Synthese (2021) 199:10135–10165
https://doi.org/10.1007/s11229-021-03239-9

ORIGINAL RESEARCH

Natural kinds of mental disorder

Sander Werkhoven[1]

Received: 11 November 2020 / Accepted: 29 May 2021 / Published online: 9 July 2021

Abstract
Are mental disorders (autism, ADHD, schizophrenia) natural kinds or socially constructed categories? What is at stake if either of these views prove to be true? This paper offers a qualified defence for the view that there may be natural kinds of mental disorder, but also that the implications of this claim are generally overestimated. Especially concerns about over-inclusiveness of diagnostic categories and medicalisation of abnormal behaviour are not addressed by the debate. To arrive at these conclusions the paper opens with a discussion of kind formation in science, followed by an analysis of natural kinds. Seven principled and empirically informed objections to the possibility of natural kinds of mental disorder are considered and rejected. The paper ends with a reflection on diagnostics of mental health problems that don't fall into natural kinds. Despite the defence of the possibility of natural kinds of mental disorder, this is likely to be the majority of cases.

Keywords Mental disorders · Natural kinds · Social constructivism · Taxonomy · Validity · Normativism · Looping effects · Multi-factorial causality · Particularism

1 Introduction

The use of diagnostic categories in psychiatry and clinical psychology (ADHD, autism spectrum disorder, schizophrenia, etc.) remains heavily contested, inside and outside of academic contexts. The target of critics is typically the *Diagnostic and Statistical Manual* (*DSM*), the taxonomy of psychiatric disorders issued by the American Psychiatric Association (2013), currently in its 5th edition.

This paper has been written in the context of the interdisciplinary research project 'Developmental Labels', funded by the Dynamics of Youth Strategic Theme of Utrecht University. I am thankful to my fellow researchers, to MA students who have commented on early drafts of this paper, and to Martin Lipman for helping my fine-tune my understanding of Kripke's work. Finally, I want to express my thanks to the helpful comments and suggestions of the anonymous reviewers.

✉ Sander Werkhoven
s.werkhoven@uu.nl

[1] Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands

The categories listed in the *DSM* are routinely criticised for their heterogeneity (wildly different conditions fall under the same category), high levels of co-morbidity (people often fall under multiple categories), over-inclusiveness and false positives (people are diagnosed with a mental disorders that shouldn't be), poor inter-rater reliability (different diagnosticians come to different conclusions about the same cases), categoricity in diagnosis rather than continuity (people are 'in or out' of a category, instead of located on a spectrum), insufficient basis in neuroscience and genetics (findings in neuro-imaging and genetics show little correlation with present diagnostic categories), as well as their stigmatising effects (people labelled with a mental disorder are often treated as inherently inferior, weird, crazy, wrong, etc.).[1] Critics of a more sociological bent protest that mental health problems are presented as individual problems, blinding us to larger societal problems behind mental health issues (poverty, racism, housing, labour conditions, drug use, neo-liberalism, complexity of modern life, etc.) that should really be the focus of our analyses and interventions.

In recent years, these familiar grievances have sparked a philosophical debate about a more specific question: whether categories of psychopathology are natural kinds or whether they should be considered conventional or socially constructed kinds.[2] Natural kinds are classes of things, properties, or processes that exist independently of us and our classifying activities—they represent the kinds present in nature and so can be discovered, typically through scientific investigation (chemical elements are the paradigm case). Socially constructed kinds (or conventional kinds), by contrast, are classes that aren't there in nature for us to discover, but groupings and categories of things that we have devised and, in a sense, impose onto nature—typically reflecting some interest or pragmatic purpose of us, the classifiers (weeds versus plants being a paradigmatic example).

The implications of the debate are generally considered to be significant. If mental disorders are natural kinds, many think we can say with confidence that people falling under them *really are* mentally disordered in the specified way. Worries about medicalisation and over-inclusion of mental health categories could then finally be cast aside. Labelling people with ADHD or autism wouldn't amount to expanding medical power or promoting the interests of psychiatrists, pharmaceutical companies, or whoever—it would just be a way of describing people as they really are, at that point in time. If categories of psychopathology were socially constructed kinds, by contrast, the general view is that the domain of psychopathology would fall prey to a host of troubling questions: *Whose* interests and *which* purposes are reflected in the use of these categories? Are we better off revising or discarding any of the categories, seeing it's all up to us anyway? Which institutions are involved in the production and maintenance of the kinds that are recognised? To what extent

[1] See Poland et al. (1994) for an early expression of these concerns.

[2] See especially BeeBee & Sabbarton-Leary (2010), Cooper (2005, pp. 45–76, 2007, pp. 44–66), Hacking (1995b, 2000, 2007a, b), Haslam (2000, 2002, 2014), Hyman (2010), Kendler et al. (2011), Kendler & Parnas (2008, 2011), Kincaid & Sullivan (2014), Pickard (2009) and Zachar (2000a, b, 2003, 2014).

does knowledge produced by medical sciences promulgate existing categories and the interests invested in them?

The debate over the status of kind-terms in psychiatric taxonomy has a descriptive part and a normative part. On the descriptive side, the question is whether there are diagnostic categories that can be considered natural kinds. This question may be asked about the current DSM/ICD taxonomy or about some future taxonomy, e.g. a diagnostic system based on the RDoC initiative by the NIMH.[3] The normative debate turns on the question whether psychiatry should aim for a taxonomy that distinguishes natural kinds and what to do with kind terms that aren't. The majority view among psychiatrist and philosophers seems to be that *current* diagnostic categories aren't natural kinds: they fail to map onto causal structures of human psychology and therefore cannot (causally) explain the relevant symptoms or predict the course of disorders and responsiveness to treatment.[4] The views on whether an *alternative* taxonomy could contain natural kinds are more divided, turning on several principled objections to the very possibility of natural kinds of mental disorder.

On this latter question, this paper offers a qualified defence of the natural kind view. On the descriptive side, the central claim will be that empirical research may well find natural kinds of mental disorder, even on a restrictive Kripkean view of natural kinds. None of the principled and empirically informed objections succeed in undermining this possibility. At the same time, however, there are good reason to think that the majority of mental health problems don't fall into natural kinds. On the normative side, the claim is that psychiatric taxonomy should aim for distinguishing natural kinds (with some important qualifications). In arguing for these claims, it will also be shown that many overestimate the implications of these points. Perhaps the most significant purpose of this paper, then, is that there is a lot less at stake in these debates than is normally presumed.

To specify these general claims, let me offer the central coordinates of the paper: (1) once natural kinds are defined as a class in which all members (and only those) share the same identity determining causal properties, then it is an empirical matter whether there are natural kinds of mental disorder—this cannot be decided in advance (2) heterogeneity of symptoms does not speak for or against the validity or naturalness of diagnostic categories (3) the (in)existence of natural kinds of mental disorder does not affect whether certain people are mentally disordered; so even if the construct for some condition can be proven to be a natural kind, this does not mean it is a disorder (4) therefore, the natural kinds debate does not resolve concerns about medicalisation and over-inclusiveness of the psychopathological domain; (5) for a disorder to be a natural kind, neuro-biological causes need not be present or discovered; higher level structures may provide an equally suitable explanatory causal basis (6) therefore, it becomes possible to defend special cases of socially

[3] See Insel et al. (2010) and Insel (2013).

[4] See for instance Craver (2009), Haslam (2000, 2002), Insel (2013), Sullivan (2014) and Tsou (2015, 2019). A quantitative study shows that both novice and practicing psychiatrists by and large don't think that (present) categories of mental disorders are real and natural, with shared underlying processes (Ahn et al., 2006).

caused conditions as natural kinds (7) psychiatry should strive for a taxonomy based on natural kinds (8) absent natural kinds of mental disorder, causal factors can be determined at an individual level, supporting a personalised psychiatry.

The paper will proceed as follows. First a positivist picture of scientific kind-formation will be presented in Sect. 2, followed by a more detailed discussion of natural kinds in Sect. 3. These sections contain the philosophical support for claims (1)–(5) and offer a *prima facie* reason to think that there may well be natural kinds of mental disorder. Section 4 addresses seven principled and empirically informed objections to the idea that there can be natural kinds of disorder. Addressing these objections will give support claim (6)–(8).

2 Fundamentals of taxonomy

A simplified, positivist picture of class formation in medicine distinguishes four stages.[5] Psychiatric taxonomy starts with the recognition that certain mental conditions, forms of behaving, or types of experience are viewed as falling short of generally accepted norms and standards of mental functioning (see Gorenstein, 1992, p. 15; Murphy, 2006, pp. 11, 151–155).[6] The starting point of psychiatric classification is therefore an *evaluative* stance vis-a-vis a set of mental conditions and ways of behaving.[7] The conditions that fall short of the relevant norms make up the complete set of those considered psychopathological, mentally disordered, mentally ill—terms I will take to be equivalent. This set may change over time as new ways of behaving become available or the relevant norms change. Importantly, the set of people considered mentally ill does not need to have anything more in common other than their shared failure to conform to the relevant norms and standards.[8]

Although the *domain* of mental illness is carved out on the basis of a type of normative judgement, this does not say much about the *particular* categories of mental illness—schizophrenia, autism spectrum disorder, ADHD, and the like. As Gorenstein puts it, particular categories denote "hypothesised properties of mental functioning" that can be supported or undermined by empirical research (1992, p. 15). The higher-level class of mentally disordered may be based on evaluative

[5] The picture has its roots in the work of Hempel (1965) and finds clear expression in two book-length works on mental illness: Gorenstein's *the Science of Mental Illness* (1992) and Murphy's more recent *Psychiatry in the Image of Science* (2006). Also note that the suggested stages are methodological rather than temporal. It may well be true that what I call stage one and two happen simultaneous, or in reverse order.

[6] Murphy, following Wakefield (1992), calls this a two-staged view. Murphy nevertheless remains hesitant about the implications of this normative starting point and brushes over it to advocate his preferred objectivism.

[7] Accepting this point does not commit one to normativism about health (see Engelhardt, 1976, 1986; Goosens, 1980; Cooper, 2002). A domain may be initially carved out by evaluative judgement but turn out to be reducible to naturalistic facts. I will return to this point in Sect. 4.7 below.

[8] It is Gorenstein's view that mentally ill people, in fact, do not have anything more in common than being viewed as mentally ill. I don't share this view. But for the present purposes this disagreement is irrelevant.

judgement, the particular classes of mental disorder aim to be descriptive, empirically founded, and explanatory in purpose (see also Pickard, 2009, p. 87; Cooper, 2005, p. 46). I shall return to this in Sect. 4.7.

