Trusting Black-Box Algorithms?
Ethical Challenges for Biomedical Machine Learning

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Georg Michael Joachim Starke

2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Bernice Elger
Dr. Eva De Clercq
Prof. Dr. Volker Roth
Prof. Dr. Georg Marckmann

Basel, 22. Februar 2022

_____
Prof. Dr. Marcel Mayor
Dekan

Real objects are always parts of institutions, trembling in their mixed status as mediators, mobilizing faraway lands and people, ready to become people or things, not knowing if they are composed of one or of many, of a black box counting for one or of a labyrinth of multitudes. And this is why the philosophy of technology cannot go very far: an object is a subject that only sociology can study – a sociology, in any case, that is prepared to deal with nonhuman as well as human actants.

Bruno Latour, *On Technical Mediation* (1994), p. 46

Table of contents

Acknowledgments

Before all else, I owe thanks to the many who made this thesis possible.

To my supervisors, for their invaluable intellectual support and personal encouragement throughout the past years. I thank Prof. Bernice Elger for her trust and professional guidance, offering me the opportunity to develop a project based on my own interests and providing the funding to complete this research. I am indebted to Dr. Eva De Clercq for her essential advice and often spontaneous availability for philosophical discussion or support in empirical methods, allowing me to delve into empirical bioethics. My sincere thanks to Prof. Volker Roth, for his interdisciplinary interest and for encouraging and enabling me to build my research on the necessary fundamental knowledge in computer science, and to Prof. Georg Marckmann for his willingness to evaluate my thesis, despite his busy schedule and professional obligations related to the pandemic.

I am immensely grateful to my colleagues and friends in Basel. To Andrea Martani, Bettina Zimmermann, and Christopher Poppe for everything from co-authoring and proof-reading papers to cooling swims in the Rhine. To Maddalena Favaretto, Lester Geneviève, Giorgia Lorenzini and Laura Arbelaez Ossa, for constructive feedback on my work and inspiring coffee-breaks, whether online and offline. To Prof. Fabrice Jotterand, for his advice in the choice of topic. To Prof. Markus Wild I am indebted for the intellectual and personal joy of attending his doctoral colloquia in the philosophy department. Prof. Thomas Vetter and Dennis Madsen provided invaluable support in learning the basics of Python and pattern recognition. I am also very grateful to PD Tenzin Wangmo for familiarizing me with empirical methods and for help in all matters organizational.

I owe thanks to everyone involved in the project, most of all to the experts who agreed to take part in the interview study, often despite additional Covid-related clinical obligations. To Benedikt Schmidt for his help with the transcription of the interviews. To my colleagues at other universities, especially Felix Gille, Thomas Grote, Pim Haselager, Philipp Kellmeyer, Stuart McLennan, Emilian Mihailovich, and Andreas Wolkenstein for valuable support at various stages during this thesis. I also thank the students I had the privilege of teaching, on AI Ethics in spring 2020, and on Ethics of Medical AI in fall 2021, who brought in new perspectives challenging me from a variety of backgrounds in humanities, medicine, and natural sciences.

Several institutions have provided invaluable support for this thesis. I am grateful to the European Association of Centers for Medical Ethics for encouraging me with the award of the Paul Schotsmans Prize, and to EUCOR, for funding a cross-border workshop on trustworthy AI in medicine. My sincere thanks also go to Fondation Brocher and their staff for providing a most inspiring research environment in Hermance in summer 2021 for working on this thesis. I am also grateful for the institutional support I have received from the University of Basel, especially the travel grants for conference attendance.

To Cécile, my ardent gratitude for emotional, moral, and intellectual support, for patience with my night-time writing, and the joy of our life together. To Bernhard, for friendship and encouragement, from joint writing sessions to hazardous cycling trips. I thank Niklas for help with higher algebra, Torben for helpful discussions on vacations, and my brother Philipp for critical proof-reading. To my parents, without whom this thesis would not have been possible.

Summary / Zusammenfassung

Summary

Based at the intersection of AI ethics and bioethics, this cumulative doctoral thesis investigates the question if and under which conditions we can and should trust black box algorithms used for medical purposes, with a specific focus on psychiatry. To do so, it sheds light on epistemic and ethical questions arising from opaque machine learning techniques in eight independent, peer-reviewed papers that form the core of this thesis and reflect its two-pronged approach: the first four chapters investigate general ethical questions of trust and trustworthiness of medical machine learning, driven by considerations from philosophy and science and technology studies (chapters 3-6), while the remaining four chapters relate abstract theoretical considerations to particular applications of machine learning in psychiatry, drawing also on empirical methods (chapters 7-10).

The stage is set in **chapter 1**, providing a brief introduction to the topic, and **chapter 2,** which introduces the theoretical and methodological framework of the thesis, embracing an integrated approach to empirically informed bioethical deliberation.

The general part of the thesis starts with **chapter 3,** which defends the notion of trust in medical machine learning against recent criticism and suggests a novel, dimensional model of trust in the spirit of Daniel Dennett. The following three chapters (4-6) investigate properties of machine learning-based appliances that make them trustworthy. **Chapter 4** scrutinizes algorithmic fairness through the lens of William James' pragmatist theory of truth. **Chapter 5** investigates the possibility of explaining and understanding medical machine learning, drawing on Karl Jaspers' *Psychopathology*. **Chapter 6** argues with Onora O'Neill why transparency as mere

disclosure it too little to generate trust in medical machine learning and suggests embracing an approach of intelligent openness instead.

The following four chapters aim to root these conceptual considerations in practice, by looking at specific applications of ML in psychiatry and neuroscience, and by engaging with relevant stakeholders through semi-structured interviews. To provide an insight into the challenges posed by machine learning in mental health, **chapter 7** systematizes ethical questions that arise from computational methods employed to diagnose, treat, and predict schizophrenia, following the principlist framework of Tom Beauchamp and James Childress. Checking these considerations against the attitudes of researchers in the field, **chapter 8** provides a unique contribution to the existing literature insofar as it is the first article that examines attitudes and expectations of experts on psychiatric machine learning towards ethical questions, drawing on a sample from Germany and Switzerland. **Chapter 9** examines these empirical findings further, exploring the impact of machine learning on psychiatric nosology. Finally, **chapter 10** gives an outlook to the future by addressing necessary changes in the training of junior doctors, arguing for the ongoing importance of an education informed by historical reflection.

Chapter 11 completes the dissertation, summarizing and discussing the different findings in light of each other. It also acknowledges its limitations and provides suggestions for further research.

Zusammenfassung

An der Schnittstelle von KI-Ethik und Bioethik untersucht diese kumulative Doktorarbeit die Frage, ob und unter welchen Bedingungen wir Black-Box-Algorithmen für medizinische Zwecke vertrauen können, mit besonderem Fokus auf Anwendungen in der Psychiatrie. Zu diesem Zweck werden epistemische und ethische Fragen, die sich durch die technische Opazität des maschinellen Lernens ergeben, in acht unabhängigen, begutachteten Beiträgen beleuchtet, die die beiden Hauptteile dieser Arbeit bilden und ihren zweigleisigen Ansatz widerspiegeln: (1) einen allgemeinen Teil, der sich auf Überlegungen aus der Philosophie und den Wissenschafts- und Technologiestudien stützt (Kapitel 3-6), und (2) einen spezifischen Teil, der die abstrakte Theorie mit konkreten Anwendungen des maschinellen Lernens in der Psychiatrie in Beziehung setzt (Kapitel 7-10).

Nach einer Einführung ins Thema in **Kapitel 1** stellt **Kapitel 2** den konzeptuellen Rahmen und die Methodologie der Arbeit vor, die ihre bioethischen Überlegungen auch auf empirische Untersuchungen stützt.

Der erste, allgemeine Teil der Arbeit beginnt mit **Kapitel 3**, welches den Begriff des Vertrauens in medizinisches maschinelles Lernen gegen jüngere Kritik verteidigt und ein neues, dimensionales Modell von Vertrauen im Geiste von Daniel Dennett vorschlägt. Daran anschliessend werden in den folgenden drei Kapiteln Eigenschaften untersucht, die für die Vertrauenswürdigkeit von Anwendungen maschinellen Lernens entscheidend sind. So beleuchtet **Kapitel 4** die Fairness von Algorithmen aus Perspektive der pragmatistischen Wahrheitstheorie von William James, während **Kapitel 5** mit Karl Jaspers' Psychopathologie auslotet, wie man medizinisches

maschinelles Lernen erklären und verstehen kann. **Kapitel 6** beschliesst den Abschnitt und argumentiert mit Onora O'Neill, warum Transparenz im Sinne einer blossen Offenlegung zu wenig ist, um Vertrauen in maschinelles Lernen in der Medizin zu generieren. Alternativ wird ein Modell intelligenter Offenheit vorgestellt.

