PhiMiSci
Philosophy and the Mind Sciences

Understanding as a bottleneck for the data-driven approach to psychiatric science

Barnaby Crook[a] (barnaby.crook@uni-bayreuth.de)

Abstract
The data-driven approach to psychiatric science leverages large volumes of patient data to construct machine learning models with the goal of optimizing clinical decision making. Advocates claim that this methodology is well-placed to deliver transformative improvements to psychiatric science. I argue that talk of a data-driven revolution in psychiatry is premature. Transformative improvements, cashed out in terms of better patient outcomes, cannot be achieved without addressing *patient understanding*. That is, how patients understand their own mental illnesses. I conceptualize understanding as the possession of adaptive mental constructs through which experience is mediated. I suggest that this notion of understanding serves as a bottleneck which any prospective approach to psychiatry must address to be efficacious. Subsequently I argue that, though the data-driven approach is undoubtedly powerful, it does not have a straightforward means of unblocking the bottleneck of understanding. I suggest that the data-driven approach must be supplemented with significant theoretical progress if it is to transform psychiatry.

Keywords
Big data · Machine learning · Mental illness · Psychiatry · Recovery · Understanding

This article is part of a special issue on "Models and mechanisms in philosophy of psychiatry," edited by Lena Kästner and Henrik Walter.

Heavenly Hurt, it gives us –
We can find no scar,
But internal difference –
Where the Meanings, are –
Emily Dickinson (1861/1998),
There's a certain slant of light

[a]University of Bayreuth, Department of Philosophy.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658
 https://philosophymindscience.org ISSN: 2699-0369

1 Introduction

Mental illness is one of the greatest sources of human suffering worldwide (Insel et al., 2015; Vigo et al., 2016). Governments, research institutions, and health organizations are leading calls for improved mental healthcare to ameliorate the devastating costs that mental illness imposes on individuals and societies (Health & Care, 2021; Trust, 2020; World Health Organization, 2019). Despite frequent claims that current approaches to tackling mental illness are inadequate, theorists are divided about how progress should be made (Bickman, 2020; Cuthbert & Insel, 2013; Hengartner & Lehmann, 2017). Recently, methods utilizing big data and artificial intelligence (AI) have been gaining increased attention. Proponents of this *data-driven approach* are hopeful that sophisticated machine learning models will "redesign the current landscape of mental illness" (Fernandes et al., 2017, p. 5) through "tailored interventions for better outcomes" (Fernandes et al., 2017, p. 3). In this paper I critically assess the claim that the data-driven approach, as currently conceived, can bring about a transformative improvement to psychiatric science. I argue that, due to the central role *patient understanding* plays in recovery from mental illness, talk of an AI revolution in psychiatry is premature.

My work follows other authors who have engaged with the philosophical issues raised by the idea of AI-driven psychiatry. In an article structured as a back-and-forth debate, Brown and colleagues (2021) discuss the possibility of future AI replacing human psychiatrists altogether. In a paper focusing on schizophrenia, Starke and colleagues (2021) discuss the ethical principles that must be upheld if machine learning is to be introduced into clinical practice. In a response to that article, Gauld and colleagues (2021) analyze the kind of change that the age of AI will presage for psychiatry and the training of psychiatrists, a topic also addressed by McCoy and colleagues (2020). Finally, Horn and Weisz (2020) tackle the question of whether AI can improve psychotherapy research. My argument covers similar ground to all of the above work. The novelty of my approach consists in the key role it affords patient understanding and how thoroughly it engages with the evidence on what makes interventions effective. Together, these ideas have important implications for how successful we can expect the data-driven approach to be.

Many practitioners and theorists worry that psychiatry is a science in crisis, with traditional frameworks insufficient to address the complexity of mental illness (Bickman, 2020; McGorry & Nelson, 2019). Advocates of the data-driven approach are optimistic that their framework is the first to be capable of tackling that complexity in a scientifically rigorous way (Durstewitz et al., 2019; Dwyer et al., 2018; Rutledge et al., 2019). I begin the paper by describing this approach. Methodologically, the data-driven approach eschews traditional hypothesis-driven inferential statistics in favor of the automated discovery of predictive patterns in large volumes of patient data. Philosophically, I note that advocates of this approach espouse *theory neutrality,* rejecting specific theories of mental illness on the grounds

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

that data are a more reliable and objective guide to clinical decision-making. I also characterize the data-driven approach as *function-oriented*, viewing psychiatry as an interdependent set of optimization problems. I argue that these commitments lead to problems for the approach once the role of patient understanding is carefully considered.

To set up my argument, I begin by clarifying what would constitute a *transformative* improvement to the status quo. I suggest that a reasonable interpretation should focus on patient outcomes, as "the amelioration of [...] mental distress and suffering" (Thornton, 2020, p. 237) is the fundamental goal of psychiatry [1]. Given such a view, I argue that for the data-driven approach to be transformative, it must bring about radical improvements in patient outcomes. My argument for tempering expectations of the data-driven approach centers on *patient understanding*. Integrating ideas about understanding from pragmatist philosophers and theorists of psychiatry (Beck & Haigh, 2014; Dilthey, 1984/1977; Wampold & Imel, 2015; Wilkenfeld, 2013), I conceptualize patient understanding as a set of *adaptive mental constructs* through which a subject's experiences are mediated. Conceptualized this way, a patient possessing understanding becomes a pre-requisite for lasting recovery from mental disorder. I suggest that this notion of understanding serves as a *bottleneck* for any approach to psychiatric science. Without addressing patient understanding, outcomes cannot be radically improved.

The above analysis leads us to a specific question: can the data-driven approach address the bottleneck of understanding? I examine three ways that it might. First, I assess what I call the *indirect path*, which consists of optimization of the clinical pipeline in terms of nosology, diagnosis, prognosis, and treatment assignment. Second, I assess what I call the *direct path*. That is, the tailoring of treatments for improved outcomes. Third, I assess whether the data-driven approach might deliver radically improved patient outcomes by discovering novel insights about mental disorders that can be used to provide adaptive explanations. I argue that there are serious problems with each of these possibilities and, therefore, that the data-driven approach has no clear route to providing transformative improvements to psychiatric science.

The structure of the paper is as follows. In section 2 I present the data-driven approach. I discuss what motivates it, how it works, and its philosophical commitments. In section 3 I make the claim that patient understanding functions as a bottleneck for any approach to psychiatric science. To do so, I present my analysis of understanding as possession of adaptive mental constructs, show how this makes understanding critical for robust recovery from mental distress, and explain in what sense this makes it a bottleneck. In section 4 I assess the potential of the data-driven approach to address the challenge of the bottleneck of understanding,

[1] To be clear then, radically improved *scientific understanding* of mental disorders, if unaccompanied by improved outcomes, would not be considered transformative by this criterion. However, one could reasonably adopt another criterion on which it would. I thank an anonymous reviewer for pushing me to clarify this point.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

arguing that, at least as currently practiced, it is ill-suited to do so. In section 5 I address possible objections to the view presented. In particular, I defend my conceptualization of understanding and justify the applicability of the bottleneck metaphor. I then conclude by suggesting that the data-driven approach can best contribute to psychiatric science in collaboration with theoretical research.

2 The data-driven approach

In this section I characterize the data-driven approach to psychiatry. I begin by describing the theoretical landscape from which the approach has emerged. I then provide an overview of the data-driven methodology, including an example (Koutsouleris et al., 2021). I conclude the section by discussing the philosophical commitments embodied by the approach.

2.1 Motivating the approach: Psychiatry in crisis

Reading the theoretical and philosophical literature on mental illness, one often comes across the view that psychiatry is a discipline in crisis. For example, Stoyanov and Maes (2021, p. 1) claim that "psychiatry remains in a permanent state of crisis [...] evidenced by the many different competing approaches and ways to understand mental and psychiatric disorders". Leonard Bickman (2020, p. 803) states that mental healthcare currently suffers from "overall poor effectiveness" caused by "poor alignment between what the patient needs, and the treatment provided". And Allen Frances (2009, p. 391) writes that "our understanding of psychopathology is fairly primitive" and "lacks the fundamental understanding of pathogenesis". These pessimistic assessments locate the precise source of the problem differently, but they all share the opinion that the *complexity* of mental disorders plays a significant role in thwarting the success of psychiatric science [2].

Mental disorders are complex in multiple ways. First, they are difficult to explain, resisting "a simple analysis in terms of biological pathways, endophenotypes, and neural mechanisms" (Borsboom et al., 2019, p. 91). Second, mental disorders are massively multifactorial (Boer et al., 2021). That is, there are many different factors, "sprinkled across multiple causal levels" (Kendler & Gyngell, 2020, p. 44), that may be relevant to any given presentation of a mental disorder. This relates to the first point. If we ought, as Engel suggested, to incorporate "all the factors contributing to both illness and patienthood" (Engel, 1977, p. 133) into our explanations of mental disorder, then these will be dense explanations indeed. However, identifying the relevant factors does not suffice to fully characterize mental disorders. This is due to a third kind of complexity: interaction among the factors. As Borsboom and colleagues put it, "mental disorders likely involve feedback loops that cross all

[2] The term complexity is used to refer to many different properties, both formal and informal (Ladyman et al., 2013). It is beyond the scope of this paper to weigh in on which formal criteria for complexity mental disorders satisfy.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

of the traditional divides between levels of explanation, none of which can claim the status of 'basis' for the others" (Borsboom et al., 2019, pp. 10–11). In summary, biological entities, psychological constructs, and social factors interact in intricate and diverse ways to produce particular instantiations of psychopathology.