Secondly, a subset of the people considered mentally ill is grouped together. This initial grouping may be based on a variety of factors, most commonly similarity and covariance of symptoms, but also aetiological factors, responsiveness to interventions, time of onset, and lots of other features. Before the DSM-III, psychoanalytic theories about aetiology informed the grouping into kinds of disorder; since the DSM-III it is covariance of symptoms that forms the basis for distinctions in kind.[9] The important point here is that such factors give reason to think that the observed phenomena reflect the presence of a particular mental property, or set of mental processes, causally responsible for the exhibited behaviour—properties or processes hypothesised to be present in all people falling under the category. These groupings will also have exemplars: idealized representations, or prototypes, of the disorder that exhibit a typical course and clinical profile that members of the kind resemble to varying degrees (Murphy, 2006, pp. 13, 205–207). The initial grouping together with its prototypical exemplars gives rise to a hypothetical construct, normally referred to with a name or label: phobia, hysteria, autism, ADHD, and antisocial personality disorder all being examples.

When hypothetical constructs are formed they begin to serve as explanations; they give answers to certain 'why' questions. When asked *why* someone lacks motivation, shows negative moods and entertains repeated thoughts of self-harm, the answer may be: 'because she suffers from depression'. When asked why someone displays flat emotions, has perceptual abnormalities and exhibits unusual immobility, one might answer 'because she has schizophrenia'. The question that immediately arises, however, is how invoking labels like depression and schizophrenia—denoting only a category of people—could ever *explain* these phenomena. If the term is only a one-word summary for features that are only found to co-vary frequently, how does applying the term explain those very same features? Saying someone behaves and feels the way she does *because* she is depressed, even though she is depressed because of how she behaves and feels, makes for an obvious circularity, carrying only an illusion of explanation.

The answer is that terms denoting hypothetical constructs have a meaning beyond the set of behavioural features that indicate their presence (Gorenstein, p. 66). That is, constructs enable *predictions,* in this case, about other aspects of a person's mental functioning, future development, response to environmental changes, and so on. The predictive power of a construct is based on the number of relevant inductive inferences it supports. In inductive reasoning, findings from particulars are generalised to a wider group. If the construct of a mental disorder proves to have significant predictive power, it is no longer merely a label that summarises previous observations: it also supports other predictions about the person. The predictive power of a construct underpins its explanatory role.

[9] See Tsou (2016) for further discussion and a helpful historical overview.

The process whereby a construct is tested and adjusted in light of its predictive success is standardly known as 'scientific validation' and may be described as a third stage of scientific kind-formation. In order to test whether a construct of like Schizophrenia or ADHD has predictive power, a criterion of application (or operational definition) is required. A criterion of application is a decision-procedure that enables the determination whether, or to what extent, the construct is applicable to an individual. Diagnostic criteria like those contained in the DSM should therefore not be confused with a *definition* of the disorder, or with a full *description* of the construct; they are only tools to determine whether the construct applies and the individual belongs to the category (see Gorenstein, p. 72). Diagnostic criteria are a subclass of observable phenomena that should enable a reliable determination of whether the construct applies. The difficulty lies in finding criteria sensitive enough to pick out the specific group to which the construct should apply, but that also aren't overly specific, making the selection too restrictive. The criteria of application also shouldn't contain too much information—certainly not everything we believe to be true about the construct—as that would render the construct useless for any sort of prediction, i.e. the very features that give the construct its validity. Finally, the criteria must be clear and precise, leaving little space for diverging interpretation, so that different diagnosticians reach the same conclusions about the same individuals, making it an overall reliable decision-procedure.

With a criterion of application, it becomes possible to separate a group of people and test predictions about the construct: can anything be said about all members of the group, and only those members of the group, other than meeting the criteria of application? The kind of predictions to be tested here are in principle unlimited: scores on cognitive and emotional tests, presence of other symptoms, levels of functioning over time, neurological properties, and so on. The process of validation, then, boils down to measuring relationships and correlations between indicators of the various constructs we recognise, and the continuous attempt to validate them in conjunction (Murphy, 2006, p. 218). The more a construct helps to predict other features of people in the category, the greater its validity.

As many have pointed out, the method of validation remains inherently limited. A construct like self-esteem, for instance, may be predictively useful for all sorts of measurable properties—like pregnancy rate, bodily height, income levels, and so on—thereby gaining a degree of validity and explanatory power (Murphy, 2006, pp. 219–220).[10] But these explanations don't cut very deep, seeing that it's only group membership and the (probabilistic) inductive inferences it supports that do all the explanatory work. Crucially, it doesn't explain *why* these correlations exist. Validated kind-terms do not provide a *causal* explanation of the observed features and established correlations. If, by contrast, the neuro-psychological mechanisms behind self-esteem were to be discovered, this would allow for an explanation why some of the correlations and predictions hold up. To arrive at constructs with genuine predictive power and explanatory import, it is therefore required to know the responsible causal mechanisms.

[10] Even more sceptically, Hacking writes: "we try to correlate autism with everything, not excluding the relative lengths of the mother's fingers and testosterone in the foetus" (Hacking 2007a, p. 209).

As Hempel already emphasised (and recommended to the APA in the 1960's): only when classifications are informed by insights into underlying causal mechanisms instead of shared observable characteristics, the scientific field moves into its mature phase (Hempel, 1965, pp. 149–151).[11] The fourth stage to be distinguished is therefore a move away from validating constructs by mere predictive power and to anchor categories on the causal structures that give rise to, and therefore causally explain, the traits and behaviours deemed pathological.

It is no surprise that many highlight the importance of this fourth stage, seeing that the current *DSM-5* is still predominantly symptom-based and has not succeeded in fixing its categories on the underlying causal mechanisms. Many authors, including those advocating of the RDoC framework, write as if the transition towards a causally informed system of categorisation requires a radical break and reorientation in psychiatric taxonomy (Insel, 2013; Insel et al., 2010). But the transition isn't all that revolutionary. Distinguishing mental disorders on causal structures is a natural outgrowth of the guiding idea of scientific validation—at least, this was Hempel's view. A symptom-based construct containing different underlying processes and mechanisms is unlikely to yield great predictive success.[12] Validity increases when categories of mental disorders are distinguished on underlying causal structures. Somewhat ironically, then, positivists like Gorenstein and Murphy plead for a return to a taxonomy of mental disorders based on causal processes, similar in principle (though not in content) to the pre-DSM-III categorisations in psychopathology.

Many people believe that the properties of the central nervous system are the relevant causal structures underpinning valid psychiatric categories. These neurological properties may be present without there being any symptoms, but in the right circumstances (e.g. some triggering event, or stressor) they will manifest their presence. The brain, as the proximate cause of human experience and behaviour, is where the relevant causal structures are supposed to be located. Murphy (2006, pp. 120–122), by contrast, together with others (e.g. Poland, 2014, p. 37) hold that the relevant causal structures of human behaviour and experience can also be localised at higher levels, like neuro-cognitive, behavioural, computational, and (inter-)personal levels, as well as inter-relations between these levels.

The question whether, or to what extent, higher level causal explanations are reducible to lower-level neurobiological explanations is a question we need not get stuck on. We can at least say that higher level explanations take the form of causal theories and display strong predictive and explanatory qualities. As the slogan goes: they may 'pull their own weight' in scientific explanation. This is the case even if we want to hold onto the intuition that higher levels are in some way grounded in neuro-biological properties (and ultimately in the microphysical world studied by

[11] Even though 'validity' often refers to predictive validity only, many think that validity in the proper sense of the term requires a causal understanding—for reasons of clarity the latter may be called 'causal validity'.

[12] Murphy (2006, p. 225) acknowledges this continuity between the process of scientific validation and causal discrimination as well. Richard Boyd also highlights this point: "Kinds useful for induction or explanation must always 'cut the world at its joints' in this sense: successful induction and explanation always require that we accommodate our categories to the causal structure of the world". (Boyd 1990, p. 139).

physics). To illustrate, it may be that at a cognitive psychological level something like Festinger's theory of cognitive dissonance predicts people's thought patterns and associated behaviours more accurately and economically than neuro-biological accounts.[13] Or at a behavioural level, learning theories may be more successful in explaining and predicting people's emotions and behavioural responses than a much more detailed and complex neuro-biological account. If higher level accounts are successful causal theories in at least this sense, they fulfil the demands of Hempel's positivist agenda: they constitute causal structures of human psychology that provide falsifiable causal explanations of human behaviour and experience. Murphy therefore seems entirely right to argue that distinguishing mental disorders based on underlying causal mechanisms *may*, but *need not,* rely on neurobiological causal mechanisms. Causal structures at higher level of organisation, including inter-relationships, can also serve as causal structures for psychiatric taxonomy.[14]

In sum, the basic, positivist picture of kind-formation in psychiatry is (1) evaluative judgement carves out the domain of psychopathology (2) grouping into kinds of mental disorders based on a variety of factors that suggest the presence of similar inner properties or processes in each member (3) construct validation through an examination of predictive success about those, and only those, falling under the kind (4) mapping distinctions between kinds of mental disorder on causal structures, thereby improving the predictive and explanatory power of constructs.

3 Natural kinds

To appreciate why this picture of scientific kind formation gives people confidence in the belief that mental disorders—once categorised on the basis of causal structures—qualify as natural kinds, more needs to be said about what is meant by natural kinds. Since Mill (1974) introduced the idea of kinds with a "real existence in nature", philosophers have defended different conceptions of what is now referred to as natural kinds.[15] It should be no surprise that different conceptions of natural kinds result in different conclusions about the status of mental disorders, with

[13] As Arpaly (2005, pp. 285–290) explains, this may be due to the causal efficacy of the *content* of mental states, like the content of beliefs and desires. This sort of efficacy (like responsiveness to reasons) are not captured by physical explanations. None of this has to result in Cartesian Dualism.

[14] Gorenstein would strongly resist this suggestion. He likens all higher-level psychological theories to speculative, idiosyncratic, archaic, and incompatible theses—reminiscent of the psychoanalytic theories that informed taxonomies before the DMS-III—only standing in the way of discovering the real material causes of human behaviour (1992, pp. 114–119). I disagree with this diagnosis: psychological explanation should be held to the same scientific standards as neuro-biological explanation.