Der zweite Teil der Arbeit zielt darauf ab, diese konzeptionellen Überlegungen in der Praxis zu verankern. Hierzu werden konkrete medizinische Anwendungen maschinellen Lernens untersucht und die Ergebnisse einer Interviewstudie vorgestellt. **Kapitel 7** systematisiert ethische Fragen, die sich aus der computergestützten Diagnose, Behandlung und Vorhersage von Schizophrenie ergeben, um einen Einblick in die Herausforderungen biomedizinischen maschinellen Lernens im Bereich der Psychiatrie zu geben. Als Referenzpunkt dienen hierbei die Prinzipien der Bioethik von Beauchamp und Childress. **Kapitel 8** stellt diese Überlegungen in Bezug zu den Einstellungen von Forscher*innen auf dem Gebiet. Dieser Beitrag ist die erste qualitative Arbeit, die die ethischen Einstellungen und Meinungen von Expert*innen für psychiatrisches maschinelles Lernen untersucht. **Kapitel 9** erörtert die Bedeutung maschinellen Lernens für die psychiatrische Krankheitslehre. **Kapitel 10** rundet den zweiten Teil mit einem Ausblick auf Curricula angehender Fachkräfte in der Psychiatrie ab und plädiert für die weiterhin grosse Relevanz der Psychiatriegeschichte, um eine ethische und verantwortungsvolle Implementierung maschinellen Lernens in der Klinik zu ermöglichen.

Kapitel 11 beschliesst die Dissertation, indem es die verschiedenen Ergebnisse der zwei Teile zusammenfasst und im Lichte der jeweils anderen diskutiert.

Chapter 1: Background and Rationale

1.1 Human trust and trustworthy machines

Trust is fundamental to human life. But are we justified to trust machines, let alone opaque machines used in medicine? It was this initial question that sparked an intellectual journey partially reflected in the eight publications collected as chapters here. In light of recent advancements in the field of Artificial Intelligence (AI) and in particular machine learning (ML) techniques employed in the medical domain, the question of their trustworthiness is more pressing than ever. Techniques such as support-vector machines (SVM), k-nearest neighbours (k-NN) algorithms and, in particular, deep learning (DL) applied on health-related data promise paradigm-shifting advances (Challen et al., 2019; Darcy, Louie, & Roberts, 2016; Esteva et al., 2019; Hinton, 2018; Topol, 2019b). More accurate and efficient diagnostic tools, personalised therapeutic regimes as well as prognostic or predictive measures seem bound to impact the treatments of patients, leading to what some call an age of "Deep Medicine" (Topol, 2019a).

While many authors have voiced general ethical concerns with view to this development (Char, Shah, & Magnus, 2018; Vayena, Blasimme, & Cohen, 2018) – many of which are, in fact, not new at all (Marckmann, 2003) – in-depth ethical analysis of modern AI driven by DL integrated into clinical care is still in its infancy. The three years in which this thesis was written have seen many advances in the ethical literature scrutinizing medical ML, focussing in particular on topics that are widely discussed in AI ethics such as fairness or explainability (Char, Abramoff, & Feudtner, 2020; London, 2019; Mittelstadt, Russell, & Wachter, 2019). As has become increasingly clear, these ethical questions can, however, not be discussed without proper philosophical reflection, examining the

epistemic conditions of our interaction with these programs (Grote & Berens, 2020; Sullivan, 2022).

Taking their cue from the *Ethics Guidelines for Trustworthy AI* published by the EU High Level Expert Group on Artificial Intelligence in 2019, the papers in this thesis shed light on the meaning of trust in the context of medical AI, and on important conditions for trustworthiness such as fairness, transparency, and explainability. To situate the abstract considerations at the practical intersection of computer science and medicine, this thesis embraces a two-pronged approach, mirrored in two parts. A more general part, comprising the chapters three to six, provides conceptual clarification and normative reasoning informed by literature from philosophy and science and technology studies. To relate these general theoretical considerations to real-life problems, the second part looks closely at particular medical applications of ML, supported by qualitative empirical research.

Before delving into specifics of human trust in trustworthy medical ML, this chapter provides a brief overview sketching the scientific background and rationale of this doctoral thesis. This will be done in three steps. The first one introduces the investigated object, i.e. opaque ML models, providing a brief overview of the terminology and a small glimpse at the architecture of DL in particular. The second step highlights the epistemic challenges posed by such opaque ML, commonly described as "black boxes". Advancing to questions of ethics, the third step motivates the specific angle chosen here, by discussing the role of trust and trustworthiness in current academic and regulatory debates about opaque ML in medicine. I conclude by giving an outlook on the scope and structure of the thesis.

1.2 Modern artificial intelligence: between success and hype

The ongoing integration of AI into everyday life has had a great impact on human life and will further reshape many aspects of society. Given the successes of modern ML, driven largely by the availability of big datasets and the advent of DL based on Artificial Neural Networks (ANNs), and the public hype revolving around the topic, recent literature on AI abounds, as even a short stroll through one's favourite bookstore will undoubtedly demonstrate. Whether it's Kazuo Ishiguro's latest novel *Klara and the Sun* (2021) or Ian McEwan *Machines Like Me* (2019), whether Kate Crawford's pointed social critique *Atlas of AI* (2021) or the similarly titled *Atlas of Anomalous AI* (2020) providing an artistic take on AI in the tradition of Aby Warburg, even offline and in print the topic seems unavoidable.

The field of AI Ethics has also seen an enormous proliferation of articles, conferences, journals, and textbooks dedicated to the ethical challenges of AI. Much of this development falls in the narrow timeframe of the three years during which this thesis was written. From Mark Coeckelbergh's *AI Ethics* (2020), Sven Nyholm's *Humans and Robots* (2020) or Julian Nida-Rümelin's *Digital Humanism* (2018) to Stuart Russell's *Human Compatible* (2019) and Erik Larson's *Myth of Artificial Intelligence* (2021), to name just a few, comprehensive and insightful books on the topic are numerous. To show what this thesis adds to these larger existing debates, a few points of clarification are in order, delineating the scope of the research presented here.

1.2.1 Artificial intelligence and machine learning

First, it is imperative to provide a short overview over the used terminology, and to distinguish between three terms that are frequently conflated in public discourses:

artificial intelligence (AI), machine learning (ML), and deep learning (DL).[1] As depicted in figure 1, among these three, AI serves as an umbrella term that is often taken as capturing all attempts to create artificial entities that think or act like humans. Notions of AI in this sense are dominant in the humanities (Dennett & Chalmers, 2019), and are also reflected in Alan Turing's famous test evaluating an AI based on an observer's ability to distinguish between human and machine (Turing, 1950).[2] However, such definitions come with the significant problem that they require, to some extent, definitions of human thought or human agency. Sidestepping this challenge, more technically oriented approaches therefore often attempt to define AI as artificial agents that "act[…] so as to achieve the best outcome or, when there is uncertainty, the best expected outcome" as measured by a predefined objective (Russell & Norvig, 2021, p. 22).

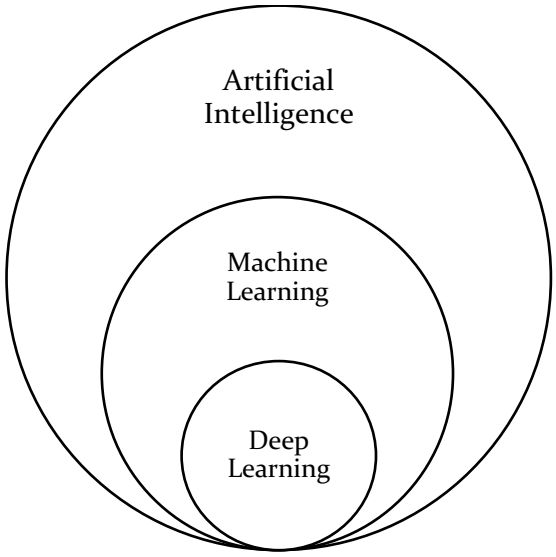

Fig. 1.1: Venn-Diagram representing the relation of AI, ML, and DL. Adapted from (Goodfellow, Bengio, & Courville, 2016, p. 9)

[1] In the following chapters, the terms are used at times interchangeably. This is largely owed to different terminological preferences in the different scientific communities and, at times, respective calls for papers. However, the focus of investigation are always opaque machine learning methods, as described below, unless explicitly noted otherwise.