The idea that psychiatry is in a crisis of complexity provides a strong rationale for adopting the data-driven approach. For example, Rutledge and colleagues (2019, p. 152) state that "big data and machine learning are uniquely placed to address [the complexities of psychiatric disorders]". Similarly, British psychiatrists Brown and Story (2021, p. 131) argue that modern technology "enables the capturing of rich, longitudinal, multimodal data, the analysis of which promises vastly improved characterization of illnesses and their trajectories". While historical approaches to psychiatry have often been stuck either focusing on a small subset of factors or falling short of scientific standards of evidence (Ghaemi, 2007), the data-driven approach leverages the availability of data and computation to integrate diverse sources of information in a rigorous way. Unlike traditional statistical methods that focus on detecting group differences, machine learning methods are designed to exploit large volumes of multivariate data to make individual predictions (Bennett et al., 2019; Gillan & Whelan, 2017). This means that clinical decisions, such as treatment recommendations, can be tailored to sophisticated statistical representations of individual patients. For this reason, the data-driven approach is sometimes termed *precision psychiatry* (e.g., Bzdok & Meyer-Lindenberg, 2018; Fernandes et al., 2017).

Some proponents of the data-driven approach have been forthright in claiming that this approach will suffice to bring about a *transformative* improvement to mental healthcare. For example, Gauld and colleagues (2021, p. 2519) proclaim that "AI will undeniably transform the future of psychiatry". Similarly, Bickman (2020, p. 803) extols the "revolutionary possibilities of artificial intelligence for improving mental healthcare". And Fernandes and colleagues (Fernandes et al., 2017, p. 1) state that the data-driven approach "promises to be even more transformative than in other fields of medicine". This view is grounded in the notion that the data-driven approach will provide models that allow clinicians "to identify the right treatment for each patient, first time around" (Gillan & Whelan, 2017, p. 34). As stated in the introduction, I interpret a transformative improvement to require radically improved patient outcomes. In order to assess whether the data-driven approach can really achieve this, we will need to take a closer look at how it works.

2.2 The methodology of the data-driven approach

I thus turn to the methodology of the data-driven approach. Readers familiar with big-data and machine learning can skip to section 2.3 where I will discuss the philosophical commitments of the approach.

The data-driven approach is defined by a methodological hallmark: *the employment of automated computational procedures for discovering predictive patterns*

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

in large volumes of data. Such approaches have become increasingly attractive as the cost of collecting, storing, and manipulating data has decreased (Chekroud et al., 2021). At a very high level, the methodology of the data-driven approach can be described as consisting of four steps. First, a problem is defined. Second, data are acquired. Third, a model is designed and trained. Fourth, the model is deployed. We will take a look at each of these steps in turn.

2.2.1 Problem definition

The first step in the data-driven approach is to define a problem. Machine learning is a technique well-suited to classification, prediction, and clustering problems. In classification, the task is choosing which of a pre-defined set of classes a particular data instance belongs to. In prediction, the task is estimating the future value of a chosen variable. In clustering, the task is organizing a dataset into self-similar groups. Models developed according to the data-driven approach have many possible applications in psychiatry. These can be coarsely divided into four (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018). First, classification models can be used for *diagnosis*. For example, a model might be constructed to determine whether subjects should be diagnosed with major depressive disorder (e.g., Sharma & Verbeke, 2020). However, applying machine learning to diagnosis need not be limited to a binary decision about a single condition. For example, in multiclass logistic regression a model will return a probability distribution over multiple possible diagnoses (Elujide et al., 2021; Qureshi et al., 2016). This makes the data-driven approach amenable to transdiagnostic studies (Pelin et al., 2021).

Second, models employing clustering algorithms can be used to aid *nosology*, the organization of mental illnesses into a set of categories (Zachar & Kendler, 2017). Within this paradigm, the objective of the model is to find a partitioning of the data that best satisfies a chosen set of statistical criteria. When models are used to find clusters *within* extant disorder categories, this is called sub-typing (Feczko et al., 2019). A third application of the data-driven approach is *prognosis*, the prediction of how an individual's condition will evolve over time. Machine learning models can be used to predict clinically relevant events, such as symptom onset (Koutsouleris et al., 2021). Recurrent neural networks, an architecture particularly well-suited to exploiting temporal regularities, can be used to compute disease trajectories (Shickel et al., 2018; Suhara et al., 2017). Finally, perhaps the most important target application of all is *treatment assignment*, using patient data to decide which therapeutic approach to pursue [3]. Studies have attempted to apply machine learning to predict responses to both pharmacological and psychotherapeutic treatments (Chekroud et al., 2021; Su et al., 2020).

[3] Initial research in this domain often involves predicting *response* to an individual treatment as opposed to assignment among treatments. Clearly, if reliable estimates of treatment response could be computed across many treatments, this would lead to a straightforward procedure for assigning treatment. Namely, predicting the response to multiple treatments and selecting the best one (indeed, the problems are treated as synonymous in Chekroud et al., 2021).

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658
 https://philosophymindscience.org ISSN: 2699-0369

2.2.2 Data acquisition

Machine learning models are typically *data hungry*, meaning that they only learn patterns that are robust and generalizable if supplied with large volumes of data. This makes data acquisition a crucial part of the approach (Dwyer et al., 2018). The reason for these onerous data requirements is known as the *curse of dimensionality*, a term referring to the fact that, as more dimensions of variation are considered, the numerosity of the data required to maintain model performance increases exponentially (Keogh & Mueen, 2017). This means that collecting the quantities of data required to train massively multivariate models of mental illness requires large-scale coordination (Bzdok & Meyer-Lindenberg, 2018).

Since variation at almost every conceivable level of organization is thought to be relevant to mental illness (Kendler & Gyngell, 2020), data can be collected in innumerable ways, from questionnaires and cognitive tasks, to biomarkers and neuroimaging, to social media and wearable electronics (Balaskas et al., 2021; García-Gutiérrez et al., 2020). The challenges involved vary across modalities. For example, collecting and analyzing neuroimaging data is time-consuming and costly (Najafpour et al., 2021). On the other hand, clinical notes in Electronic Health Record (EHR) data are abundant and accessible, but unstructured (Shickel et al., 2018). Pre-processing diverse data modalities such that they can be integrated by a single model is a technical challenge. Additionally, it is often unclear whether different operationalizations of psychological concepts actually measure the same latent construct (Poldrack & Yarkoni, 2016). Given these difficulties, Shickel and colleagues (2018, p. 24) describe developing a unified representation of such disparate sources of information as the "holy grail of clinical deep learning research".

2.2.3 Model design and training

Given a rich and voluminous dataset, the next step is to design a model. In the machine learning paradigm this consists of specifying the objective function, architecture, and learning rule of the model. The objective may be to predict remission rates, responsiveness to specific medications, or symptom response trajectories (Bzdok & Meyer-Lindenberg, 2018; Chekroud et al., 2021; Rutledge et al., 2019). Translating the natural language expression of the goal into a mathematically specified objective function is a crucial part of the design phase. The architecture of a model consists of its components and how they are organized. For example, a deep neural network consists of multiple layers of neuron-like computational units equipped with activation functions (Durstewitz et al., 2019), while a decision tree consists of hierarchically organized decision nodes (Feczko et al., 2019). Machine learning researchers have developed a veritable zoo of model architectures with a variety of mathematical and representational properties (Bronstein et al., 2021). In principle, architectures can be designed to exploit the kinds of regularities in the sampled data domain. In the case of high-dimensional, multimodal patient data, however, design principles guiding architectural choices are not yet well estab-

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

lished (Si et al., 2021), reflecting the lack of scientific consensus on the nature of mental illness.

Once an objective and architecture are chosen, training takes place. A machine learning model is usually initialized with random parameter values (Chollet, 2021, p. 46). What the model outputs is a function of the input and these parameter values. In a deep neural network, for example, the parameter values are the connection weights between the neurons and each individual neuron's bias. The goal of learning is to find a set of parameter values that optimizes the objective. For example, if the objective is to diagnose depression, the model might take neuroimaging and physiological data associated with an individual and output a binary classification reflecting the presence or absence of a diagnosis (Arbabshirani et al., 2017; Sharma & Verbeke, 2020). The model's output is then compared with the true label for the data point. In this case, the label would be provided by a clinician's diagnostic judgement about the patient. Having access to the ground truth label makes this an example of *supervised learning*. Should the machine's output be incorrect, an error signal, often called the *loss*, is computed. This error signal, in conjunction with a learning rule, like stochastic gradient descent, is used to update the parameter values [4]. After many iterations of this process, a model converges on a stable set of parameter values that encode its best approximation to the true function relating the input data to the output labels. In our example, the parameters can be considered as encoding a hypothesized relationship between biological variables and depression. Crucially, to avoid overfitting on the particularities of the data sample used for training, models must always be validated on novel datasets (Bzdok & Meyer-Lindenberg, 2018; Rutledge et al., 2019).

The simple example above is the tip of the iceberg when it comes to applications of the data-driven approach. While a thorough review is out of scope for this paper, it is worth noting the broad categories of *unsupervised* and *semi-supervised* methods, which do not rely on labels (Bengio et al., 2013; Bronstein et al., 2021; Domingos, 2012). In our depression diagnosis case, ground truth labels were provided by human judgement. In addition to being costly and time-consuming to generate, the use of human judgment in producing labels limits the potential value data-driven models can provide (Bickman, 2020). Durstewitz and colleagues (2019) point out that the value of data-driven diagnosis is upper-bounded by how well extant diagnostic labels capture the underlying nature of mental illnesses. For this reason, the data-driven approach "needs to go beyond the mere prediction of symptoms by current diagnostic schemes, but rather has to help refining our diagnoses" (Durstewitz et al., 2019, p. 1591). In line with this idea, researchers have employed unsupervised machine learning methods to discover novel clusters of patients that can be used to predict outcomes, stratify treatments, and increase understanding (Gould et al., 2014; Miranda et al., 2021; Pelin et al., 2021). The crucial point is that,

[4] In practice, one iteration of a training loop usually involves a batch of data instances, rather than just one.

while supervised learning is limited by human experts' ability to provide useful labels, the data-driven approach as a whole need not be.

