PhiMiSci
Philosophy and the Mind Sciences

Machine learning and its impact on psychiatric nosology

Findings from a qualitative study among German and Swiss experts

Georg Starke[a,b] (georg.starke@unibas.ch)
Bernice Simone Elger[a,c](b.elger@unibas.ch)
Eva De Clercq[a](eva.declercq@unibas.ch)

Abstract

The increasing integration of Machine Learning (ML) techniques into clinical care, driven in particular by Deep Learning (DL) using Artificial Neural Nets (ANNs), promises to reshape medical practice on various levels and across multiple medical fields. Much recent literature examines the ethical consequences of employing ML within medical and psychiatric practice but the potential impact on psychiatric diagnostic systems has so far not been well-developed. In this article, we aim to explore the challenges that arise from the recent use of ANNs for the old problems of psychiatric nosology. To enable an empirically supported critical reflection on the topic, we conducted semi-structured qualitative interviews with Swiss and German experts in computational psychiatry. Here, we report our findings structured around two themes, namely (1) the *possibility* of using ML for defining or refining of psychiatric classification, and (2) the *desirability* of employing ML for psychiatric nosology. We discuss these themes by relating them to recent debates about network theory for psychiatric nosology and show why empirical research in the field should critically reflect on its contribution to psychopathology research. In sum, we argue that beyond technical, regulatory, and ethical challenges, philosophical reflection is crucial to harness the potential of ML in psychiatry.

Keywords

Machine learning · Nosology · Psychiatry · Ethics · Expert interviews

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

[a]Institute for Biomedical Ethics, University of Basel, Switzerland

[b]College of Humanities, École Polytechnique Fédérale de Lausanne, Switzerland

[c]University Center of Legal Medicine, University of Geneva, Switzerland

Starke, G., Elger, B. S., & De Clercq, E. (2023). Machine learning and its impact on psychiatric nosology. Findings from a qualitative study among German and Swiss experts. *Philosophy and the Mind Sciences*, *4*, 4. https://doi.org/10.33735/phimisci.2023.9435

1 Introduction

Deep Learning (DL) based on Artificial Neural Networks (ANNs) is at the heart of many recent success stories in the field of Machine Learning (ML). Within psychiatry, DL also promises useful tools for the diagnosis and treatment of psychiatric disorders (Durstewitz et al., 2019; Jacobson & Bhattacharya, 2022; Quaak et al., 2021; Walter et al., 2019). The recent first approval of a DL-based program by the US Food and Drug Administration to aid with the diagnosis of autism spectrum disorder in young children bears witness to this potential (Dattaro, 2021). Beyond diagnosis, DL-based programs could also provide complementary offers of digital psychotherapy (Lui et al., 2017; Martinez-Martin & Kreitmair, 2018), predict individual treatment outcomes (Chekroud et al., 2021) or give prognostic estimates, for instance concerning psychosis (Salazar de Pablo et al., 2021).

Responding to long-standing nosological debates within the discipline and dissatisfaction with existing diagnostic criteria (Cuthbert & Insel, 2013; Insel & Cuthbert, 2015; Kendler, 2016), DL is also increasingly discussed as a potential technique to arrive at novel or refined psychiatric classifications (Brunn et al., 2020; Eitel et al., 2021). DL-based clustering promises to provide a data-driven approach that can subdivide groups of patients automatically based on neurobiological and behavioural data, finding novel modes of representation (Karim et al., 2021; Schulz et al., 2020). In principle, such clustering can draw on many different kinds of data, including functional and structural neuroimaging data, EEG measurements, genetic and epigenetic data as well as clinical and neurocognitive observations (Huys et al., 2016). To give an example for a neuroscience-focused approach, harnessing the advantages of DL, Chang et al. recently reported to have identified subgroups of patients with major psychiatric disorders such as bipolar depression, major depressive disorder, and schizophrenia that are characterized by a frontal–posterior functional imbalance and seem to respond differently to psychopharmacological interventions (Chang et al., 2021). While such findings require validation and replication, they could improve existing diagnostic criteria and provide hypothesis for future research (Eitel et al., 2021).

In parallel to the rise of neuroscientific and psychiatric research endeavours driven by DL, there has also been a blossoming of theoretical approaches that define psychiatric disorders in terms of clusters or networks. Such approaches have been especially prominent among nonessentialist theories, i.e., theories that do not espouse a mind-independent understanding of psychiatric disorders as given natural kinds that share intrinsic natural properties. Among these nonessentialist approaches, Denny Borsboom's suggestion that mental disorders could best be described as complex networks of causally-linked, interconnected symptom components has been particularly influential (Borsboom, 2017). Symptom network theory promises to provide a non-reductionist link between biological and psychological features of mental disorders (Borsboom et al., 2018) and is, as highlighted by a recent review, also supported by a large corpus of empirical results (Robinaugh et al., 2020). Similarly, Peter Zachar's description of psychiatric disorders

Starke, G., Elger, B. S., & De Clercq, E. (2023). Machine learning and its impact on psychiatric nosology. Findings from a qualitative study among German and Swiss experts. *Philosophy and the Mind Sciences*, *4*, 4. https://doi.org/10.33735/phimisci.2023.9435

as "imperfect communities" represents an influential nonessentialist approach, describing mental disorders as clusters of symptoms that are historically grown and reflect pragmatical interest (Zachar, 2014, pp. 115–136).

