Computational Psychiatry and the Evolving Concept of a Mental Disorder

Konstantin Genin, Thomas Grote, Thomas Wolfers

Abstract: As a discipline, psychiatry is in the process of finding the right set of concepts to organize research and guide treatment. Dissatisfaction with the status quo as expressed in standard manuals has animated a number of computational paradigms, each proposing to rectify the received concept of mental disorder. We explore how three different computational paradigms: normative modeling, reinforcement learning and network theory, reconceptualize mental disorders. Although each paradigm borrows heavily from machine learning, they differ significantly in their methodology, their preferred level of description, the role they assign to the environment and, especially, the degree to which they aim to assimilate psychiatric disorders to a standard medical disease model. By imagining how these paradigms might evolve, we bring into focus three rather different visions for the future of psychiatric research. Although machine learning plays a crucial role in the articulation of these paradigms, it is clear that we are far from automating the process of conceptual revision. The leading role continues to be played by the theoretical, metaphysical and methodological commitments of the competing paradigms.

Keywords: mental disorders, machine learning, concepts in psychiatry, computational methods, conceptual change

1. Introduction

Psychiatric research and practice presuppose, to a large degree, a conceptual taxonomy that allows mental disorders to be identified and distinguished from each other on the basis of observable symptoms. The currently prevailing taxonomy is expressed in the International Classification of Disease (WHO, 2021) and the Diagnostic and Statistical Manual of Mental Disorders (DSM) (APA, 2022). While the classifications outlined in these standard manuals are important contributions to our conceptualization of mental disorder, there is significant dissatisfaction with the prevailing taxonomy. The concepts are often criticized as insufficiently fine-grained to capture the high degree of heterogeneity within, and comorbidity between, mental disorders (van Loo & Romeijn, 2015; Petrolini & Vicente, 2022). The clinical consequences of this situation is that a DSM diagnosis does not clearly indicate any precise course of treatment. Given this state of affairs, it is natural to suspect that beneath the DSM categories lies a great deal of biological heterogeneity, a suspicion seemingly confirmed by the fact that psychiatric research has so far failed to identify reliable biomarkers for mental disorders (Tabb & Lemoine, 2021). Some amount of conceptual house-cleaning seems to be in order.

As a discipline, psychiatry is rather in the process of finding the right set of concepts to organize research and guide treatment (see Aftab & Ryznar (2020) for an excellent history of psychiatric nosology). Dissatisfaction with the status quo has animated alternative research frameworks, with the Research Domain Criteria (RDoC) project by the National Institute of Mental Health (NIMH) being one of the most prominent examples (Insel et al., 2010). Research initiatives such as RDoC suggest that mental disorders should be investigated across cognitive domains and levels of biological description. Mapping genetic, neurological, and cognitive data is suggested to contribute to a better understanding of the full spectrum of mental disorders (Tabb, 2019). The methodological guidance provided by RDoC is, however, still unclear. The main message for researchers is that they should continue to raid the methodological armory of modern clinical science until they find some way to ground psychiatric concepts in biological substrata.

Under the banner of computational psychiatry[1], several approaches have been proposed to answer this call for conceptual and methodological reform. Flush from a wave of victories in other fields, machine learning (ML) and data science methods promise a way to leverage large datasets to tame the heterogeneity within disease concepts and deliver more precise diagnoses and treatment recommendations. Normative modeling (Marquand et al., 2016; Rutherford et al., 2022) is a heterogeneity-mapping strategy aimed primarily at the refinement of existing disease concepts. Others propose more radical revision. Network models (Borsboom & Cramer, 2013; Borsboom, 2017) demur from the search for biological substrata and suggest that we re-conceptualize psychiatric disorders as, at least so far as clinical practice is concerned, nothing more than causal networks of interacting and potentially self-reinforcing behavioral symptoms. These approaches suggest targeted therapeutic interventions intended to shift individuals, conceived of as complex dynamical systems, out of pathological equilibria. Reinforcement learning approaches (Montague et al., 2012; Huys et al., 2015; Colombo, 2022) reconceptualize mental disorders as the outcome of a (largely rational) learning process in which the learning mechanism was either poorly calibrated or encountered unfortunate initial conditions. This suggests interventions aimed at recalibrating the individuals' learning strategies.

Each of these approaches suggest revisions to the existing menagerie of psychiatric concepts. Strikingly, one encounters nearly as much methodological heterogeneity in this set of novel approaches as we lamented within the prevailing concepts of psychiatric disease. In this paper, we attempt to answer two questions. How, exactly, do the different modeling techniques re-conceptualize mental disorders? And what promise do these conceptual changes hold for more precise and reliable clinical practice? These questions are pressing, because computational psychiatry is an inherently pluralist field, where the interrelations between different approaches are inadequately understood. Hence, investigating how machine learning models re-conceptualize or refine mental disorders is important for two reasons: (i) it enables a better understanding of how developments in computational psychiatry can translate into clinical applications, and (ii) it facilitates coordination between different computational methods.

2. Psychiatry and the Need for Conceptual Change

A mental disorder concept fulfills a number of different functions: it helps to systematize research and guide the design of experiments; it plays an important communicative role for patients and their family members; it serves a key organizing role in the various bureaucracies concerned with medical payment and reimbursement. Ideally, correctly attributing an individual's clinical symptoms to a specific disorder also enables effective treatment (Van Loo et al., 2019). According to the traditional two-part model of pathology, dysfunctional organic processes internal to the individual manifest themselves in various observable signs. Together, the internal processes and the complex of observable signs jointly determine a disease concept (Radden, 2023). Typically, diagnoses are suggested by symptoms apparent in gross examination but only confirmed or disconfirmed by the presence of *biomarker*s. Paradigm cases of biomarkers include the presence of certain molecules in bodily tissues, organs or fluids. For example, the WHO compiles height and weight velocity charts for children. If pediatricians record statistically abnormally low height velocities, adjusted for age and sex, they are advised to start endocrinological investigation. So statistically abnormal height velocity, apparent in gross examination, is taken as suggestive of malfunction in endocrine processes, confirmed or disconfirmed by biomarkers present in blood samples.