2.2.4 Validation and deployment

Once data have been collected and a model has been trained, the next steps are validation and deployment. As of yet, the majority of studies implementing the data-driven approach are proof-of-concept and most trained models have not been clinically deployed (Cearns et al., 2019; Chekroud et al., 2021). In an insightful review, Cearns and colleagues (2019) provide a checklist that any machine learning model should comply with to demonstrate readiness for clinical application. Models should show: 1) an improvement on whatever approach constitutes the current state-of-the-art for the clinical application in question, 2) validation with a large, external dataset, and 3) clear specification of the scope of the model (i.e., the population for which its validity has been rigorously demonstrated). Once deployed, a machine learning model must be continually evaluated to ensure that its accuracy remains robust, that it is secure, and that it does not lead to unacceptable algorithmic bias. Practically, deployment also requires coordination with clinicians to ensure that models are well-utilized (Tonekaboni et al., 2019). As the data-driven approach matures, further concrete issues relating to clinical deployment are sure to emerge.

2.2.5 The data-driven approach in practice: An example

An example will help to elucidate how the data-driven approach plays out in practice. In a large-scale study, Koutsouleris and colleagues (2021) used multimodal machine learning to predict whether patients with clinical high-risk states and recent-onset depression would transition to psychosis. In an extensive data acquisition process spanning five countries and more than three years, Koutsouleris and colleagues collected data from many domains including sociodemographic, questionnaires, self-reports, structured interviews, cognitive and behavioral tasks, neuroimaging, and genetics. This highlights the importance of collecting spatially and temporally distributed data for constructing generalizable models that will be robust to variation in social and cultural factors (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018). To make use of their avalanche of information, Koutsouleris and colleagues built a *stacked* model, a kind of architecture suitable for automatically learning how to best integrate disparate data sources [5]. The researchers' multimodal predictive model was able to outperform all unimodal competitors as well as human clinical raters in prognostic accuracy. In addition to the predictive value of their model, the authors were also able to extract clinical insights into psychosis by analyzing their findings. For example, they found high-risk patients predicted

[5] Stacked models consist of multiple *base* models, each exploiting a particular data modality, and a *meta-model* which learns how to weight the predictions of the base models.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

not to transition to psychosis had increased temporo-occipital brain volume relative to healthy controls, suggesting a neural basis for mechanisms of resilience. Further, they found certain neurocognitive factors, such as impaired facial affect recognition, were markers of poor psychosis outcomes in both clinical high-risk and recent-onset depression groups.. This highlights another important feature of the data-driven approach. Investigating a transdiagnostic population, that is, incorporating data from groups which are typically investigated separately, allows the data-driven approach to find and exploit patterns which transcend the group differences captured by traditional inferential approaches (Hengartner & Lehmann, 2017; McGorry & Nelson, 2019).

2.3 The philosophical commitments of the data-driven approach

Now that we have covered the methodological basis of the data-driven approach, we can turn to its philosophy. I raise two points. First, the data-driven approach purports to be *theory neutral*, relying on automated statistical procedures rather than theoretical assumptions to cut through the complexity of mental illness. Second, the data-driven approach is *function-oriented*, focusing on building models to optimize each step of the clinical decision-making pipeline. These features will turn out to be important for assessing whether the approach can transform psychiatry by radically improving patient outcomes.

2.3.1 Theory neutrality

Different schools of thought have long disagreed about the *nature* of mental illnesses (Kendler, 2016; Shorter, 1997; Tsou, 2021; see also Dembic, 2023; Leder & Zawidzki, 2023, this volume). Should we think of them as brain disorders defined by neural dysfunction, socially constructed labels for deviant behavior, irreducibly subjective pathologies of psychology, or an intricate mixture of these ingredients? The data-driven approach chooses to wash its hands of the issue entirely, preferring to remain *theory neutral* (Paulus, 2015). Unlike the attempt to develop mechanistic models of disease processes, the data-driven approach to psychiatry makes no strong assumptions about the nature or causal structure of mental illness (Bennett et al., 2019). Practitioners take this feature to be a strength of the approach. For example, Chekroud and colleagues (2021, p. 154) claim that machine learning offers a set of "powerful hypothesis-free approaches" to predicting treatment outcomes. Similarly, in a review of techniques applying deep learning to Electronic Health Record data, Shickel and colleagues (2018, p. 17) describe the philosophy as "letting the data speak for itself by discovering latent relationships and hierarchical concepts from the raw data, without any human supervision or prior bias". From the data-driven perspective, then, theoretical assumptions about mental illness do more harm than good, and we would be better off relying on the presumed objectivity of models trained with automated procedures.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

It is important to note that the data-driven approach is not, and indeed no approach can be, truly atheoretical (Joober & Tabbane, 2019). In practice, even huge quantities of data are useless without significant constraints on learning (Domingos, 2012). Researchers developing machine learning models make many choices about which data to collect, which architecture to use, how to specify their objective, and so on (see section 2.2). Each of these decisions biases a model towards learning some structures rather than others, a phenomenon often discussed in machine learning under the term *inductive bias* (Battaglia et al., 2018). Despite this caveat, the intention to remain as neutral as possible with respect to theory still constitutes a significant philosophical commitment of the data-driven approach. In the argument to come, I will suggest that this is a mistake. Without theory-driven research to improve understanding of how treatments can improve outcomes, the data-driven approach has limited potential.

2.3.2 Function orientation

Powerful methodological tools alter how domains of inquiry are perceived by their users. Cichy and Kaiser (2019, p. 312) express this point thusly, "once technological artefacts are commonly used, they are not mere tools to realize predefined scientific goals but begin to shape social reality in a way that affects the user's desires and interests". The data-driven approach is no exception. As Dwyer and colleagues put it (2018, p. 95), "modern machine learning methods can contribute greatly to clinical psychology and psychiatry *by changing the way that problems are considered* [emphasis added]". I introduce the term *function orientation* to describe the perspective on psychiatry that the data-driven approach imposes. The term *function* here should be interpreted in the mathematical sense of mappings from inputs to outputs. Note that the function-oriented perspective is largely implicit in the way data-driven research is carried out, rather than being explicitly endorsed by its practitioners.

The function-oriented perspective decomposes psychiatric science into a set of statistical problems, each addressable through the development of the right machine learning model. It formulates these problems in mathematical terms, with solutions consisting of trained models judged by accuracy metrics. The overall goal is to optimize the entire pipeline of clinical decision making, from diagnosis to treatment assignment. As Dwyer and colleagues (2018, p. 94) put it, "the ultimate aim of translational machine learning is to generate procedures that would be beneficial for clients, general practitioners, and in specialized hospital settings to improve patient outcomes". This is a noble aim. However, the overall effect of viewing psychiatry this way is that discrete clinical decisions are foregrounded, while anything not amenable to being quantified and incorporated into a statistical learning problem is neglected. For example, the role of the psychiatrist as an empathic listener and the idiosyncratic content of mental distress, resistant to concise mathematical representation, can easily be obscured in this formulation (Horn & Weisz, 2020). Of course, these are relative effects. Function orientation does not

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

make one blind to the interpersonal nature of psychiatry. However, to the extent that one develops expertise in a way of seeing that relies on noting and measuring formalizable properties, one naturally becomes less attuned, on average, to those properties which cannot be easily formalized. This is what Cichy and Kaiser (2019, p. 312) mean when they stress that "models are not neutral tools." In section 4.1, I will argue that function orientation is a flawed perspective. Without addressing the limitations of currently available treatments, optimizing clinical decision making cannot be transformative.

3 The bottleneck of understanding

Having characterized the data-driven approach, I now introduce the notion of *understanding* that I claim to be of critical importance to treating mental illness [6]. While this is a general account of understanding, when it is applied to how patients understand their own illnesses it takes on a crucial role for psychiatric science. In particular, I argue that, for patients suffering from mental illnesses, developing this form of understanding is necessary for recovery. According to my argument, patient understanding serves as a *bottleneck* which any effective approach to psychiatric science must address to improve patient outcomes.

3.1 Understanding as possession of adaptive mental constructs

I propose that understanding can be usefully conceived of as a set of adaptive mental constructs through which experience is mediated. I derive this definition by integrating ideas from theorists of psychiatry and the pragmatist tradition in philosophy. The notion of *understanding* has a rich history in psychiatry. Perhaps best known is Karl Jaspers' account (1959/1997), on which understanding is a psychiatrist's intuitive, empathic means of interpreting a patient's mental world. However, while the interpersonal relationship between practitioner and patient is of immense importance to psychiatry, understanding can also be isolated and considered at an individual level. For example, consider German polymath Wilhelm Dilthey's account (Dilthey, 1984/1977). Central to Dilthey's notion of understanding is what he calls the "acquired nexus of psychic life" (1984/1977, p. 42). This nexus includes our values, habits, ideas, and goals and influences "every single act of consciousness" (1984/1977, p. 59). In Dilthey's view, these mental constructs form a coherent whole that "determines the nature of our understanding of ourselves and of others" (1984/1977, p. 55). Following Dilthey, then, we will take

[6] Note that I am not arguing that this is the *correct* or *best* way to conceptualize understanding, either across all contexts or even within the domain of psychiatry. There are other notions of understanding that may be valuable and important in different contexts. See section 5.1 for more on this point.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

the basis of understanding to consist of a set of mental constructs through which experience is mediated. However, we focus our attention specifically on patients' understanding of their own inner lives, as this is what is of particular relevance to the goals of psychiatric science.

