Ludwig, D. (2016). "Ontological Choices and the Value-Free Ideal", *Erkenntnis* 81(6), 1253–1272.

Luhrmann, T. (2001). *Of two minds: An anthropologist looks at American psychiatry*. New York: Vintage Books.

Luhrmann, T. (2011). "Hallucinations and Sensory Overrides", *Annual Review of Anthropology* 40(1), 71–85.

Luhrmann, T. (2012). "Beyond the Brain". The Wilson Quarterly 36(3), 28–34. DOI:10.2307/41933919

Luhrmann, T., Padmavati, R., Tharoor, H. & Osei, A. (2015). "Difference in Voice-Hearing Experiences of People with Psychosis in the USA, India and Ghana: Interview-Based Study", *The British Journal of Psychiatry* 206, 41–44.

Luhrmann, T. & Marrow, J. (eds.) (2016). *Our Most Troubling Madness: Case Studies in Schizophrenia Across Cultures*. Oakland, CA: University of California Press.

Lutz, C. (1998). *Unnatural Emotions – Everyday Sentiments on a Micronesian Atoll & Their Challenge to Western Theory*. Chicago: The Chicago University Press.

Machamer, P., Darden, L. & Craver, C. (2000). "Thinking about Mechanisms", *Philosophy of Science* 67, 1–24.

Machery, E. (2009). *Doing without Concepts*. Oxford: Oxford University Press.

MacKenzie, D. (2008). *An Engine, Not a Camera: How Financial Models Shape Markets.* Cambridge, MA: The MIT Press

Mallon, R. (2003). "Social Construction, Social Roles, and Stability". In F. F. Schmitt (ed.): *Socializing Metaphysics: The Nature of Social Reality*. Lanham, MD: Rowman & Littlefield, 327–354.

Mallon, R. (2007). "Human Categories Beyond Non-essentialism", *The Journal of Political Philosophy* 15, 146–168.

Mallon, R. (2016). *The Construction of Human Kinds*. Oxford: Oxford University Press.

Marchionni, C. (2008). "Explanatory Pluralism and Complementarity: From Autonomy to Integration", *Philosophy of the Social Sciences* 38, 314–333.

Martinez, E. J. (2017). “Stable Property Clusters and Their Grounds”, *Philosophy of Science*, 84, 944–955.

Martínez, M. L. (2009). “Ian Hacking’s Proposal for the Distinction between Natural and Social Sciences”, *Philosophy of the Social Sciences* 2, 221–234.

Maung, H. H. (2016). “Diagnosis and causal explanation in psychiatry”, *Studies in History and Philosophy of Biological and Biomedical Sciences* 60, 15–24.

Mawson, A., Berry, K., Murray, C. & Hayward, M. (2011). “Voice Hearing Within the Context of Hearers’ Social Worlds: An Interpretative Phenomenological Analysis”, *Psychology and Psychotherapy: Theory, Research and Practice* 84, 256–272.

Mazza, G. (2016). “*Work and Respect in Chennai*”. In T. Luhrmann & J. Marrow (eds.): *Our Most Troubling Madness: Case Studies in Schizophrenia Across Cultures*. Oakland, CA: University of California Press.

McGuire, C. (2020). *Measuring difference, numbering normal*. Manchester University Press.

McLorg, P. A. & Taub, D. E. (1987). “Anorexia Nervosa and Bulimia: The Development of Deviant Identities”, *Deviant Behavior* 8, 177–189.

Menzies, P. & Price, H. (1993). “Causation as a Secondary Quality”. *The British Journal for the Philosophy of Science* 44, 187–203. DOI: https://doi.org/10.1093/bjps/44.2.187

Merton, R. K. (1968). “The Matthew Effect in Science. The reward and communication systems of science and considered.” *Science* 159, 56–63.

Mewlem, F., Agnew-Blais, J., Taylor, E. & Asherson, P. (2019). “Do different factors influence whether girls versus boys meet ADHD diagnostic criteria? Sex differences among children with high ADHD symptoms”, *Psychiatry research* 272, 765–773. DOI: https://doi.org/10.1016/j.psychres.2018.12.128

Mill, J. S. (1843). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation.* London, England: Longmans, Green, Reader, and Dyer.

Millikan, R. G. (1989). “In Defence of Proper Functions”, *Philosophy of science,* 56(2), 288–302.

Millikan, R. (1999). “Historical Kinds and the “Special Sciences””, *Philosophical Studies* 95, 45–65.

Millikan, R. (2005). “Why Most Concepts Aren’t Categories”. In H. Cohen and C. Lefebvre (eds.): *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier, 305–315.

Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz, Ł., Wachowski, W., Stjernberg, F., Loughlin, V. & Hohol, M. (2018). “From Wide Cognition to Mechanisms: A Silent Revolution.” *Frontiers in Psychology*, *9*, 2393–2393. DOI: https://doi.org/10.3389/fpsyg.2018.02393

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Mitchell, S. (2008). ”Explaining complex behaviour”. In K. Kendler & J. Parnas (eds.) *Philosophical Issues in in Psychiatry: Explanation, Phenomenology, and nosology*. Baltimore: Johns Hopkins University Press, 19–38.

Morris, R. (2018). “Stranger in a strange land: an optimal-environments account of evolutionary mismatch*”. Synthese* 197(9), 4021–4046.

Mowlem, F et. al. (2019). ”Do different factors influence whether girls versus boys meet ADHD diagnostic criteria? Sex differences among children with high ADHD symptoms.” *Psychiatry Research* 272, 765–773.

Murphy, D. (2001). “Hacking’s Reconciliation: Putting the Biological and Sociological Together in the Explanation of Mental Illness”, *Philosophy of the Social Science* 31, 139–161.

Murphy, D. (2006). *Psychiatry in the Scientific Image.* Cambridge, MA: The MIT Press.

Murphy, D. (2008). “Health and Disease.” In Plutynski, A. & Sarkar, S.: *A companion to the philosophy of biology.* Malden, MA: Blackwell Pub.

Murphy, D. (2011). “Conceptual Foundations of Biological Psychiatry”, In Gifford (ed.) *Philosophy of Medicine: Volume 16 in Handbook of the Philosophy of Science*, 425–451.

Murphy, D. (2013). “The Medical Model and the Philosophy of Science”, In Fulford et al. (eds.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press, 966–986.

Murphy, D. (2014). “Natural Kinds in Folk Psychology and in Psychiatry”. In Kincaid and Sullivan (ed.), *Classifying Psychopathology: Mental Kinds and Natural Kinds*. Cambridge, MA: The MIT Press.

Murphy, D. (2015). “’Deviant Deviance’: Cultural Diversity in DSM-5”. In S. Demazeux & P. Singy (eds): *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel*. Dordrecht: Springer, 97–110.

Murphy, D. (2017a). “Can psychiatry refurnish the mind?” *Philosophical Explorations*, 20, 160–174.

Murphy, D. (2017b). “Description and explanation of the culture-bound syndromes”, In K. Kendler & J. Parnas (eds.): *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry*. Oxford: Oxford University Press, 152–165.

Murphy, D. & Stich, S. P. (1999). “Griffiths, elimination and psychopathology”, *Metascience* 8, 13–15.

Murphy, D. & Woolfolk, R. L. (2000). “The Harmful Dysfunction Analysis of Mental Disorder”, *Philosophy, Psychiatry & Psychology*, 7(4), 241–252.

Murphy, J. M. (1976). “Psychiatric labeling in cross-cultural perspective”. *Science*, 191(4231), 1019–1028. DOI: https://doi.org/10.1126/science.1251213

Mäki, U. (2008). “Putnam’s Realisms: A View from the Social Sciences.” In S. Pihlström, P. Raatikainen and M. Sintonen (eds): *Approaching Truth: Essays in Honour of Ilkka Niiniluoto*. *Tributes*, 5, 295–306. London: College publications.

Nichter, M. (1981). ”Idioms of Distress: Alternatives in the Expression of Psychosocial Distress: A Case from South India.” *Culture, Medicine, and Psychiatry* 5, 379–408.

Niiniluoto, I. (2019). "Scientific Progress", In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), URL = <https://plato.stanford.edu/archives/win2019/entries/scientific-progress/>.

Nisbett, R. E., & Norenzayan, A. (2002). ”Culture and cognition.” In H. Pashler & D. Medin (Eds.): *Steven's handbook of experimental psychology: Memory and cognitive processes*. John Wiley & Sons Inc, 561–597.

Nisbett, R. E. (2003). *The Geography of Thought: How Westerners and Asians Think Differently… And Why*. London: Nicholas Brealey Publishing.

Nordenfelt, L. (1986). “Health and disease. Two Philosophical Perspectives”. *Journal of Epidemiology and Community Health* 41, 281–284.

Nordenfelt, L. (1995). ”Om holistiska hälsoteorier.” In Kristian Klockars and Berndt Österman (eds), *Begrepp om hälsa*. Stockholm: Liber utbildning, 29–42.

Nordenfelt, L. (1997). *Talking About Health: A Philosophical Dialogue*. Amsterdam: Rodopi.

National Institute of Mental Health (NIMH) (2011): "NIMH Research Domain Criteria, Draft 3.1: June, 2011" *https://www.nimh.nih.gov/research-funding/nimh-research-domain-criteria-rdoc*

Ortner, S. (2016). "Dark anthropology and its others: Theory since the eighties." *HAU: Journal of Ethnographic Theory* 6(1), 47–73.