Definitions of psychiatric disorders that are based on neuroscience more frequently represent essentialist views, i.e., theories that take reality to be mind-independent and attempt to carve nature at its joints. Such definitions are, for instance, rooted in an understanding of psychiatric disorders as brain disorders (Insel & Cuthbert, 2015) or point to harmful impairment of natural functioning (Faucher & Forest, 2021; Horwitz & Wakefield, 2007). However, it is crucial to distinguish in this context between the ontological question what psychiatric disorders are, and the more practical question how to classify them. As Zachar has noted with regard to the Diagnostic and Statistical Manual of Mental Disorders (DSM), "a careful reading of the introduction to both the DSM-IV and the DSM-5 indicates that alongside the de facto essentialism about the nature of psychiatric disorders there is also a de facto nonessentialism about classification" (Zachar, 2014, p. 128). Distinguishing between viewpoints about the nature of psychiatric disorders and beliefs about classificatory systems, which in turn fulfil multiple functions (Reed et al., 2011), is therefore important to understand how neuroscience-based essentialist views can be seen as compatible with a dimensional approach to psychiatric classification, as endorsed in the DSM-5 (Regier et al., 2013).

Surprisingly, despite the individual prominence of each topic in recent literature, the impact of ML techniques and in particular of DL on psychiatric nosology has so far not received much systematic consideration. Many authors have hinted at the potential of DL for nosology (Brunn et al., 2020; Durstewitz et al., 2019) and some have called for increased attention to the conceptualization of psychiatric disorders in the context of AI-based methods (Winter et al., 2021). Yet, the relation of a DL-based clustering of disorder subtypes to the competing models of psychiatric disorders remains to be investigated in depth. An exception to this is the paper by Wanja Wiese and Karl Friston, who have provided an insightful philosophical discussion of the transformative effects of computational methods on psychiatric nosology and warned against an unintended marginalisation of subjective experience (Wiese & Friston, 2021).

To gain a better understanding whether this worry is shared by other researchers from neuroscience and psychiatry and in which ways ML may have an impact on psychiatric nosology, it seems crucial to explore the explicit and implicit knowledge of scholars in the field (Döringer, 2021). Reporting the findings from semi-structured qualitative interviews with researchers from Germany and Switzerland, we present the opinions and attitudes of experts in computational psychiatry with regard to the impact of ML on psychiatric nosology. To our knowledge, while there have been some qualitative findings investigating the attitudes of psychiatrists and psychologists towards AI methods (Blease et al., 2020, 2021), this is the first interview-based study looking at nosology in particular. In addition, we relate our findings to debates from the philosophy of science, arguing for a non-reductionist view of mental disorders that allows for methodological pluralism. Based on these considerations, we point to further lines of research that seem warranted.

2 Methods

We recruited Swiss and German experts on the use of ML in psychiatry. Participants were identified by systematically searching on the websites of psychiatric university hospitals in Switzerland and Germany for clinicians and researchers engaging with artificial intelligence or machine learning. Within our narrow recruitment criteria, we aimed to include as diverse a sample as feasible, with view to the respective career stages and gender. Once identified, we invited experts to participate in our study via e-mail and sent a reminder after a week in case we did not receive a response. We limited the field of experts to scholars who held at least a doctorate in a relevant field, i.e., medicine, neuroscience, or computer science.

The interviews took place between April 2020 and July 2021 and were conducted by the first author, a German physician (MD) with an additional degree in philosophy, research and working experience in neuroscience and psychiatry, and basic knowledge of programming and ML. The interviews formed part of his PhD in bioethics, which included intensive training and supervision in qualitative data collection. To fine-tune the interview guide and review the interview quality, the first three interviews with experts served as pilots. Based on their transcripts, GS and EDC revised the interview guide critically, resulting in minor changes.

Due to the constraints of the pandemic, interviews were conducted via phone (10) or online video call (5), in German (13) or English (2), depending on the participants' individual preferences. Interviews lasted between 25 and 66 minutes. All interviews were transcribed verbatim by the first author, with help from a medical master student (see acknowledgments). All quotes used for the purpose of this paper were translated by GS and checked by EDC. The interviewer was familiar with three of the participants prior to conducting the interviews, owing to earlier research activities.

To identify important themes relating to psychiatric nosology, we analysed the data from our sample using reflexive thematic analysis (Braun & Clarke, 2006, 2019). Individual codes were given to each segment of each transcribed interview, with one segment representing a unit of meaning, consisting of one or more sentences. Initially, the authors conducted the coding jointly for the first four interviews, supported by a master student (see acknowledgments). After agreeing on a coding tree structure, comprising themes and subthemes, the remaining transcripts were coded by the first author, using MaxQDA software. This data analysis accompanied the data collection, also to monitor data saturation, conceptualized as thematic redundancy indicated by recurrent coding (Given, 2015).

A full description of our study design, including the informed consent sheet and the interview guide, was submitted for review to the responsible research ethics committee (Ethikkommission Nordwest- und Zentralschweiz, EKNZ), prior to any data acquisition. The ethics committee determined that our project did not fall under restrictions that the Swiss legal framework imposes on research with human subjects and issued a statement of non-objection (Req-2019-00920). Notwithstanding this decision, we adhered to high ethical standards, by obtaining informed consent and by ensuring confidentiality and data security: (1) Prior to their participation in our study, we asked participants for their written informed consent, and confirmed this again orally at the beginning of the interview. (2) Furthermore, we omitted identifying information such as names and places already at the stage of transcribing, (3) and stored the data separately from identifying data on our university servers in Switzerland.