A further property that plays a role in conceptualizations of understanding from both inside and outside of psychiatry is that of *adaptivity*. In general, something is adaptive if it enables goals to be achieved under a set of environmental constraints. Within psychiatry, this notion appears in Aaron Beck's cognitive model of mental illness (Beck, 1985; Beck & Haigh, 2014). On this view, a patient's psychological distress is a downstream effect of their maladaptive *cognitive schemas*, which are the "beliefs, expectancies, evaluations, and attributions" that "serve to order everyday experience" (Beck & Haigh, 2014, p. 12). Note the similarity to Dilthey's *acquired nexus*, described above. Under the cognitive model, interventions involve *cognitive restructuring*, which is the active modification of a patient's maladaptive beliefs and schemas, to make them adaptive. However, one need not adopt the specific conceptual lexicon of the cognitive model in order to subscribe to the view that an adaptive notion of understanding is crucial to psychiatry. For example, Bruce Wampold defines adaptive explanations as those that "provide a means to overcome or cope with [...] difficulties" (Wampold & Imel, 2015, p. 58). Wampold and Budge suggest that "patients typically present with a maladaptive explanation for their disorder" and "a primary therapeutic activity of the therapist is to provide an adaptive explanation" (Wampold & Budge, 2012, p. 612). Wampold's work on the *common factors* of efficacious psychotherapy shows that the delivery of such adaptive explanations is a necessary feature of *all* effective therapies (Wampold, 2015; Wampold & Imel, 2015). [7] While this line of work focuses on adaptive *explanations*, from a patient's perspective, accepting and internalizing such an explanation *just is* coming to understand their experience in an adaptive way.

The idea that understanding is usefully conceptualized as adaptive, affording particular abilities or skills, also resonates with recent currents of thought in the pragmatist philosophy of understanding (Grimm, 2019; Hills, 2016; Regt, 2019; Wilkenfeld, 2013). Consider Daniel Wilkenfeld's (2013) account:

> Understanding is at root the possession of the right sort of mental representations of that which is understood; [...] a mental representation counts as being "of the right sort" in virtue of the fact that possession of it enables one to perform [...] feats relevant in that context. (Wilkenfeld, 2013, p. 1000)

[7] The idea that diverse methods of psychotherapy owe their efficacy to a set of shared common factors was first proposed by Saul Rosenzweig (Rosenzweig, 1936) before being explored in greater depth by Jerome Frank (Frank & Frank, 1993). Various common factors, including the therapeutic alliance, empathy, and shared expectations, have been proposed (Wampold, 2015). While there is no consensus on a particular conceptual model, there is very strong evidence that the common factors account for most of the variance in clinical outcomes (Peterson, 2019).

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

Notice that this view equates *understanding* with being empowered to *do* certain things. Namely, whatever is relevant to a particular context. Wilkenfeld's account is general but can be readily applied to psychiatric patients. The mental representations relevant to patients may involve aspects of self-understanding, life narratives, memories of and beliefs about aberrant experiences, and so on (Glover, 2020,; McConnell, 2020). The relevant feats which the afflicted hope to be enabled to perform may be things like engagement in autonomous, competent behavior, the pursuit of goals, and effective self-regulation, particularly in the face of stressors (Ivanov & Schwartz, 2021; Jacob, 2015).

So then, putting the ideas above together, patient understanding can be cast as the patient's possession of a set of adaptive mental constructs. Following Beck and Dilthey, I use the term *mental constructs* to refer to things like beliefs, habits, goals, expectancies, evaluations, values, and so on [8]. Following Wampold and Wilkenfeld, I use the term *adaptive* to mean enabling one to do relevant things in a given context. Notice two things about this notion of understanding. First, it covers both intuitive and intellectual senses of the word. By intuitive, I mean related to early, subconscious aspects of information processing, like the biased emotional processing hypothesized to be important to depression (Harmer et al., 2009). By intellectual, I mean involving articulable propositional content, like a belief that one is incapable (Beck, 1985). Second, the notion of understanding we are using is not *factive*. That is, it does not require conformation to any particular standard of truth (Doyle et al., 2019). This is suitable for our context for two reasons. First, in the study of effective psychotherapy there is no empirical evidence that the truth value of the explanation is related to outcome [9] (Frank & Frank, 1993; Wampold et al., 2007). Second, for some of the mental constructs relevant to this notion of understanding, such as evaluations, it is not clear whether or how a truth value could be ascribed. I further clarify and defend this analysis of understanding in section 5.1.

3.2 Understanding as a bottleneck to recovery

Above, we conceptualized understanding as possession of adaptive mental constructs. For those suffering from mental illness, gaining understanding might involve developing beliefs, expectancies, and values which enable the self-regulation of affective experience and behavior. I now make the claim that *patient understanding is a bottleneck for robust recovery.*[10]

[8] The argument does not turn on *which particular* mental constructs play the role of organizing ongoing experience. See section 5.1 for further elaboration on this point.

[9] Of course, such evidence may be forthcoming. There are theoretical reasons to expect some relationship between truth and efficacy, all else being equal (see Baker et al., 2008; Laska et al., 2014, for discussion of this issue).

[10] I follow defenders of the *recovery model* in viewing recovery from mental illness as describing *a process towards a goal of living a self-efficacious and satisfying life* (Anthony, 2000; Jacob, 2015; Ramon et al., 2007).

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

A bottleneck is a component within a broader system that limits the rate at which something is produced. For example, a shortage of semiconductor chips limits the rate at which cars can be manufactured. I suggest that patient understanding is a bottleneck for recovery in the following sense: the long-term rate at which a patient recovers is limited by the rate at which that patient develops understanding of their illness. In other words, the outcomes psychiatric science wants to bring about cannot be achieved without finding a way to reliably increase understanding. Including the 'long-term' qualifier here captures the idea that short-term improvements in an outcome measure could reflect random fluctuations in environmental factors such as the number and intensity of stressors. Such fluctuations are one of the reasons why recovery is typically a non-linear process that happens in "fits and starts" (Jacob, 2015, p. 118). Casting understanding as a bottleneck is a way of framing the idea that, given how we have defined understanding, it is *essential* for a robust recovery. If one lacks adaptive mental constructs, one cannot perform relevant feats in one's own context. As previously stated, these feats may include setting and seeking personal goals, regulating affective experience, and engaging in a community. If one lacks the mental constructs that enable the performance of such feats, it is difficult to see how one could live a satisfying and self-efficacious life, conditions constitutive of recovery[11].

I see the value of the bottleneck framing as being threefold. First, the concept of a bottleneck implies embeddedness in a system with multiple components. In this case, the system consists of psychiatric science as a whole, including disease classification, treatment development, therapeutic best practice, and so on. This makes the metaphor apt for sharpening thinking about the relationship between understanding and other levels of organization which may be pertinent to improving outcomes. For example, variables specifiable at the genetic, behavioral, and social levels may each be critical intervention targets for specific instantiations of psychiatric disorder. That is, inducing the adaptive mental constructs that constitute understanding may, in some cases, be most effectively achieved through interventions that are primarily medicinal or behavioral, rather than psychological[12]. However, the bottleneck framing highlights that in order to engender a robust recovery, *any* intervention must induce understanding. Beck (1985, pp. 333–334) makes a related point, noting that both psychotherapy and pharmacotherapy must alter defective cognitive content to be efficacious. Second, framing the challenge of psychiatry as addressing the bottleneck of understanding foregrounds subjective experience and psychopathology. Why is this a good thing? Because psychopathology is the very subject matter of psychiatry (Stanghellini & Broome, 2014). Neurobiolog-

[11] While I have framed this in terms of the recovery model (see fn.10), the argument also applies, albeit less strongly, to symptom-focused models of recovery. To see this, consider that some of the feats relevant to patients will be constitutive of their diagnoses. For example, a patient with generalized anxiety disorder may hope to concentrate effectively on tasks and avoid becoming irritable (both symptoms of the disorder in the DSM 5, American Psychiatric Association, 2022).

[12] I thank an anonymous reviewer for pressing me to make this point clear.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

ical, genetic, and sociocultural approaches are necessarily studying the correlates of what is, fundamentally, *mental* disorder (see Thornton, 2020, p. 235, for a more detailed version of this argument). Finally, focusing on a person-centered, intentional, mental ability like understanding when conceptualizing the goals of psychiatry may be of pragmatic value. Both theory and empirical evidence suggest that a sense of agency and self-efficacy are important for recovery from mental illness (Glover, 2020; McConnell & Snoek, 2018). Framing the challenge of recovery as the development of understanding emphasizes the power of the human subject to actively negotiate an alleviation of their own suffering (Anthony, 2000; Coulombe et al., 2016).

4 The data-driven approach and the bottleneck of understanding

If the view presented in section 3 is correct, a transformative improvement to psychiatric science must involve unblocking the bottleneck of understanding. In this section I argue that the data-driven approach lacks a clear means of achieving this. First, I argue that, because extant treatments lack specificity, optimizing their allocation will not address the bottleneck of understanding. Then, I argue that using the data-driven approach to target patient outcomes directly is fraught with theoretical challenges. Finally, I argue that there are conceptual difficulties plaguing the translation of data-driven insights into therapeutically valuable information that can aid patient understanding.

4.1 Limits of a function-oriented approach

Recall that the data-driven philosophy being *function-oriented* refers to how it conceptualizes psychiatry as a collection of machine learning problems. As discussed in section 2.3.2, methodological tools act like filters, rendering particular aspects of phenomena salient. In this case, the data-driven approach focuses on clinical decision-making under the implicit assumption that optimizing these decisions will dramatically improve outcomes. Since clinical decisions can only be made with respect to existing interventions, the improvements the data-driven approach hopes to bring about must be realized through negotiating a more effective *allocation* of interventions (Bzdok & Meyer-Lindenberg, 2018, p. 225). I call this the *indirect path* to transforming patient outcomes, since it does not involve targeting patient understanding directly.