[15] See Hacking (1991, 2007a) for two accessible overviews of thinking about natural kinds. I have drawn extensively from these, without sharing Hacking's skeptical conclusion that "there are so many radically incompatible theories of natural kinds now in circulation that the concept itself has self-destructed" (2007a, p. 205).

many authors picking a conception of natural kinds that serves their preferred standpoint.[16] We should be able to do better than that. A brief historical sketch and some recent developments in philosophical thinking about natural kinds should offer a decent grip on what is meant by the term, enabling a non-question begging discussion about the ontology of mental disorders.

When Mill wrote about kinds with a "real existence in nature", he worked with a straightforward distinction between two kinds of classes (Mill, 1974, p. 122). Every general name or attribute makes it possible to form a class: e.g. a class of white things, a class of square things, a class of things made in China, etc. If we examine those classes to see whether particulars have anything more in common than their one shared attribute, the result is negative: they have no further shared, common properties—except those that are implied by, or a direct consequence of, the attribute the class was based on (e.g. all square things also have four sides, four corners, etc.). By contrast, another kind of class will have much more in common: species of plants and animals, or elements like sulphur and phosphorus have many common properties, some of which not yet discovered. Mill thought that the second of these "answers to a much more radical distinction in the things themselves than the other does"; these latter classes of things are "made by nature" (Mill, 1974, p. 123). A Millian conception of natural kinds, then, trades on all particulars in the class sharing a large, if not infinite number of common properties. These kinds are of scientific interest, because we can draw conclusions about all the particulars by only investigating a small sample; that is, natural kinds support inductive reasoning.

On a Millian account it is not difficult to argue that mental disorders are natural kinds. In the previous section we saw that mental disorders are standardly (at least since the DSM-IV) differentiated by co-variance of symptoms, and that the process of scientific validation attempts to fine-tune categories and their application criteria such that each person in the class (or at least a great majority) shares numerous properties with, and only with, people in that class. This would be enough to qualify as a natural kind on Mill's view: the classes of mental disorder aren't arbitrary (they don't include a random collection of conditions), and they're also not single attribute classes. Current categories of mental disorder would make for a natural kind (although, to be fair, Mill's natural kinds wouldn't allow for too much heterogeneity in symptomatic manifestation).

Mill's conception of natural kinds was heavily criticised, however, most powerfully by C.S. Pierce (see Hacking, 1991, p. 119ff, 2007a, p. 222ff). When Mill insisted that common properties of natural kinds shouldn't follow from "some mode of implication" or "be derivable … by some law of causation" (Mill, 1974, pp. 122–123)—so as to keep single-attribute kinds out of the picture—he overlooked the scientific efforts to discover underlying properties that do in fact account for many, if not all, of the observable commonalities (Hacking, 1991, p. 119). When science

[16] Haslam (2014) provides an overview of different conceptions of natural kinds and their consequences for our thinking about mental disorders. He points out that on a strict essentialist conception virtually no mental disorder qualifies as natural kind, whereas on a lose conception virtually all disorders count as natural kinds.

discovers the atomic structure of sulphur as having 16 protons it can explain (at least in part) why all sulphurous things have the common properties that they have. Those commonalities are the result of a single attribute and what Mill called "some law of causation" (Mill, 1974, p. 123), which on Mill's view would rule it out as a kind existing in nature. Peirce's critique of Millian natural kinds, then, draws attention to the fact that natural kinds have their numerous common properties *because* of their shared underlying properties: inner structures that are at first hidden, but that can be exposed by the sciences.

As Hacking points out, Putnam (1975) and Kripke (1980) radicalised Peirce's critique of Mill (Hacking, 1991, pp. 121–122). They did so by further prying apart the observable common properties of things in a class (their 'superficial' properties) from their underlying scientific properties (their 'inner structures') that cause and explain them. In Putnam's famous example, the superficial properties of water include that is fluid, drinkable, transparent, tasteless, freezes and evaporates at certain temperatures, etc. The underlying structure, science revealed, is that water is H2O. Putnam's arguments attempt to show that if, in some parallel world, a substance would have all the superficial qualities of water but doesn't consist of H2O, the substance would not be water. For natural kind terms like water, Putnam concluded that underlying scientific properties, like consisting of H2O, are *necessary*; superficial properties, by contrast, are by themselves not *sufficient* for membership of a natural kind. Because all members of a natural kind share some underlying scientific property, they will share many other properties as well (against stable background conditions), like water always being fluid at certain temperatures and pressures. That is why correct inductive inferences can be made about members of natural kinds: their common observable features are *lawful* consequences of their shared, underlying, 'hidden' properties.

Kripke (1980) offers a slightly different take on natural kinds, focusing more on how we discover them. On Kripke's view, natural kind terms have their reference fixed through an act of "initial baptism" (Kripke, 1980, pp. 78, 135): some sample of things is given a name or label on the basis of shared superficial properties, under the supposition that they share inner structural features. This assumption is what makes it a natural kind term rather than a term denoting superficial properties. The term gold (on an artificial reconstruction) was first applied to a sample of things that were all reddish-yellow, heavy, ductile, etc. This initial baptism fixed the reference of the term gold. Natural science revealed the inner structure of things in the sample as having atomic number 79, confirming the supposition of a shared inner structure and establishing a posteriori what gold really is.

Like Putnam, Kripke thinks there might be possible worlds in which gold things were not reddish-yellow or heavy (e.g. worlds in which there aren't perceptive creatures like us) but in which they would still be gold; therefore, manifest properties, again, aren't *sufficient* for being gold. And vice versa, there is no possible world in which what we named 'gold' does not have N79 and still be gold; therefore having N79 is metaphysically *necessary* for being gold. Kripke doesn't just side with Peirce in highlighting that shared inner structure explain superficial properties; he argues that inner structures form the essence of the kind, while collections of common superficial properties are by themselves not sufficient for natural kind membership.

It is therefore up to science to find out which natural kinds there are, depending on the inner structures it discovers.[17] Particulars sharing a great deal of manifest properties is not enough to establish that they form a natural kind.

Putnam and Kripke's view of natural kinds yield clear instructions about which kinds to accept and when a particular can be grouped under it. Kripke offers five such instructions: (1) a class picked out by some description may have particulars that share superficial properties but don't have the relevant underlying scientific properties (that explain superficial properties in the majority of other cases). Kripke's view is that such particulars do not belong to the kind. Fool's gold may look like gold, but it isn't (Kripke, 1980, pp. 124–125, 137). (2) Conversely, particulars that share the underling scientific properties but don' exhibit the superficial properties do belong to the same kind: something may not quite look like gold, but if it shares the same inner structure (N79) it belongs to the kind. (3) It may also turn out that a class consists of two or more different natural kinds, sharing a great deal of superficial features but two or more different inner structures. Jade, for instance, looked like it was one kind of thing, but it turned out to be two different kinds of minerals, now distinguished as jadeite and nephrite (Putnam, 1975, p. 160). In these cases, we should accept that there are in fact two different natural kinds of things (Kripke, 1980, p. 136). (4) If it turns out that the underlying properties of the sample are the same as that of a kind we already recognised, then we may drop the newly introduced kind and kind-term and include the particulars under the already existing class (Kripke, 1980, p. 136). Finally (5) if the initial sample is more radically diverse, Kripke suggest we may abandon the kind-term altogether or realize that it does not refer in the way we thought: the class was picked out under the assumptions that it shared underlying scientific properties explaining superficial sameness, but it turns out it doesn't—all the term did, was denote superficial qualities (Kripke, 1980, p. 136).

It should be clear that Kripke's account of distinguishing natural kinds matches closely with the fourth stage of scientific inquiry described in Sect. 2. When positivists argue that psychiatric taxonomy should be based on insights into causal mechanisms, they express a view congruous with the Putnam/Kripke thesis that only scientific investigation into inner structures can tell us which kinds there are and what their necessary properties are. Co-variance of shared, observable features may help us to an initial grouping (reference fixing of the kind-term for Putnam/Kripke; stage two of taxonomy-formation described above) and give us reason to *assume* the sample is a natural kind. But it is research into the underlying causal structures that establishes whether they do, in fact, form a natural kind. What was described as the fourth stage of scientific inquiry, then, is precisely what enables the discovery of natural kinds on Putnam/Kripke's understanding of natural kinds.

In the context of categories of mental disorder, the Putnam/Kripke view entails that some of the currently recognised categories may pick out a group of people that share the same inner structure: causal mechanisms and processes that explain the

[17] On their view we therefore cannot show with thought experiments whether mental disorders are natural kinds, as Pickard seems to suggest (Pickard, 2009, p. 87).

frequent co-occurrence of observable symptoms. It may also be that current taxa comprise multiple causally responsible inner structures, like a recent neuro-imaging study suggesting that children diagnosed with ADHD actually fall into four distinct kinds (Lecei et al., 2019). In such cases the Putnam/Kripke view would be to abandon ADHD (like with jade), at least in scientific contexts, and to continue with the newly found kinds. Causal structures may also be exposed that do not respect current diagnostic categories, like one study finding three distinct causal pathways involved in psychosis—indifferent to whether psychosis occurs in schizophrenia, bipolar disorder or schizo-affective disorder (Clementz et al., 2016). This would give a strong reason to drop the original categories in favour of categories based on these causal discriminations.

Further, it might also be that other people share the same inner structural features, but don't quite exhibit the characteristic symptoms. Putnam and Kripke's arguments imply that these people would belong to the same kind: these are non-symptomatic cases of a disordered mind (like non-gold-looking gold still being gold).[18] The scope for such cases, however, is extremely limited: inner structures against a stable background *must* give rise to similar observable symptoms; that's why these inner structures explain the observable symptoms. People who exhibit the characteristic symptoms but do not share the relevant causal structures, by contrast, would not fall under the same category (like fool's gold not being gold). Finally, it may turn out that the sample doesn't share any causal structures, in which case the conclusion is that the name doesn't pick out a natural kind. All the term did was denote superficial properties shared by a number of people. The term may still be perfectly useful in a descriptive and communicative way, but it should no longer be thought of as corresponding to a real division in nature and does not support inductive generalisation.

The examples used by Mill, Putnam and Kripke (water, gold) may give the impression that essences always have to be single, causally effective, microphysical properties. This is how authors in the literature on psychopathology often portray essentialism about natural kinds.[19] But Kripke and his followers are much more open about what 'inner structures' could be, even going as far as accepting historical or genealogical essences for persons. This has led some philosopher to claim that it is possible to view species and other biological taxa as natural kinds with genealogical essences as well, enabling conclusions like guinea pigs not being rodents and whales not being fish (see e.g. Laporte, 2004, pp. 63–65). Now, with historical essences the 'inner structure' of the kind becomes extrinsic rather than intrinsic

[18] Pickard considers this possibility as well in the context of schizophrenia and accepts that "it seems at least possible that" such a person should still be considered schizophrenic (Pickard, 2009, p. 87).