A detailed analysis of our main findings concerning the ethical dimension of using ML in psychiatry is provided elsewhere (Starke et al., 2022). In this manuscript, we focus on the impact of ML on psychiatric nosology, allowing for a more in-depth conceptual reflection. Questions from the interview guide that are relevant to the current manuscript are provided in Table 1.

- For which applications of machine learning do you see the greatest potential in future psychiatry?
  - For which particular clinical objectives?
  - For which psychiatric disorders?
- Are there, in your opinion, challenges of using medical machine learning that are specific to psychiatry?
- As you know, some authors argue that machine learning, and Deep Learning in particular, promise a way to divide psychiatric disorders objectively into natural types and thus solve the old problems of psychiatric nosology. Where would you stand on this?
- How should one best deal with cases of impaired judgement, for example when it comes to a potential program to recommend a particular antipsychotic medication during a psychotic episode?

Table 1: Relevant questions from the interview guide.

3 Results

Semi-structured interviews were conducted with 15 participants out of 26 invited experts (57,6%; 2 women and 13 men). Three experts declined due to time constraints, one did not consider themself an expert, and four did not reply. We stopped recruiting additional participants after reaching saturation on the main themes of our study, i.e., once participants reiterated ideas that had already been present in similar form in previously conducted interviews (Saunders et al., 2018). All participants held at least a doctorate (MD and/or PhD), covering career stages between postdoc and retired professor (mean years since doctorate 14.4a, sd ±10.8) and were affiliated with German or Swiss academic institutions pursuing research on psychiatric disease. Ten participants were licensed physicians, five had degrees in psychology or neuroscience, and eight participants reported additional multidisciplinary training in mathematics, physics, engineering, and philosophy. Analysing our interviews with particular focus on nosology, we related our findings to two large themes, namely (1) the *possibility* of using ML for defining psychiatric classifications, and (2) the *desirability* of employing ML to design psychiatric classificatory systems.

3.1 On the possibility of using ML for defining psychiatric classifications

With view to psychiatric classification, the desire to improve current systems was shared unanimously among the interviewees. However, participants' views on the *possibility* of using ML for this purpose diverged. Some participants embraced an optimist outlook, hoping for new classifications through the use of DL on large data samples comprising biological and behavioural data as well as self-reported symptoms:

> "I think that if we manage to put together large amounts of data, which you can do with these [neural] networks, that we will then also have another possibility to find groups, subgroups in psychiatry, or perhaps new forms of groupings. I believe that this requires a lot of data that we do not yet have formatted accordingly [...], but in principle I think it is possible, yes." (P2)

Also others considered ML as particularly useful for psychiatric nosology since it could contribute to mapping different features of psychiatric disorders in a higher dimensional space, taking into account the complex and contingent forms of mental disorders, shaped by history, culture, and language. Some participants were therefore optimistic concerning ML, if it incorporated a turn towards a dimensional diagnostic system.

> "I think we would have to find a dimensional system to describe psychiatric illnesses in the best possible way, similar to the way we describe personality. [...] Instead of dividing people somehow into diagnostic classes, one could simply describe them with a profile on these different dimensions. And if you then have to decide somehow whether you should treat someone with antidepressants or something, then you could also define a cut-off on the dimension of depressiveness." (P13).

The majority of interviewees however regarded ML for nosological purposes more sceptically. Some experts insisted that if we were to aim at new classifications, we would need to move beyond mapping symptoms to specific biomarkers, and turn to the underlying mechanisms instead, rooted in neurobiology.

> "So, if you try to do that at the level of symptoms, I think it's hopeless. Because you know only too well that with prominent examples, – that a certain symptom can be caused by completely different neurological mechanisms. And that's why a parcellation or a delimitation of diseases can generally, in my view, not be done on the symptom level, but always only on the level of mechanisms and causes. All our claims are not at the level of data, but at the level of possible mechanisms that can explain the data we have observed." (P4)

At the same time, other sceptics frequently pointed to the lack of success in identifying univocal associations between neurobiological data and psychiatric disorders in research so far, even after decades searching for psychiatric biomarkers.

> "I am very suspicious, having worked in the field for quite a few years, as to whether it will really be possible – whether [machine learning] will prove so helpful to arrive at diagnostic classifications. That isn't possible at the moment because there are no unequivocal correlations, for example, between certain brain-structural changes and a diagnosis of some kind. You do not have this for a single disorder in psychiatry. You can't say, for example, frontal lobe grey matter reduction means someone suffers from depression. No: they might suffer from depression, or maybe schizophrenia and so on. There are no unequivocal correlations." (P1)

Some interviewees stressed the additional difficulty of arriving at suitable ML models, in light of the fact that current psychiatric diagnostic classifications are not built on biological observations but on the reported phenomenological symptoms of patients.

> "I mean, in psychiatry in general it is also a methodological problem. Because, as I said, the classifications are phenomenological, they have nothing to do with neurobiology, I think we still know far too little about it. And this whole psychiatric classification system has to do with that. That would have to be fundamentally questioned if we were to imagine a greater significance for AI." (P7)

One interviewee reasoned that our current classificatory approaches are reflected in the training data to a degree that makes it impossible to arrive at a new classificatory system.