The problem with the indirect path is the lack of *specificity* of extant treatments. A treatment would be specific if it had an active ingredient that rectified a particular dysfunction (Wampold & Imel, 2015, p. 60). Unfortunately, there is little evidence that treatments from either of the two major modalities, psychotherapy or pharmacotherapy, operate through specific ingredients that directly tar-

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

get pathology (Ivanov & Schwartz, 2021; Middleton & Moncrieff, 2019; Sathyanarayana Rao & Andrade, 2016; Wampold & Imel, 2015).[13] In the case of psychotherapy, Wampold and Imel summarize their comprehensive review of the subject thusly, "there is no compelling evidence that the specific ingredients of any particular psychotherapy or specific ingredients in general are critical to producing the benefits of psychotherapy" (Wampold & Imel, 2015, p. 253). When it comes to pharmacotherapies, Middleton and Moncrieff state that "there is no evidence that antidepressants work by correcting a chemical imbalance or other identifiable abnormality" (Middleton & Moncrieff, 2019, p. 52). While there are many theories attempting to get a firmer handle on therapeutic mechanisms of action (e.g., Beck & Haigh, 2014; Harmer et al., 2017), these have not yet led to the development of specific treatments.

This lack of specificity in treatments is concerning for the data-driven approach. If it were the case that extant treatments were specific but misallocated, correcting their allocation would be a clear path to radically improved outcomes. However, given the lack of evidence that extant treatments are specific (Budd & Hughes, 2009; Wolpert et al., 2021), the effects of improved allocation may be small. The problems associated with mental disorders are, as Leichsenring and colleagues (2022, p. 141) put it when summarizing their findings from a survey of meta-analyses across major mental disorders, "not sufficiently addressed by the available treatments". By analogy, imagine a restaurant full of diners randomly assigned variously shaped spoons to cut thick steaks. An optimal reallocation of the spoons according to hand size would be unlikely to bring about a transformative improvement in the average carving ability of the patrons. The aim of data-driven psychiatry can be summarized as the prescription of the right treatment to the right person at the right time (Trivedi, 2016). However, if a specific treatment has not been devised yet, no algorithm can recommend its prescription.

4.2 Limits of a theory-neutral approach

Above we questioned the *indirect path* by which the data-driven approach could revolutionize patient outcomes. What about a *direct path?* The idea here would be that the data-driven approach could improve treatments by optimizing for patient understanding itself. In other words, machine learning could be used to tailor treatments to achieve better outcomes. Research in this area is already underway (see Chekroud et al., 2021; Su et al., 2020, for reviews). In one study, researchers coded therapist utterances in cognitive behavioral therapy sessions and trained a model to determine the association between therapeutic content and clinical outcomes (Ewbank et al., 2020). In this case, the quantity of utterances classed as cognitive and behavioral change methods in therapy sessions was found to correlate with better chances of symptom improvement. While such early results are valu-

[13] As Wampold and Imel state (2015, p. 47), the necessity of some form of exposure for treating specific phobias may be an exception to this rule.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

able, notice the theoretical challenges involved in pushing such research programs further. Ewbank and colleagues used a cognitive behavioral therapy framework to define twenty-four classes of utterance (e.g., *mood check, set goals, therapeutic empathy*). The coding of these utterances runs into thorny theory-laden issues of variability and interpretation (Horn & Weisz, 2020). What are the right categories? How many categories can a single utterance belong to? Should semantically distinct ways of expressing similar sentiments be considered therapeutically equivalent? Far from the theory-neutral philosophy of the data-driven approach, applying the methodology to improving treatment demands introducing significant theoretical assumptions. This should not be read as a criticism of this research direction. Rather, it is an attempt to show that for the data-driven approach to tackle understanding, interdisciplinarity and theory are required. The hypothesis-free number crunching espoused by some proponents of the data-driven approach is not a viable option.

There is a different kind of problem for the strategy of developing pharmacological treatments to overcome the lack of specificity mentioned in section 4.1. In Chekroud and colleagues' review of data-driven approaches to predicting treatment outcomes (2021) they find that factors at the psychological and social levels "consistently offer more meaningful and generalizable predictions" (2021, p. 165) than those at the neurobiological and genetic levels. In other words, when it comes to treatment outcomes, most of the predictive signal is not to be found at a level that can be targeted directly by pharmacological interventions. This means that, without a well-validated mapping between the predictive psychological constructs and the neurobiological entities that implement them, pharmacological interventions cannot be targeted directly at understanding (Ivanov & Schwartz, 2021; Poldrack & Yarkoni, 2016). As Ivanov and Schwartz (2021, p. 5) say, "we are still very far indeed" from possessing such a mapping. This critique is not meant to trivialize the important work of using the data-driven approach to improve the allocation of extant treatments for conditions like depression (Chekroud et al., 2016; Nie et al., 2018). However, these success stories are of the modest, indirect variety. They do not give any reason to think that the data-driven approach can optimize novel pharmacological treatments to directly target psychopathology.

4.3 Translation problems for data-driven insights

Another avenue by which the data-driven approach might transform mental healthcare is by extracting therapeutically actionable insights from trained models. Indeed, a common assertion among proponents of the data-driven approach is that these methods might discover subtle and complex patterns which aid mechanistic understanding of mental disorders. For example, Durstewitz and colleagues (2019, p. 1592) suggest that as well as using deep neural networks for prediction, we may be able "to employ them to also gain insight into physiological, computational, and cognitive mechanisms". Along the same lines, Koppe and colleagues (2021,

p. 187) state that using visualization techniques on machine learning models "may help uncover interpretable multi-modal biomarkers of psychiatric disease".

One reason to think this might be so is the *convergence logic* that characterizes data-driven success stories in science. I introduce the term convergence logic to refer to the idea that, because accurate prediction in complex, high-dimensional domains is exceedingly difficult, any model that attains high accuracy under tight constraints *must* do so in virtue of learning robust patterns which capture core features of the data domain being modeled (Schrimpf et al., 2020). This line of reasoning is behind the neuroscientific research program using deep neural networks to model the primate ventral visual stream (Yamins & DiCarlo, 2016). In that case, the conjecture is that in order to be able to accurately classify images, a neural network *has to* converge upon mechanisms that approximate its biological target[14]. The convergence logic is also manifest in the linguistic structure that is learned by neural network language models (Manning et al., 2020). In order for a model to accurately predict language tokens, it *has to* internalize aspects of linguistic structure. Moving back to psychiatry, if large computational models incorporating multiple modalities are able to generate accurate prognoses or reliably pick out novel disorder subtypes, one might think *it must be* because they are converging on robust structure in the data[15]. On this view, machine learning models might be the right tools to discover the kind of "fuzzy, cross-level mechanisms" that Kendler suggests might be the "true nature of psychiatric disorders" (Kendler, 2012, p. 18).

How would such discovery relate to patient understanding? Kendler and Campbell (2014, p. 1) suggest that scientific knowledge might "expand the domain of the understandable" by providing models of psychopathological processes that render them intelligible (note that Kendler and Campbell are referring to Jaspers' clinician-centered notion of understanding here). As an example, the authors discuss the pathological misattribution of meaning that leads to delusional beliefs in schizophrenia, considered paradigmatically un-understandable by Jaspers. Modern neuroscience posits the aberrant behavior of dopaminergic neurons encoding motivational salience to explain how these beliefs are formed. Kendler and Campbell suggest that this form of dopamine system dysfunction *makes sense of* the psychopathological experiences. If the data-driven approach could provide scientific knowledge about a range of psychiatric disorders, those insights might be used to provide adaptive explanations to patients and induce the understanding required to radically improve outcomes[16]. Of course, there may be fundamental limits to this enterprise in certain cases. For example, it is difficult to see how scientific understanding could translate into an understanding

[14]Naturally, this holds only for a given level of abstraction.

[15]We are assuming here that rigorous cross-validation rules out models achieving accuracy by overfitting on spurious correlations.

[16]Other authors have argued that bringing neuroscience into the clinic could have negative consequences, inducing fatalism by conveying a deterministic world, stripped of human agency (Hyman & McConnell, 2020).

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

of psychopathological experiences which lack internal coherence[17]. This is an avenue that future research must explore. However, even if the convergence logic holds and novel multilevel patterns which can be mapped onto features of mental disorder are discovered[18], there are still important barriers for translating data-driven insights into improved clinical outcomes.

The first is the issue of intelligibility. The flexibility and expressivity of the machine learning paradigm come with a cost. *How* these models transform input into output is, *prima facie*, unintelligible (Zednik, 2021). In their review of machine learning applications in psychiatry, Cearns and colleagues (2019, p. 9) claim forcefully that "knowing every single one of the hundreds of millions of parameter values in a given ML model would fail to provide even a spec of practically useful insight into the inner workings of a trained model". However, this assessment is overly pessimistic. While it is clear that no human could infer meaning directly from hundreds of millions of parameter values, functional characterization of trained models at a higher-level of abstraction, such as layers or pathways in a neural network, may be possible. In the domain of psychiatry, this would require translating the components of trained machine learning models into psychiatrically meaningful terms (Sheu, 2020). Strategies for interpreting the internal structure of trained models are being developed (Elhage et al., 2021; Koppe et al., 2021; Olah et al., 2018; Sheu, 2020). Researchers following such strategies, such as Anthropic's *mechanistic interpretability* approach (Elhage et al., 2021), use a combination of mathematics and experimental techniques to probe machine learning models in an attempt to characterize their functional organization. While further research is clearly required, applying interpretability techniques to large predictive models is a promising option for discovering the multi-level structure of psychiatric disorders.

Unfortunately, even the discovery of novel insights into mental illness would not guarantee a transformative impact for the data-driven approach. This is due to what I call the *translation problem.* Namely, that the understanding gleaned by psychiatrists and the scientific community still needs to be translated into therapeutic techniques that induce *patient* understanding. If a patient cannot internalize an explanation, then that explanation will not be adaptive for them. Explanations invoking fuzzy and complex multi-level mechanisms, even if understood by psychiatrists, may not be communicable to the average patient. In order for these hard-won insights to be clinically relevant, they need to be translated into language that resonates with the "assumptive worlds" of specific patients (Frank & Frank, 1993, p. 30).[19] Supporting this point, research has shown that having

[17]I thank an anonymous reviewer for raising this point.