Papineau, D. (1994). "Mental Disorder, Illness and Biological Disfunction." *Royal Institute of Philosophy Supplement* 37, 73–82.

Parnas, J. & Sass, L. A. (2008). "Varieties of "Phenomenology": On Description, Understanding, and Explanation in Psychiatry", In K. Kendler & J. Parnas (eds.): *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. Baltimore, Maryland: Johns Hopkins University Press, 239–278.

Parnas, J. & Bovet, P. (2015). "Psychiatry made easy: operation(al)ism and some of its consequences", K. Kendler & J. Parnas (eds.): Philosophical Issues in Psychiatry III: *The nature and sources of historical change*. Oxford: Oxford University Press, 190–212.

Parnas, J. & Urfer-Parnas, A. (2017). "The ontology and epistemology of symptoms: The case of auditory verbal hallucinations in schizophrenia", K. Kendler & J. Parnas (eds.): *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry*. Oxford: Oxford University Press, 201–216.

Patil, T. & Giordano, J. (2010). ”On the ontological assumptions of the medical model of psychiatry: Philosophical considerations and pragmatic tasks.” *Philosophy, Ethics, and Humanities in Medicine*, 5(3). DOI: https://doi.org/10.1186/1747-5341-5-3

Pereira, A. (2020). “Specific Phobia is an Ideal Psychiatric Kind”, *Philosophy, Psychiatry, & Psychology* 27(3), 299–315.

Petryna, A., Lakoff, A. & Kleinman, A. (ed.) (2006). *Global Pharmaceuticals: Ethics, Markets, Practice.* Durham, N.C.: Duke University Press.

Pietikäinen, P. (2013). *Hulluuden historia.* Helsinki: Gaudeamus Helsinki University Press.

Pike, K. M. & Dunne, P. E. (2015). “The Rise of Eating Disorders in Asia: A Review”, *Journal of Eating Disorders,* 3(33). DOI: 10.1186/s40337-015-0070-2

Pillard, R. C. and Bailey, J. M. (1998). “Human Sexual Orientation Has a Heritable Component”, *Human Biology,* 70(11).

Pinto, R., Ashworth, M. & Jones. R. (2008). “Schizophrenia in black Caribbeans living in the UK: an exploration of underlying causes of the high incidence rate”. *British journal of general practice*, 58(551), 429–434.

Pober, J. M. 2013. ”Addiction Is Not a Natural Kind”, *Frontiers in Psychiatry* 4, 123.

Poland, J. (2014). “Deeply Rooted Sources of Error and Bias in Psychiatric Classification”, In Kincaid and Sullivan (ed.), *Classifying Psychopathology: Mental Kinds and Natural Kinds,* 29–64. Cambridge, MA: The MIT Press.

Potter, N. N. (2016). *The Virtue of Defiance and Psychiatric Engagement*. Oxford: Oxford University Press.

Psillos, S. (1999). *Scientific Realism – How Science Tracks Truth.* London: Routledge.

Psillos, S. (2002). *Causation and explanation*. Durham: Acumen.

Putnam, H. (1975). “The Meaning of Meaning”. In *Mind, Language and Philosophy – Philosophical Papers Volume 2*, 215–271. Cambridge: Cambridge University Press.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Pörn, I. (1995). ”Vad är hälsa?” In K. Klockars & B. Österman (eds): *Begrepp om hälsa*, 14–28. Stockholm: Liber utbildning.

Pöyhönen, S. (2010). “Natural Kinds with Extended Mechanisms, Rough Draft”, https://www.ed.ac.uk/files/atoms/files/ppig_extended_mechanisms.pdf, visited on 1 October 2018.

Pöyhönen, S. (2013a). “Carving the Mind by its Joints: Culture-Bound Psychiatric Disorders as Natural Kinds”. In K. Talmont-Kaminsky &

M. Milkowski (eds.): *Regarding the Mind, Naturally: Naturalist Approaches to the Science of the Mental*. Newcastle upon Tyne: Cambridge Scholars Publishing, 30–48.

Pöyhönen, S. (2013b). *Chasing Phenomena: Studies on classification and conceptual change in the social and behavioral sciences.* Philosophical Studies from the University of Helsinki 39. Helsinki: Theoretical Philosophy and Social & Moral Philosophy.

Pöyhönen, S. (2014). "Explanatory Power of Extended Cognition", *Philosophical Psychology* 27, 735–759.

Quine, W. V. O. (1960). *World and Object*. Cambridge, MA: The MIT Press.

Quinn, P. O., & Madhoo, M. (2014). "A review of attention-deficit/hyperactivity disorder in women and girls: uncovering this hidden diagnosis". *The primary care companion for CNS disorders*, *16*(3), PCC.13r01596. DOI: https://doi.org/10.4088/PCC.13r01596

Radden, J. (2003). "Is This Dame Mellancholoy?: Equating Today's Depression and Past Melancholia", *Philosophy, Psychiatry, & Psychology* 10, 37–52.