> "That's where the dragon bites its own tail. [...]. At the end of the day, we feed our algorithms with pre-assumptions and pre-allocations. [...] And machine learning, which forms certain substructures through

> deep learning, so to speak, must always be mapped to the outcome at the end of the day, otherwise it can't be used. [...] This is why we will fail to introduce new psychiatric dimensions now. At the end of the day, [Deep Learning] may provide us with hypotheses, make us reconsider certain labels and, in particular, reconsider the response to medication in the context of our diagnosis. But I don't think machine learning itself will miraculously give us any true entities." (P5)

3.2 On the desirability of using ML for defining psychiatric classifications

The second recurring theme was whether it would be *desirable* to use ML to arrive at novel psychiatric classifications. Optimist stances emphasized the methodological benefits of an ML-based classificatory system, allowing for hypothesis-free or more objective approaches, whereas others delineated conditions which such approaches should respect.

In the view of optimists, ML could enable new ideas and move beyond existing hypotheses:

> "I have always been of the opinion, even before ML existed, that we need much more hypothesis-free thinking and not these prefabricated pigeonholes that we have in psychiatry. And that, in my opinion, is one of the great possibilities of such methods, that one can really recognise completely new associations, and perhaps also connections of symptoms, patterns of brain changes, patterns of other endocrine changes, patterns of causes, and thereby generate new causal ideas." (P15)

As a potential result of such hypothesis-free methods, several scholars named the ambition of moving beyond subjective symptoms and gaining a more objective model of psychiatric disorders through an automated approach.

> "Especially in psychiatry there is the problem that many symptoms are subjective and retrospective. This already plays a big and problematic role in clinical care but also in assessments. Because many things a patient says cannot be objectively affirmed or denied. It would be interesting if there were possibilities to have more objective access to the inner world of the patient. That would be of great importance for the patient." (P8)

More sceptical voices mentioned the danger that defining psychiatric disorders based on ML models could imply ignoring the history of psychiatry and may contribute to impoverishing the discipline as such.

> "What is not good is to postulate, as some authors do, and say: in 5 years we will have reached the point with computing power that we can simply put this 19th century thinking, schizophrenia, bipolar etc.,

> in the museum, and that' s it. I think that's wrong. And not because of the terms. You can abolish the terms if you like. I can also do psychiatry without the schizophrenia term, no problem. But behind the concept of schizophrenia there is a very rich tradition of thought. Key words: Jaspers, Kurt Schneider ... If all that were to be stirred away because it is old, I would consider that a substantial loss for the discipline." (P3)

Another objection to an ML-based nosology was raised by several experts who tied the desirability of a refined classificatory system to its clinical usefulness, providing a prognosis or predict therapeutic response for individual patients.

> "You can determine a lot after you have talked to the person for two minutes, because everything may already be clear. Or if you just see them walking down the corridor. This means that it is certainly not so much a question of finding a diagnosis and classification, but rather the important thing is to give a prognosis or a therapy response. I think these are the important areas of application." (P7)

On a related note, several clinicians also called for a focus on the subjective perspective of the individual patient when asked about the desirability of an ML-based classificatory system.

> "I am convinced that the diagnosis *itself* is not relevant. It's about how the person is doing; can I make them feel better? I don't need the diagnosis for that if I have a treatment right away. Diagnosis is just a vessel to get to treatment. If the biomarker says this person has depression, but the person laughs, can sleep well and says,"I am not depressed", then he is not depressed. I.e., the diagnosis is always in the eye of the beholder – what the psychiatrist defines, what the patient feels." (P9)

Another participant embedded their scepticism in a historical context, linking the history of different ML techniques and the history of modern biologically oriented psychopathology. Reflecting on long-standing failures to provide a biologically grounded classification of psychiatric disorders, they were convinced, that although helpful, ML could not resolve the problem of nosology and that investing too much hope in such a project might even be harmful, by leading to another AI winter.

> "If Kraepelin had had Deep Learning, he would have been using that to classify the patients. But he couldn't. So, he just classified them with his sorting cards and everything. And then, you know, k-means and clustering algorithms came up in 1958. It was the first – one of the first introductions of the techniques. And then by the 1960s and 1970s they were already using it for psychiatry. But it hasn't worked. And you know, it's just an overstatement that it will solve all the problems and

> define objective groups. We have been going after that for a hundred and something years, and it hasn't happened yet. It certainly may help. I am not denying that. [...] But saying that Deep Learning is going to solve all these problems is exactly like what happened in the 1960s, and then the first AI winter came after that, because the claims were so ridiculously inflated." (P14)

Finally, on a more clinical level, several experts reported concern that moving towards an ML-based classificatory, diagnostic system may also alter clinical symptoms. Given that the themes of delusions often mirror aspects of a particular age, these clinicians reasoned that such a shift would likely also result in an increase of ML-related delusions.