[2] The Nobel laureate Herbert Simon's (in)famous prediction from 1956 points in the same direction, claiming that within 20 years "machines will be capable of doing any work Man can do" (Simon 1956, cit. in Larson 2021, p. 52).

Within the overarching area of AI, machine learning (ML) provides a subset of methods that improve their performance with experience. In contrast to, for instance, knowledge bases, ML methods *learn* in the sense that they improve their performance over time. Put formally, ML can be defined as a program that improves its performance in a class of tasks T as measured by a performance measure P with experience E concerning said class of tasks T (Mitchell, 1997). This broad definition encompasses a plethora of different statistical approaches, from support vector machines (SVM) and random forests to k-nearest neighbours or logistic regression.

1.2.2 A short primer on deep learning

Deep learning in turn represents a small but significant subset of ML methods that have only risen to fame in the past two decades. In DL, programs can find their own, multi-layered representations based on vast training data, allowing the program to find novel patterns in the data (LeCun, Bengio, & Hinton, 2015). While the fundamental theoretical underpinnings of these systems date back to the 1960s (Goodfellow et al., 2016, p. 221), it was only the exponentially increased computational power of modern processors that enabled its recent successes. Particularly well-known to the general public are its successes in image or voice recognition (Thompson, Greenewald, Lee, & Manso, 2020), or AlphaGo's widely noted victory in the board game Go (Silver et al., 2016).

Publicly less well-known than the advertised success stories of DL are its underlying statistical methods. While I will leave details to pertinent textbooks on the matter (Goodfellow et al., 2016), it is important for the rationale of this thesis to understand the epistemic problems posed by this ML approach in particular. DL is said to be "deep" in the sense that it comprises different techniques relying on artificial neural networks with multiple groups of units called layers. Commonly, multiple layers are then in turn

arranged in a chain structure, where each layer can be considered a function of the layer preceding it. A very simple example of a fully connected ANN with three hidden layers is depicted in figure 1.2 and may be useful to elucidate this architecture.[3]

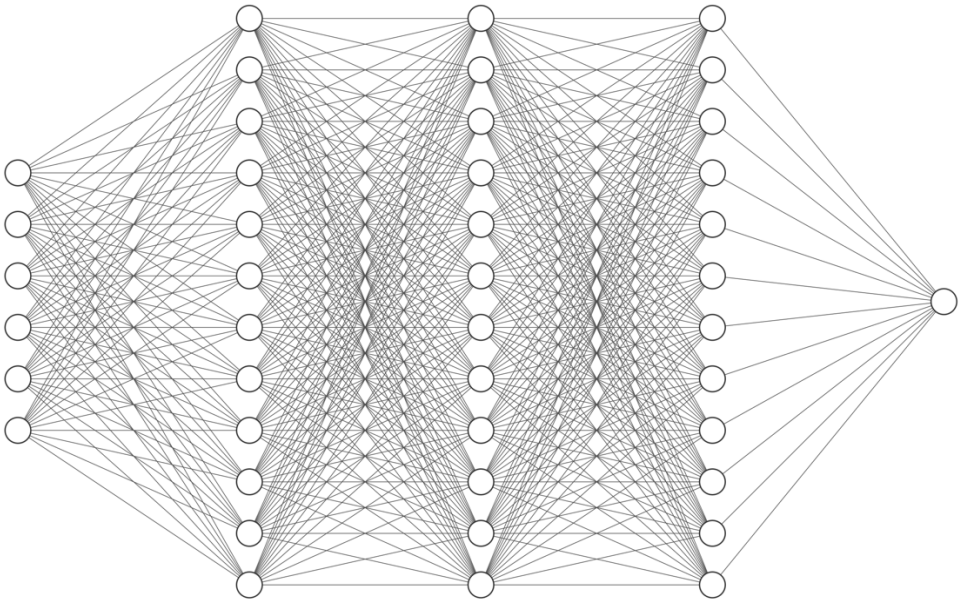

Fig. 1.2: Schematic of a simple fully-connected artificial neural network with three hidden layers. Created with NN-SVG (LeNail, 2019).

Let us assume that the output values of each layer y are computed using a non-linear function σ (e.g., with a sigmoidal function or a Rectified Linear Unit (ReLU) activation function), weighted by a learnable parameter w_i. If we define the input as $x \in \mathbb{R}^m$ and the output as $y \in \mathbb{R}^n$, then this output can be computed as follows:

$$y = \begin{pmatrix} \sigma\,(w_{1,1}\,x_1 + \cdots + w_{1,m}\,x_m \\ \sigma\,(w_{n,1}\,x_1 + \cdots + w_{n,m}\,x_m \end{pmatrix}$$

[3] There are of course many other popular network architectures such as convolutional neural networks (CNN).

Let us now assume that the ANN depicted here is intended to classify individual patients with a major depressive episode as responders and non-responders to treatment with selective serotonine reuptake inhibitors (SSRIs) based on multiple factors, covering e.g. (neuro-) biological data, psychometric scores, time since first depressive symptoms, or previously received medication, all of which might enter the input layer (Durstewitz, Koppe, & Meyer-Lindenberg, 2019; Lin et al., 2018). To train the network for the purpose of identifying (non-) responders, a large dataset from a comparable patient group is required from known responders and non-responders. Based on (more or less random) model parameters, the ANN can calculate a first prediction from the input data and compare this prediction against the ground-truth of the labelled cases. Using techniques such as gradient descent, the network can then be trained to minimize an error function concerning its prediction. To do so, the error is backpropagated through the network to update the weights (cf. Theodoridis & Koutroumbas, 2009, pp. 162-169). This implies calculating the respective gradients $\frac{dJ}{dw}$ of the error function J (w, x) with input x and learnable weights w. These gradients can then be used to update the weights for each iteration t+1 based on the previous iteration t and the learning rate λ:

$$w_{t+1} = w_t - \lambda \frac{dJ}{dw}$$

This procedure is repeated until the predictions of the model approach fit the training data.[4] If the DL model can then also pass various measures of quality control, most importantly its validation in an independent sample, it could then potentially be employed to predict treatment response in individual patients.

[4] For the sake of simplicity, I omit further details here such as hyperparameter tuning, regularization or comparison of different optimization algorithms.

1.3 Black-box algorithms: the problem of opacity

For what follows, it is crucial to note that, owing to the complexity of typical ANNs and the number of parameters, their decision making on an individual basis is typically opaque to human understanding. After all, as we have seen in the previous section, ML methods such as DL can result in exceedingly complex models, rendering them both inscrutable and nonintuitive to the human observer (Selbst & Barocas, 2018). For instance, already the very simple example above would comprise close to 400 updated weights, so a real-life example, drawing on multi-dimensional clinical data may easily comprise tens or hundreds of millions of weights. In such cases, we therefore understand how the program was designed and trained but we may not fully understand why it arrives at a particular decision, classifying for instance a particular patient as non-responder. In the literature, such programs are therefore typically called black-box algorithms (Durán & Jongsma, 2021).

The notion "black box" and its history seem, ironically, somewhat of a black box themselves (Geitz, Vater, & Zimmer-Merkle, 2020: 7). The term, widely used in different contexts, serves its respective purpose well, and yet we know rather little about how it works, or how it came about (ibid.). One attempt to trace the black box terminology to its origins has identified the electrical engineer Harold Stephen Black as its name giver (Vater, 2020), who designed the *Integrated Feedback Amplifier* in 1934 while working at the Bell Laboratories to automatically improve the signal-to-noise-ratio in telecommunication networks through feedback instead of filtering. In this narrative, the black box would actually be "Black's Box" that moves towards a kind of machine learning *avant la lettre* (Vater, 2020). Others attribute the "blackness" of the box to the 1939 development of a flight recorder by François Hussenot that shared similarity with a

camera, requiring total darkness in its insides (Engber, 2014), or to a plumbed black suitcase in which the physicist Edward Bowen transported an early cavity magnetron from the UK to the US, providing the allied forces with a decisive advantage in radar technology (Von Hilgers, 2010).[5]

Wherever its origin may be, the concept of the black box quickly grew in popularity, not least in the context of behaviourism, where it was prominently employed by B.F. Skinner (1985), describing the human brain linking sensorial inputs with behavioural outputs. Given the wide-ranging metaphorical use of black boxes, it is therefore crucial to be precise in one's terminology. As Jenna Burrell has noted, their opacity can take three different forms (Burrell, 2016). It can occur (1) as an opacity that is intended, for instance to safeguard secrecy on a corporate or state level, (2) as opacity due to users' technical illiteracy, and (3) as opacity that results from the very characteristics of ML (ibid.). This thesis is almost exclusively concerned with the third form of opacity, such as opacity resulting from DL architectures.[6] In clinical contexts, such opacity poses particular ethical challenges. How can we address so-called responsibility gaps, created by complex interactions between human agents and black-box algorithms, if a program's recommendation is erroneous and endangers patients (Matthias, 2004)? How can informed consent be obtained to use a program if it is by principle incomprehensible to both patients and health care professionals? How can we avoid discrimination against

[5] While the term is still highly present in current debates and despite its racially innocent history, it has recently also drawn criticism within a discourse that strives towards a racially more neutral terminology in technology (Cope & Gurung, 2020). In this vein, the UK National Cyber Security Centre decided to substitute the widely used term "blacklist" in 2020 with the term "deny list" (National Cyber Security Center, 2020). For the time being, the notion of black-box systems remains common in the literature, but it may make sense to substitute it with a more descriptive term such as "opaque box" in the future.