Strikingly, psychiatry is one of the few remaining areas of medicine in which biomarkers remain to be developed. Although biomarkers play a marginal role in some cases—for example, in ruling out brain tumors—diagnoses are still made largely on the basis of gross symptoms. As codified by the DSM, the standard approach is to diagnose mental disorders by consulting a checklist of ostensibly distinct, clinically meaningful, cognitive, emotional, or behavioral symptoms associated with distress or social disability. A patient can be diagnosed with a given mental disorder if they meet a significant subset of these symptoms. For roughly half of the disease categories in the DSM, it is possible for two individuals to cross the diagnostic threshold without sharing any of the same symptoms (Olbert et al., 2014). On the traditional two-part model of disease, DSM categories underdetermine a disease concept since they are silent about the internal processes which give rise to the cluster of symptoms. More dramatically, they are agnostic between conventionalist and realist interpretations of the diagnostic clusters: are mental disorders *simply* conventional clusters of co-occurring symptoms, or are these symptoms the outward expressions of a malfunction in some underlying mechanism as yet undiscovered?[2]

This agnosticism may sidestep difficult philosophical disputes, but it has not helped to forestall dissatisfaction with the reliability and validity of DSM categories. Three problems that stand out are heterogeneity, comorbidity, and the lack of causal support. Consider each in turn:

- *Heterogeneity*: DSM categories suggest a certain degree of homogeneity: patients that share the same diagnosis are meant to be sufficiently similar to support clinical generalizations. However, in light of the variability in symptoms and the different developmental pathways in patients diagnosed with the same mental disorder, this homogeneity assumption seems to be misguided (Petrolini & Vicente, 2022). DSM categories do not seem to be sufficiently fine-grained to facilitate reliable treatment recommendations for individual patients. The presumption of homogeneity is also reflected in biological psychiatry, where case-control comparisons seek to identify average biological differences between a group of patients and "healthy" controls. If DSM categories obscure a great deal of biological heterogeneity, these studies would not be expected to identify meaningful average differences.
- *Comorbidity*: It is common that different mental disorders, such as major depressive disorder and major anxiety disorder co-occur in a patient. The different mental disorders, in turn, interact with each other – resulting in more severe courses of

disease and the appearance of novel features that are not characteristic of basal conditions of the individual disorders. Aside from being another source of heterogeneity, the common occurrence of comorbidity casts doubt on realist interpretations of DSM categories (van Loo & Romeijn, 2015). If, as Petrolini and Vicente (2022) argue, comorbidity is (1) rather the rule than the exception; (2) leads to novel symptoms that differ from paradigm instances of the constituent diseases and (3) requires therapeutic approaches that differ from those usually suggested for the constituent diseases, then it is no longer clear that the constituent disease concepts are really fit for guiding diagnostic and therapeutic decision-making.

- *Lack of Causal Support*: Like other branches of medicine, psychiatry has embraced the goal of precision medicine, which is to provide "the right treatment to the right patient at the right time". The main driver for precision medicine in fields like oncology has been the discovery of fine-grained biomarkers. These are crucial not only for facilitating reliable diagnosis but also for discovering specific causal pathways that can be targeted by novel drug therapies (Tabb & Lemoine, 2021). However, neither do DSM categories provide any explanations regarding the etiology of mental disorders, nor has psychiatry been successful so far in identifying reliable biomarkers or conclusively implicating any physiological mechanisms. One possible explanation for the inability of psychiatry to detect biomarkers is the daunting complexity of the human brain, containing billions of neurons and supporting cells, and huge networks that interact in complex ways with the brain's environment over time.

The picture that emerges against this backdrop is that DSM categories are not fit for the purpose of modern psychiatry. Emerging frameworks, such as RDoC, are in part an expression of frustration with the degree to which psychiatry persistently deviates from the traditional model of disease. Systematic research across neurological, genetic and cognitive data is meant to identify the biological causes of mental disorders, assimilating them to the traditional two-part model of pathology (Insel et al., 2010). The hope of RDoC is that psychiatry can thus be made a “normal” field of medicine. Although RDoC encompasses many different computational approaches, not all approaches in computational psychiatry are similarly motivated. Indeed, approaches like network modeling and, perhaps to a lesser extent, reinforcement learning, also represent a departure from the traditional two-part model of disease. Understanding whether, and to what extent, these different approaches depart from the two-part model is a prerequisite for getting a grip on how mental disorders are being reconceptualized by emerging methodologies.

Of particular interest in this context are machine learning techniques, which are becoming increasingly prominent in the tool-kit of psychiatric research. In recent years, there has been a fundamental shift in the understanding of the promise of machine learning. Before the advances in deep learning, machine learning models were understood as continuous with traditional statistical techniques and given a strongly instrumentalist interpretation (Anderson, 2008): these were merely predictive machines whose internal representations–-insofar as they could be said to have any—concealed no interesting conceptual content. In contrast, the “latent space” of highly predictive neural networks is now frequently claimed to express a novel perspective on the world. If this perspective could only be understood, it might enrich the conceptual treasury of the sciences in stunning and unpredictable ways (Boge, 2022; Buckner 2020; Yarkoni & Westfall, 2017).

On this latter “optimistic” perspective, machine learning methods might yield conceptual change in psychiatry simply by nature of the novel perspectives those methods contribute to the scientific debate. However, while these methods have shown to be useful in identifying predictive features that can refine existing scientific concepts, the promise of machine learning to induce conceptual shifts all on its own is still largely hypothetical. A “pessimistic” perspective, by contrast, emphasizes that researchers’ paradigms and presuppositions constrain both the development of machine learning models and the interpretation of their output—we will not break free from existing conceptual paradigms merely by the application of machine learning methods alone (Ratti, 2020).

The approaches discussed here lie mostly on the pessimistic side of this spectrum. Normative modelers estimate individual-level deviations from a "healthy" norm and suggest that stratifying by these patterns of statistical deviation will yield apt psychiatric categories. Since individuals are recruited on the basis of standard psychiatric diagnoses, this is more likely to result in a subtle *conceptual refinement* rather than in a *conceptual revision* that profoundly changes our understanding of mental disorders. Other approaches, such as reinforcement learning and network modeling, make theory-driven conceptual revisions that are supported by, but do not emerge from, their computational methods. These reconceptualizations are made by the theorists themselves—it is their concepts which frame the results of the computational methods and not vice-versa. 'Refinement' and 'revision' also operate within a spectrum; what is at stake for the latter is that it leads to a reorganization of classification structures (Thagard, 1990). Both developments may ultimately give rise to more effective psychiatric treatments.

Psychiatric researchers tend to be externally cautious in framing their preferred research paradigm. However, the rhetorical caution of their perspective papers belies the methodological boldness of their actual research. To understand how their research refigures psychopathological nosology, it is crucial to attend to the methodological details. For that reason, we attempt throughout to give a critical review of the competing methodologies proposed by the varied approaches in computational psychiatry.

In what follows, we turn to a detailed study in how different computational paradigms bear on the conceptualization of mental disorders. We focus especially on whether, and to what extent, these approaches diverge from the traditional two-part model of pathology on which a disease concept is determined by (i) a malfunction in some internal biological system and (ii) the complex of observable signs and symptoms arising from this malfunction. By considering to what extent machine learning models can induce conceptual revision, our paper also contributes to the understanding regarding the boundary conditions for machine learning methods in the sciences.