Raerinne, J. (2011): *Generalizations and Models in Ecology: Lawlikeness, Invariance, Stability, and Robustness.* Doctoral dissertation.

Rashed, M. (2019). *Madness and the Demand for Recognition*. Oxford: Oxford University Press.

Rashed, M. & Bingham, R. (2014). “Can Psychiatry Distinguish Social Deviance from Mental Disorder?”, *Philosophy, Psychiatry, & Psychology* 21, 243–255.

Reicher, S., Spears, R. & Haslam, A. S. (2010). “The Social Identity Approach in Social Psychology.” In M. Wetherell, & C. T. Mohanty (eds.): *Sage Identities Handbook*. London: Sage, 45–62. DOI: http://dx.doi.org/10.4135/9781446200889.n4

Revonsuo, A. (2000). ”The reinterpretation of dreams: An evolutionary hypothesis
of the function of dreaming.” *Behavioral and Brain Sciences*, 23(6), 877–901.

Reydon, T. A. C. (2009): “How to Fix Kind Membership: A Problem for HPC Theory and a Solution”, *Philosophy of Science* 76, 724–736.

Reydon, T. A. C. and Ereshefsky, M. (2022). “How to Incorporate Non-Epistemic Values into a Theory of Classification”, *European Journal for Philosophy of Science* 12. DOI: https://doi.org/10.1007/s13194-021-00438-6

Risjord, M. (2007). “Scientific Change as Political Action. Franz Boas and the Anthropology of Race.” *Philosophy of the Social Sciences,* 37(1), 24–45.

Risjord, M. (2014). *Philosophy of Social Science. A Contemporary Introduction*. London: Routledge.

Robbins, J. (2004). *Becoming Sinners – Christianity and Moral Torment in a Papua New Guinea Society.* Los Angeles: University of California Press.

Root, M. (2000). "How We Divide the World", *Philosophy of Science* 67, 628–639.

Rosenberg, A. (2016). *Philosophy of Social Science*. Boulder, CO: Westview Press.

Rosenberg, C. (2002). "The Tyranny of Diagnosis: Specific Entities and Individual Experience". *The Milbank Quarterly* 80(2), 237–260.

Rosenhan, R. (1973). "On being sane in insane places", *Science* 179, 251–258.

Rubinstein E. (2016). "Emplotting Hikikomori: Japanese Parents' Narratives of Social Withdrawal." *Culture, medicine and psychiatry* 40(4), 641–663. DOI: https://doi.org/10.1007/s11013-016-9495-6

Rudner, R. (1953). "The Scientist Qua Scientist Makes Value Judgments". *Philosophy of science* 20(1), 1–6.

Sadler, J. Z. (2013). "Values in Psychiatric Diagnosis and Classification", In Fulford et al. (eds.): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press.

Sadler, J. Z. (2015). "Ethics and Values in Diagnosing and Classifying Psychopathology", In Sadler, J., Fulford, K. & van Staden, C. (eds.), *The Oxford Handbook of Psychiatric Ethics*. Oxford: Oxford University Press.

Sahlins, M. (1985). *Islands of History.* Chicago: University of Chicago Press.

Salmela, M. (2004). ”Terveyden filosofia tehokkaan terveydenhuollon kätilönä”, *Niin & Näin* 3/2004.

Sankey, H. (1994). *The Incommensurability Theses*. Aldershot: Avebury.

Sarkia, M., Kaidesoja, T. & Hyyryläinen, M. (2020). ”Mechanistic explanations in the cognitive social sciences: Lessons from three case studies”, *Social Science Information* 59, 580–603.

Schaffner, J. (2006). “Reduction: The Cheshire cat problem and a return to roots“, *Synthese* 151(3), 377–402.

Schaffner, J. (2008). “Etiological models in psychiatry: Reductive and nonreductive.” In K. Kendler & J. Parnas (Eds.): *Philosophical issues in psychiatry: Natural kinds, mental taxonomy and causation.* Baltimore: Johns Hopkins University Press, 48–90.

Schaffner, J. (2013). “Reduction and Reductionism in Psychiatry”. In Fulford et al. (eds): *The Oxford Handbook of Philosophy and Psychiatry.* Oxford: Oxford University Press, 1003–1022.

Schaffner, J. (2015). “Contrastive Causation”, *The philosophical Review* 114, 297–328.

Schaffner, J. & Tabb, K. (2015). “Hempel as a critic of Bridgman’s operationalism: lessons for psychiatry from the history of science”, In

K. Kendler and J. Parnas (eds.) Philosophical Issues in Psychiatry III: *The nature and sources of historical change*. Oxford: Oxford University Press, 213–220.