[6] Minor exceptions are chapter 6 on transparency, that touches on the topic of secrecy, and chapter 10 on education, addressing questions of technical literacy.

socially salient groups and protect vulnerable populations from systematic bias without understanding the underlying computational processes?

As noted at the beginning, many authors have recently suggested that trust could provide a model to deal with such challenges ensuing from opaque ML techniques. I discuss different theoretical approaches to trust at length in chapters 3 and 6. Here, I will therefore only briefly reflect on the role of trust and trustworthiness in recent regulatory debates and embed my work in current scholarly discussions on the ethics of medical AI.

1.4 Trust and trustworthy AI

In March 2019, the EU Commission's High Level Expert Group on Artificial Intelligence published their widely received *Ethics Guidelines for Trustworthy AI* ("Ethics guidelines for trustworthy AI," 2019). In these guidelines, 52 experts (among whom only four were trained ethicists (Metzinger, 2019)) formulated conditions for lawful, ethical, and robust AI that they took to be crucial for trustworthy systems (European Commission, 2021).[7] The recommended conditions for ethical AI were structured around four principles: respect for human autonomy, prevention of harm, fairness, and explicability.

These four principles were, in turn, largely derived from the AI4people framework (Floridi et al., 2018), that synthesized 47 recommendations from six international regulatory suggestions into five principles. These five principles represent the four classical principles of bioethics, as laid out by Tom Beauchamp and James Childress, namely beneficence, non-maleficence, respect for autonomy, and justice, and

[7] The first draft of the guidelines was amended after a public consultation with over 500 contributions from companies, associations, academic scholars, and private individuals (European Commission, 2021).

complement them with an AI-specific principle of explicability (ibid.). This resemblance to principlist approaches in bioethics does not seem entirely surprising, given that Luciano Floridi, lead author of the AI4people framework and one of the EU Commission's experts, regards bioethics as the field that reflects most closely the requirements of digital ethics, with novel challenges, agents, and environments (Floridi, 2013; Floridi et al., 2018).

Following the lead of the EU guidelines, trust has taken centre stage in academic debates about the ethics of AI, and medical AI in particular. While some have strongly opposed it as being too anthropomorphist a notion that lends itself to ethics-washing (Bryson, 2018; DeCamp & Tilburt, 2019; Hatherley, 2020; Metzinger, 2019; Ryan, 2020), others have defended its contribution to current debates when understood properly. Drawing on earlier work on trust in robots and e–trust (Coeckelbergh, 2012; Taddeo, 2010; Taddeo & Floridi, 2011), the different suggested defences of trust in AI share the conviction that for trust to be ethically meaningful, it needs to be bound to certain conditions of trustworthiness (Braun, Bleher, & Hummel, 2021; Durán & Jongsma, 2021; Ferrario, Loi, & Viganò, 2021; Gille, Jobin, & Ienca, 2020; Hartmann, 2020; Starke, van den Brule, Elger, & Haselager, 2021).

While this stance is not new and in line with broader theories of trust (Baier, 1986, 2013; Hardin, 2002; Luhmann, 1979; Misztal, 1996; O'Neill, 2002a, 2002b), there is some disagreement as to which conditions of trustworthiness should be considered in the particular context of medicine, and how trust should be conceptualized (Gille et al., 2020). For instance, Juan Durán and Karin Jongsma have defended trust in medical AI based on "computational reliabilism" (2021) that is evaluated based on four criteria, namely on verification and validation methods, robustness analysis, a history of

(un)successful implementations and expert knowledge (Durán & Formanek, 2018). Andrea Ferrario, Michele Loi, and Eleonora Viganò have suggested a layered model with simple, reflective, and paradigmatic trust as incremental forms building upon each other (2021). This thesis adds to the expanding theoretical literature on trust in opaque medical ML by suggesting a novel dimensional model of trust that allows to independently combine technical reliability measures of ML with ethical and societal evaluations, allowing for a flexible and fine-grained assessment.[8]

1.5 Structure of this thesis

In its structure, the thesis moves from general questions of trust and trustworthiness to particular challenges posed by ML in psychiatry. The first chapters are dedicated to questions of trust and trustworthiness on a general, conceptual basis. After a short overview of the employed methodology in chapter 2, chapter 3 lays the groundwork with a defence of trust in medical AI. There, I introduce a dimensional model inspired by Daniel Dennett that allows to evaluate the trustworthiness of a black-box algorithm from three independent angles. However, the focus of ethics should lie on promoting conditions of trustworthiness, not fostering potentially unwarranted, blind trust as Onora O'Neill has convincingly argued (O'Neill, 2013). Therefore, the following chapters are predominantly concerned with different aspects of trustworthiness. Hence, I next address questions of fairness and transparency, which are commonly considered to constitute the two most vital ethical concerns of clinically applied ML (Vayena et al., 2018).

[8] It may be worth noting that our paper was written and submitted before the two other papers mentioned here were published, which is why they are not discussed in chapter 3 itself.

With view to fairness, in chapter 4, I draw on the pragmatist framework developed by Hasok Chang following William James to argue why in many cases, instead of aiming at a supposedly objective truth, therapeutic usefulness should serve as guiding principle for assessing the fairness of ML applications in medicine (Chang, 2017). The following two chapters are concerned with transparency and the related concept of explainability. This focus seems warranted since explainability is commonly assumed to build, gain, or increase trust in AI (Braun, Hummel, Beck, & Dabrock, 2020; Markus, Kors, & Rijnbeek, 2021; Miller, 2019). I defend a slightly more sceptical view, in line with other recent publications (Ferrario & Loi, 2021). Chapter 5 discusses medical ML in light of Karl Japers' distinction between explaining and understanding (Jaspers, 1948). Expanding on the problem of often unknown causal relations in the realm of biomedicine, I argue how understanding could provide a useful complimentary model to explaining. Chapter 6 concludes the first part of the thesis, arguing with Onora O'Neill against a model of transparency as mere disclosure and in favour of intelligent openness, aimed at successful communication with the relevant stakeholders (Manson & O'Neill, 2007; O'Neill, 2002a, 2018).

In the following chapters, the thesis then situates these conceptual considerations in the messy reality of opaque ML employed in clinical settings, integrating theoretical examination and empirical data collection. Looking at a specific field of potential applications, I focus exclusively on applications in psychiatry and neuroscience. Beyond reasons of convenience, namely my prior experience with computational methods in psychiatric neuroimaging (Mulej Bratec et al., 2020; Starke, 2020), this focus has also substantive motives. On the one hand, neuroscience has often provided inspiration to the development of AI – not least in the form of ANNs –, rendering the two fields

inherently closer than, e.g. computer science and rheumatology (Ullman, 2019). On the other hand, psychiatry and neuroscience have long dealt with an archetypical black-box – namely the human brain (Clark, 2013). While analogies between ML and the human brain are necessarily limited, psychiatry in particular has developed strategies to pragmatically act under uncertainty that can also be informative to debates in AI ethics.

To provide readers with an introduction to ethical challenges of psychiatric ML, I first map the field by identifying different ethical issues related to its potential application. Using a well-established example from the field of biological psychiatry, namely real-life applications of ML for patients with schizophrenia, chapter 7 offers a systematic overview based on the principlist framework by Tom Beauchamp and James Childress. I then present the results of an empirically informed study in bioethics, drawing on interviews with experts on ML in psychiatry, in two chapters. Chapter 8 reports the attitudes and ethical expectations of researchers towards psychiatric ML, shedding further doubt on the ascribed role of explainability. Chapter 9 addresses a topic particular to the context of psychiatry by reporting the views and attitudes of researchers towards the impact of ML on psychiatric nosology. Chapter 10 concludes this part by discussing possible changes to psychiatric curricula, arguing that an ethically responsible implementation of machine learning in the clinic still requires attention to history.