3. Normative Modeling

Researchers have been using statistical clustering methods to search for homogeneous subgroups of psychiatric categories since at least Paykel (1977) and Farmer et al. (1983). Most attempts at subtype clustering are driven by symptoms alone. Even when biological measures are available to the clustering algorithm, these methods have trouble distinguishing between nuisance variation and relevant biological variation and are sensitive to rather arbitrary decisions about the desired number of subtypes. The resulting clusters are highly variable across studies and are subject to little if any external variation. (See Marquand (2016) for a review of clustering methods in psychiatry). With few exceptions (perhaps Zhang et al., 2021) these clustering approaches have not identified any clinically relevant subtypes. Normative modeling is an attempt to circumvent the difficulties that attend unsupervised clustering approaches.

We begin with a high-level overview of the methodology of normative modeling. First, a suitable "normative" reference population is recruited, usually a healthy, non-clinical population. We record baseline demographic features of this population and measure some interesting, contextually relevant biological variables. For example, we might record baseline features like age and sex and, using whole brain imaging, collect measurements of gray matter volume in a large number of loci. Using the data we have collected from the control population, we then train a "normative" model that predicts the biological features from the baseline covariates: in the running example, it predicts gray matter volumes from baseline covariates such as age and sex. Next, we collect a population of patients that have been diagnosed with some psychiatric disorder, for example ADHD. We collect the same baseline demographic data and biological data for the clinical population. Then, for each patient in the clinical group, we compare the gray matter volumes predicted by the normative model with the gray matter volumes observed in imaging and record the degree of deviation. The latter gives us a sense of how each individual varies from the norm in each of the measured brain regions. When it makes sense, we can also use the normative model to predict gray matter volumes for an "average" patient. For example, if the average (male) with ADHD is 33.33 years old, we can use the model to predict gray matter volume for this "synthetic" individual. This gives us a sense of how the "average (male) clinical patient" differs from the norm.

In its broad strokes, this procedure has been repeated for autism spectrum disorder (Zabihi et al., 2019), attention deficit hyperactivity disorder (Wolfers et al., 2019), bipolar disorder and schizophrenia (Wolfers et al., 2018). The pattern of results emerging from these studies is striking: although there is a discernible deviation in gray matter volumes between the "average patient" and healthy controls, there is a surprising amount of heterogeneity in the way in which individuals deviate from the norm. For example, although it is relatively easy to find brain regions in which *some* patients with ADHD significantly deviate from the norm, it is hard to find a single brain region for which more than 2% of patients deviate significantly. In other words, although most patients deviate from the norm in *some* way, there is no *typical* way in which most patients deviate from the norm. It seems that none of the patients is well approximated by an "average" patient.

Normative modelers take these results to confirm the suspicion that our received disease categories conceal a great deal of biological heterogeneity. Hence, as a negative result, they highlight shortcomings of DSM categories of mental disorder. Moreover, as normative modelers argue, these results complement traditional case-control studies, which fish for average group-level differences between patients and controls with traditional statistical tests. Contrary to case-control studies, the aim is to develop a nuanced view of the variability of "norm-deviations" at the individual level. One natural idea is to apply clustering algorithms to the deviations from normative models to find subtypes of deviation (see e.g., Zabihi et al., 2020). This may allow clustering algorithms to focus on biologically relevant, rather than nuisance, variation.

Normative modelers argue that their approach provides valuable information to refine psychiatric nosology. But there are reasons to think that this may be premature. For one, it is not clear how much heterogeneity we should expect from a good disease category. It is true that no two ADHD patients are likely to deviate significantly from the norm in precisely the same loci.. It is unclear whether deviation in individual loci, rather than in the organization of functionally meaningful clusters, is the most biologically relevant level of granularity. For example, Segal et al. (2022) find a great deal of heterogeneity at the level of loci, but these deviations were embedded within common functional circuits and networks in up to 56% of cases.[3] Criticizing psychiatric nosology on the basis of extremely fine-grained heterogeneity might be like criticizing the concept of astigmatism on the grounds that no two astigmatic cornea deviate from perfect rotational symmetry in precisely the same way. Until we know how much heterogeneity to expect from a "normative" disease category, we do not know whether psychiatric disorders meaningfully deviate from the norm. Are psychiatric disorders more biologically heterogeneous than epilepsy, Alzheimer's, Parkinson's, or migraine? Luckily for normative modelers, this question can be answered with more normative modeling.

Normative modeling answers many of our demands for a reconstructed psychiatric nosology. It provides a framework for integrating symptoms and biological substrata and offers a methodological tool-kit for investigating the heterogeneity of mental disorders.

Although it is not fated that patterns of biological deviation from the norm will be clinically informative, normative modeling suggests subtyping strategies that could lead to refined disease concepts and more effective treatment. Moreover, it assimilates psychiatric disorder to the biostatistical notion of disease (Boorse, 1975/77), according to which disease is largely a matter of deviation from statistically normal functioning. It therefore inherits the strengths of the biostatistical theory: if disease is a matter of statistical deviation from a norm, then nosology can be perfected with a sufficiently subtle statistical methodology. Moreover, this methodology will look a lot like one familiar from other areas of medicine and natural science. Accordingly, normative modeling inherits the weaknesses of the biostatistical theory: the question of how the normative population is to be selected is left largely unanswered and it becomes difficult to distinguish disorder from mere deviance (Aftab & Rashed, 2021). Moreover, the results of any normative modeling exercise will depend delicately on who precisely is taken as the normative reference point (Kingma, 2007). Thus, like all statistical enterprises, normative modeling suffers from a variant of the reference class problem, which is likely to resist a complete solution (Reichenbach, 1949; Hájek, 2007). Normative modelers can no doubt answer these challenges. However, normative modeling is not likely to yield a radical revision of our psychiatric categories—it rather presupposes the standard clinical categories with a view to charting the heterogeneity within them. Thus normative modeling is one of the most sophisticated expressions of the hope that the rough material of our received psychiatric categories can be chiseled into the apt concepts of a future psychiatry.