Scheff, T. J. (1966). *Being Mentally Ill*. Chicago: Aldine.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T. & Vogeley, K. (2013). ”Toward a second-person neuroscience”, *The Behavioral and brain sciences*, 36(4), 393–414. DOI: https://doi.org/10.1017/S0140525X12000660

Schilbach, L. (2016). ”Towards a second-person neuropsychiatry.” *Phil. Trans. R. Soc*, B371(1686). DOI: http://doi.org/10.1098/rstb.2015.0081

Schuur, R. (2019). “Mental Health and Illness: Past Debates and Future Directions”. In Tekin & Bluhm (ed.): *The Bloomsbury Companion to Philosophy of Psychiatry.* London: Bloomsbury Academic.

Schwartz, P. H (2007a). ”Decision and Discovery in Defining “Disease””. In H. Kincaid & J. McKitrick (eds.): *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Biomedical Science*. Springer 47–63.

Schwartz, P. H. (2007b). “Defining dysfunction: Natural selection, design, and drawing a line”, *Philosophy of Science* 74, 364–385.

Schwartz, P. H. (2014). “Reframing the disease debate and defending the biostatistical theory”. *The Journal of Medicine and Philosophy* 39(6), 572–589.

Schwitzgebel, E. (2002). “Why Did We Think We Dreamed in Black and White?”, *Studies In History and Philosophy of Science Part A*, 33, 649–660. DOI: 10.1016/S0039-3681(02)00033-X.

Searle, J. R. (1996). *The Construction of Social Reality*. London: Penguin.

Sedgwick, P. (1973). “Illness: Mental and Otherwise”, *The Hastings Center Studies* 1, 19–40

Segall, M., Campbell, D. & Herskovits, M. J. (1966). *The influence of culture on visual perception.* New York: Bobbs-Merrill.

Seligman, R. & Kirmayer, L. J. (2008). “Dissociative Experience and Cultural Neuroscience: Narrative, Metaphor and Mechanism”, *Culture, Medicine, and Psychiatry* 32, 31–64.

Semin, G. & Echterhoff, G. (Eds.) (2010). *Grounding sociality: Neurons, mind, and culture.* London: Psychology Press.

Siegel, J. (1997). *Fetish, Recognition, Revolution.* Princeton: Princeton University Press.

da Silva Carneiro, S. C, Pirmez, R., de Hollanda, T. R., Cuzzi, T. & Ramos-e-Silva, M. (2013). “Syphilis Mimicking Other Dermatological Diseases: Reactive Arthritis and Mucha-Habermann Disease”, *Case Reports in Dermatology* 5, 15–20. DOI: 10.1159/000346745

Simons, R. (1996). *Boo! Culture, Experience, and the Startle Reflex*. Oxford: Oxford University Press.

Slater, M. H. (2015). “Natural Kindness”, *British Journal for the Philosophy of Science* 66, 375–411.

Solomon, M. (2017). “On the appearance and disappearance of Asperger’s syndrome”, In Kendler, K. & Parnas, J. (eds.): *Philosophical Issues in Psychiatry IV. International Perspectives in Philosophy and Psychiatry*. Oxford: Oxford University Press, 176–186.

Solomon, M. (2022). ”On Validators for Psychiatric Categories”, *Philosophy of Medicine*, 3(1). DOI: https://doi.org/10.5195/pom.2022.74

Spitzer, R, L. & Endicott, J. (1978). “Medical and mental disorder: Proposed definition and criteria”, In R. L. Spitzer and D. F. Klein (Eds.), *Critical Issues in Psychiatric Diagnosis*. New York, NY: Raven Press, 15–40.

Spitzer, R. L. & Wilson, P. (1975). “Nosology and the official psychiatric nomenclature”. In A. Freedman, H. Kaplan & B. Sadock (eds.): *Comprehensive textbook of psychiatry* 1, 826–845. Baltimore, MD: Williams & Wilkins.

Steel, D. (2006). “Methodological Individualism, Explanation, and Invariance”, *Philosophy of the Social Sciences* 36, 440–463.

Steel, D. (2008). *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Steel, D. (2010). “Cartwright on Causality: Methods, Metaphysics and Modularity”. *Economics and Philosophy* 26, 77–86.

Stegenga, J. (2018). *Medical nihilism*. Oxford: Oxford University Press.

Sterelny, K. (2010). “Minds: extended or scaffolded”, *Phenomenology and Cognitive Science* 9, 465–481.

Stich, S. (1992). ”What is a theory of mental representation?” *Mind* 101(402), 243–61.

Stich, S. & Bishop, M. (1998): “The Flight to Reference, or How Not to Make Progress in the Philosophy of Science”, *Philosophy of Science* 65, 33–49.

Stich, S. P. & Mallon, R. (2000). ”The odd couple: The compatibility of social construction and evolutionary psychology”. *Philosophy of Science* 67(1), 133–154.