[18]There is not space here to comment on whether such patterns would necessarily constitute mechanisms. The argument does not turn on whether or not they would.

[19]Wampold and colleagues flesh this notion out by suggesting that explanations "should be proximal to the client's currently held explanation or expectation, and should not create dissonance with the attitudes and values of the client" (Wampold et al., 2007, p. 125). This coheres with the

a *shared rationale* with a practitioner is important to patients and contributes to improved outcomes (Cuevas et al., 2014; Johansson & Eklund, 2003; Wampold & Budge, 2012). Further, non-specific components of psychiatric treatments, such as clinical communication and treatment framing, also play a significant role in patient outcomes (Priebe et al., 2020). This suggests that how explanations are presented is important and that the inferential distance between practitioner and patient should not be stretched too far. These points highlight the difficulty of the translation problem. Discovering the multi-level structure of psychiatric disorder is not enough. The data-driven approach still requires important theoretical work to be done before the bottleneck of understanding can be addressed.

4.4 Implications

Overall, then, while the data-driven approach is a fascinating and powerful addition to the landscape of psychiatric science, it has no straightforward means of unblocking the bottleneck of understanding. As such, the suggestion that this approach will bring about a transformative improvement to patient outcomes is not warranted. This is not to claim that the data-driven approach has no role to play. Though I have argued it will not be transformative, increasing the efficiency of the allocation of extant treatments is still an important goal. Further, the data-driven analysis of pharmacological and psychotherapeutic interventions is a promising line of research from which important principles and insights regarding treatment strategies can be derived (Ewbank et al., 2021). Finally, discoveries about how fine-grained patterns of variation across multiple levels relate to psychopathology and the course of mental illness will surely contribute to how we understand and treat psychiatric disorder. To be clear then, the conclusion of the argument is that the contribution of the data-driven approach ought to be incremental and interdisciplinary, rather than discontinuous and technocratic. Proponents take pains to acknowledge the technical, ethical, and practical challenges facing their endeavor (Bzdok & Meyer-Lindenberg, 2018; Cearns et al., 2019). However, the function-oriented and theory-neutral philosophy often obscures the deep *conceptual* challenges pervading a data-driven approach to psychiatric science. The problem of inducing understanding in patients suffering mental distress cannot be solved by data alone.

5 Responses to objections

I now anticipate and respond to possible objections to two of the core concepts invoked in my argument.

point, explained in section 3.1, that adaptive explanations, and by extension understanding, need not be true.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

5.1 On understanding

I cast *understanding* as the possession of adaptive mental constructs that enable a subject to perform feats relevant to them. However, critics might protest that this pragmatic notion has not been adequately defended. I will briefly address three potential worries. First, there are many competing conceptualizations of understanding out there, so why pick this one in particular? Second, does this notion really capture the views of the scholars I cited? Third, does this conceptualization really capture anything unified at all?

On the first issue, as to why I choose this notion of understanding rather than another, I make a clarification rather than supplying an argument. I endorse a form of conceptual pluralism whereby concepts can be invoked or defined stipulatively for the sake of argument so long as this practice is not abused to willfully obfuscate. The concept I am invoking was inspired by notions of understanding from pragmatist philosophy and theoretical work in psychiatry. I take this to suffice to show that my usage is neither without precedent nor needlessly idiosyncratic. Ultimately, for the purposes of this paper, whether *understanding* is the right term to label the concept I have described is less important than whether or not that concept plays the role I have suggested in limiting the potential of the data-driven approach to psychiatry.

A more serious worry is that I have erroneously conflated distinct views and presented them under a common label. Along these lines, one might worry that the *acquired nexus* invoked by Wilhelm Dilthey (1984/1977) in his discussion of understanding and the *cognitive schemas* central to Aaron Beck's (Beck & Haigh, 2014) theorizing are rather distinct from the *adaptive explanations* that Wampold (Wampold & Budge, 2012) describes. The former two notions involve subconscious aspects of mentality that organize ongoing perception and experience, whereas the latter notion seems to be talking about explicit and propositional explanations. Are these authors not pointing to distinct concepts? I think a close reading shows that they each have something broader in mind. First, both Dilthey and Beck include propositional entities like beliefs when spelling out their ideas. They do not limit their concepts to the unconscious. Second, Wampold takes a very liberal stance on what constitutes an explanation. For example, he describes an inability to recognize one has a problem as a dysfunctional explanation, even though this would clearly not be an articulable proposition (Wampold et al., 2007, p. 123). As such, despite their different framings, these authors are capturing a broadly co-extensive set of mental constructs. Namely, those that play a critical role in structuring, organizing, and influencing ongoing experience.

Of course, even if I am right that *understanding* captures what Dilthey, Beck, and Wampold have in mind, it may still be the case that it does not play the role I am claiming it does. A third objection that one might have along these lines is that this notion of understanding is both too broad and too vague to be useful. It is broad in the sense described above, covering both subconscious and conscious constructs. It is vague in the sense that I have not specified precisely what set of

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

things is supposed to be captured by *mental constructs*. In response to this worry, I stress that both the broadness of the notion and the vagueness with respect to what kind of entities are being referred to are features of the account, not bugs. Much of my argument turns on the fact that we have an inadequate theory of *how exactly* these influences – beliefs, values, goals, expectations, and the like – give rise to psychopathological experience. If we knew which constructs were relevant, how they interact with one another, and how to specify and measure them, the prospects of the data-driven approach addressing the bottleneck of understanding would be significantly improved (see Poldrack & Yarkoni, 2016 for an exposition of the view that there is unlikely to be any single 'correct' way to carve psychology into a set of mental constructs). As such, I gladly acknowledge that my account is underspecified. In my view, this reflects an appropriate degree of epistemic uncertainty with respect to the details.

5.2 On the bottleneck metaphor

One possible objection to casting understanding as a bottleneck is that the mental states that make up understanding do not bear the appropriate relation to the other components in the system to justify the metaphor. In particular, a proponent of the biological approach to psychiatry might argue that a suitable pharmacological intervention not only alters a patient's brain states, but also alters their mental states. This interlocuter could suggest that treating the underlying neurobiology and the mental dysfunction as independent smacks of Cartesian dualism. On this view, regardless of the messy details, as long as subjective psychopathology is *implemented* by neural wetware, it stands to reason that intervening on that wetware *constitutes* intervening on the patient's relevant mental constructs. This threatens the applicability of the bottleneck metaphor, which implies that pharmacological interventions are causally upstream of the mental constructs that feature in understanding.

It is certainly true that the complex relation holding between brain and mind makes psychiatry significantly trickier to reason about than other systems which experience bottlenecks, such as factory production lines. However, for the bottleneck metaphor to be useful, all that is required is a dissociation between the neural changes known to be induced by pharmacological agents and the development of understanding. Notice that such a dissociation is a function of our state of knowledge, and thus not a strong claim about the relation between brain and mind. The utility of treating the cognitive and chemical as dissociable can be seen in the cognitive neuropsychological theory of how antidepressants work (Harmer et al., 2017, 2009). On this view, the early biological effects of selective serotonin reuptake inhibitors lead to a positive shift in the processing of emotionally salient information, which then leads to improved mood. This case highlights how scholars still find it useful to consider the chemical and the cognitive as functionally distinct. Of course, Harmer and colleagues are well aware that the cognitive effects

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

they point to are implemented by further biological changes (such as increased neural plasticity). The point is that the precise details of these changes are not yet known and hence cannot be targeted directly by pharmacological intervention (see also section 4.2). Further, it has been shown that no specific feature of brain biology is sufficient to demarcate healthy and pathological populations (First et al., 2018; Holmes & Patrick, 2018). This implies that the dissociation should hold for any pharmacological agent which targets a particular biomarker, at least to the extent that there is uncertainty regarding the agent's downstream effects. Of course, it is possible that future technology will enable fundamentally different kinds of neurobiological intervention. However, this does not threaten the bottleneck framing, which is specifically intended to characterize the challenge psychiatric science faces *today*, given our current epistemic and technological situation. In sum, I maintain it is reasonable to consider patient understanding a bottleneck to robust recovery.

6 Conclusion

The focus of this article is the question: Does the data-driven approach have transformative potential for psychiatric science? In providing an answer to this question, I focused on a core limitation. Namely, that *patient understanding* functions as a bottleneck for recovery, and the data-driven approach is ill-suited to address that bottleneck. If this view is correct, talk of an AI revolution in mental healthcare is premature. Given the current state of psychiatric science, the efficient allocation of treatments, which the data-driven approach can deliver, is likely to yield only modest improvements to mental health outcomes. In order to unlock additional benefits, further knowledge of the specificities of psychopathology and how treatments can be tailored to induce the understanding necessary for recovery are needed. This may involve engaging with research on features of psychiatry that are difficult to formalize, such as empathy, non-verbal communication, and the framing of treatments (Priebe et al., 2020), the role of placebo effects in treatment outcomes (Enck & Zipfel, 2019), and the practitioner-patient relationship (Horn & Weisz, 2020). As explained in section 4.2, this line of research involves many researcher degrees of freedom and cannot proceed in a theory-neutral way. While the data-driven approach can contribute to these endeavors, it will do so best in collaboration with a theory-driven attempt to understand mental disorders.