> "Paranoid experiences, delusions often reflect the times, the *zeitgeist*. In the past, delusions were often caused by religion. Since religion no longer plays such a role, at some point this idea of being bugged came up, or of being irradiated by rays, and now the delusional contents are changing more and more in the direction of the computer." (P1)
>
> "We very often see psychotic patients whose delusions have a lot to do with this topos, i.e., computers, artificial intelligence, who's listening to me, is there a CIA guy sitting around the corner and so on. And I could imagine that for this group of patients, for chronically psychotic people, it would [...] become an issue if psychiatry were to become more and more algorithmised and mechanised. Because that would somehow strengthen their suspicions, which they have due to their illness. In concrete terms, if I'm sitting here at my desk and the patient is sitting opposite of me and I have 10 computers on the table that are constantly printing out something and beeping, then you don't have to be schizophrenic to become a bit suspicious." (P3)

4 Discussion

The present study aimed to explore experts' attitudes on the role of ML for psychiatric nosology. To our knowledge, this is the first study that reports the viewpoints of researchers in the field on this topic. With regard to both the possibility and the desirability of using ML to define mental disorders and refine classificatory system, we found optimist and sceptical stances. In the following, we draw on our findings to argue in favour of a methodologically pluralist, non-reductive approach to psychiatric disorders. In particular, we highlight how engaging with conceptual theories such as the network theory of mental disorders could help to advance research in the field, and we show how the reflexive impact of ML-based diagnostics on patients' symptoms described by our interviewees further supports a non-reductionist approach if seen in the light of Hacking's notion of *human kinds*.

Concerning the possibility of employing ML methods to solve problems of psychiatric nosology, we found conflicting voices among our interviewees. Optimist stances were embraced by few scholars, pointing out potential benefits of using hypothesis-free, data-driven approaches. Yet, despite interviewing only experts pursuing research in the very field, the majority of interviewees questioned such promises on a methodological basis. They stressed that available data already mirror current nosological assumptions, leading to feedback effects that prevent advancing beyond current conceptual frameworks. They also referenced the historically poor track record of searching for clinically useful biomarkers in psychiatry as well as our incomplete understanding of causal connections between neurobiology and mental phenomena, between mind and brain.

This polyphony of our interviewees' positions constitutes one of the main findings of our study. The diverse stances mirror longstanding scholarly debates, for instance whether research in psychiatry should be data-driven or theory-driven (Huys et al., 2016; Itani & Rossignol, 2020) or how to bridge the gap between neurobiological mechanisms and phenomenological symptoms (Borsboom et al., 2018). The variety of positions also seemed to reflect fundamental metaphysical disagreement about the nature of mental disorders. Many of our interviewees seemed to implicitly endorse an understanding of psychiatric disorders as brain disorders that can and should be objectified, whereas others highlighted the limits of DL, stressing phenomenological and historically contingent aspects of mental disorders. Wiese and Friston (2021) have recently highlighted how research in computational psychiatry, while in theory metaphysically neutral, often tends to place its focus on brain function (Friston et al., 2014; Montague et al., 2012; Stephan & Mathys, 2014) and less on genetic mechanisms (Rødevand et al., 2021) or clinical predictors (Koutsouleris et al., 2021). Our sample seems therefore quite reflective of the nosological debates that have vexed psychiatry since its inception (Aftab & Ryznar, 2021), and to mirror questions how to conceptualise the relation between neurobiology and mental phenomena that remain unsolved for biological psychiatry (Walter, 2013).

While this result is already interesting in itself as an overview of current attitudes and opinions in the field, we believe that our findings can also inform the philosophical debate on using machine learning for psychiatric nosology. In particular, the various perspectives raised by the interviewed experts highlight the multi-faceted and complex way in which mental disorders present themselves, ranging from the biological and chemical to the social and phenomenological. If some form of unsupervised ML is supposed to advance research towards a more complete account of mental disorders, it would therefore need to integrate these varying levels of explanations. A helpful model for thinking about the integration of such levels has been proposed by Lena Kästner (2018) in the context of mechanistic explanations: Instead of conceptualizing different levels of an explanation in a hierarchical or layered manner, it may prove beneficial to our scientific understanding of complex phenomena if we assume a dimensional view of explanatory levels (Kästner, 2018). Such dimensions can account for the diverging epistemic perspectives of the involved research domains and preserve the respective richness of their descriptions, allowing for complementary and pluralist accounts (ibid.).

Appreciating and integrating diverging epistemic perspectives, as presented in this paper, seems also very well-suited for the analysis and conceptualisation of mental disorders: It helps to avoid forms of reductionism that promise overly simplistic explanations of psychiatric disorders but do not appreciate the complexity of the phenomenon. For as Ludwik Fleck provokingly admonished in his 1927 *Some Specific Features of the Medical Way of Thinking*, "the worse the physician the 'more logical' his therapy" (Fleck, 1986, p. 42). The worry expressed here, that in medical practice overly simple explanations are hardly a sign of an experienced clinician, resonates well with the opinions of the interviewed experts that put the benefit to the patient front and center. These positions are also in line with the comprehensive literature criticising psychiatric practice for its focus on assigning labels (Brinkmann, 2017; Callard et al., 2013) and with positions that favour more pragmatic definitions of mental disorders (Kendler et al., 2011; Zachar, 2014).

One proposed and much-discussed system of mental disorders that offers a non-reductionist view, accommodating different dimensions of explanations, is the symptom network theory (Borsboom, 2017; Borsboom et al., 2018; Oude Maatman, 2020). As mentioned in the introduction, this theory takes causally connected symptoms as its focal point, satisfying the call by practitioners to focus on clinically relevant features. At the same time, it allows for appreciating biological as much as social determinants of mental disorders by situating them in a complex network that can be described from different epistemic perspectives. Engaging with these philosophical debates will therefore also prove useful to empirical researchers, as it provides a framework for the integration of empirical research from different research domains, to make use of the "growing body of empirical research and move the field toward its fundamental aims of explaining, predicting, and controlling psychopathology" (Haslbeck et al., 2021).