[19] For instance: Kendler et al. (2010, p. 1144) "according to essentialism, other key properties of the disorder are consequences of the underlying essence, which is taken to be something relatively simple and unifying, such as a single DNA mutation or a single infectious agent", and "essentialism assumes a single and simple causal agent"; or Haslam (2014, p. 16) "natural kinds are the subset of discrete kinds whose basis is a single cause that is common to all category members and that directly gives rise to the kind's properties"; or Zachar (2014, p. 82) "using the criteria of essentialism to define natural kinds … [involves] a single underlying property [as] necessary and sufficient determinant of a kind." These formulations of the view make it weaker than necessary.

to the individuals included, which is problematic for several reasons (see Okasha, 2002 for discussion). The main point here, however, is that there is no reason why essences must be singular, physical or microphysical properties; they may well involve multiple properties or mechanisms, complex systems, also at higher levels of organisation. We may simply defer to science to find out what the inner structure of kinds are. This point is often missed, and the reason that natural kind essentialism has received a bad reputation in discussions in biology and medicine, with many authors (e.g. Cooper, 2005; BeeBee & Sabbarton-Leary, 2010; Kendler et al., 2011; Murphy, 2006; Zachar, 2014) favouring the Homeostatic Property Cluster (HPC) theory of natural kinds developed by Boyd (1989, 1990).

A few words on Boyd's HPC then. Like everyone in the natural kind tradition, Boyd starts with the observation that properties often cluster together in nature, which he describes as a form of homeostasis. Following Peirce, Putnam and Kripke, Boyd recognises that homeostatic clustering of properties results from underling causal mechanisms (Boyd, 1989, p. 16). Natural kind terms pick out classes in which these homeostatic clusters of properties occur. Not all particulars picked out need to have all of the observable properties—members may vary to a certain degree. The clustering of properties, in any case, is due to causal mechanisms that ensure properties get instantiated more or less together. There may be multiple causal mechanisms at work at various layers of organisation to maintain, and at times restore, the homeostatic cluster of surface properties. Boyd's HPC conception of natural kinds is on these points not an inch removed from the Putnam/Kripke view.

The real difference is that Boyd allows particulars in a natural kind to have some, but not all of the underlying causal mechanisms. Particulars in a class may vary not only in surface properties, then, but also in the causal mechanisms that underpin these differences. This enables Boyd to say that his conception of natural kinds does not make use of *any* necessary conditions and therefore isn't essentialist (1990, p. 142). It follows that natural kinds will have fuzzy boundaries: there will be "extensional vagueness", as it is no longer fixed which, and how many, of the causal mechanisms need to be present to fall under a natural kind (1989, p. 17). At this point, Boyd's homeostatic view differs from Kripkean essentialism. It enables Boyd to consider many groupings as natural kinds, notably biological species. But on the flipside, it generates a level of arbitrariness when distinguishing kinds, as it is no longer clear where one natural kind begins and another starts. There is no specification as to which, or how many, of the relevant causal structures need to be present for kind membership. In the end, this will create difficulties in determining which kinds there are in the world, and what's included. For this reason, I think we should be less enthusiastic about an HPC view of natural kinds than most people are in the literature.

To conclude, it should be clear why the stages of scientific kind formation sketched in the previous section supports claims about natural kinds. Science aims for valid kinds with high predictive value and explanatory power. That is best achieved by distinguishing kinds on by shared causal structures exclusive to the kind. Natural kinds are groupings based on a shared (or on HPC: sufficiently similar) underlying causal structures that explain clusters of co-occurring properties. Scientific inquiry, then, should give us natural kinds.

The upshot of this discussion is that we have no a priori way of telling whether mental disorders form natural kinds, and if there were any which ones they'd be. Pointing to co-variance in symptoms (or the relative lack thereof) won't suffice as a defence of either side of the debate. Heterogeneity in in symptomatology, we can now clearly see, does not plead in favour or against a natural kind view. It is an a posteriori matter, to be determined by genetics, neuroscience, psychology, and other special sciences, whether there are natural kinds of psychopathology. Yet, some authors have principled reasons to resist the idea that there might exist natural kinds of mental disorders. In the remainder of this paper I shall turn to these reasons and try to offer a response.

4 Objections

4.1 Complexity and continuity

Perhaps the biggest source of scepticism about natural kinds of psychopathology stems from genetic and neuro-biological research into *DSM* categories. Decades of fundamental research into abnormal psychology have resulted in a negative sentiment, expressed well by Derek Bolton: "there is not much prospect that the science of the aetiology of psychiatric conditions will deliver a single, optimal classification scheme—the reason being that the last few decades of research has uncovered systemic complexity, rather than reductionist simplicity" (Bolton, 2012, p. 6).[20]

The response is that we just need more insight into causal structures—so that complexity will eventually make way for causal simplicity—has become dissatisfying. An equally dissatisfying response would be to emphasise that this only shows that currently recognised *DSM* categories are useless for research purposes, and that all we need is a new psychiatric taxonomy. Both reactions don't do justice to the depth of the concern. To illustrate why, consider just one example: findings about the onset of psychosis in the context of schizophrenia (Broome et al., 2005):

- *Neurochemically*, there is evidence for a link between dysregulation of the mesolimbic dopamine system and the positive symptoms of psychosis. Heightened dopamine release facilitates "the formation of 'meaningful connections'" and may result from a malfunctioning of the cortical-limbic circuitry (2005, p. 26). This circuitry consists of the interplay between the hippocampus (regulating the impact of experiences on mesolimbic dopamine release) and amygdala (causing emotional overrides of the hippocampus's regulatory activity)—with the prefrontal cortex able to override input to both systems, tempering their reactions to

[20] Spitzer, the chief architect of the DSM III, and First, an editor of DSM IV, express a similar sentiment: "Despite the considerable advances in psychiatric research, disappointingly, little progress has been made toward understanding the pathophysiological processes and aetiology of mental disorders. If anything, the research has shown that the situation is even more complex than initially imagined, and we believe that not enough is known to structure the classification of psychiatric disorders according to aetiology" (Spitzer & First, 2005, p. 1898).

stimuli. Patients diagnosed with schizophrenia show deficits in prefrontal executive functions, and neuro-imaging confirms that the hippocampus and amygdala have lower volumes.

- *Genetically,* the risk of schizophrenia is increased by several genes that affect the glutamate system, which in turn regulates dopamine release. Other genes associated with schizophrenia are genes involved in the breakdown of prefrontal dopamine, thereby affecting cognition. A polymorphism in the COMT gene, determining the rate of catabolism of frontal dopamine, strongly influences psychosis induced by cannabis.
- *Developmentally*, the collections of involved genes, as well as environmental infractions early in life (especially obstetric complications), affect neurodevelopment. Those children bound to develop schizophrenia fail to learn new cognitive skills during childhood and increase the likelihood of developing minor quasi-psychotic symptoms as early teenagers, which increase in strength, frequency, and associated distress. Pre-schizophrenic children often suffer from depression and anxiety.
- *Psychologically,* are caused by mechanisms that integrate information into a temporal-spatial context, and mechanisms involved in self-monitoring of intentional action (making actions feel alien). Perceptual experiences are also affected by attention directing systems, with stimuli entering awareness being perceived as a threat, resulting in responses (emotional, cognitive, and behavioural) that, in turn, stimulate more intrusions. Both information integration and attention-direction are strongly correlated with dopamine regulation. Abnormal belief formation is hypothesised to arise from efforts at explaining the bombardment of salient stimuli and the anxiety this results in—delusions are efforts to conjure up explanations.
- *Socially*, it has been found that urban upbringing is a risk factor, as well as the quality of maternal-child relationship. Migration increases risk for schizophrenia, probably due to the social isolation, lack of social support, and experiences of social defeat, to a point where increasingly deviant development ensues.

The complexity described here is still an oversimplification, and there is no reason to think aetiology will get any less complex as research into genetics, neuroscience, cognitive science, and social science progresses. Moreover, multi-factorial aetiology can be expected in virtually all forms of mental disorder. To liken mental disorders to the inner structures of gold and water, with law-like consequences that explain and support universal and accurate predictions, may seem a non-starter.

A few points should be observed about the above example. The first is that at genetic and neurochemical levels *multiple* mechanisms appear to be involved: several regulatory and overriding functions play a part within a larger complex of neural circuitry. Furthermore, cross-level causal relations are suggested to be at play, for instance between the cognitive-behavioural responses to overstimulation and an increase in frequency of intrusions.[21] The most important point for now, however,

[21] Some have suggested we should therefore think of causal factors as *a network*, where interactions between various layers and feedback loops are causally involved in mental disorders and their character-

is that a larger psycho-neurological causal structure *can* be discerned. The inner structure, or underlying mechanism, we saw, need not be a single gene or neurotransmitter deficiency: it may be a larger complex mechanistic structure comprised of various sub-processes that together bring about the manifest properties. The inner structure of a kind of disorder, may itself be a complex dynamic whole.

How about the social and environmental aspects than? Is the way of thinking I'm suggesting not too reductionist? Here I think it is helpful to introduce a distinction between proximate causes and distal causes (Gorenstein, 1992, p. 106). A causal explanation of psychosis involves (as we have seen) neurochemical and psychological mechanisms, directly responsible for the manifested symptoms. These mechanisms themselves also have causes; those are distal causes. Distal causes can be genetic, like the genes involved in the neurodevelopment of the mesolimbic dopamine system and information-integration processes later on.[22] Distal causes may also be environmental, like periods of social isolation and repeated experiences of defeat (influencing dopamine regulation), with other social factors (and their geopolitical causes) being further down the causal chain.[23] The proximate causes, by contrast, constitute what the disorders is: the properties or mechanisms internal to the individual that are directly responsible for the observable phenomena or symptoms.

In this case, the proximate causal structure is the cortical-limbic circuitry and its role in dopamine regulation combined with cognitive psychological control mechanisms. If psychosis in schizophrenia were to be a natural kind, it looks that the interplay between the cortical-limbic circuitry and psychological control mechanisms would constitute its inner, identity-determining structure.[24] To be clear: this structure doesn't *cause* the disorder, it *is* the disorder—we call it a causal structure only because it causes the manifest properties.[25] As this inner structure would the determine identity of the disorder, the same kind of disorder couldn't be caused by a different mechanistic structure (just as the same mechanistic structure can't be the cause of some other disorder). Further, this inner structure may be brought about, or

Footnote 21 (continued)
istic forms of expression. See especially (Borsboom, 2017; Borsboom & Cramer, 2013; Borsboom et al., 2018).