Acknowledgments
Work on this paper was funded by the Volkswagen Foundation grant "Explainable Intelligent Systems" (EIS) (grant numbers 9B 830 and 98 509). The author would like to thank all the participants of the Minds, Models, and Mechanisms workshop (funded by DFG grant number 446794119) for stimulating discussions of these issues as well as the anonymous reviewers for many useful suggestions. Thanks also to Theresa Waclawek, Chiara Caporuscio, and Timo Speith for their comments on a draft. Finally, thanks to all members of the Explainable Intelligent Systems research group for helpful feedback on the initial ideas from which this article grew.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders*. https://doi.org/10.1176/appi.books.9780890425787

Anthony, W. A. (2000). A recovery-oriented service system: Setting some system level standards. *Psychiatric Rehabilitation Journal*, *24*(2), 159–168. https://doi.org/10.1037/h0095104

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*(Pt B), 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current status and future prospects of clinical psychology. *Psychological Science in the Public Interest*, *9*(2), 67–103. https://doi.org/10.1111/j.1539-6053.2009.01036.x

Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. *PLOS ONE*, *16*(3), e0248152. https://doi.org/10.1371/journal.pone.0248152

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [Cs, Stat]*. http://arxiv.org/abs/1806.01261

Beck, A. T. (1985). Cognitive therapy, behavior therapy, psychoanalysis, and pharmacotherapy. In M. J. Mahoney & A. Freeman (Eds.), *Cognition and psychotherapy* (pp. 325–347). Springer US. https://doi.org/10.1007/978-1-4684-7562-3_14

Beck, A. T., & Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual Review of Clinical Psychology*, *10*, 1–24. https://doi.org/10.1146/annurev-clinpsy-032813-153734

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/tpami.2013.50

Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The two cultures of computational psychiatry. *JAMA Psychiatry*, *76*(6), 563–564. https://doi.org/10.1001/jamapsychiatry.2019.0231

Bickman, L. (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. *Administration and Policy in Mental Health*, *47*(5), 795–843. https://doi.org/10.1007/s10488-020-01065-8

Boer, N. S. de, Bruin, L. C. de, Geurts, J. J. G., & Glas, G. (2021). The network theory of psychiatric disorders: A critical assessment of the inclusion of environmental factors. *Frontiers in Psychology*, *12*, 221. https://doi.org/10.3389/fpsyg.2021.623970

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*, e2. https://doi.org/10.1017/S0140525X17002266

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478 [Cs, Stat]*. http://arxiv.org/abs/2104.13478

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, *218*(3), 131–134. https://doi.org/10.1192/bjp.2019.245

Budd, R., & Hughes, I. (2009). The dodo bird verdict–controversial, inevitable and important: A commentary on 30 years of meta-analyses. *Clinical Psychology & Psychotherapy*, *16*(6), 510–522. https://doi.org/10.1002/cpp.648

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*. https://doi.org/10.1016/j.bpsc.2017.11.007

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 1–12. https://doi.org/10.1038/s41398-019-0607-2

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet. Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chollet, F. (2021). *Deep learning with python, second edition*. Simon; Schuster.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Coulombe, S., Radziszewski, S., Meunier, S., Provencher, H., Hudon, C., Roberge, P., Provencher, M. D., & Houle, J. (2016). Profiles of recovery from mood and anxiety disorders: A person-centered exploration of people's engagement in self-management. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00584

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

Cuevas, las C. D., Peñate, W., & Rivera, de L. (2014). To what extent is treatment adherence of psychiatric patients influenced by their participation in shared decision making? *Patient Preference and Adherence*, *8*, 1547–1553. https://doi.org/10.2147/PPA.S73029

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 126. https://doi.org/10.1186/1741-7015-11-126

Dembic, S. (2023). Mental disorder: An ability-based view. *Philosophy and the Mind Sciences*, *4*. https://doi.org/10.33735/phimisci.2023.9630

Dilthey, W. (1977). Ideas concerning a descriptive and analytic psychology (1894). In W. Dilthey (Ed.), *Descriptive psychology and historical understanding* (pp. 21–120). Springer Netherlands. https://doi.org/10.1007/978-94-009-9658-8_2 (Original work published 1984)

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Doyle, Y., Egan, S., Graham, N., & Khalifa, K. (2019). Non-factive understanding: A statement and defense. *Journal for General Philosophy of Science*, *50*(3), 345–365. https://doi.org/10.1007/s10838-019-09469-3

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, *24*(11), 1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Elujide, I., Fashoto, S. G., Fashoto, B., Mbunge, E., Folorunso, S. O., & Olamijuwon, J. O. (2021). Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Informatics in Medicine Unlocked*, *23*, 100545. https://doi.org/10.1016/j.imu.2021.100545

Enck, P., & Zipfel, S. (2019). Placebo effects in psychotherapy: A framework. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00456

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science (New York, N.Y.)*, *196*(4286), 129–136. https://doi.org/10.1126/science.847460

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, *77*(1), 35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, *31*(3), 300–312. https://doi.org/10.1080/10503307.2020.1788740

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The heterogeneity problem: Approaches to identify psychiatric subtypes. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of "precision psychiatry." *BMC Medicine*, *15*(1), 80. https://doi.org/10.1186/s12916-017-0849-x

First, M. B., Drevets, W. C., Carter, C., Dickstein, D. P., Kasoff, L., Kim, K. L., McConathy, J., Rauch, S., Saad, Z. S., Savitz, J., Seymour, K. E., Sheline, Y. I., & Zubieta, J.-K. (2018). Clinical applications of neuroimaging in psychiatric disorders. *The American Journal of Psychiatry*, *175*(9), 915–916. https://doi.org/10.1176/appi.ajp.2018.1750701

Frances, A. (2009). Whither DSM-V? *The British Journal of Psychiatry: The Journal of Mental Science*, *195*(5), 391–392. https://doi.org/10.1192/bjp.bp.109.073932

Frank, J. D., & Frank, J. B. (1993). *Persuasion and healing: A comparative study of psychotherapy*. JHU Press.

García-Gutiérrez, M. S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A., & Manzanares, J. (2020). Biomarkers in psychiatry: Concept, definition, types and relevance to the clinical reality. *Frontiers in Psychiatry*, *11*, 432. https://doi.org/10.3389/fpsyt.2020.00432

Gauld, C., Micoulaud-Franchi, J.-A., & Dumas, G. (2021). Comment on starke et al.: "Computing schizophrenia: Ethical challenges for machine learning in psychiatry": From machine learning to student learning: Pedagogical challenges for psychiatry. *Psychological Medicine*, *51*(14), 2509–2511. https://doi.org/10.1017/S0033291720003906

Ghaemi, S. N. (2007). Existence and pluralism: The rediscovery of Karl Jaspers. *Psychopathology*, *40*(2), 75–82. https://doi.org/10.1159/000098487

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42. https://doi.org/10.1016/j.cobeha.2017.07.003

Glover, J. (2020). Psychiatry, folk psychology, and the impact of neuroscience - a response to Steve Hyman's Loebel lectures. In J. Savulescu, R. Roache, W. Davies, & J. P. Loebel (Eds.), *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine* (pp. 301–319). Oxford University Press.

Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *NeuroImage: Clinical*, *6*, 229–236. https://doi.org/10.1016/j.nicl.2014.09.009

Grimm, S. R. (2019). Varieties of Understanding. In S. R. Grimm (Ed.), *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (pp. 2–13). Oxford University Press. https://doi.org/10.1093/oso/9780190860974.001.0001

Harmer, C. J., Duman, R. S., & Cowen, P. J. (2017). How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet. Psychiatry*, *4*(5), 409–418. https://doi.org/10.1016/S2215-0366(17)30015-9

Harmer, C. J., Goodwin, G. M., & Cowen, P. J. (2009). Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *The British Journal of Psychiatry*, *195*(2), 102–108. https://doi.org/10.1192/bjp.bp.108.051193

Health, D. of, & Care, S. (2021). Mental health recovery plan backed by £500 million. In *GOV.UK*. https://www.gov.uk/government/news/mental-health-recovery-plan-backed-by-500-million

Hengartner, M. P., & Lehmann, S. N. (2017). Why psychiatric research must abandon traditional diagnostic classification and adopt a fully dimensional scope: Two solutions to a persistent problem. *Frontiers in Psychiatry*, *8*, 101. https://doi.org/10.3389/fpsyt.2017.00101

Hills, A. (2016). Understanding why. *Noûs*, *50*(4), 661–688. https://doi.org/10.1111/nous.12092

Holmes, A. J., & Patrick, L. M. (2018). The myth of optimality in clinical neuroscience. *Trends in Cognitive Sciences*, *22*(3), 241–257. https://doi.org/10.1016/j.tics.2017.12.006

Horn, R. L., & Weisz, J. R. (2020). Can artificial intelligence improve psychotherapy research and practice. *Administration and Policy in Mental Health*, *47*(5), 852–855. https://doi.org/10.1007/s10488-020-01056-9

Hyman, S. E., & McConnell, D. (2020). Mental illness: The collision of meaning with mechanism. In *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine* (pp. 263–289). Oxford University Press. https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697

Insel, T. R., Collins, P. Y., & Hyman, S. E. (2015). Darkness invisible: The hidden global costs of mental illness. *Foreign Affairs*, *94*(1), 127–135.