I conclude the thesis with a discussion that critically relates the results of both parts to each other: The empirical findings of the interviews are confronted with the theoretical ethical considerations, whereas the interviews themselves also challenge previous assumptions grounded in moral theory (Molewijk, Stiggelbout, Otten, Dupuis, & Kievit, 2004).

In 2019, the American philosopher Robert Brandom published his *magnum opus* on Hegel's Phenomenology of Spirit, entitled *A Spirit of Trus*t (Brandom, 2019). In this book, Brandom suggests an "ethics of trust" (Knappik, 2020), drawing on the idea that both truth and linguistic meaning are normative matters (Sartwell, 2020). While Brandom's reading of Hegel goes far beyond the scope of trust discussed in this thesis, the insight that communication, whether in medical or scientific contexts, requires some form of trust (Manson & O'Neill, 2007; Shapin, 1995), is reflected at various stages in this thesis. It may therefore also be read as an attempt to promote a "spirit of trust", for as Brandom puts it:

> A proper understanding of ourselves as discursive creatures obliges us to institute a community in which reciprocal recognition takes the form of forgiving recollection: a community bound by and built on trust. (Brandom, 2019: 635)

1.6 References

Baier, A. (1986). Trust and Antitrust. *Ethics, 96*(2), 231-260. doi:10.1086/292745

Baier, A. (2013). What is trust? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 175-185). Oxford: Routledge.

Brandom, R. B. (2019). *A Spirit of Trust*. Cambridge, MA: Harvard University Press.

Braun, M., Bleher, H., & Hummel, P. (2021). A Leap of Faith: Is There a Formula for "Trustworthy" AI? *Hastings Center Report, 51*(3), 17-22. doi:10.1002/hast.1207

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics, 47*(e3). doi:10.1136/medethics-2019-105860

Bryson, J. (2018). No One Should Trust AI. *AI & Global Governance*. Retrieved from https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. doi:10.1177/2053951715622512

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety, 28*, 231-237. doi:10.1136/bmjqs-2018-008370

Chang, H. (2017). Operational Coherence as the Source of Truth. *Proceedings of the Aristotelian Society, 117*(2), 103-122.

Char, D. S., Abramoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *American Journal of Bioethics, 20*(11), 7-17. doi:10.1080/15265161.2020.1819469

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *New England Journal of Medicine, 378*(11), 981-983. doi:10.1056/NEJMp1714229

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204. doi:10.1017/S0140525X12000477

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*(1), 53-60. doi:10.1007/s10676-011-9279-1

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press.

European Commission. Proposal for a Regulation of the European Parliament and of the Council of laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, COM/2021/206 (2021)

Cope, C., & Gurung, S. (2020). Not a black and white issue: using racially neutral terms in technology. *Home Office Digital, Data and Technology*. Retrieved from https://hodigital.blog.gov.uk/2020/07/23/not-a-black-and-white-issue-using-racially-neutral-terms-in-technology/

Crawford, K. (2021). *Atlas of AI : power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine Learning and the Profession of Medicine. *JAMA, 315*(6), 551-552. doi:10.1001/jama.2015.18421

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *Lancet Digital Health, 1*(8), E390. doi:10.1016/S2589-7500(19)30197-9

Dennett, D., & Chalmers, D. (2019). Is Superintelligence Impossible? On Possible Minds: Philosophy and AI. *Edge*. Retrieved from https://www.edge.org/conversation/david_chalmers-daniel_c_dennett-is-superintelligence-impossible

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645-666. doi:10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329-335. doi:10.1136/medethics-2020-106820

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

Engber, D. (2014, April 6, 2014). Who Made That Black Box? *The New York Times Magazin,* p. 18.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24-29. doi:10.1038/s41591-018-0316-z

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, (2019). Retrieved from https://data.europa.eu/doi/10.2759/177365

Ferrario, A., & Loi, M. (2021). The Meaning of "Explainability Fosters Trust in AI". *Available at SSRN*. doi:10.2139/ssrn.3916396

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: it is possible to trust medical AI. *Journal of Medical Ethics, 47*(6), 437-438.

Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Geitz, E., Vater, C., & Zimmer-Merkle, S. (2020). Einleitung: Black Boxes. In E. Geitz, C. Vater, & S. Zimmer-Merkle (Eds.), *Black Boxes–Versiegelungskontexte und Öffnungsversuche* (pp. 3-18). Berlin: De Gruyter.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine, 1-2*, 100001. doi:10.1016/j.ibmed.2020.100001

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: The MIT Press.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205-211. doi:10.1136/medethics-2019-105586

Hardin, R. (2002). *Trust and trustworthiness*. New York: Russell Sage Foundation.

Hartmann, M. (2020). *Vertrauen : Die unsichtbare Macht*. Frankfurt am Main: Fischer.

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478-481. doi:10.1136/medethics-2019-105935

Hinton, G. (2018). Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA: The Journal of the American Medical Association, 320*(11), 1101-1102. doi:10.1001/jama.2018.11100

Ishiguro, K. (2021). *Klara and the Sun*. New York: Knopf.

Jaspers, K. (1948). *Allgemeine Psychopathologie* (5th ed.). Heidelberg: Springer.

Knappik, F. (2020). Brandom on postmodern ethical life. Moral and political problems. In G. Bouché (Ed.), *Reading Brandom. On A Spirit of Trust* (pp. 184-197). New York: Routledge.

Larson, E. J. (2021). *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge, MA: Harvard University Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi:10.1038/nature14539

LeNail, A. (2019). Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software, 4*(33), 747.

Lin, E., Kuo, P.-H., Liu, Y.-L., Yu, Y. W.-Y., Yang, A. C., & Tsai, S.-J. (2018). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in Psychiatry, 9*, 290.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report, 49*(1), 15-21. doi:10.1002/hast.973

Luhmann, N. (1979). *Trust and power* (English ed.). Chichester: Wiley.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Marckmann, G. (2003). *Diagnose per Computer? Eine ethische Bewertung medizinischer Expertensysteme* Köln: Deutscher Ärzte-Verlag.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183.

McEwan, I. (2019). *Machines like me: A novel*. New York: Knopf.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1-38.

Misztal, B. A. (1996). *Trust in modern societies: the search for the bases of social order*. Cambridge: Polity Press.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *FAT* '19: Proceedings of the conference on fairness, accountability, and transparency*, 279-288. doi:10.1145/3287560.3287574

Molewijk, B., Stiggelbout, A. M., Otten, W., Dupuis, H. M., & Kievit, J. (2004). Empirical data and moral theory. A plea for integrated empirical ethics. *Medicine, Health Care and Philosophy, 7*(1), 55-69. doi:10.1023/b:mhep.0000021848.75590.b0

Mulej Bratec, S., Betram, T., Starke, G., Brandl, F., Xie, X., & Sorg, C. (2020). Your presence soothes me: A neural process model of human aversive emotion regulation via social buffering. *Social Cognitive and Affective Neuroscience, 15*(5), 561-570. doi:10.1093/scan/nsaa068

National Cyber Security Center. (2020). Terminology: it's not black and white. Retrieved from https://www.ncsc.gov.uk/blog-post/terminology-its-not-black-and-white

Nida-Rümelin, J., & Weidenfeld, N. (2018). *Digitaler Humanismus: eine Ethik für das Zeitalter der künstlichen Intelligenz*. München: Piper.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. London: Rowman & Littlefield Publishers.

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

O'Neill, O. (2002b). *A Question of Trust. The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

O'Neill, O. (2018). *From Principles to Practice: Normativity and Judgement in Ethics and Politics*. Cambridge: Cambridge University Press.

O'Neill, O. (2013). Trust before trustworthiness? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 237-238). Oxford: Routledge.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. New York: Viking.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Hoboken: Pearson.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749-2767.

Sartwell, C. (2020). Systems of Philosophy: On Robert Brandom's "A Spirit of Trust". *Los Angeles Review of Books*. Retrieved from https://lareviewofbooks.org/article/systems-of-philosophy-on-robert-brandoms-a-spirit-of-trust/

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*, 1085.

Shapin, S. (1995). Trust, Honesty, and the Authority of Science. In R. Bulger, E. Meyer Bobby, & H. V. Fineberg (Eds.), *Society's Choices: Social and Ethical Decision Making in Biomedicine* (pp. 388-408). Washington, DC: National Academy Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., & Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489.

Skinner, B. F. (1985). Cognitive science and behaviourism. *British Journal of Psychology 76*(3), 291-301.