4. Reinforcement Learning

The family of approaches collected here under the heading of "reinforcement learning" encompasses techniques drawing from economics and economics-adjacent fields such as Bayesian epistemology, game theory and decision theory as well as reinforcement learning proper. Nevertheless, these approaches are sufficiently similar in methodology that they admit of a common high-level description. First, a battery of cognitive tasks is selected and the performance of "healthy" controls and clinical populations on these tasks is recorded. The tasks typically involve some kind of sequential decision making and may or may not also involve strategic interactions with other subjects. Ideally, the task is sufficiently simple and well-studied that the researchers can appeal to a body of theory about its neurological underpinnings. The researchers select a mathematical model that characterizes the essential features of the decision-making process. The model may be selected on the basis of theory, or by data-driven exploration. Crucially, the model must be both (1) sufficiently flexible to faithfully reconstruct the behavior of participants from both populations and (2) sufficiently simple that the decision-making behavior can be captured with a small number of interpretable model parameters. The parameters thus estimated for each patient constitute their "computational phenotype" and, ideally, capture some essential feature of their decision-making disposition.

These computational phenotypes are precise, low-dimensional characterizations of individual cognitive dispositions at an intermediate level of description. They are more precise (but also more theory-laden) than phenomenological symptoms, but also closer to behavior, and therefore more interpretable, than neurophysiological data (Montague et al., 2012). Many researchers in this area hope that clusters of computational phenotypes would yield important, clinically-relevant subtypes of existing categories of psychiatric disorders. However, these phenotypes tend to be "trans-diagnostic," i.e. the same computational phenotype may be shared across a number of standard disorder categories. Finally, there is often an effort to corroborate computational phenotypes with differential patterns of neurophysiological functioning. Researchers hold out hope that sub-typing by computational phenotypes will be the key to understanding why some pharmacological interventions work for some patients and fail for others.

Approaches drawing on reinforcement learning have been widely applied in the study of compulsive disorders such as problem gambling, binge eating or substance abuse. The characteristic features of compulsive disorders is that problem behavior persists even though patients know that it leads to unwanted consequences and despite their expressed desire to abstain. A common explanation for this phenomenon is that decisions can arise from two distinct systems of control, the first "goal-directed" and deliberative and the second "habitual". The goal-directed system recommends choices based on their likely outcomes as predicted by a model of the decision-making task. The habitual system recommends choices that were previously rewarded. Voon et al. (2015), summarize the distinction with a pithy slogan: "goal-directed choices are *prospective*, whereas habitual choices are *retrospective* " (p. 345).

These two different decision making systems are modeled with different learning strategies adopted from the reinforcement learning literature. Habitual decision making is operationalized as a computationally inexpensive "model-free" update, wherein the value of an state-action pair is increased when it leads to a reward and decreased otherwise. Goal-directed decision making is operationalized as a computationally expensive "model-based" update, wherein the learner maintains a model of the transition structure of the decision making task and computes expected rewards by simulating possible trajectories through the states.

Reinforcement learning approaches to psychiatry have developed a relatively standardized sequential decision-making task to distinguish the degree to which an individual relies on model-based rather than model-free decision-making (Dew et al., 2005). For this task, model-based learners would distinguish the "reason" for which an action-state pair is rewarded or unrewarded whereas model-free learners would not. In the first stage, participants are presented with two choices A and B. Participants are instructed that their choice in the first stage will lead to either a blue or a green screen, where they will be presented with choices C and D or E and F, respectively. Each of the choices at the second stage will lead to a reward with its own probability evolving independently according to a random process. Crucially, participants are told that A is much more likely to lead to the green screen whereas B is much more likely to lead to the blue screen. With high probability, a purely model-free learner will stay with their previous first-stage choice if it leads to a reward and switch if it does not, irrespective of whether they saw a probable or an improbable transition. On the other hand, the stay/switch behavior of a model-based learner depends on whether their first-stage choice led to a probable or improbable transition. If their choice is rewarded after a probable transition, they are likely to stay; if it is rewarded after an improbable transition, they are likely to switch. Conversely, if their choice is not rewarded after an improbable transition, they are likely to stay; and if it is not rewarded after a probable transition, they are likely to switch. By observing the stay/switch behavior of participants, researchers can estimate a parameter, or computational phenotype, characterizing the degree to which they more resemble the purely model-based or model-free learner.

Applying this strategy, researchers claim to have identified impaired model-based decision making in patients with binge eating disorder, obsessive-compulsive disorder and methamphetamine dependence (Voon et al., 2015), schizophrenia (Culbreth et al., 2016), and problem gambling (Wyckmans et al., 2019). Although extensively studied, the results for patients with alcohol use disorder are highly equivocal: Sebold et al. (2014) show impaired model-based decision making for patients with alcohol use disorder, whereas Voon et al. (2015) and Sebold et al. (2017) find no difference between patients with alcohol use disorder and healthy controls. Interestingly, Sebold et al. (2014) and Wyckmans et al. (2019) show that patients only deviate significantly from model-based decision-making after unrewarded trials, suggesting that goal-directed decision-making is inhibited in the context of disappointment or frustration. This is consistent with studies showing that stress also inhibits goal-directed decision making (Soares et al, 2012). This raises the possibility that clinical populations are simply more stressed than others. Furthermore, it is widely accepted that working memory is an important factor in learning performance and that standard designs may confound the influence of working memory with that of learning dispositions (Collins & Frank, 2012). This illustrates what Wiecki et al. (2015) call the "task-impurity" problem: that no cognitive task measures just one construct but rather a mixture of distinct cognitive processes. In the case of compulsive disorders, the task impurity problem raises the worrying possibility that, rather than identifying their causes, researchers are merely studying the neurotoxic effects of substance disorders. This worry is allayed somewhat by similar patterns of results in problem gambling (Wyckmans et al., 2019), which presumably does not have the same neurotoxic effects.

The picture emerging from the reinforcement literature on compulsive disorders is that the diverse set of symptoms and disorders are to be unified as the outer manifestations of a deficiency in executive function characterized by an over-reliance on model-free decision making, especially in the wake of disappointments and setbacks.[4] Researchers hypothesize neurophysiological origins in the dopamine system (Voon et al., 2015) and genetic risk factors (Doll et al., 2016). On the basis of these hypothesized physiological causes, researchers propose plausible biomarkers and clinical interventions targeting executive function (Wyckmans et al, 2019). Thus, this research program can be seen as hewing straightforwardly to the traditional two-part model of disease, where symptoms are the outer manifestation of dysfunction in internal systems.