Stich, S., Mallon, R., Nichols, S. & Machery, E. (2004). “Semantics, Cross-Cultural Style”, *Cognition* 92(3):B1-B12. DOI: 10.1016/j.cognition.2003.10.003.

Sullivan, J. (2014). “Stabilizing Mental Disorders: Prospects and Problems”, In H. Kincaid & J. Sullivan (eds.): *Classifying Psychopathology: Mental Kinds and Natural Kinds.* Cambridge, MA: The MIT Press, 257–281.

Sullivan, J. (2017). “Models of Mental Illness”. In Solomon, S. & Kincaid (eds.): *The Routledge Companion to Philosophy of Medicine*. New York: Routledge, 455–464.

Sun, R. (2012). “Prolegomena to Cognitive Social Science”. In R. Sun (ed.): Grounding Social Sciences in Cognitive Sciences. Cambridge, MA: The MIT Press, 3–32.

Szasz, T. (1961). *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*. New York, NY: Harper & Row.

Szasz, T. (1999). *Insanity*. New York: Syracuse University Press.

Tabb, K. (2019). “Philosophy of psychiatry after diagnostic kinds”, *Synthese* 196, 2177–2195.

Tabb, K. & Schaffner, K. F. (2017). ”Causal pathways, random walks and tortuous paths: moving from the descriptive to the etiological in psychiatry.” In K. S. Kendler & J. Parnas (eds): *Philosophical Issues in Psychiatry IV: Nosology.* Oxford: Oxford University Press, 332–340.

Taylor, C. (1971). “Interpretation and the Sciences of Man”. *Review of Metaphysics* 25, 1–51.

Tekin, Ş. (2014). “The Missing Self in Hacking’s Looping Effects”. In H. Kincaid & J. A. Sullivan (eds.): *Classifying Psychopathology: Mental Kinds and Natural Kinds.* Cambridge, MA: The MIT Press, 227–256.

Tekin, Ş. (2016). “Are Mental Disorders Natural Kinds?: A Plea for a New Approach to Intervention in Psychiatry”. *Philosophy, psychiatry & psychology* 23(2), 147–163.

Telakivi, P. (2020). *Extending the Extended Mind: From Cognition to Consciousness*. Philosophical Studies from the University of Helsinki 49. University of Helsinki.

Thagard, P. (2000). *How Scientists Explain Disease.* Princeton: Princeton University Press.

Thakker, J. & Ward, T. (1998). "Culture and Classification: The Cross-Cultural Application of the DSM-IV". *Clinical Psychology* 18, 501–529.

Thomas, N., Hayward, M. & Peters, E. et al. (2014). "Psychological Therapies for Auditory Hallucinations (Voices): Current Status and Key Directions for Future Research", *Schizophrenia Bulletin* 40. Suppl. 4, 202–212.

Tikkinen, K. A. O., Rutanen, J., Frances, A., Perry, B. L., Dennis, B. B., Agarwal, A., Maqbool, A., Ebrahim, S., Leinonen, J. S., Järvinen, T. L. N., & Guyatt, G. H. (2019). "Public, health professional and legislator perspectives on the concept of psychiatric disease: a population-based survey". *BMJ Open*, *9*(6), e024265–e024265. https://doi.org/10.1136/bmjopen-2018-024265

Tiles, M. (1993). "The Normal and Pathological: The Concept of a Scientific Medicine". *The British Journal for the Philosophy of Science.* 44(4), 729–742.

Tobin, E. (2018). "Mechanisms and Natural Kinds", In S. Glennan & P. Illary (eds): *Routledge Handbook of Mechanisms and Mechanical Philosophy*. New York: Routledge, 198–210.

Tsou, J. (2007). Hacking on the Looping Effects of Psychiatric Classifications: What Is an Interactive and Indifferent Kind? *International Studies in the Philosophy of Science* 3, 329–344.

Tsou, J. Y. (2020). "Social Construction, HPC Kinds, and the Projectability of Human Categories". *Philosophy of the social sciences* 50(2), 115–137.

Tsou, J. Y. (2021). *Philosophy of Psychiatry*. Cambridge: Cambridge University Press.

Tuomela, R. (1977). *Human Action and Its Explanation: A Study on the Philosophical Foundations of Psychology*. Dordrecht: Reidel.

Van Bouwel, J., Weber, E. & De Vreese, L. (2011). "Indispensability arguments in favour of reductive explanations", *Journal for General Philosophy of Science*, 42, 33–46.

Van Bouwel, J. (2014). "Pluralists About Pluralism? Different Versions of Explanatory Pluralism in Psychiatry" In M. C. Galavotti, D. Dieks, W. Gonzalez, S. Hartmann, T. Uebel & M. Weber (eds.): *New Directions in Philosophy of Science (The Philosophy of Science in a European Perspective Series).* Springer, 105–119.