Starke, G. (2020). *Der ventromediale Hypothalamus, Furcht und ihre Regulation im Menschen.* Technische Universität München, München.

Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2021). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics*. doi:10.1111/bioe.12891

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

Taddeo, M. (2010). Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust. *Minds and Machines, 20*(2), 243-257. doi:10.1007/s11023-010-9201-3

Taddeo, M., & Floridi, L. (2011). The case for e-trust. *Ethics and Information Technology, 13*(1), 1-3. doi:10.1007/s10676-010-9263-1

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4 ed.). London: Academic Press.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Topol, E. J. (2019a). *Deep medicine : how artificial intelligence can make healthcare human again*. New York: Basic Books.

Topol, E. J. (2019b). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56. doi:10.1038/s41591-018-0300-7

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, LIX*(236), 433-460. doi:10.1093/mind/LIX.236.433

Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science, 363*(6428), 692-693.

Vater, C. (2020). Turings Maschine und Blacks Box–Mechanische Intelligenz nach dem Feedback. In E. Geitz, C. Vater, & S. Zimmer-Merkle (Eds.), *Black Boxes–Versiegelungskontexte und Öffnungsversuche* (pp. 323-350). Berlin: De Gruyter.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. doi:10.1371/journal.pmed.1002689

Vickers, B., & Mcdowell, K. (2020). *Atlas of Anomalous AI*. London: Ignota.

Von Hilgers, P. (2010). Ursprünge der Black Box. In A. Ofak & P. Von Hilgers (Eds.), *Rekursionen: Von Faltungen des Wissens* (pp. 135-153). München: Fink.

Chapter 2: Methodology

2.1 Bioethics at the intersection of values and facts

In his *Lectures on Logic*, Immanuel Kant famously summarized the scope of philosophy in four questions: "1) What can I know? 2) What ought I to do? 3) What may I hope? 4) What is man?" (AA IX: 25). Normative ethics, as opposed to descriptive ethics, strives for answers to the second question, not investigating the state of moral phenomena but how they should be, reappraising them in light of moral desiderata critically and systematically (Marckmann, 2022, p. 3f.). Put briefly, ethical theory can therefore "be thought of as a set of reasons and interconnected arguments, explicitly and systematically articulated, with some degree of abstractness and generality that gives directions for ethical practice" (Nussbaum, 2000, p. 56f.). While the practice of ethics is not the privilege of a chosen few but of anyone reflecting systematically on their actions, ethical theory is rooted traditionally in academic philosophy.

In many ways, this has long kept ethics as a systematic investigation of the normative apart from the empirical sciences. At least since Hume postulated in his 1739 *Treatise on Human Nature* postulated in 1739 that we cannot logically infer an *ought* from an *is* (Hume, 1739 [1896], p. 469f.), philosophers have been wary of committing such fallacy. In fact, Kant's moral philosophy can be read as one attempt to avoid this conundrum. Within his distinction between theoretical reason, concerned with laws of nature, and practical reason, concerned with what we ought to do, Kant differentiated between two kinds of practical reason: an empirically determined and a pure form of practical reason (Höffe, 2007, p. 211f.). Since the morally good is defined in terms of pure practical reason,

without recourse to empirical experiences, one can arrive at statements about moral obligations, about *oughts*, without recourse to an empirical *is* (ibid).[9]

While Kant's system provides moral principles that can guide action, his non-empirical construction of normative obligations rivals with many other non-consequentialist theories such as virtue ethics, and consequentialist theories such as utilitarianism. Reflecting societal pluralism and the lack of agreement on fundamental high-level ethical principles such as the categorical imperative, current bioethics therefore often relies on mid-level principles, upon which scholars from different traditions can agree and that are designed to reflect common morality (Beauchamp & Childress, 2019).[10] Ethical reasoning based on mid-level principles has long constituted a dominant paradigm in medical ethics and equally shapes recent approaches in AI ethics (Beauchamp & Childress, 2013; Floridi et al., 2018).

However, over the past decades, a supposed disregard of principle-oriented ethics for the empirical has drawn substantive criticism, suggesting for instance casuistry or narrative ethics as alternative approaches (e.g., Nussbaum (1990); Toulmin (1981); see Flynn (2021) for a comprehensive overview). In addition, feminist authors also have pointed out how bioethical inquiry stressing liberal individualist principles such as autonomy has not paid sufficient attention to gendered social contexts (Mackenzie & Stoljar, 2000; Scully, Baldwin-Ragaven, & Fitzpatrick, 2010; Wolf, 1996). In medical

[9] Even if one agrees with Kantian principles, the challenge remains of how to act on them in particular situations. Here, judgement is required, a "peculiar talent which can be practiced only but cannot be taught" (Kant AA III: 172; cf. O'Neill 2018 , 110-112).

[10] The *Principles of Biomedical Ethics* can also be seen as providing primarily a compromise between utilitarianism and deontology though, with comparatively little focus on e.g., virtue ethics, with the authors representing a rule-utilitarian (Beauchamp) and Christian deontologist (Childress) tradition respectively (Arras 2017: 3). For a more comprehensive discussion of the coherentism defended by Beauchamp and Childress see also Marckmann (2003, pp. 7-9).

ethics, both voices from outside the discipline as well as ethicists themselves have therefore called for more attention to context by grounding research in empirical work (Bruchhausen, 2001; Musschenga, 2009). Taking a slightly different form, criticism of general ethical principles enshrined in guidelines has recently also come to the forefront in AI ethics. In particular, critics surmise that such ethical guidelines may eclipse more pressing challenges such as underlying social conditions, power structures and environmental costs of AI (Crawford, 2021; Hao, 2021), which also need to be investigated from a social science perspective, involving relevant stakeholders. To address the ethical challenges posed by medical ML, it seems therefore imperative to engage both with abstract philosophical thought and its empirical reality.

In this thesis, I embrace a two-pronged approach, relating general, abstract reasoning to particular applications of medical ML, supported by empirical investigation. The results of both parts are related to each other in the sense of integrated empirical bioethics which recognizes an interdependent relation between values and facts (Molewijk, Stiggelbout, Otten, Dupuis, & Kievit, 2004). This chapter introduces and justifies this methodology in two steps. In the first section (2.2), I offer a very short introduction to empirical bioethics and argue why an integrated approach is the most appropriate framework to investigate ethical questions posed by medical ML. In the second subsection, I then provide a brief description of the methods of this thesis, motivating both its conceptual as well as its empirical methodology.

2.2 Towards integrated empirical bioethics

Responding to the aforementioned critics of armchair ethics, calls for a larger role of social sciences in bioethics have resulted in vast increases of empirical research published in key bioethical journals (Borry, Schotsmans, & Dierickx, 2006; Wangmo et

al., 2018). Following Pascal Borry and colleagues, this "empirical turn" is thought to have it origins in three complementary developments: (1) theoretical discontent of bioethicists with top-down approaches of applied ethics, (2) practical exposure of clinical ethicists to qualitative methods due to their professional integration into medical departments, and (3) the focus on evidence-based approaches in medicine since the 1990s (Borry, Schotsmans, & Dierickx, 2005).

However, the challenging question remains how to combine empirical and normative investigation with appropriate methodological rigour (Hurst, 2010; Ives et al., 2018; Marckmann, 2013) and without bioethicists becoming jacks of all trades, masters of none (Dunn, Gurtin-Broadbent, Wheeler, & Ives, 2008). In a highly received paper, Bert Molewijk and colleagues (2004) have distinguished four types of ethicists that use empirical data in different ways and form a kind of spectrum.

> (1) The "prescriptive applied ethicist" reasons deductively from a fixed moral theory and uses empirical findings for arriving at particular judgements, e.g., a consequentialist who investigates the likely outcomes of an action empirically.
>
> (2) The "theorist" also assumes that moral theory takes precedence over empirical findings but allows a refinement of moral theory based on empirical findings. Frances Kamm's non-consequentialist approach to an ethics refined by empirically traceable intuitions to thought experiments may come to mind here (Kamm, 2008).
>
> (3) The critical applied ethicist does not give precedence to either theory or empirical data, allowing for both to criticise each other. Many current approaches to empirical bioethics fall into this camp, and an explicit defence is offered, for instance, by Leget, Borry, and De Vries (2009).

> (4) The particularist, finally, does away with moral theory altogether, making the empirically traceable morality present in specific social contexts the only arbiter of ethics – casuistry can serve as an example here.