In contrast to work on compulsive disorders, the reinforcement learning program in the study of depression is harder to straightforwardly assimilate to the two-part disease model. Adopting a broadly Bayesian framework, Huys et al. (2015) hypothesize that depression results from subjective priors assigning high probability to states leading to low rewards, irrespective of which act is performed. Pessimistic priors are proposed as a unification of a variety of behaviors observed in depressed patients. If all acts are considered to be equally futile, it is rational to avoid exertion, resulting in symptoms of psychomotor retardation and subjective feelings of helplessness, fatigue and lack of energy. If negative outcomes are inevitable, then aversive information is more informative and can prepare the agent for the worst, predicting rumination on negative information. Pessimistic priors can also lead to a vicious feedback loop: the low expected value of exploration reduces the tendency to act and, consequently, the rate at which potentially corrective information is acquired. Low expectations may arise from cognitive biases or defective learning mechanisms. Crucially, however, they may also arise if an otherwise faultless learner encounters many aversive experiences (Maier & Watkins, 2005) or is situated in a truly low-utility environment (Lewinsohn et al., 1979). By directing the search for causes away from malfunctioning internal organic processes and toward the environment, reinforcement learning can deviate dramatically from the traditional disease model.

Lloyd et al. (2022) investigate how adverse childhood experiences affect learning behavior, especially in dispositions to balance exploration and exploitation. The canonical task used to study how agents manage the tradeoff between exploration and exploitation is a foraging task, in which agents decide whether to exploit a known patch whose propensity to generate rewards is relatively well-known, or explore a new patch with a different and unknown propensity to generate rewards. An important theorem of reinforcement learning characterizes the optimal strategy in such tasks: the forager should explore when the reward rate expected from the current patch is below the expected average reward rate in the environment, and exploit otherwise (Charnov, 1976). In line with the discussion of Huys et al.

(2015), environments with low expected average rewards favor exploitation, whereas environments with high average rewards favor exploration. Moreover, in rapidly-changing environments, older rewards are less informative in predicting future outcomes, therefore learners should favor more recent feedback (Behrens et al., 2007). Based on these theoretical results, Lloyd et al. predicted that compared to participants with fewer adverse childhood experiences, individuals with more adverse childhood experiences should (1) explore less; (2) weigh recent evidence more; (3) perform better in low-reward environments and (4) worse in high-reward environments. The empirical results of Lloyd et al. (2022) confirm only their first prediction. Although individuals with more difficult childhoods explore less than others, they do not seem to weigh recent evidence more, or perform better in more adverse environments. The explanation seems to be that individuals with difficult childhoods overweight negative outcomes and underweight positive ones, leading them to deviate from optimal learning even in adverse environments. The findings of Lloyd et al. suggest that disorders like depression are not merely rational strategies calibrated to adversity, but caused by learning dispositions impaired by excessive adversity.

The picture emerging from the reinforcement learning literature is not straightforwardly assimilable to the traditional two-part model of disease. Its emphasis on developmental learning history (see e.g., Giron et al., forthcoming) points to adversity in the environment, whereas its emphasis on miscalibrated learning dispositions points to malfunctioning in internal cognitive systems. Reinforcement learning synthesizes both "etiological" explanations, focusing on antecedent causes, and "constitutive-mechanistic" explanations, focusing on the functional organization of internal cognitive mechanisms that give rise to observed behavior (Piccinini, 2020; Colombo, 2022). While the latter tendency is orthodox and amenable to the search for intrinsic neurophysiological causes, the former is heterodox and suggests extrinsic social causes. Depending on the emphasis placed on intrinsic or extrinsic causes, the reinforcement learning program is either a confirmation of, or a challenge to, the traditional model of disease. Hence, there are interesting parallels with enactivist approaches in psychiatry, which challenge the causal primacy of the brain, and emphasize the importance of socio-cultural factors in explanations of mental disorders (de Haan, 2020). This could inspire therapeutic interventions that put increased emphasis on the (re-)design of background conditions (the specific environment but also socio-structural factors) for patients suffering from mental disorders. Moreover, while some reinforcement learners claim that computational phenotypes are candidate biomarkers (Wiecki et al., 2015) these biomarkers are highly unfamiliar: they are model parameters estimated from observed problem-solving behavior, not the results of tests for the presence of certain molecules in bodily tissues or fluids.

Reinforcement learning also stands in an equivocal relationship to the distinction between conceptual refinement and revision. On the one hand, Wiecki et al. (2015) suggest that computational phenotypes could be used as inputs to clustering algorithms and therefore support sub-typing strategies similar to those envisioned by normative modelers. From that perspective, reinforcement learning could support a project of conceptual refinement. On the other hand, the trans-diagnostic nature of computational phenotypes also suggests that the blooming variety of compulsive disorders could be unified as the outward manifestations of a disturbed learning disposition. From this perspective, comorbidity is a pseudo-problem emerging from an excessive pre-theoretical fine graining of disorders all stemming from the same common cause. For example, one reading of the literature suggests that the differences between compulsive disorders are largely superficial and that they can all receive a unified explanation grounded in overreliance on model-free learning. This seems like a significant revision of our received categories of mental disorder.

5. Network Theory

Both normative modeling and reinforcement learning approaches can be seen as searches for the hidden causes of the symptom clusters associated with mental disorder. Normative modeling hunts for causes in neurophysiology, whereas reinforcement learning looks for dysfunction in cognitive systems for learning and decision-making. In a striking departure, the network theory of mental disorder rejects the search for hidden causes altogether. According to network theory, "we cannot find central disease mechanisms for mental disorders because no such mechanisms exist" (Borsboom, 2017, p.5). The network theory holds that, rather than being effects of a common cause, psychiatric symptoms cause each other. Mental disorders arise when psychiatric symptoms enter into a pathological feedback-loop, where the presence of some symptoms promotes and sustains the presence of others. Thus a psychiatric disorder is nothing more than a pathological equilibrium state of a network of causally interrelated symptoms.

Network theorists represent patterns of symptom interaction in a network structure. Nodes in the network represent either symptoms, adopted from the DSM or other standard psychological scales and inventories, or "external factors". External factors may include adverse life events, or other non-symptom variables like abnormal brain function. Symptom nodes are connected if, and only if, they directly activate each other. The symptom network thus generates a topology in which some symptoms are more tightly connected than others, giving rise to mental disorders, i.e. groups of symptoms that arise together. According to network theory, what separates mental health from mental disorder is the topology of the symptom network: if the network is "strongly connected" then, when a symptom is activated by an external factor, its strong connections with the other symptoms send the network into a pathological state wherein symptoms activate and sustain each other, even when the activating external factor is no longer present; if the network is "weakly connected" then, although symptoms may be activated by external factors, the symptom network returns to a state of deactivation so long as the external factor is removed. Accordingly, mental health is defined as the “stable state of a weakly connected [symptom] network” and mental disorder as the “stable state of a strongly connected [symptom] network” (Borsboom, 2017, p. 9).