Machine learning, and deep learning in particular, should therefore not be seen as a remedy in itself to the challenges of nosology, but rather as a computational tool that may support scientific progress by allowing an improved modelling of complexity, integrating vast amounts of different data types that represent different dimensions of a phenomenon. In this context, at least three caveats though seem crucial.

First, a diagnostic system based on ML should not be mistaken to provide an objective "view from nowhere", to borrow Nagel's phrase (1986). On the one hand, any computational model will be shaped by the type of data selected for its training, and by the context of their acquisition, as repeatedly stressed by our interviewees. In addition, insofar as computational psychiatry draws on a concept of *miscomputation*, it employs a value-laden and perspectival notion of normalcy for its explanations (Colombo, 2021). Also with the support of ML, it will therefore remain crucial to be mindful of the epistemic perspectives informing classificatory systems in psychiatry.

A second caveat concerns the limited possibility of arriving at causal structures with deep learning techniques. While DL may provide researchers with new hypotheses or inspiration through its ability to detect correlations in large datasets (Davies et al., 2021), it usually does *not* provide causal scientific explanations, with very few exceptions such as explicit causal modelling (Parascandolo et al., 2018). This constitutes an important difference to symptom network theory, which demands causal links between different nodes in the symptom network (Borsboom, 2017). Deep neural network models therefore only provide one step in the generation of scientific knowledge, offering "first steps to determining which causal mechanisms or dependency relations should be explored further" (Sullivan, 2022), or as P15 put it: first steps to "recognise completely new associations, and perhaps also connections of symptoms, patterns of brain changes, patterns of other endocrine changes, patterns of causes, and thereby generate new causal ideas."

A third caveat is that also with the use of ML, psychiatric classificatory systems will not carve nature at its joints but will remain dynamic and open to change. Evidence for this claim can be found in the anecdotal clinical reports of psychotic symptoms being shaped by the real or feared integration of ML into psychiatry that came up repeatedly in many of our interviews, despite not corresponding to any item in our interview guide. Assuming that these reports are not isolated concerns, this unintended impact of ML on psychiatric diagnostic seems to fit well with what Ian Hacking has described as the looping effect of human kinds where human classifications and their social environment are causally intertwined through feedback mechanisms (Hacking, 1999).

Hacking's work on natural and human kinds has informed the past decades of debate in psychiatric research. In Hacking's view, natural kinds are supposed to offer a unique taxonomy "that represents nature as it is, and reflects the network of causal laws" (Hacking, 1991, p. 111), whereas human kinds are the subject of the social sciences, providing "classifications that could be used to formulate general truths about people" (Hacking, 1996, p. 352). While the debate about this distinction's conceptual bearings is vast and controversial (Bird & Tobin, 2022; Cooper, 2004; Craver, 2009; Tsou, 2007; Van Riel, 2016), some authors have also used it to design empirical research, investigating for instance the way in which young adolescents interact and transform psychiatric concepts (Lindholm & Wickström, 2020). Here, our point is much more modest though: If the use of an ML-based diagnostic regime does indeed shape the symptom of patients, and if said symptoms are used as training data to the diagnostic model, this would imply the need to regularly update the classificatory model. This observation alone may therefore be seen as a reason to not harbour a machine-learning based "aspiration to automatically segregate brain disorders into natural kinds" (Bzdok & Meyer-Lindenberg, 2018).

Our study has several limitations. Since our purposive sampling was highly targeted on a specific research field within psychiatry in Germany and Switzerland, our results are not representative, neither for psychiatry in general nor for other cultural contexts. As is the case for all qualitative research, our results are therefore not generalizable. For this reason and to safeguard the anonymity of our participants, we can therefore not provide insights into quantifiable relations between, e.g., the experts' years of experience or their success in publishing and their attitudes towards ML, but believe that such inquiry would constitute a valuable route for future research. In addition, the close involvement of the interviewer in the field as well as his medical background may have influenced his interactions with the interviewees. Yet, since our study aimed at exploring different facets of an emerging research field, not at representative descriptions, we believe that these limitations do not diminish the value of our findings.

5 Conclusion

This study provides the first qualitative insights into the impact of ML on psychiatric nosology. It highlights how ML and DL in particular does seemingly not provide a solution to problems of defining psychiatric disorders but instead mirrors existing disagreements. Our findings should therefore be read as an exhortation to scholars working in the field of computational psychiatry to engage more deeply with philosophical debates and bridge the gaps between research employing DL and the philosophy of mind. Doing so may support the development of non-reductionist research programs that appreciate the complexity of mental disorders by integrating empirical findings from different research domains.

Acknowledgments
First and foremost, we would like to thank our interviewees for their time and willingness to participate in our study, despite their multiple obligations during the pandemic. We are also grateful for valuable comments on an earlier draft by Christopher Poppe and two anonymous reviewers and indebted to Benedikt Schmidt for help with transcribing the interviews. GS would also like to acknowledge funding from the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation's Grant Number: 32NE30_199436) to work on the revision of this paper.