Van Fraassen Bas, C. (1980). The Scientific Image. Oxford, England: Oxford University Press.

Van Riel, R. (2017). "Mental Disorders and the Indirect Construction of Social Facts". *Journal of Social Ontology,* 3, 27–48.

Van Loo, H., Romeijin, J.-W., de Jonge, P., & Schoevers, R. (2013). "Psychiatric Comorbidity and Causal Disease Models", *Preventive Medicine,* 57, 748–752.

van Wynsberghe, A. (2016). *Healthcare robots: ethics, design and implementation*. Routledge, Taylor & Francis Group.

Vesterinen, T. (2021). "Identifying the Explanatory Domain of the Looping Effect: Congruent and Incongruent Feedback Mechanisms of Interactive Kinds", *Journal of Social Ontology*, 6(2), 159–185. DOI: https://doi.org/10.1515/jso-2020-0015

Wakefield, J. C. (1992a). "The Concept of Mental Disorder: On the Boundary Between Biological Facts and Social Values". *The American psychologist* 47 (3), 373–388.

Wakefield, J. (1992b) "Disorders as harmful dysfunction: A conceptual critique of D.S.M-II-R's definition of mental disorder". *Psychological Review* 99, 232–247.

Wakefield, J. (1999). "Disorder as a black box essentialist concept". *Journal of Abnormal Psychology*, 108, 465–471.

Wakefield, J. (2021). *Defining mental disorder: Jerome Wakefield and his critics*.

Walter, H. (2013). "The third wave of biological psychiatry". *Frontiers in Psychology*, 4, 1–8.

Washington, N. (2016). “Culturally Unbound: Cross-Cultural Cognitive Diversity and the Science of Psychopathology”, *Philosophy, Psychiatry, & Psychology* 23(2), 165–179.

Watters, E. (2011). *Crazy Like Us. The Globalization of the Western Mind*. London: Constable & Robinson Ltd.

Werkhoven, S. (2021). Natural kinds of mental disorder. *Synthese* 199, 10135–10165. https://doi.org/10.1007/s11229-021-03239-9

Westermeyer, J. & Wintrob, R. (1979). “”Folk” Criteria for the Diagnosis of Mental Illness in Rural Laos: On Being Insane in Sane Places”, *American Journal of Psychiatry* 136, 755–761.

Wilkinson, S. (2023). *Philosophy of Psychiatry. A Contemporary Introduction.* New York: Routledge.

Williams, R. (1978). M*arxism and Literature.* Oxford: Oxford University Press.

Winch, P. (1958). *The Idea of Social Science*. London: Routledge & Kegan Paul Ltd.

Winzeler, R. L. (1995). *Latah in Southeast Asia: The History and Ethnography of a Culture-Bound Syndrome.* Cambridge: Cambridge University Press.

Winzeler, R. L. (1999). “Is *Latah* Always Fakery and Deception?”, *Transcultural Psychiatry*, 36, 385–390.

Wittgenstein, L. (1953). *Philosophical Investigations.*

Woods, A., Jones, N., Bernini, M. and Callard, F. et al. (2014). “Interdisciplinary Approaches to the Phenomenology of Auditory Verbal Hallucinations”, *Schizophrenia Bulletin* 40, Suppl. 4, 246–254.

Woodward, J. (1984). ”A theory of singular causal explanation.” *Erkenntnis* 21(3), 231–262.

Woodward, J. (2000). “Explanation and Invariance in the Special Sciences”, *British Journal for the Philosophy of Science* 5, 197–254.

Woodward, J. (2002). “What is a Mechanism? A Counterfactual Account”, *Philosophy of Science* 69, p. 366–377.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, J. (2006). “Sensitive and Insensitive Causation”, *The Philosophical Review* 115, 1–50.

Woodward, J. (2013), ”II—Mechanistic Explanation: Its Scope and Limits”. *Aristotelian Society Supplementary Volume*, 87, 39–65. https://doi.org/10.1111/j.1467-8349.2013.00219.x

Woodward, J. (2015). “Cause and Explanation in Psychiatry: An Interventionist Perspective”. In Kendler, K. S. & Parnas, J. (eds.): *Philosophical Issues in Psychiatry: Explanation, Phenomenology, and Nosology*. Baltimore, MD: John Hopkins University Press, 132–195.

Woodward, J. (2020). “Levels: What Are They and What Are They Good For?”. In K. Kendler, J. Parnas & P. Zachar (Eds.): *Levels of*

Analysis in Psychopathology: Cross-Disciplinary Perspectives. Cambridge: Cambridge University Press. DOI:10.1017/9781108750349

World Health Organization (1992). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization, Division of Mental Health.