Network-theoretic strategies for diagnosis and treatment follow naturally from the basic commitments. Diagnosis involves identifying which symptoms are present and which interactions sustain them. Treatment involves intervening either on external factors, symptoms, or on the connections between symptoms. For example, antipsychotics intervene on symptoms of delusion, whereas cognitive behavioral therapy aims to weaken connections between symptoms by providing coping strategies. The dream of network theory is to couple a library of interventions to a set of network structures. Crucially, this pins the clinical hopes of network theory to the possibility of inferring individual network structures.

Of all the frameworks we have discussed, network theory receives perhaps the clearest and most explicit philosophical motivation. But how does it work in practice? We give a high-level overview of a typical network-theoretic application. First, symptom-level information is collected from one or more clinical populations. These symptoms, taken from standard diagnostic scales and inventories, correspond to the nodes in the network model. Then, for each pair of symptoms, researchers compute partial correlations. Roughly speaking, the partial correlation between two symptoms is a measure of the degree of association between them, controlling for all other symptoms in the network. Typically, some sparsity-bias is imposed to report only those partial correlations that are significantly different from zero. The resulting pattern of (sparsity-corrected) partial correlations is summarized in a graph where the strength of connection between two nodes is proportional to their (sparsity-corrected) partial correlation. Network theorists favor *partial* correlations, rather than simple correlations, because, in paradigmatic cases, the partial correlation can distinguish between the situation in which an association arises between two symptoms due to (i) a direct causal relationship between them and (ii) a shared common cause. In the latter case of “spurious” causation, the partial correlation will (ideally) be zero. Finally, network theorists search for systemically important symptoms using various measures of *centrality.* For example, a symptom with high *node strength* has many strong connections with many other symptoms. A symptom with high *betweenness* centrality lies on the shortest path connecting many other symptoms. Finally, a symptom with high *closeness* centrality is directly connected to many other symptoms. Since partial correlations between symptoms are typically positive, these measures of centrality capture different ways in which a symptom may be systemically important for maintaining the symptom network in a pathological state (Borsboom & Cramer, 2013; Borsboom, 2017).

In a paradigmatic example, Rhemtulla et al. (2016) collect responses to a standard battery of questions administered to patients diagnosed with various substance-use disorders. For example, questions measured (i) the tendency to use more than planned: "did you find that when you find that when you started using it, you ended up taking much more than you had planned?" (ii) tolerance: "did you find that you needed to use a lot more in order to (get high/feel its effects) than you did when you first started using it?" (iii) hazardous use: "did you ever use it in a situation in which it might have been dangerous?" and (iv) legal consequences: "did you have legal problems or traffic accidents because you were using it?" (p. 231). Separate symptom networks were estimated for users of cannabis, stimulants, opioids, sedatives, cocaine and hallucinogens. These various symptom networks reveal interesting similarities and differences. For example, the connection between inability to stop and hazardous use is strong across all substances; but the connection between hazardous use and legal consequences is absent for opioids, cocaine and hallucinogens but strong for sedatives. Moreover, regardless of the chosen measure, the tendency to use more than planned has high centrality for most substances, whereas tolerance is central for sedatives but not for hallucinogens.

Like reinforcement learning, network theory suggests that different types of substance-use disorder share a common underlying structure (as for each type, the same symptoms are involved). However, unlike reinforcement learning, network theory highlights that depending on the respective type of substance-use disorders, the node strength of the symptoms varies. In turn, network theory vindicates the existence of transdiagnostic criteria for substance-use disorders, while at the same time, accounting for fine-grained differences.

On first impression, network theory is a highly heterodox paradigm. By positing that mental disorders are best understood as patterns of causal interaction between symptoms, rather than as the hidden common causes of symptoms, network theorists not only depart from the standard two-part model of disease, they suggest that attempts to assimilate psychiatry to this traditional medical model are fundamentally misguided. On the other hand, network theory is rather orthodox in its attitude toward received DSM categories. Indeed, one of the fundamental assumptions of this approach is that the psychopathological symptoms found in standard diagnostic manuals are already defined at the right level of granularity and identify the relevant candidates for inclusion in network models. Moreover, diagnostic manuals standardly require practitioners to code not only the presence of symptoms but the connections between them. Thus, the received wisdom of the diagnostic manuals emerges as a kind of proto-network theory. Mental disorders, therefore, are not "ill-understood ephemeral entities, the nature of which will have to be uncovered by future psychological, neuroscientific or genetic research"; rather the fact that "we have the set of basic symptoms, and also understand many of the relations between them, means that we already have a quite reasonable working model of what disorders are and how they work" (Borsboom 2017, p. 11). Network theory therefore proposes a methodological impetus to organize the available empirical facts and revitalize the erstwhile degenerating DSM paradigm.

In their work on depression, Fried et al. (2015) offer an *in-paradigm* revision to the standard DSM criteria. They estimate a network for depression, where the nodes include both

DSM and non-DSM symptoms. Estimating measures of centrality, they find that both DSM and non-DSM symptoms are among the ten most central nodes. Core DSM symptoms like sadness and diminished pleasure and interest are highly central, but non-DSM symptoms like anxiety and panic are as well. Measures of centrality suggest revisions to the received DSM criteria emphasizing severe and central, and de-emphasizing peripheral, symptoms. The lesson to be learned is that network-theory can inform us about the validity of individual symptoms for a given mental disorder.

Network theory offers a promising framework for revitalizing the symptom-based paradigm. However its ambition seems often to exceed its methodological capabilities. Borsboom (2017) gives connections between nodes a strongly causal interpretation: nodes are connected if, and only if, they "directly activate each other" (p. 6). Crucially, however, connection strengths are typically estimated via partial correlation methods which, although reasonably capable of detecting spurious correlations due to common causes, are notoriously misleading when conditioning on common effects. For example, suppose that caffeine consumption and relationship problems both directly activate anxiety symptoms, but that neither directly causes the other. Controlling for symptoms of anxiety, caffeine consumption and relationship problems will be correlated, leading a naive partial correlation method to posit a spurious direct connection between them.[5] Indeed, the pitfalls of partial correlation were part of the inspiration for the development of methods for causal discovery from observational data, which might be fruitfully applied to symptom networks.[6]

Moreover, network theorists have pinned their hopes for clinical applications on the possibility of inferring individual symptom networks. Of course it is always difficult to make inferences about individuals with statistical methods, but the difficulty also arises at an earlier stage. Most methods for inferring network structure, including standard partial correlation methods, presume that the data are all sampled from a single structure. However, it is a presupposition of the network-theoretic approach that there is heterogeneity in the network structure, even among a single clinical population. Therefore, researchers must either (i) estimate separate network structures for populations that they have antecedently separated on the basis of supposed heterogeneity or (ii) apply some method for disentangling mixtures of network structures. The latter is an active area of research, the fruits of which might also be usefully applied for inferring symptom networks (e.g. Thiesson et al. 1997; Saeed et al. 2020).