Ivanov, I., & Schwartz, J. M. (2021). Why psychotropic drugs don't cure mental illness—but should they? *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.579566

Jacob, K. S. (2015). Recovery model of mental illness: A complementary approach to psychiatric care. *Indian Journal of Psychological Medicine*, *37*(2), 117–119. https://doi.org/10.4103/0253-7176.155605

Jaspers, K. (1997). *General psychopathology*. JHU Press. (Original work published 1959)

Johansson, H., & Eklund, M. (2003). Patients' opinion on what constitutes good psychiatric care. *Scandinavian Journal of Caring Sciences*, *17*(4), 339–346. https://doi.org/10.1046/j.0283-9318.2003.00233.x

Joober, R., & Tabbane, K. (2019). From the neo-Kraepelinian framework to the new mechanical philosophy of psychiatry: Regaining common sense. *Journal of Psychiatry & Neuroscience : JPN*, *44*(1), 3–7. https://doi.org/10.1503/jpn.180240

Kendler, K. S. (2012). Levels of explanation in psychiatric and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry*, *17*(1), 11–21. https://doi.org/10.1038/mp.2011.70

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15*(1), 5–12. https://doi.org/10.1002/wps.20292

Kendler, K. S., & Campbell, J. (2014). Expanding the domain of the understandable in psychiatric illness: An updating of the Jasperian framework of explanation and understanding. *Psychological Medicine*, *44*(1), 1–7. https://doi.org/10.1017/S0033291712003030

Kendler, K. S., & Gyngell, C. (2020). Multilevel interactions and the dappled causal world of psychiatric disorders. In *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine* (pp. 25–45). Oxford University Press. https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697

Keogh, E., & Mueen, A. (2017). Curse of dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192

Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, *46*(1), 176–190. https://doi.org/10.1038/s41386-020-0767-z

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., … Meisenzahl, E. (2021). Multimodal machine learning workflows for prediction

of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry*, *78*(2), 1–15. https://doi.org/10.1001/jamapsychiatry.2020.3604

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67. https://doi.org/10.1007/s13194-012-0056-8

Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy*, *51*(4), 467–481. https://doi.org/10.1037/a0034332

Leder, G., & Zawidzki, T. (2023). The skill of mental health: Towards a new theory of mental health and disorder. *Philosophy and the Mind Sciences*, *4*. https://doi.org/10.33735/phimisci.2023.9684

Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: An umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry*, *21*(1), 133–145. https://doi.org/10.1002/wps.20941

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

McConnell, D. (2020). The proper place of subjectivity, meaning, and folk psychology in psychiatry. In *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine* (pp. 290–303). Oxford University Press. https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697

McConnell, D., & Snoek, A. (2018). The importance of self-narration in recovery from addiction. *Philosophy, Psychiatry, and Psychology*, *25*(3), 31–44. https://doi.org/10.1353/ppp.2018.0022

McCoy, L. G., Nagaraj, S., Morgado, F., Harish, V., Das, S., & Celi, L. A. (2020). What do medical students actually need to know about artificial intelligence? *Npj Digital Medicine*, *3*(1), 1–3. https://doi.org/10.1038/s41746-020-0294-7

McGorry, P. D., & Nelson, B. (2019). Transdiagnostic psychiatry: Premature closure on a crucial pathway to clinical utility for psychiatric diagnosis. *World Psychiatry*, *18*(3), 359–360. https://doi.org/10.1002/wps.20679

Middleton, H., & Moncrieff, J. (2019). Critical psychiatry: A brief overview. *BJPsych Advances*, *25*(1), 47–54. https://doi.org/10.1192/bja.2018.38

Miranda, L., Paul, R., Pütz, B., Koutsouleris, N., & Müller-Myhsok, B. (2021). Systematic review of functional MRI applications for psychiatric disease subtyping. *Frontiers in Psychiatry*, *12*, 665536. https://doi.org/10.3389/fpsyt.2021.665536

Najafpour, Z., Fatemi, A., Goudarzi, Z., Goudarzi, R., Shayanfard, K., & Noorizadeh, F. (2021). Cost-effectiveness of neuroimaging technologies in management of psychiatric and insomnia disorders: A meta-analysis and prospective cost analysis. *Journal of Neuroradiology = Journal De Neuroradiologie*, *48*(5), 348–358. https://doi.org/10.1016/j.neurad.2020.12.003

Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PloS One*, *13*(6), e0197268. https://doi.org/10.1371/journal.pone.0197268

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), 10.23915/distill.00010. https://doi.org/10.23915/distill.00010

Paulus, M. P. (2015). Pragmatism instead of mechanism: A call for impactful biological psychiatry. *JAMA Psychiatry*, *72*(7), 631–632. https://doi.org/10.1001/jamapsychiatry.2015.0497

Pelin, H., Ising, M., Stein, F., Meinert, S., Meller, T., Brosch, K., Winter, N. R., Krug, A., Leenings, R., Lemke, H., Nenadić, I., Heilmann-Heimbach, S., Forstner, A. J., Nöthen, M. M., Opel, N., Repple, J., Pfarr, J., Ringwald, K., Schmitt, S., … Andlauer, T. F. M. (2021). Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *46*(11), 1895–1905. https://doi.org/10.1038/s41386-021-01051-0

Peterson, B. S. (2019). Editorial: Common factors in the art of healing. *Journal of Child Psychology and Psychiatry*, *60*(9), 927–929. https://doi.org/10.1111/jcpp.13108

Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology*, *67*(1), 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Priebe, S., Conneely, M., McCabe, R., & Bird, V. (2020). What can clinicians do to improve outcomes across psychiatric treatments: A conceptual review of non-specific components. *Epidemiology and Psychiatric Sciences*, *29*. https://doi.org/10.1017/S2045796019000428

Qureshi, M. N. I., Min, B., Jo, H. J., & Lee, B. (2016). Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical extreme learning machine: Structural MRI study. *PLOS ONE*, *11*(8), e0160697. https://doi.org/10.1371/journal.pone.0160697

Ramon, S., Healy, B., & Renouf, N. (2007). Recovery from mental illness as an emergent concept and practice in Australia and the UK. *International Journal of Social Psychiatry*, *53*(2), 108–122. https://doi.org/10.1177/0020764006075018

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

Regt, H. W. de. (2019). From explanation to understanding: Normativity lost? *Journal for General Philosophy of Science*, *50*(3), 327–343. https://doi.org/10.1007/s10838-019-09477-3

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6*, 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: Toward clinical applications. *Current Opinion in Neurobiology*, *55*, 152–159. https://doi.org/10.1016/j.conb.2019.02.006

Sathyanarayana Rao, T. S., & Andrade, C. (2016). Classification of psychotropic drugs: Problems, solutions, and more problems. *Indian Journal of Psychiatry*, *58*(2), 111–113. https://doi.org/10.4103/0019-5545.183771

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Sharma, A., & Verbeke, W. J. M. I. (2020). Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers dutch dataset (n = 11,081). *Frontiers in Big Data*, *3*. https://doi.org/10.3389/fdata.2020.00015

Sheu, Y.-H. (2020). Illuminating the black box: Interpreting deep neural network models for psychiatric research. *Frontiers in Psychiatry*, *11*, 551299. https://doi.org/10.3389/fpsyt.2020.551299

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, *22*(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

Shorter, E. (1997). *A history of psychiatry: From the era of the asylum to the age of Prozac*. John Wiley & Sons.

Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., & Roberts, K. (2021). Deep representation learning of patient data from electronic health records (EHR): A systematic review. *Journal of Biomedical Informatics*, *115*, 103671. https://doi.org/10.1016/j.jbi.2020.103671

Stanghellini, G., & Broome, M. R. (2014). Psychopathology as the basic science of psychiatry. *The British Journal of Psychiatry*, *205*(3), 169–170. https://doi.org/10.1192/bjp.bp.113.138974

Starke, G., Clercq, E. D., Borgwardt, S., & Elger, B. S. (2021). Computing schizophrenia: Ethical challenges for machine learning in psychiatry. *Psychological Medicine*, *51*(15), 2515–2521. https://doi.org/10.1017/S0033291720001683

Stoyanov, D., & Maes, M. H. (2021). How to construct neuroscience-informed psychiatric classification? Towards nomothetic networks psychiatry. *World Journal of Psychiatry*, *11*(1), 1–12. https://doi.org/10.5498/wjp.v11.i1.1

Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, *10*(1), 1–26. https://doi.org/10.1038/s41398-020-0780-3

Suhara, Y., Xu, Y., & Pentland, A. 'Sandy'. (2017). DeepMood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *Proceedings of the 26th International Conference on World Wide Web*, 715–724. https://doi.org/10.1145/3038912.3052676

Thornton, T. (2020). Psychiatry's inchoate wish for a paradigm shift and the biopsychosocial model of mental illness. In *Psychiatry reborn: Biopsychosocial psychiatry in modern medicine* (pp. 229–239). Oxford University Press. https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Trivedi, M. H. (2016). Right patient, right treatment, right time: Biosignatures and precision medicine in depression. *World Psychiatry*, *15*(3), 237–238. https://doi.org/10.1002/wps.20371

Trust, W. B. (2020). A unified vision for the future of mental health, addiction, and well-being in the united states. In *Well Being Trust*. https://wellbeingtrust.org/press-releases/ceos-from-14-top-mental-health-organizations-join-together/

Tsou, J. Y. (2021). Philosophy of psychiatry. *Elements in the Philosophy of Science*. https://doi.org/10.1017/9781108588485

Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet. Psychiatry*, *3*(2), 171–178. https://doi.org/10.1016/S2215-0366(15)00505-2

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wampold, B. E., & Budge, S. L. (2012). The 2011 Leona Tyler award address: The relationship—and its relationship to the common and specific factors of psychotherapy. *The Counseling Psychologist*, *40*(4), 601–623. https://doi.org/10.1177/0011000011432709

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work, 2nd ed.* Routledge/Taylor & Francis Group.

Wampold, B. E., Imel, Z. E., Bhati, K. S., & Johnson-Jennings, M. D. (2007). Insight as a common factor. In *Insight in psychotherapy* (pp. 119–139). American Psychological Association.

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, *190*(6), 997–1016. https://doi.org/10.1007/s11229-011-0055-x

Wolpert, M., Pote, I., & Sebastian, C. L. (2021). Identifying and integrating active ingredients for mental health. *The Lancet Psychiatry*, *8*(9), 741–743. https://doi.org/10.1016/S2215-0366(21)00283-2

World Health Organization. (2019). *Special initiative for mental health (2019–2023).* https://www.who.int/publications-detail-redirect/special-initiative-for-mental-health-(2019-2023)

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Zachar, P., & Kendler, K. S. (2017). The philosophy of nosology. *Annual Review of Clinical Psychology*, *13*, 49–71. https://doi.org/10.1146/annurev-clinpsy-032816-045020

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, *34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, *4*, 5. https://doi.org/10.33735/phimisci.2023.9658