World Health Organization (2022). *The ICD-11 Classification of Mental and Related Health Problems.* Geneva: World Health Organization, Division of Mental Health.

von Wright, G. H. (1971). *Explanation and Understanding*. London: Routledge & Kegan Paul Ltd.

Wrigley, A. (2007). "Realism and Anti-Realism about Mental Illness", *Philosophical Papers* 36, 371–397.

Zachar, P. (2000): "Psychiatric Disorders Are Not Natural Kinds", *Philosophy, Psychiatry, and Psychology* 7, 167–182.

Zachar, P. (2008). "Real kinds but no true taxonomy: An essay in psychiatric systematics." In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry: Explanation, phenomenology, and nosology*. Johns Hopkins University Press, 327–367.

Zachar, P. (2014a). "Beyond Natural Kinds: Toward a "Relevant" "scientific" Taxonomy in Psychiatry." In Kincaid and Sullivan (ed.), *Classifying Psychopathology: Mental Kinds and Natural Kinds.* Cambridge, MA: The MIT Press.

Zachar, Peter (2014b): *A Metaphysics of Psychopathology*. Cambridge, MA: The MIT Press.

Yap, P. M. (1951) (2000). "Mental Diseases Peculiar to Certain Cultures: a Survey of Comparative Psychiatry", In R. Littlewood & S. Dein (eds.): *Cultural Psychiatry & Medical Anthropology*: *An Introduction and Reader*. London: The Atholone Press.

Yap, P. M. (1952). "The Latah Reaction: Its Pathodynamics and Nosological Position", *The Journal of Mental Science* 98, 515–564.

Yap, P. M. (1966). The Culture-bound Reactive Syndromes", In W. Caudill and Tsung-Yi Lin (eds.), *Mental Health Research in Asia and The Pacific*. Honolulu: East-West Center Press, 35–53.

Ylikoski, P. (2001). *Understanding Interests and Causal Explanation.* Ph.D. thesis. Department of Social and Moral Philosophy, University of Helsinki. https://helda.helsinki.fi/bitstream/handle/10138/21811/understa.pdf?sequence=, visited on 1 September 2018.

Ylikoski, P. (2003). "Ian Hacking and ilmiöiden synty", *T&E* 1, 12–30.

Ylikoski, P. (2007). "The Idea of Contrastive Explanandum". In J. Persson & P. Ylikoski (eds.): *Rethinking Explanation. Boston studies in the philosophy of science,* 252, 27–42. Dordrecht: Springer.

Ylikoski, P. K. (2012). "Micro, Macro, and Mechanisms." In H. Kincaid (Ed.), *The Oxford Handbook of Philosophy of Social Science.* Oxford University Press, 21–45.

Ylikoski, P. K. (2013). Causal and Constitutive Explanation Compared. *Erkenntnis*, 78(2), 277–297. DOI: https://doi.org/10.1007/s10670-013-9513-9

Ylikoski, P. (2017). "Social Mechanisms". In Glennan, S. & Illari, P. (eds), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*, 30. Routledge Handbooks in Philosophy, Routledge, Abingdon, 401–412.

Ylikoski, P. (2018). "Selittäminen, ymmärtäminen ja kausaaliset mekanismit", In Kaidesoja, T. Kankainen, T. & Ylikoski, P. (eds): *Syistä selityksiin: Kausaalisuus ja selittäminen yhteiskuntatieteissä*. Helsinki: Gaudeamus.

Ylikoski, P. (2021). "Understanding the Coleman Boat". In Manzon, G. & Elgar, E. (eds.), *Research Handbook on Analytical Sociology,* 49–63. Cheltenham: Edward Elgar Publishing Limited.

Ylikoski, P. & Kokkonen, T. (2010). *Evoluutio ja ihmisluonto*. Helsinki: Gaudeamus.

Ylikoski, P. & Kuorikoski, J. (2010): "Dissecting Explanatory Power", *Philosophical Studies* 148, 201–219.

Ylikoski, P. K., & Kuorikoski, J. (2012). "Explanatory relevance across disciplinary boundaries." In C. Marchionni, & J. Vromen (Eds.), *Neuroeconomics: Hype or Hope?* Routledge, 119–128.

Ylikoski, P. & Pöyhönen, S. (2015). "Addiction-as-a-Kind-Hypothesis", *International Journal of Alcohol and Drug Research,* 4(1), 21–25.

Young, A. (1995). *The Harmony of Illusions. Inventing Post-Traumatic Stress Disorder*. Princeton: Princeton University Press.

Young, A. & Breslau, N. (2016): "What is "PTSD"? The Heterogeneity Thesis". In D. E. Hinton & B. J. Good (Eds.): *Culture and PTSD: Trauma in global and historical perspective*. University of Pennsylvania Press, 135–154. https://doi.org/10.9783/9780812291469-004