As an addition to these four types, Molewijk and colleagues have suggested a fifth approach, "integrated empirical bioethics", which seems the most adequate for investigating biomedical ML. In contrast to the other attempts of empirical bioethics mentioned above, integrated empirical bioethics is guided by the belief that facts and values are, contrary to Hume, inextricably interwoven, in a triple sense (Molewijk et al., 2004). This model assumes firstly that epistemic values specific to the respective disciplines are deeply embedded in the findings of empirical sciences, it considers secondly moral theory as based on "empirical background assumptions" such as particular assumptions concerning anthropology. Thirdly, it respects, as the authors put it, "that 'ought' implies 'can'" (ibid., p. 59), so that bioethics aimed at providing guidance for action needs to consider the factual possibilities of the relevant agents.

There are at least two main reasons why integrated empirical ethics constitutes the most appropriate methodology for the topic of this thesis. First, the fact-value dichotomy that is already problematic with regard to biological phenomena seems even more dubious when applied to technological artifacts (Latour, 1987),[11] including medical ML. After all, the very object of investigation here is not something given, a *datum*, but a *factum*, something human-made, arising from particular social practices. It is therefore inseparably linked to the epistemic and non-epistemic values and practices of the different involved scientific communities. For instance, in the design of a binary

[11] In fact, Latour even compares the "building" of scientific facts with the building of a black-box automaton, with both processes enlisting human and non-human actors (pp. 130-132.).

classification tool for a medical diagnosis, ML developers will be guided by epistemic values such as simplicity of the model but also consider social, non-epistemic values when determining the error costs in optimization, to arrive at an acceptable inductive risk linked to misclassification (Karaca, 2021).

Second, an integrated approach promises the most fruitful results – fruitful, at least, if we assume that bioethics should strive for a positive impact on society (Lindemann, 2019). As the last years have shown, conceptual starting points, developing guidelines based on abstract bioethical principles, are prone to ethics washing (Metzinger, 2019). At the same time, relying on empirical methods in AI ethics to foster interpretability without reflecting on the role of non-human actors in structural power imbalances, has been empirically shown to yield ethically highly questionable models (John-Mathews, 2022). Contrary to others, I do therefore not believe that keeping the empirical and the normative separate "as in a good friendship or marriage" is warranted here (Leget et al., 2009), but that instead, bioethical methodology should reflect the mutual dependency of fact and values when considering the ethics of medical ML.

2.3. Methodology of this thesis

The topic of this thesis is inherently interdisciplinary, drawing not only on literature in bioethics and philosophy, but also from history, medicine, psychology, neuroscience, human-computer-interaction studies, and computer science, aiming towards an integrated approach in its fullest sense. Consequently, to do justice to its object of inquiry, this thesis needs to engage with theoretical and empirical approaches that suit the multifaceted and multidisciplinary aspects of ethical questions posed by medical ML. Since methodological details are also discussed in the following individual chapters, I will only focus on my choice of conceptual and integrative framework here.

2.3.1 Conceptual approaches to trust and trustworthiness of medical ML

In the next four chapters, I turn to philosophers who embody interdisciplinary approaches, using their theoretical framework to answer specific questions related to trust and trustworthiness. For normative arguments, I draw mostly on non-consequentialist lines of reasoning by authors whose work can provide valuable additions to the existing literature, motivated by two deliberations. First, this tradition offers extensive discussions of trust and trustworthiness in the context of bioethics, most notably in the work of Onora O'Neill (Baier, 2013; Manson & O'Neill, 2007; O'Neill, 2002a, 2002b; O'Neill, 2013). Second, while there are of course also myriad contributions to the ethics of medical ML from consequentialist authors (Afnan et al., 2021; D'Hotman, Loh, & Savulescu, 2021; Savulescu, Kahane, & Gyngell, 2019), John Rawls, incidentally the PhD supervisor of O'Neill, is not without reason considered "artificial intelligence's favorite philosopher" (Procaccia 2019, cit. in Lundgard, 2020: 3). His theory of justice provides operationalizable constraints of fairness on an otherwise outcome-driven scientific endeavour, and his method of reflective equilibrium is widely accepted as a standard approach in bioethics, endorsed, for instance, also by Beauchamp and Childress (Arras 2017: 182, cit. in Flynn (2021)).

More importantly though, my conceptual approach is rooted in the methodological approach of integrated empirical bioethics. Highlighting its aforementioned skepticism about the fact-value-dichotomy, chapter 3 is heavily inspired by Bruno Latour (Latour, 2000), while chapter 4 draws on pragmatist philosophy of science (Chang, 2017; James, 1907 [1922]) – a tradition that has long been concerned with the interplay of facts and values (Putnam, 2004). Chapter 5 draws on the writings of Karl Jaspers, psychiatrist and philosopher, whose system has also been read as bridging the ethical divide between

values and facts in his existentialist philosophy (Dege, 2020). Finally, my reading of Onora O'Neill's views of trust and transparency, representing one of the most prominent Kantian voices in contemporary bioethics, concludes the first part, drawing explicitly on empirical findings from sociology (Beck, 2016).

2.3.2 Integrative approaches to the implementation of medical ML

The second part of my thesis, concerned with the more practical implementation of ML in psychiatry, is even more clearly grounded in an integrated approach to bioethics. Chapters 7 and 10, written for a medical audience, bridge principlist ethics and history of science with the reality of current psychiatric ML applications, while chapters 8 and 9 represent and discuss the empirical findings of a qualitative interview study. Here, the thesis explores the attitudes and beliefs of researchers involved in creating and using medical black boxes, to enable ethical and conceptual reflection that pays close attention to context and can thereby help to fill "blind spots in AI ethics" (Hagendorff, 2021).

For this purpose, I draw on qualitative interviews with experts on ML in psychiatry. The aim of these interviews was to understand the attitudes and beliefs towards medical ML within this community, with their respective educational backgrounds, professional cultures and at times even incommensurable terminologies (Turilli & Floridi, 2009), to tackle the challenges which arise at the intersection of computer science, psychiatry, and ethics. Due to the lack of empirical research in this field, this qualitative interview study was of explorative nature. Hence, research questions were not guided by preconceived hypotheses but aimed to inform further ethical analysis of medical ML by (1) identifying key ethical challenges for medical ML, (2) addressing specific conditions of trustworthiness such as fairness, transparency and explainability and (3) explore

expert's perspectives on current regulatory frameworks. The study followed the COREQ check-list (Consolidated Criteria for Reporting Qualitative Research, see appendix) (Tong, Sainsbury, & Craig, 2007). A detailed description of the study's methods is provided in chapters 8 and 9.

2.4. Methodological outlook

As will become increasingly clear throughout this dissertation, Kant's clear separation of four different domains of philosophical investigations is largely untenable when investigating ethical challenges posed by medical ML. Any attempt of evaluating what we ought to do requires at the very least some knowledge about what we can know. Or, put differently, the ethical question whether we should trust a particular ML model will depend largely on the epistemic question of what we can know about it, and thereby about its trustworthiness. Before turning to trustworthiness, let us first examine though whether trust can constitute an adequate way of dealing with black-box algorithms employed in medical contexts at all.

2.5 References

Afnan, M. A. M., Rudin, C., Conitzer, V., Savulescu, J., Mishra, A., Liu, Y., & Afnan, M. (2021). Ethical implementation of artificial intelligence to select embryos in In Vitro Fertilization. *arXiv preprint arXiv:2105.00060*.

Arras, J. D. (2017). *Methods in bioethics: The way we reason now*. Oxford: Oxford University Press.

Baier, A. (2013). What is trust? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 175-185). Oxford: Routledge.

Beauchamp, T., & Childress, J. (2019). Principles of biomedical ethics: marking its fortieth anniversary. *American Journal of Bioethics, 19*(11), 9-12.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Beck, U. (2016). *Risikogesellschaft: Auf dem Weg in eine andere Moderne*. Frankfurt am Main: Suhrkamp Verlag.

Borry, P., Schotsmans, P., & Dierickx, K. (2005). The birth of the empirical turn in bioethics. *Bioethics, 19*(1), 49-71.

Borry, P., Schotsmans, P., & Dierickx, K. (2006). Empirical research in bioethical journals. A quantitative analysis. *Journal of Medical Ethics, 32*(4), 240-245.

Bruchhausen, W. (2001). Medizin und Moral ohne Kontext. *Ethik in der Medizin, 13*(3), 176-192.

Chang, H. (2017). Operational Coherence as the Source of Truth. *Proceedings of the Aristotelian Society, 117*(2), 103-122.

Crawford, K. (2021). *Atlas of AI : power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

D'Hotman, D., Loh, E., & Savulescu, J. (2021). AI-enabled suicide prediction tools: ethical considerations for medical leaders. *BMJ Leader, 5*(2), 102-107.