6. Discussion

The three paradigms we have discussed represent strikingly different, and perhaps incompatible, responses to the crisis in psychiatry. By imagining how they will evolve, we can bring into focus three rather different visions for the future of psychiatric research. It is clear that the three differ on their preferred level of description. While network theorists stay at the symptom level, normative modelers are so far mostly interested in underlying neurophysiology. Meanwhile, reinforcement learners stake out a middle ground at the level of abstract cognitive/computational mechanisms. An irenic reading of the situation might imagine a peaceful division of labor: network theorists systematize the symptoms, reinforcement learners uncover their decision-theoretic causes, and normative modelers map out the underlying biological heterogeneity. However, these paradigms do not merely represent accidental differences of focus—there is a deeper metaphysical disagreement between them that, as a surface phenomenon, expresses itself in a preference for different levels of description. Network theory is, in its most natural interpretation, an anti-reductionist paradigm (Rathkopf, 2018). The network-theoretic preference for the symptom level is an expression of the conviction that looking for underlying common causes, and thereby assimilating psychiatry to the standard medical paradigm, is fundamentally misguided. In their current manifestations, reinforcement learning and normative modeling are, each in their way, reductionist paradigms: the former hoping to reduce psychiatry to neurophysiology, and the latter hoping for a reduction to an intermediate, cognitive level of description. That puts network theorists in an antagonistic posture toward the received two-part model of pathology that the others do not share. Conversely, network theorists are the most vocal defenders of the foundering DSM paradigm. While network theorists hope to right the ship by computational means, the others hope ultimately to decommission it.

These paradigms also differ on how they conceptualize the boundaries of mental disorder. Here it is reinforcement learning that is the odd-man-out: by emphasizing the effects of prior learning experiences, reinforcement learners open the door to social and environmental explanations of mental disorder. On the other hand, normative modelers and network theorists localize mental disorder to the individual: the latter implicate the connection structure of an individual's symptom network[7], while the former emphasize neurophysiological differences.

The different paradigms can also be individuated by different explanatory strategies. Reinforcement learning falls squarely into the mechanistic camp, which it supplements with an etiological perspective. The aim is to identify computational phenotypes that (mechanistically) explain the occurrence of mental disorders. The etiological perspective, by contrast, appeals to unfortunate learning histories to explain how decision-making mechanisms come to be miscalibrated. Normative modeling charts the underlying neurophysiology of mental disorders. While this can animate the hunt for causes or constitutive biomarkers, it is left unspecified how the neurological substrata are to be causally connected to the symptomology. Although it provides a sophisticated account of the dynamics of interaction between symptoms, network theory rejects mechanistic explanation, since it posits no "underlying" mechanisms.

From a different perspective, we are witnessing a paradigmatic dispute between conceptual lumpers and conceptual splitters. Reinforcement learners propose a unification of mental disorders as pathologies of a small number of learning and decision-making faculties. Beneath the blooming variety of psychiatric symptoms lie a small number of culpable learning mechanisms. Thus, comorbidity is to be expected, since many of these diseases are common effects of a single disordered mechanism, whereas heterogeneity is attributed to subtle differences in the small number of parameters characterizing learning and decision-making phenotypes. On the other hand, normative modelers propose that clustering clinical populations by their pattern of deviation from the healthy norm will reveal homogeneous, and consequently disentangled, sub-types of mental disorder. Network theorists reject both lumping and splitting. The disorder concepts are fine as they are: comorbidity is the expected consequence of the existence of crucial bridge symptoms that connect the symptom networks of different disorders (Borsboom, 2017), whereas heterogeneity is attributed to subtle topological differences in individual symptom networks.

7. Conclusion

Science must proceed with the concepts that it has. Attempts at conceptual revision are no different: they rely crucially on the very concepts that they aim to replace. The received metaphor is the appropriate one: researchers are in the position of the sailors on Neurath's boat, who have to stand on an adjacent plank while replacing another. The fact that researchers must repeatedly use DSM categories while recruiting participants, performing psychiatric evaluations, and running their data analyses, underscores how any attempt to overcome the status quo is constrained by received concepts. But while these concepts constrain the design of the replacement, they do not determine it. The variety of current proposals for conceptual change illustrate just how underdetermined the future shape of psychiatry is by the received concepts. The leading role continues to be played by the theoretical, metaphysical and methodological commitments of the competing paradigms.

Although machine learning plays a crucial role in the articulation of these paradigms, it is clear that we are far from automating the process of conceptual revision.

Notes

1. The moniker "computational psychiatry" is, in some circles, closely identified with Bayesian models of the brain (see Friston, 2022). We use it as a generic term for psychopathology with computational methods.

2. For more on the historical development of the DSM categories see Aftab and Ryznar (2020).

3. See also Rutherford et al. (2023) for a normative modeling approach applied at a more structural level, taking into account functional connectivity.

4. Reinforcement learners are therefore closer to the unitarian pole of the unitarian-separatist dyad identified by Aftab and Ryznar (2021).

5. That information "flows" between causally unrelated factors when conditioning on a common effect is the hallmark phenomenon of so-called collider structures (see Hausman, 1998, p. 83-4).

6. See Glymour et al. (2019) for a relatively up-to-date review of causal discovery methods.

7. Although network theory acknowledges the importance of external factors, these are conceived as triggers that are not constitutive of mental disorders. While it might be possible to create network structures that account for social factors or neurophysiology, this may make unrealistic demands on existing methodologies of causal discovery (see also Kästner, 2022).

Acknowledgements

TW gratefully acknowledges starter funding from the medical faculty at the University of Tübingen and Emmy Noether Funding from the DFG (513851350). KG, TG & TW acknowledge funding from the Cluster of Excellence - Machine Learning: New Perspectives for Science (390727645).

References

Aftab, A., & Rashed, M. A. (2021). Mental disorder and social deviance. *International Review of Psychiatry*, 33(5), 478-485.

Aftab, A., & Ryznar, E. (2021). Conceptual and historical evolution of psychiatric nosology. *International Review of Psychiatry,* 33(5), 486-499.

American Psychiatric Association. (2022). Diagnostic and statistical manual of mental disorders (5th ed., text rev.). https://doi.org/10.1176/appi.books.9780890425787

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16-07.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, *10*(9), 1214-1221.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43-75.

Boorse, C. (1975). On the distinction between disease and illness. *Philosophy & public affairs*, 49-68.

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

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence*, *2*(12), 731-736.

Borsboom, D. (2017). A network theory of mental disorders. *World psychiatry*, *16*(1), 5-13.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, *9*, 91-121.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, *9*(2), 129-136.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024-1035.