References

Aftab, A., & Ryznar, E. (2021). Conceptual and historical evolution of psychiatric nosology. *International Review of Psychiatry*, *33*(5), 486–499. https://doi.org/10.1080/09540261.2020.1828306

Bird, A., & Tobin, E. (2022). Natural kinds. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy, spring 2022 ed.* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2022/entries/natural-kinds/

Blease, C., Kharko, A., Annoni, M., Gaab, J., & Locher, C. (2021). Machine learning in clinical psychology and psychotherapy education: A mixed methods pilot survey of postgraduate students at a swiss university. *Frontiers in Public Health*, *9*, 623088. https://doi.org/10.3389/fpubh.2021.623088

Blease, C., Locher, C., Leon-Carlyle, M., & Doraiswamy, M. (2020). Artificial intelligence and the future of psychiatry: Qualitative findings from a global physician survey. *Digital Health*, *6*, 2055207620968355. https://doi.org/10.1177/2055207620968355

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D., Cramer, A., & Kalis, A. (2018). Brain disorders? Not really… Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 1–54. https://doi.org/10.1017/S0140525X17002266

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Brinkmann, S. (2017). Perspectives on diagnosed suffering. *Nordic Psychology*, *69*(1), 1–4. https://doi.org/10.1080/19012276.2016.1270404

Brunn, M., Diefenbacher, A., Courtet, P., & Genieys, W. (2020). The future is knocking: How artificial intelligence will fundamentally change psychiatry. *Academic Psychiatry*, *44*, 461–466. https://doi.org/10.1007/s40596-020-01243-8

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Callard, F., Bracken, P., David, A. S., & Sartorius, N. (2013). Has psychiatric diagnosis labelled rather than enabled patients? *BMJ*, *347*, f4312. https://doi.org/10.1136/bmj.f4312

Chang, M., Womer, F. Y., Gong, X., Chen, X., Tang, L., Feng, R., Dong, S., Duan, J., Chen, Y., & Zhang, R. (2021). Identifying and validating subtypes within major psychiatric disorders based on frontal–posterior functional imbalance via deep learning. *Molecular Psychiatry*, *26*, 2991–3002. https://doi.org/10.1038/s41380-020-00892-3

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

Colombo, M. (2021). (Mis)computation in computational psychiatry. In F. Calzavarini & M. Viola (Eds.), *Neural mechanisms. Studies in brain and mind* (pp. 427–448). Springer. https://doi.org/10.1007/978-3-030-54092-0_18

Cooper, R. (2004). Why Hacking is wrong about human kinds. *British Journal for the Philosophy of Science*, *55*(1), 73–85.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594. https://doi.org/10.1080/09515080903238930

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*(1), 1–8. https://doi.org/10.1186/1741-7015-11-126

Dattaro, L. (2021). *Green light for diagnostic autism app raises questions, concerns*. https://www.spectrumnews.org/news/green-light-for-diagnostic-autism-app-raises-questions-concerns/

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., & Juhász, A. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, *600*(7887), 70–74. https://doi.org/10.1038/s41586-021-04086-x

Döringer, S. (2021). "The problem-centred expert interview." Combining qualitative interviewing approaches for investigating implicit expert knowledge. *International Journal of Social Research Methodology*, *24*(3), 265–278. https://doi.org/10.1080/13645579.2020.1766777

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

Eitel, F., Schulz, M.-A., Seiler, M., Walter, H., & Ritter, K. (2021). Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. *Experimental Neurology*, *339*, 113608. https://doi.org/10.1016/j.expneurol.2021.113608

Faucher, L., & Forest, D. (2021). *Defining mental disorder: Jerome Wakefield and his critics*. MIT Press.

Fleck, L. (1986). Some specific features of the medical way of thinking [1927]. In T. S. Robert S. Cohen (Ed.), *Cognition and fact: Materials on Ludwik Fleck* (pp. 39–46). Springer Netherlands. https://doi.org/10.1007/978-94-009-4498-5_2

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158. https://doi.org/10.1016/S2215-0366(14)70275-5

Given, L. M. (2015). *100 questions (and answers) about qualitative research*. SAGE publications.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *61*(1/2), 109–126. https://doi.org/10.1007/BF00385836

Hacking, I. (1996). The looping effects of human kinds. In P. Sperber D. (Ed.), *Causal cognition: A multi-disciplinary debate* (pp. 351–383). Oxford Academic. https://doi.org/10.1093/acprof:oso/9780198524021.003.0012

Hacking, I. (1999). *The social construction of what?* Harvard University Press.

Haslbeck, J., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*. https://doi.org/10.1037/met0000303

Horwitz, A. V., & Wakefield, J. C. (2007). *The loss of sadness: How psychiatry transformed normal sorrow into depressive disorder*. Oxford University Press.