[22] Genes may also be a direct cause. Psychosis induced by cannabis, for instance, has been shown to be strongly influenced by the COMT gene, determining more directly the rate of catabolism of frontal dopamine. In such cases, it seems to me that genetic causes are proximate causes.

[23] In my understanding, the new Network Approach to mental disorder muddles this distinction, even though the main contributors do speak of "more direct" forms of causation (see Borsboom & Craver, 2013, p. 106). Even if symptoms themselves play a higher-order causal role in the overall network by affecting lower-level processes, there is still a meaningful distinction to be drawn between direct and indirect causation, the former being the most important for distinctions between categories.

[24] None of this implies that insights into distal causes is insignificant, either for scientific purposes or therapeutic reasons. If, for example, certain genetic dispositions in combination with cannabis use significantly increases the risk of psychotic episodes, this obviously has important therapeutic and policy implications. The same holds for the development of psychotic symptoms in immigrant populations due to long-term isolation and social defeat. But if we are concerned with the inner structures of those diagnosed as psychotic, it is the proximate causes that matter. They are the underlying mechanisms that make the disorder what it is.

[25] In the same way that water *is* H2O (H2O doesn't cause water), while the chemical structure H2O is causally responsible for various manifest properties.

maintained, by various distal causes (genetic, developmental, social), each one also having their respective causes. This may give the impression of an impenetrable, endlessly complex, and multi-factorial causality. But a distinction between proximate and distal causes will make it possible (at least in some areas of psychopathology) to identify the inner structure that makes that pathology the kind of pathology that it is. Complexity and multifactorial aetiology, then, don't have to be an obstacle and may be fully compatible with a causally informed taxonomy.[26] In areas of psychopathology where it turns out impossible to identify proximal causal mechanisms, or where causal mechanisms differ among individuals, however, it must be concluded that the conditions under consideration don't fall into a natural kind of disorder.

Another concern to mention here is that the causal structures involved in psychopathologies typically fall on a continuum, or spectrum. The cortical-limbic circuitry, just as psychological control operations, generate outputs that fall on a continuous scale—a scale that also includes outputs corresponding with healthy functioning. That doesn't seem to fit well with the idea of discrete categories of disorder, with an identity-determining inner causal structure.[27] Now, even if all relevant mechanistic structures and their outputs fall on some continuum, the question is whether that threatens the idea of a shared inner structure that is causally responsible for other shared features, determining the kind. I don't think it does, for two reasons. First, if the causal mechanism or its output are such that at a certain point they cause a large number of further features (in all and only those sharing it), then that's the point on the where 'nature has a joint'—so to speak. This is the case also if that point lies on a continuous scale. If a certain level of dopamine release (together with other factors) result in anomalous perceptual experiences and delusions—with everything this implies—then that is where the cut-off point lies. Continuity in output is, at least in this sense, compatible with categoricity. Second, it may be permitted that natural kinds have some level of vagueness at their boundaries; there may be contentious cases. That doesn't threaten the existence of the kind, nor make the boundaries arbitrary (as I levelled against Boyd's HPC account). That said, if the outputs of mechanistic structures only generate a gradual change of surface features, without a point or area of predictive significance, then we must conclude that there aren't natural kinds to be discerned. If empirical research finds that many forms of psychopathology exist only on such continua (see e.g. Haslam et al,. 2020), many won't fall into natural kinds.

4.2 No neurobiological basis

It may turn out that a kind of mental disorder does not involve any neurochemical abnormalities or difference in brain structure, unlike what was suggested about the onset of psychosis in schizophrenia. Should the lack of any 'material' basis mean

[26] The same is true for somatic diseases, of course. Distally, cancer will have all sorts of social, material, lifestyle, and dietary related causes; proximally, its inner structure consists of the cellular changes resulting in uncontrolled cell division—that makes cancer what it is.

[27] I am thankful to a reviewer for pointing this out.

they aren't natural kinds? Based on the previous section, the answer is a resounding 'no!'. The relevant inner structures underpinning psychiatric categories need not be neurochemical or genetic, they may equally be psychological or even historical. The causal structure may be a kind of shared genealogy, but also some abiding cognitive pattern or emotional mechanism, existing in what is otherwise a perfectly normal brain. The causal structure may also involve interactions between various levels, like an interplay between cognitive patterns and neurochemical processes. As long as there is an inner structure that is shared with, and only with, other people that make up the kind, explaining the relevant symptoms and supporting predictions, this will suffice.

There are many who emphasise this point, but their message falls too often on deaf ears. Graham, for instance, also argues that mental disorders may fall into objective categories with significant predictive power but exist in an "unbroken brain" (2010, p. 24). Natural kinds of psychopathology, Graham writes elsewhere, do not require "the brain being disordered (damaged, neurologically impaired, dysregulated, cells gone bad, etc.)" (Graham, 2014, p. 127).[28] Similarly, Poland recognises that "behavioural skills deficits, severe depression or anxiety, deficient social cognitive capacities, and even delusions and hallucinations all might involve violations of some sort of norm of high-level functioning without implying a brain pathology is present" (Poland, 2014, p. 55). The only requirement on causal structures is that they are exclusively present in all members of the kind, that they explain the symptoms of individual and support inductive inferences about members of the kind. These conditions could be met by inner structures located at a psychological level, as well as cross-level processes. Mental disorders don't need to be brain diseases to qualify as natural kinds.

4.3 Which laws? Predictive success inherently limited

Another source of principled resistance to natural kinds of mental disorders may arise from the emphasis on *predictions*, and the difficulties involved in predicting human behaviour with a meaningful level of accuracy (see Gorenstein, 1992, pp. 108–109). The examples from chemistry to which discussions of natural kinds typically resort make the role of predictions look quite innocuous. Water will always evaporate at certain temperatures and pressures, gold will always melt at certain points, so predictions can be made with high levels of accuracy. This is a lot harder in other scientific domains and virtually impossible in the life and social sciences. Even if there would be clear and isolable neural-psychological mechanism behind disorders like schizophrenia, shared exclusively by the population picked out with a suitable selection criterion, how much could be predicted about an individual's behaviour, experiences, the natural course of development, future levels of functioning, etc.? Could it be predicted which delusions a person will have, with what frequency and intensity? Will the person speak in a disorganised manner? What will be

[28] See especially Arplay (2005), Graham & Stephens (2007), and Poland (2014).

said, when, and how? To what extent will the individual be withdrawn, from whom and in what way? It seems a stretch of the imagination that predictions can be made at this level of specificity. Predictive power explained by shared causal mechanisms is the hallmark of natural kinds, so shouldn't this be a reason to abandon the idea?

One way to address this concern is in terms of intervening variables, by which I mean a variable that impacts the relationship between an independent and dependent variable—in this case the relationship between underlying mechanisms and symptomatic expression. Everyone suffering from schizophrenia will have other mental properties (both stable and transient) as well as a material and social context that affect how symptoms get manifested. These variables may intervene on the processes that proximally cause the symptomatic expression and thereby bring about considerable variation between people's clinical manifestations. Intervening variables may even determine whether the disorder gets expressed at all (think of an alcoholic living on an island without alcohol). The implication here is that if a mental disorder is a natural kind, it is possible to have the disorder without the characteristic (or any) clinical manifestation. Due to intervening or masking factors, a-symptomatic natural kinds of mental disorder are perfectly possible (see Pickard, 2009, p. 87). Note the contrast with descriptive kind terms that do not refer to natural kinds: those do not apply in the absence of the relevant symptoms. Causal structures should therefore be thought of as dispositional: only in certain contexts and situations, with certain triggering events, will the relevant behaviours be manifested. Predictions should, at the very least, be qualified by such intervening variables and manifestation conditions.

Even then, however, some may insist that predictions about particular cases remain nearly impossible. Generally, it seems that too high a burden is placed on neuroscience and psychology when this is raised as a genuine concern. Even in physics and chemistry it can be extremely difficult to predict some individual event (e.g. the rolling of a dice, the way smoke twirls up) despite knowing all the causally relevant properties and the natural laws that apply. That is why *controlled experiments* are required for any sort of prediction, eliminating the effects of other variables. Against a set of stable background conditions, the same causal structures should produce similar effects. In psychology and neuroscience, controlled settings are hard, if not impossible, to achieve. It should suffice, therefore, if certain *patterns* of behaviour can be predicted under certain environmental conditions (see Gorenstein, 1992, p. 109). A degree of abstraction is thereby permitted: schizophrenics will form *some* delusional beliefs if certain regulatory processes fail; children with ADHD will struggle concentrating on *some* tasks and will do better in *certain* tasks in different social environments. Prediction at this level of generality is all one can hope for and should suffice as predictive success.

4.4 Making up people

The author who has done most to challenge the view that mental disorders are natural kinds is Ian Hacking. His work presents more than one challenge to the suggestion that there could exist natural kinds of mental disorders. I focus on what I take to be two distinct challenges, starting with what he calls "making up people" (1986).

The central idea of 'making up people is that by introducing human kind-terms, it becomes possible *to be* a certain person. Being of a certain kind shapes people's experiences, self-understanding, and possibilities for action. When the kind term is introduced one can, in a sense, *be* a homosexual, a pervert, a high-functioning autistic person, a hyperactive child, a multiple, etc. Hacking's point is that once the kind-terms are introduced and adopted, this radically changes how someone comes to think about oneself. Before the kinds were introduced these "were not a possible kind of experience to have had" (Hacking, 1995b, p. 169). With human kind-terms, then, we 'make up' the possibility of being a certain type of people, with real-world effects.

On a strong reading, what Hacking suggest is that individuals come to act in accordance with the kind-description *as a result* of the introduction and propagation of a human kind-term, like the wave of people with multiple personality disorder in the 1970's and 80's. The introduction of the category offers "a way for troubled people to express their difficulties; the role is one of many that awaits" (Hacking, 2007a: p. 368). On this this reading, Hacking points to a *causal* process, explaining *why* large groups of people came to exhibit similar behaviours and to report similar experiences. This offers an alternative explanation for why people exhibit co-varying behaviours associated with a disorder. We might be fooled into thinking that co-varying behaviours are the result of shared inner causal processes; studies like Hacking's demonstrate that the introduction, dissemination and legitimation of a kind-term, by itself, can do most of the causal work.