Colombo, M. (2022). Computational Modelling for Alcohol Use Disorder. *Erkenntnis*, 1-21.

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of abnormal psychology*, *125*(6), 777.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704-1711.

De Haan, S. (2020). An enactive approach to psychiatry. *Philosophy, Psychiatry, & Psychology*, *27*(1), 3-25.

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *Journal of Neuroscience*, *36*(4), 1211-1222.

Farmer, A. E., McGuffin, P., & Spitznagel, E. L. (1983). Heterogeneity in schizophrenia: a cluster-analytic approach. *Psychiatry Research*, *8*(1), 1-12.

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O., ... & Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of abnormal psychology*, *124*(2), 256.

Friston, K. (2023). Computational psychiatry: from synapses to sentience. *Molecular psychiatry*, *28*(1), 256-268.

Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (forthcoming). Developmental changes resemble stochastic optimization. *Nature Human Behaviour*.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual review of neuroscience*, *38*, 1-23.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, *156*, 563-585.

Hausman, D. M. (1998). *Causal asymmetries*. Cambridge University Press.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., ... & 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.

Kästner, L. (2022). Modeling psychopathology: 4D multiplexes to the rescue. *Synthese*, *201*(1), 9.

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

Lewinsohn, P. M., & Talkington, J. (1979). Studies on the measurement of unpleasant events and relations with depression. *Applied Psychological Measurement*, *3*(1), 83-101.

Lloyd, A., McKay, R. T., & Furl, N. (2022). Individuals with adverse childhood experiences explore less and underweight reward feedback. *Proceedings of the National Academy of Sciences*, *119*(4), e2109373119.

Maier, S. F., & Watkins, L. R. (2005). Stressor controllability and learned helplessness: the roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. *Neuroscience & Biobehavioral Reviews*, *29*(4-5), 829-841.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72-80.

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: theoretical framework and empirical application. *Journal of abnormal psychology*, *123*(2), 452.

Paykel, E. S. (1977). Depression and appetite. *Journal of Psychosomatic Research*, *21*(5), 401-407.

Petrolini, V., & Vicente, A. (2022). The challenges raised by comorbidity in psychiatric research: The case of autism. *Philosophical Psychology*, *35*(8), 1234-1263.

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Van Loo, H. M., & Romeijn, J. W. (2015). Psychiatric comorbidity: fact or artifact?. *Theoretical Medicine and Bioethics*, *36*, 41-60.

Radden, Jennifer (2023), "Mental Disorder (Illness)", *The Stanford Encyclopedia of Philosophy*

Rathkopf, C. (2018). Network representation and complex systems. *Synthese*, *195*, 55-78.

Ratti, E. (2020). What kind of novelties can machine learning possibly generate? The case of genomics. *Studies in History and Philosophy of Science Part A*, *83*, 86-96.

Reichenbach, H. (1949, January). Philosophical foundations of probability. In *Proceedings of the [first] Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 1-21). University of California Press.

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and alcohol dependence*, *161*, 230-237.

Rutherford, S., Kia, S. M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., ... & Marquand, A. F. (2022). The normative modeling framework for computational psychiatry. *Nature protocols*, *17*(7), 1711-1734.

Rutherford, S., Barkema, P., Tso, I. F., Sripada, C., Beckmann, C. F., Ruhe, H. G., & Marquand, A. F. (2023). Evidence for embracing normative modeling. *Elife*, *12*, e85082.

Saeed, B., Panigrahi, S., & Uhler, C. (2020, November). Causal structure discovery from distributions arising from mixtures of dags. In: *International Conference on Machine Learning* (pp. 8336-8345). PMLR.

Sebold, M., Deserno, L., Nebe, S., Schad, D. J., Garbusow, M., Hägele, C., ... & Huys, Q. J. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, *70*(2), 122-131.

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., ... & Heinz, A. (2017). When habits are dangerous: alcohol expectancies and habitual decision making predict relapse in alcohol dependence. *Biological psychiatry*, *82*(11), 847-856.

Tabb, K. (2019). Philosophy of psychiatry after diagnostic kinds. *Synthese*, *196*(6), 2177-2195.

Tabb, K., & Lemoine, M. (2021). The prospects of precision psychiatry. *Theoretical Medicine and Bioethics*, *42*(5-6), 193-210.

Thagard, P. (1990). Concepts and conceptual change. *Synthese*, *82*, 255-274.

Thiesson, B., Meek, C., Chickering, D. M., and Heckerman, D. (1997). Learning mixtures of DAG models. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pages 504–513.

Van Loo, H. M., Romeijn, J. W., & Kendler, K. S. (2019). Changing The Definition of The Kilogram: Insights For Psychiatric Disease Classification. *Philosophy, Psychiatry, & Psychology*, *26*(4), E-97.

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., ... & Bullmore, E. T. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular psychiatry*, *20*(3), 345-352.

Wiecki, T. V., Poland, J., & Frank, M. J. (2015). Model-based cognitive neuroscience approaches to computational psychiatry: clustering and classification. *Clinical Psychological Science*, *3*(3), 378-399.

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., ... & Marquand, A. F. (2018). Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA psychiatry*, *75*(11), 1146-1155.

Wolfers, T., Floris, D. L., Dinga, R., van Rooij, D., Isakoglou, C., Kia, S. M., ... & Beckmann, C. F. (2019). From pattern classification to stratification: towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neuroscience & Biobehavioral Reviews*, *104*, 240-254.

Wyckmans, F., Otto, A. R., Sebold, M., Daw, N., Bechara, A., Saeremans, M., ... & Noël, X. (2019). Reduced model-based decision-making in gambling disorder. *Scientific reports*, *9*(1), 19625.

World Health Organization (2021). International statistical classification of diseases and related health problems (11th ed.). https://icd.who.int/

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Zabihi, M., Oldehinkel, M., Wolfers, T., Frouin, V., Goyard, D., Loth, E., ... & Marquand, A. F. (2019). Dissecting the heterogeneous cortical anatomy of autism spectrum disorder using normative models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(6), 567-578.

Zabihi, M., Floris, D. L., Kia, S. M., Wolfers, T., Tillmann, J., Arenas, A. L., ... & EU-AIMS LEAP Group. (2020). Fractionating autism based on neuroanatomical normative modeling. *Translational psychiatry*, *10*(1), 384.

Zhang, Y., Wu, W., Toll, R. T., Naparstek, S., Maron-Katz, A., Watts, M., ... & Etkin, A. (2021). Identification of psychiatric disorder subtypes from functional connectivity patterns in resting-state electroencephalography. *Nature biomedical engineering*, *5*(4), 309-323.