Starke, G., Elger, B. S., & De Clercq, E. (2023). Machine learning and its impact on psychiatric nosology. Findings from a qualitative study among German and Swiss experts. *Philosophy and the Mind Sciences*, *4*, 4. https://doi.org/10.33735/phimisci.2023.9435

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. https://doi.org/10.1038/nn.4238

Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science*, *348*(6234), 499–500. https://doi.org/10.1126/science.aab2358

Itani, S., & Rossignol, M. (2020). At the crossroads between psychiatry and machine learning: Insights into paradigms and challenges for clinical applicability. *Frontiers in Psychiatry*, *11*, 1029. https://doi.org/10.3389/fpsyt.2020.552262

Jacobson, N. C., & Bhattacharya, S. (2022). Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. *Behaviour Research and Therapy*, *149*, 104013. https://doi.org/10.1016/j.brat.2021.104013

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, *22*(1), 393–415. https://doi.org/10.1093/bib/bbz170

Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, *68*, 68–79. https://doi.org/10.1016/j.shpsa.2018.01.011

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., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150. https://doi.org/10.1017/S0033291710001844

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., & Weiske, J. (2021). Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry*, *78*(2), 195–209. https://doi.org/10.1001/jamapsychiatry.2020.3604

Lindholm, S. K., & Wickström, A. (2020). "Looping effects" related to young people's mental health: How young people transform the meaning of psychiatric concepts. *Global Studies of Childhood*, *10*(1), 26–38. https://doi.org/10.1177/2043610619890058

Lui, J. H., Marcus, D. K., & Barry, C. T. (2017). Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. *Professional Psychology: Research and Practice*, *48*(3), 199–210. https://doi.org/10.1037/pro0000122

Martinez-Martin, N., & Kreitmair, K. (2018). Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent. *JMIR Mental Health*, *5*(2), e32. https://doi.org/10.2196/mental.9423

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/j.tics.2011.11.018

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Oude Maatman, F. (2020). Reformulating the network theory of mental disorders: Folk psychology as a factor, not a fact. *Theory & Psychology*, *30*(5), 703–722. https://doi.org/10.1177/0959354320921464

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018). *Learning independent causal mechanisms*. *80*, 4036–4044. https://doi.org/10.48550/arXiv.1712.00961

Quaak, M., Mortel, L. van de, Thomas, R. M., & Wingen, G. van. (2021). Deep learning applications for the classification of psychiatric disorders using neuroimaging data: Systematic review and meta-analysis. *NeuroImage: Clinical*, *30*, 102584. https://doi.org/10.1016/j.nicl.2021.102584

Reed, G. M., Correia, J. M., Esparza, P., Saxena, S., & Maj, M. (2011). The WPA-WHO global survey of psychiatrists' attitudes towards mental disorders classification. *World Psychiatry*, *10*(2), 118–131. https://doi.org/10.1002/j.2051-5545.2011.tb00034.x

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The DSM-5: Classification and criteria changes. *World Psychiatry*, *12*(2), 92–98. https://doi.org/10.1002/wps.20050

Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Rødevand, L., Bahrami, S., Frei, O., Lin, A., Gani, O., Shadrin, A., Smeland, O. B., O'Connell, K. S., Elvsåshagen, T., & Winterton, A. (2021). Polygenic overlap and shared genetic loci between loneliness, severe mental disorders, and cardiovascular disease risk factors suggest shared molecular mechanisms. *Translational Psychiatry*, *11*(1), 1–11. https://doi.org/10.1038/s41398-020-01142-4

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., & Steyerberg, E. W. (2021). Implementing precision psychiatry: A systematic review of individualized prediction models for clinical practice. *Schizophrenia Bulletin*, *47*(2), 284–297. https://doi.org/10.1093/schbul/sbaa120

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, *52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8

Schulz, M.-A., Chapman-Rounds, M., Verma, M., Bzdok, D., & Georgatzis, K. (2020). Inferring disease subtypes from clusters in explanation space. *Scientific Reports*, *10*(1), 1–6. https://doi.org/10.1038/s41598-020-68858-7

Starke, G., Schmidt, B., De Clercq, E., & Elger, B. S. (2022). Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00177-1

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92. https://doi.org/10.1016/j.conb.2013.12.007

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, *73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Tsou, J. Y. (2007). Hacking on the looping effects of psychiatric classifications: What is an interactive and indifferent kind? *International Studies in the Philosophy of Science*, *21*(3), 329–344. https://doi.org/10.1080/02698590701589601

Van Riel, R. (2016). What is constructionism in psychiatry? From social causes to psychiatric classification. *Frontiers in Psychiatry*, *7*, 57. https://doi.org/10.3389/fpsyt.2016.00057

Walter, H. (2013). The third wave of biological psychiatry. *Frontiers in Psychology*, *4*, 582. https://doi.org/10.3389/fpsyg.2013.00582

Walter, M., Alizadeh, S., Jamalabadi, H., Lueken, U., Dannlowski, U., Walter, H., Olbrich, S., Colic, L., Kambeitz, J., Koutsouleris, N., Hahn, T., & Dwyer, D. B. (2019). Translational machine learning for psychiatric neuroimaging. *Progress in Neuropsychopharmacology & Biological Psychiatry*, *91*, 113–121. https://doi.org/10.1016/j.pnpbp.2018.09.014

Wiese, W., & Friston, K. J. (2021). AI ethics in computational psychiatry: From the neuroscience of consciousness to the ethics of consciousness. *Behavioural Brain Research*, 113704. https://doi.org/10.1016/j.bbr.2021.113704

Winter, N. R., Cearns, M., Clark, S. R., Leenings, R., Dannlowski, U., Baune, B. T., & Hahn, T. (2021). From multivariate methods to an AI ecosystem. *Molecular Psychiatry*, *26*, 6116–6120. https://doi.org/10.1038/s41380-021-01116-y

Zachar, P. (2014). *A metaphysics of psychopathology.* MIT Press.