A simple response to Hacking's account of 'making up people' is to point out that it is an empirical question whether inner properties or the social role-model explanation holds up best for any given kind. This is roughly the answer given in the literature by (e.g. Sullivan, 2014, p. 260; Tsou, 2007, p. 340). Perhaps the rise of transient illnesses like multiple personality disorder can be explained best on a social model like Hacking's 'making up people'. Hacking has offered a useful framework to analyse these phenomena, especially the clustering and co-occurrence of symptoms in people that do not share any further underlying genetic or neuro-psychological properties. At the same time, disorders like autism spectrum disorder and schizophrenia might be best explained by inner causal mechanisms.[29] This response should suffice and keeps the possibility of natural kinds of disorders very much alive.

It may seem obvious that 'making up people' generates categories that we shouldn't think of as natural kinds. But as the discussion in Sect. 3 showed, the matter is more complicated. *Recall that a natural kind-term is applied to a class of things under the supposition that it shares a similar inner structure. 'Inner structure' is a place-holder for whatever the sciences discover as the underlying properties, shared exclusively by members of the class, explanatory of their superficial properties and others class-specific commonalities, supporting predictions and*

[29] Hacking gestures in that direction in (1999, pp. 108–109). A year earlier, he wrote "some mental disorder are, in my opinion, real. In the case of schizophrenia, for example, despite conflicting claims, I hope that within 20 years we shall have a grip on one or two or perhaps three fundamental types of schizophrenia" (1998, p. 98).

inductive generalisation. Inner structures might be microphysical properties, but also features like shared ancestry and higher-level properties like cognitive-behavioural patterns. One could argue that what Hacking describes in his studies of multiple personality disorder is itself a higher-level causal structure. All members, in the end, share a similar causal process: they have adopted a certain social role that became available with the introduction of the kind-term, within a certain time and socio-cultural niche. This analysis explains, in a causal sense, the commonalities in behavioural manifestations and a great deal of predictions that could be made about them (how they'd respond to certain types of therapy, for instance, including renaming the disorder). If one accepts that 'inner structures' can be filled in with whatever the special sciences find as the underlying cause, sociological causal analyses could be included. In that event, one might go on to claim that the inner structure of transient mental disorders like multiple personality disorder *is* the social dynamic that led people to adopt the role.

Following this line of thinking, one could go as far as to claim that classificatory activity could give rise to a natural kind of person—really there—explained by the social causes. This wouldn't make all social categories natural kinds, though, as it all depends on the predictions the category supports and whether those predictions are sufficiently explained by the inner causal (social) structure. But sometimes, as Hacking seems to show, the introduction of kind terms appears to have the power to create joints in the world of people.

Hacking would no doubt resist this line of thinking. For Hacking, and many others, the key feature of what philosophers were after with the notion of 'natural kinds', *if anything*, is independence from classificatory activity. If we follow that line, then multiple personality disorder is not a natural kind, even if it concerns naturally occurring events (people's behaviours and experiences) with a clear, shared, scientifically exposed inner structure. If we follow Boyd (1991, pp. 413–146), however, and permit sociological explanation as genuine explanation, then we should permit that our classificatory activity can bring natural kinds of people into the world (and push them out of the world again).

In the end, this issue comes down to the question whether sociological explanation is a genuine form of scientific explanation and whether sociological causes can form the 'inner structure' of natural kinds. This question won't be solved in this paper. If not, socially caused kinds of mental disorder should never be thought of as natural kinds. If it is, then classification-caused 'made up' people could qualify as natural kinds, so that there will be more natural kinds of mental disorder than expected.

4.5 looping effects and interactive kinds

A second thesis of Hacking's—expressed by the slogan 'looping effects' (1986, 1999, 2007b)—poses another objection to the possibility of natural kinds of mental disorder. Here too the starting point is that people respond to the way they are classified. Such responses can amount to a type of self-fulfilling prophesy, as "people classified in a certain way tend to conform to or grow into the way that they are

described" (1995a, p. 21). The response may also go in the opposition direction, though, resisting certain stereotypical behaviours (1995b, p. 370). In any case, the classified *change* due to them being classified, meaning that over time the kind itself changes. An updated kind-description will provoke further behavioural responses, and so feedback loops emerge that will make kinds of mental disorders *drift*—they are moving targets. As a result, Hacking argues, there is no stable object of knowledge: members of the class continually change in light of new knowledge about them. This circle is inescapable, Hacking suggests at times, so the hope of arriving at stable classifications should be given up. Things like water and gold, by contrast, are "indifferent kinds" as they irresponsive to how they are classified; mental disorders are "interactive kinds" and do not permit of stable classification (1999, pp. 103–106).

Hacking's account of looping effects is a radical thesis, as it is not just supposed to take place at an individual level, but on the level of the class or kind (see Tsou, 2007). It is no doubt true that individuals feel and behave differently when categorised as having a certain mental disorder. But Hacking's claim is that the whole class comes to feel and act differently after being categorised. And not just that, they must do so in a broadly similar and concerted fashion if it is going to affect the kind and the kind-description.[30] If people diagnosed with multiple personality disorder start behaving differently in all sorts of diverging ways, there wouldn't be any direction in which the kind is drifting, so no renewed kind-description would appear and no looping effect would occur. In short, seems quite unlikely that individuals interacting with their categorisation all come to behave differently in *the same fashion* so that the *kind* changes.

But there is another way the objection can be reconstrued, namely as a procedural problem. We have seen that on a natural kinds view outward manifestation may diverge and change. At some point this may force the adoption of new inclusion criteria. If these new inclusion criteria pick out a different sample of people, the referent of the kind term will have changed. This poses a procedural problem, as science won't be able to keep up with investigating underlying causal structures. Whenever a research programme investigates whether a group of people share any inner structural properties that could explain behavioural symptoms, the social world would have already moved on: the kind-description has changed, the inclusion criteria got adjusted, and a different sample of people is now picked. Before research into this new sample could be carried out, members once again responded to being categorised, requiring a change in description and inclusion criteria, and so on. There wouldn't be a stable object, scientific research would be clutching at straws. If this is the real objection presented by looping effects, the concern is again exaggerated. Even with categories that have changed so significantly that the kind-term has come to pick out a different sample of people, it has hardly stopped geneticist, neurologists and psychologist to study potential underlying causal structures. The suggestion that looping effects impede the very project of investigating underlying causal structures of mental disorder because the object drifts off before it has even come into focus just doesn't seem to match the reality of scientific research.

[30] Tsou (2007, pp. 339–340) calls the former "weak" and the latter "strong" implications of looping.

In another sense, shifts in the reference of diagnostic categories do pose an impediment to scientific research. When a kind-description or inclusion criteria changes, a different set of people is picked out that inevitably have different inner structures. This has happened with autism, which began as an extremely rare phenomenon and symptom of schizophrenia and now refers to a spectrum of behavioural and communicative (in)competencies (Eyal, 2017). The object of study of an autism researcher has changed, so findings from the past may not hold for this new sample. Now, whether we should think of this as the result of looping effects isn't clear: is it the behavioural change of *those people* diagnosed with autism that led to alterations in kind description? It seems not: it concerns a mere shift in reference, so that different individuals get included. But even if this were to be counted as some form of looping or conceptual drift, it doesn't undermine the possibility of discovering natural kinds of mental disorder. By investigating the causal structures of those grouped together—even if these groups change—inner structures may be discovered that enable distinctions between different kinds of disorder. Looping effects and shifts in diagnostic categories therefore don't pose any fundamental or principled objection to the possibility or likelihood of exposing natural kinds of mental disorder.

4.6 Perspectivalism at the mechanistic level

Craver (2009) has formulated objections arising from the difficulties with singling out the mechanisms and causal structures that should be tracked by natural kinds of psychopathology. Craver's concerns stem from the fact that the identification and separation of mechanisms—or mechanistic structures—involves human decisions and practical considerations. If analysis at the mechanistic level is shot through with human values, then so is a taxonomy of mental disorders based on it. Craver identifies three separate problems in identifying and separating mechanistic structures:

(1) *Which mechanisms* Multiple mechanisms are bound to be involved in any putative psychiatric kind. This forces a choice on the taxonomist: split the kind into many kinds, each depended on the involvement of some mechanisms (leading to promiscuity in kinds), or lump together into one kind whenever any of several mechanisms is involved (leading to causal heterogeneity).[31] Craver claims that the taxonomist's choice is bound to be informed by practical considerations, e.g. whether one is primarily interested in therapy or prevention. In selecting the mechanisms underpinning a psychiatric kind, then, human perspective and interests play a decisive role. (Craver, 2009, pp. 582–585)

[31] This problem is particularly pressing for a HPC view of natural kinds. On the HPC, there are always multiple mechanistic processes involved in sustaining a cluster of superficial properties. This makes it possible to lump together several mechanistic structures into one kind. On an essentialist view, lumping will be more difficult: some mechanistic structures have to be present to constitute a kind, so diverging mechanisms cannot be lumped together so easily. An essentialist view is therefore likely fall on the splitting side of Craver's dilemma.

(2) *Level of abstraction* Every mechanistic structure consists of sub-mechanisms at a lower level of abstraction: macro physiological structures consists of cells, cells consist of sub-mechanistic structures like mitochondria, mitochondria consist of things like membranes, and so on. Differences at the level of sub-mechanisms amount to different causal structures, so could warrant the addition of another kind. What is the level of abstraction that a natural kind of psychopathology should track? This appears to be a human decision, again informed by the interests we have in introducing a kind. (Craver, 2009, pp. 585–589)

(3) *Boundaries of Mechanisms* It is not always clear where one mechanism ends and another starts. It is crucial to define boundaries of mechanisms if taxonomical terms are to track causal mechanisms. Given that there are many ways to divide up the causal structure of the world, in practice it will be the property cluster we seek to explain that determines the boundaries of mechanisms. Craver's point here is that the natural kind view has it the wrong way around: it wants to split property clusters into kinds depending on causal mechanisms, whereas the individuation of causal mechanism takes place on the basis of the property clusters it seeks to explain (Craver, 2009, pp. 589–591).

The first and second concern highlight that psychiatric kinds can be split or lumped together, depending on (1) which sub-mechanisms are included and (2) the level of abstraction at which the mechanism is analysed. Craver is right that taxonomists are forced to make these choices once categories are based on causal structures. But why should such a choice be made on pragmatic grounds, reflecting practical interests? Sometimes it clear that practical interests motive the choice: in certain public health context we tend to lump people together (e.g. all cardiovascular diseases), while in therapeutic contexts groups are split more finely (e.g. the varieties of coronary restrictions). But taxonomists and psychometricians are primarily interested in construct validity. Why not take validity as the decisive criteria when considering whether to lump or split? If it makes no significant difference for the explanatory and predictive power of the category if there are marginal variations in underlying causal mechanisms, lumping wins the day. If predictive powers increase significantly when categories are split further, then splitting wins the day. The same holds for the level of analysis: if a lower level analysis does not increase explanatory and predictive power, one doesn't have to go to that level of analysis. Water is H2O: no need to delve further into variances of subatomic physics. Many authors, it seems to me—especially those advocating the notion of a 'practical kind'—jump to conclusions when they claim that practical interests play a role (or should play a role) in kind formation.[32]

The third objection is also less of a concern than Craver suggests. A consensus view of mechanisms, he writes, is that "mechanisms are entities and activities organized together such that they do something" (Craver, 2009, p. 582). Individuating mechanisms, therefore, inevitable involves a reference to what the mechanisms

[32] See especially (Zachar 2000a, b, 2003, 2014). There the view is defended that kinds should be informed not only by causal structures and predictive success, but also goals like therapeutic sensitivity, economic priorities and social-political priorities (like reducing stigmatisation).

does, more so than their spatio-temporal features. Craver's concern is that the choice of a property cluster will determine the boundaries of the underlying mechanisms; hence, the search for underlying mechanisms will add little over and above the usefulness of the property cluster with which we started. This last conclusion, however, doesn't seem to follow. As we saw in Sect. 3, clusters of behavioural properties (or manifest properties) lead to a fixing of the class, allowing scientists to investigate any underlying structures and mechanisms. But once those mechanisms are exposed, they may radically upset the classes and property clusters we started with.

This becomes clear when we consider an example offered by Gorenstein (1992, pp. 103–104). ADHD, anti-social personality disorder and alcoholism involve enormously diverse property clusters. Investigating the underlying causal structures revealed that each is associated with an inability to inhibit dominant response tendencies. When this impairment occurs in children it may result in impulsive behaviour, when occurrent in adults it may result in the inability to inhibit gratifying behaviour even if it violates moral and legal codes, and in other contexts the same impairment may result in alcoholism. Perhaps these disorders should therefore be subsumed one category, given their shared causal structure—the proposed label here was 'spectrum disorder' (ibid). This example shows that the original property clusters helped to identify the relevant mechanisms, but that doesn't mean categories of mental disorder cannot be radically revised in light of causal mechanisms. If Craver's third objection were right, the original property cluster would determine the individuation of the relevant mechanism and drastic revisions would be ruled out.

4.7 Normativism

A final principled reason to resist the natural kinds view is because psychiatric kind-terms are value-laden. Recall that the starting point of psychiatric classification was an evaluative stance vis-à-vis a set of mental conditions and ways of behaving. Supposing that the objective world is free from norms and values, how could any further subdivisions of the domain of mental health amount to natural kinds?[33] If the genus is evaluative and therefore constructed, how could any categories falling under it be any different?

This objection to the possibility of natural kinds of mental disorder rests on a persistent mistake. It is perfectly possible for the domain of (psycho)pathology to be a social construction, while further subdivisions are natural kinds. To see why, consider a paradigm case of social construction: plants versus weeds. This distinction evidently depends on the sort of plants we prefer and want to have in our gardens and streets versus those that we don't. But the domains of (non-weedy) plants could very well comprise a large number of natural kinds of plants, just as weeds may

[33] Normative realists need not worry about this line of objection. They can accept that our evaluative judgements correspond to norms existing in the objective world. I don't explore this line of thinking here, as a natural kind view of mental disorders can be defended independently of a commitment to normative realism. Reductive normative naturalists also do not worry about these complaints. They can accept that health and pathology are evaluative concepts, but that they can be reduced to naturalistic concepts. I discuss these issues elsewhere (anonymous 2018, 2019, forthcoming).

comprise natural kinds of weeds. The same is true for mental disorders. Even if the domain is a social construction, this does not carry any implications for particular kinds of mental disorder. Particular kinds of disorder may be defined in terms of some inner structure and support prediction and explanation. A similar point has been made by Pickard (2009, p. 87) and Cooper (2005, p. 46).

Be that as it may, critics bringing up the value-laden nature of categories of mental disorders do point to something important. Even if a category of psychopathology picks out a group of people with a uniquely shared way of mental functioning, this does not yet make it a *disorder.* There may be numerous 'joints' in the world of human minds, but this does not tell which portion of that world counts as disordered. For this sort of conclusion evaluative judgement is required. The parallel with weeds and plants applies here again: if a type of vegetation makes for a natural kind, this doesn't emply anything about whether it's a weed or a plant, that's determined by our evaluative judgement. Similarly with forms of mental functioning: if some form of mental functioning makes for a natural kind, this doesn't say anything about whether it is a disorder. This means we should accept the following sort of paradoxical formulation: if ADHD is a natural kind, then ADHD *really* is a form of mental functioning. If someone falls into the kind, that person indeed really 'has' ADH(D). But whether this person has a mental *disorder* is a different question, settled by our answer to the question whether we think of ADHD-functioning as falling short of the norms associated with mental health.

The significance of this point is not to be underestimated. A common reaction to contested categories of mental disorder is to defer the issue to neuropsychology. The motivation behind seems to be that a condition like autism would only 'really' be a disorder once it has been shown to have a neuropsychological basis; without such a basis it would not really be a disorder. This common line of thinking involves a serious mistake. Even if categories are found to have clear, singular, uniquely shared, fully explanatory neurobiological basis, this does not mean that it is a mental *disorder*. It only means that it is a naturally occurring form of mental functioning, i.e. a natural kind. Whether it is a disorder is another matter, determined by an *evaluation* of the form of mental functioning. And inversely, if a category of mental disorder isn't a natural kind and does not have any neuro-psychological basis, this does not imply that people falling into the kind aren't disordered. If they don't meet certain norms of functioning, they count as mentally disordered.

If it is disputed whether a form of mental functioning is a mental disorder, then, science alone cannot settle the dispute. First a norm is required, only then does it become an empirical question whether that norm is violated. If ordinary functioning is taken as the norm (Fulford, 1989), then a significant decline from this level of functioning is the relevant norm violation. If a certain objective level of functional capacity is the norm (anonymous 2019), a failure to achieve this level amounts to a norm violation. In both events, the norm isn't set by science, so science alone cannot settle the matter.

Deferring to neuropsychology when there is discussion if some condition really is a disorder is therefore a serious mistake: science can settle whether it's a natural kind, not whether it's a natural kind *of disorder*. Or put differently: whether one is mentally disordered does not turn on whether the kind-term used to describe the

disorder is a natural kind.[34] Hence, the natural kinds discussion does not help, or even touch, concerns about medicalisation and the over-inclusiveness of the psychopathological domain.

In sum, the fact that mental disorders are value-laden and dependent on evaluative judgement, does not mean they can't be natural kinds. But when a form of mental functioning turns out to be a natural kind, this does not mean it is a disorder. And when a form of functioning turns out not to be a natural kind, it may very well still be a disorder. These points are subtle and get mixed up too often.

5 Conclusions

The natural kinds discussion about mental disorders has a descriptive side and a normative side. On the descriptive side, we may conclude that the reported causal heterogeneity and poor predictive power of current diagnostic categories makes them unlikely candidates for natural kinds.[35] This does not mean that there aren't any natural kinds of mental disorder. Research in underlying causal structures may well reveal natural kinds of mental disorder, either as subdivision of current categories or as kinds that cut right through current taxonomical divisions. All depends on what causal structures the sciences discern, the degree to which they are exclusively shared by groups of people, and the explanatory and predictive power they afford. Principled reasons to oppose the possibility of the sciences exposing natural kinds of mental disorders proved unconvincing.

On the normative side, it seems clear that psychiatry and other mental health disciplines would benefit from a taxonomy comprising natural kinds. Being able to make predictions about the person merely based on kind membership enables scientific research into treatment options and offers a degree of control. These kinds should be privileged over other, more descriptive kinds for just that reason. Note, however, that the distal causes of the disorder may still vary between people, and that interventions might be most successful when targeting distal causes. If, for instance, social isolation and experiences of defeat are part of the distal causes of some forms of schizophrenia, therapy may be most successful when directed at those factors (even if schizophrenia is a natural kind with a neurobiological inner structure). So even a taxonomy of natural kinds wouldn't have to result in a fully protocolised or standardised clinical practice. The benefits of a psychiatric taxonomy of natural kinds should therefore also not be overestimated.

An important qualification here relates to natural kinds of disorders with social causes, if we were indeed to recognise those as natural kinds. If our categorisations

[34] One might think that the set of mental disorders is *exhausted* by those falling into natural kinds of disorder. I don't see any reasons for holding such a view; if anything, it should be recognised as a persistent and pernicious confusion: why would there have to be (many) others with the same abiding causally relevant neurological or psychological properties for one to be mentally disordered?

[35] Other common complaints about the DSM do *not* support this conclusion: heterogeneity in symptomatology, high levels of co-morbidity, lack of genetic and neurological basis, and poor inter-reliability are fully compatible with a natural kind view of mental disorders.

have brought about a natural kind of mental disorder, the answer to the normative question becomes less clear, as maintaining it as a taxonomical term is precisely what makes people act the way they do, and in a sense, keeps them disordered in the described way. Removing the kind or renaming it (as was done with multiple personality disorder), might be therapeutically effective and the right course of action. The scientific agenda and the clinical agenda may run apart at this point. For scientific research it is important to hold onto the categories, if only to discover that it has a social and self-fulfilling inner structure. For therapeutic reasons, they are probably best renamed or removed.

This is the degree to which a natural kind view of mental disorders can be defended. I will end with a few words about what it would mean if the science find that at a causal level heterogeneity and complexity is irreducible—i.e. if empirical research finds that there are no natural kinds of disorders. Then what? The first thing to note is that people can still be disordered (by falling short of norms of mental functioning), and that their disorderdness is not any less real. A second point is that the kind-terms would prove not to mean all that much: the diagnostic categories would just be a shorthand, or summary term, for a set of exhibited behaviours. The taxonomical terms would be purely descriptive, instead of explanatory. They don't mean what everyone thought they meant (pointing to shared underlying causes). Psycho-diagnosis would no longer be a matter of placing someone under a category with the help of application criteria, with evidence-based knowledge about effective interventions.[36] This approach would proven largely mistaken.

This feeds into the final point: also in the absence of natural kinds, there will always be causes at an individual level (proximal and distal) that explain the experience problems and relevant symptoms. The only implication of people not falling into a natural kind is that those causal structures aren't present in other people in a sufficiently similar way. A well-trained psychiatrist or psychologist may still get an idea about how the condition has come about in individual cases, and on that basis devise an effective intervention strategy. Knowledge of other cases with similar symptomatology, however, would in this scenario be of comparatively limited value. Diagnostic practice would consist in trying to form an understanding of how the problematic way of mental functioning has come about *in this unique* case and devising therapeutic interventions on that basis of it. Personalised medicine is nothing new in areas of somatic medicine; it may also hold the future for large parts of psychiatry and clinical psychology. Even though I have argued that the sciences may expose natural kinds of mental disorder, a particularistic and individualised approach may, in the end, be best suited for this area of health care.

[36] The purely descriptive sense of taxonomical terms is intended by the *DSM-IV* and *DSM-5*. In a sense, it is too early to tell whether the terms indeed are purely descriptive. In the meantime, however, it seems clear that many—lay people and professionals alike—take them to be natural kind terms.

References

Ahn, W., Flanagan, E., Marsh, J., & Sanislow, C. (2006). Belief about essences and the reality of mental disorders. *Psychological Science, 17*(9), 759–766.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, fifth edition: DSM-5*. American Psychiatric Association.

Arpaly, N. (2005). How it is not "just like diabetes": Mental disorders and the moral psychologist. *Philosophical Issues, 15*, 282–298.

Beebee, H., & Sabbarton-Leary, N. (2010). 'Are psychiatric kinds 'real'? *European Journal of Analytic Philosophy, 6*(1), 11–27.

Bolton, D. (2012). Classification and causal mechanisms: A deflationary approach to the classification problem. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 6–11). Oxford: Oxford University Press.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*, 5–13.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91–121.

Borsboom, D., Cramer, A., & Kalis, A. (2018). 'Brain disorders? Not Really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences, 24*, 1–54.

Boyd, R. (1989). 'What realism implies and what it does not. *Dialectica, 43*(1–2), 5–29.

Boyd, R. (1990). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies, 61*, 127–148.

Broome, M. R., Woolley, J. M., Tabraham, P., Johns, L. C., Bramon, E., Murray, G. K., Pariante, C., McGuire, P. K., & Murray, R. M. (2005). What causes the onset of psychosis? *Schizophrenia Research, 79*, 23–34.

Clementz, B. A., Sweeney, J. A., Hamm, J. P., Ivleva, E. I., Ethridge, L. E., Pearlson, G. D., Keshavan, M. S., & Tamminga, C. A. (2016). Identification of distinct psychosis biotypes using brain-based biomarkers. *The American Journal of Psychiatry, 173*(4), 373–384.

Cooper, R. (2002). Disease. *Studies in History and Philosophy of Biological and Biomedical Sciences, 33*, 263–282.

Cooper, R. (2005). *Classifying madness: A philosophical examination of the diagnostic and statistical manual of mental disorders*. Springer.

Cooper, R. (2007). *Psychiatry and philosophy of science*. Acumen Publishing Ltd.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology, 22*(5), 575–594.

Engelhardt, H. T. (1976). Ideology and etiology. *Journal of Medicine and Philosophy, 1*, 256–268.

Engelhardt, H. T. (1986). *The foundations of bioethics*. Oxford University Press.

Eyal, G. (2017). Autism looping. In *Routledge international handbook of critical mental health* (pp. 141–149).

Fulford, K. W. M. (1989). *Moral theory and medical practice*. Cambridge University Press.

Goosens, W. (1980). Values, health and medicine. *Philosophy of Science, 47*, 100–115.

Gorenstein, E. (1992). *The science of mental illness*. Academic Press.

Graham, G. (2010). *The disordered mind: An introduction to the philosophy of mind and mental illness*. Routledge.

Graham, G. (2014). Being a mental disorder. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying psychopathology: Mental kinds and natural kinds* (pp. 123–144). MIT Press.

Graham, G., & Stephens, G. L. (2007). Psychopathology: Minding mental illness. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science* (pp. 339–367). Amsterdam: Elsevier.

Hacking, I. (1986). Making up people. In S. Heller & D. E. Wellbery (Eds.), *Reconstructing individualism: Autonomy, individuality, and the self in western thought* (pp. 222–236). Stanford: Stanford University Press.
Hacking, I. (1991). A tradition of natural kind. *Philosophical Studies, 61*, 109–126.
Hacking, I. (1995a). *Rewriting the soul: Multiple personality and the science of memory*. Princeton University Press.
Hacking, I. (1995b). The looping effect of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). New York: Oxford University Press.
Hacking, I. (1998). *Mad travelers: Reflections on the reality of transient mental illnesses*. Harvard University Press.
Hacking, I. (1999). *The social construction of what?* Harvard University Press.
Hacking, I. (2007a). Natural kinds: Rosy dawn, scholastic twilight. *Royal Institute of Philosophy, 61*(Supplement), 203–239.
Hacking, I. (2007b). Kinds of people: Moving targets. *Proceedings of the British Academy, 151*, 285–318.
Haslam, N. (2000). Psychiatric categories as natural kinds: Essentialist thinking about mental disorders. *Social Research, 67*(4), 1031–1058.
Haslam, N. (2002). Kinds of kinds: A conceptual taxonomy of psychiatric categories. *Philosophy, Psychiatry, & Psychology, 9*(3), 203–217.
Haslam, N. (2014). Natural kinds in psychiatry: Conceptually implausible, empirically questionable, and stigmatizing. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying psychopathology: Mental kinds and natural kinds* (pp. 11–28). MIT Press.
Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine, 50*, 1418–1432.
Hempel, C. G. (1965). Fundamentals of taxonomy. In C. G. Hempel (Ed.), *Aspects of scientific explanation: And other essays in the philosophy of science* (pp. 137–154). New York: The Free Press.
Hyman, S. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology, 6*, 155–179.
Insel, T. (2013). Transforming diagnosis. Retrieved October, 2019, from https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2013/transforming-diagnosis.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. *American Journal of Psychiatry, 167*(7), 748–751.
Kendler, K. S., & Parnas, J. (Eds.). (2008). *Philosophical issues in psychiatry: Explanation, phenomenological, and nosology*. John Hopkins University Press.
Kendler, K. S., & Parnas, J. (Eds.). (2012). *Philosophical issues in psychiatry II: Nosology*. Oxford University Press.
Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine, 41*(6), 1143–1150.
Kincaid, H., & Sullivan, J. A. (Eds.). (2014). *Classifying psychopathology: Mental kinds and natural kinds*. MIT Press.
Kripke, S. (1980). *Naming and necessity*. Basil Blackwell.
LaPorte, J. (2004). *Natural kinds and conceptual change*. Cambridge University Press.
Lecei, A., van Hulst, B., de Zeeuw, P., van der Pluijm, M., Rijks, Y., & Durston, S. (2019). Can we use neuroimaging data to differentiate between subgroups of children with ADHD symptoms: A proof of concept study using latent class analysis of brain activity. *NeuroImage: Clinical, 21*, 101601.
Mill, J. S. (1974) [1884]. A system of logic: Ratiocinative and inductive. In: J. M. Robson (Ed), *Collected works of John Stuart Mill Volume VII*. Toronto: University of Toronto Press.
Murphy, D. (2006). *Psychiatry in the scientific image*. MIT Press.
Okasha, S. (2002). Darwinian metaphysics: Species and the question of essentialism. *Synthese, 131*, 191–213.
Pickard, H. (2009). Mental illness is indeed a myth. In M. Broome & L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience* (pp. 83–101). Oxford University Press.
Poland, J. (2014). Deeply rooted sources of error and bias in psychiatric classification. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying psychopathology: Mental kinds and natural kinds* (pp. 29–64). MIT Press.

Poland, J., Von Eckardt, B., & Spaulding, W. (1994). Problems with the DSM approach to classifying psychopathology. In G. Graham & G. L. Stephens (Eds.), *Philosophical psychopathology* (pp. 235–260). MIT Press.

Putnam, H. (1975). The meaning of "meaning." *Minnesota Studies in the Philosophy of Science, 7*, 131–193.

Spitzer, R. L., & First, M. B. (2005). Classification of psychiatric disorders. *Journal of the American Medical Association, 294*, 1898–1900.

Sullivan, J. A. (2014). Stabilizing mental disorders: Prospects and problems. In H. Kincaid & J. A. Sullivan (Eds.), *Classifying Psychopathology: Mental kinds and natural kinds* (pp. 257–281). MIT Press.

Tsou, J. Y. (2007). Hacking on the looping effects of psychiatric classifications: What is an interactive and indifferent kind? *International Studies in the Philosophy of Science, 21*(3), 329–344.

Tsou, J. Y. (2015). DSM-5 and psychiatry's second revolution: Descriptive vs. theoretical approaches to psychiatric classification. In S. Demazeux & P. Singy (Eds.), *The DSM-5 in perspective: Philosophical reflections on the psychiatric babel* (pp. 43–62). Springer.

Tsou, J. Y. (2016). Natural kinds, psychiatric classification, and the history of the *DSM*. *History of Psychiatry, 27*(4), 406–424.

Tsou, J. Y. (2019). Philosophy of science, psychiatric classification, and the DSM. In Ş Tekin & R. Bluhm (Eds.), *The bloomsbury companion to philosophy of psychiatry* (pp. 177–196). Bloomsbury.

Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist, 47*(3), 373–388.

Zachar, P. (2000a). *Psychological concepts and biological psychiatry: A philosophical analysis*. John Benjamins Publishing Company.

Zachar, P. (2000b). Psychiatric disorders are not natural kinds. *Philosophy, Psychology and Psychiatry, 7*(3), 167–182.

Zachar, P. (2003). The practical kinds model as a pragmatist theory of classification. *Philosophy, Psychology and Psychiatry, 9*(9), 219–227.

Zachar, P. (2014). Beyond natural kinds: Towards a "relevant" "scientific" taxonomy in psychiatry. In H. Kincaid & J. A. Sullivan (Eds.), *(2014), Classifying psychopathology: Mental kinds and natural kinds* (pp. 75–104). MIT Press.