Dege, C. L. (2020). Diversity in unity in post-truth times: Max Weber's challenge and Karl Jaspers's response. *Philosophy & Social Criticism, 46*(6), 703-733.

Dunn, M. C., Gurtin-Broadbent, Z., Wheeler, J. R., & Ives, J. (2008). Jack of all trades, master of none? Challenges facing junior academic researchers in bioethics. *Clinical Ethics, 3*(4), 160-163. doi:10.1258/ce.2008.008035

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Flynn, J. (2021). Theory and Bioethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)* (Spring 2021 ed.). doi:10.1007/s43681-021-00122-8

Hagendorff, T. (2021). Blind spots in AI ethics. *AI and Ethics*, 1-17. doi:10.1007/s43681-021-00122-8

Hao, K. (2021). Stop talking about AI ethics. It's time to talk about power. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/2021/04/23/1023549/kate-crawford-atlas-of-ai-review/

Höffe, O. (2007). *Immanuel Kant*. München: Beck.

Hume, D. (1739 [1896]). *A Treatise of Human Nature*. Oxford: Clarendon Press.

Hurst, S. (2010). What 'empirical turn in bioethics'? *Bioethics, 24*(8), 439-444.

Ives, J., Dunn, M., Molewijk, B., Schildmann, J., Bærøe, K., Frith, L., Huxtable, R., Landeweer, E., Mertz, M., & Provoost, V. (2018). Standards of practice in empirical bioethics research: towards a consensus. *BMC Medical Ethics, 19*(1), 1-20.

James, W. (1907 [1922]). *Pragmatism: A New Name for Some Old Ways of Thinking*. New York: Longmans, Green & Co.

John-Mathews, J.-M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change, 174*, 121209.

Kamm, F. M. (2008). *Intricate ethics: rights, responsibilities, and permissable harm*. Oxford: Oxford University Press.

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *European Journal for Philosophy of Science, 11*(4), 1-27.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Latour, B. (2000). The Berlin key or how to do words with things. In P. M. Graves-Brown (Ed.), *Matter, materiality and modern culture* (pp. 10-21). London: Routledge.

Leget, C., Borry, P., & De Vries, R. (2009). 'Nobody Tosses a Dwarf!' The relation between the empirical and the normative reexamined. *Bioethics, 23*(4), 226-235.

Lindemann, H. (2019). Bioethicists to the Barricades! *Bioethics, 33*(8), 857-860.

Lundgard, A. (2020). Measuring justice in machine learning. *arXiv preprint arXiv:2009.10050*.

Mackenzie, C., & Stoljar, N. (2000). *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford: Oxford University Press.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Marckmann, G. (2003). *Diagnose per Computer? Eine ethische Bewertung medizinischer Expertensysteme* Köln: Deutscher Ärzte-Verlag.

Marckmann, G. (2013). Wann ist eine ethische Analyse eine gute ethische Analyse? Ein Plädoyer für die Methodenreflexion in der Medizinethik. *Ethik in der Medizin, 25*(2), 87-88.

Marckmann, G. (2022). Grundlagen ethischer Entscheidungsfindung in der Medizin. In G. Marckmann (Ed.), *Praxisbuch Ethik in der Medizin* (pp. 3-13). Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Molewijk, B., Stiggelbout, A. M., Otten, W., Dupuis, H. M., & Kievit, J. (2004). Empirical data and moral theory. A plea for integrated empirical ethics. *Medicine, Health Care and Philosophy, 7*(1), 55-69. doi:10.1023/b:mhep.0000021848.75590.b0

Musschenga, B. A. (2009). Was ist empirische Ethik? *Ethik in der Medizin, 21*(3), 187-199.

Nussbaum, M. C. (1990). *Love's knowledge: Essays on philosophy and literature*. Oxford: Oxford University Press.

Nussbaum, M. C. (2000). Why Practice Needs Ethical Theory: Particularism, Principle, and Bad Behavior. In S. J. Burton (Ed.), *The Path of the Law and its Influence: The Legacy of Oliver Wendell Holmes, Jr* (pp. 50-86). Cambridge: Cambridge University Press.

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

O'Neill, O. (2002b). *A Question of Trust. The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

O'Neill, O. (2013). Trust before trustworthiness? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 237-238). Oxford: Routledge.

O'Neill, O. (2018). *From Principles to Practice: Normativity and Judgement in Ethics and Politics.* Cambridge: Cambridge University Press.

Putnam, H. (2004). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. *Nature Human Behaviour, 3*(12), 1241-1243.

Scully, J. L., Baldwin-Ragaven, L. E., & Fitzpatrick, P. (2010). *Feminist Bioethics. At the Center, on the Margins*. Baltimore: The John Hopkins University Press.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International journal for quality in health care, 19*(6), 349-357.

Toulmin, S. (1981). The tyranny of principles. *Hastings Center Report, 11*(6), 31-39. doi:10.2307/3560542

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105-112. doi:10.1007/s10676-009-9187-9

Wangmo, T., Hauri, S., Gennet, E., Anane-Sarpong, E., Provoost, V., & Elger, B. S. (2018). An update on the "empirical turn" in bioethics: analysis of empirical research in nine bioethics journals. *BMC Medical Ethics, 19*(1), 1-9.

Wolf, S. (1996). *Feminism and Bioethics: Beyond Reproduction*. Oxford: Oxford University Press.

Chapter 3: Intentional Machines: A Defence of Trust in Medical AI

Intentional Machines: A Defence of Trust in Medical AI

Georg Starke[1], Rik van den Brule[2,3], Bernice Simone Elger[1,4], Pim Haselager[2]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland, [2] Donders Centre for Brain, Cognition, and Behaviour, Radboud University Nijmegen, the Netherlands, [3] Behavioral Science Institute, Radboud University Nijmegen, the Netherlands, [4]University Center of Legal Medicine, University of Geneva, Switzerland.

Acknowledgements: The authors would like to thank Ron Dotsch, Gijsbert Bijlstra, and Daniel Wigboldus for their crucial input to an earlier version of the manuscript. GS would also like to thank Emilian Mihailov as well as the participants of a virtual workshop on the philosophy of medical AI held at the University of Tübingen in October 2020 for their critical and constructive comments.

Abstract

Trust constitutes a fundamental strategy to deal with risks and uncertainty in complex societies. In line with the vast literature stressing the importance of trust in doctor-patient relationships, trust is therefore regularly suggested as a way of dealing with the risks of medical AI. Yet, this approach has come under charge from different angles. At least two lines of thought can be distinguished: (1) that trusting AI is conceptually confused, i.e. that we *cannot* trust AI, and (2) that it is also dangerous, i.e. that we *should not* trust AI – particularly if the stakes are as high as they routinely are in medicine. In this paper, we aim to defend a notion of trust in the context of medical AI against both charges. To do so, we highlight the technically mediated intentions manifest in AI systems, rendering trust a conceptually plausible stance for dealing with them. Based on literature from Human-Robot Interaction, psychology and sociology, we then propose a novel model to analyse notions of trust, distinguishing between three different aspects: reliability, competence, and intentions. We discuss each aspect and make suggestions how medical AI may become worthy of our trust.

Keywords: trust, trustworthiness, artificial intelligence, healthcare

3.1 Introduction

Trust is crucial for human actions and interactions in technologically advanced societies. It constitutes, as the sociologist Niklas Luhmann famously put it, "a mechanism to reduce social complexity" (Luhmann, 1968) and enables us to act in situations characterised by uncertainty. Across different areas of life, our actions are increasingly shaped by technological means and computational operations characterised by such uncertainty – a development which the rise of Artificial Intelligence (AI) is bound to accelerate further. In medicine, this progress has already spawned diagnostic, predictive and therapeutic programs, and will likely leave no medical specialty untouched (Topol, 2019). Potential applications range from automated pathological evaluations of cancer biopsies (Arvaniti et al., 2018) to AI-assisted surgery planning (Knoops et al., 2019), from predicting acute circulatory failure in intensive care units (Hyland et al., 2020) to choosing an appropriate medication for psychiatric disorders (Starke, De Clercq, Borgwardt, & Elger, 2021).

Yet, the impending employment of these complex systems for clinical decision-making necessitates an appropriate attitude of dealing with the lack of explainability often inherent in these programs. In the past years, both public and private research bodies have repeatedly suggested trust as a possible way to deal with the uncertainty posed by applications of AI (Desai & Kroll, 2017; Hatherley, 2020). In medicine, calls for trust in AI seem particularly tempting, given the importance ascribed to trust for the complex relations between patients, physicians and other agents in healthcare settings (O'Neill, 2002a), and consequently for the functioning of the health care system in general (Gille, Smith, & Mays, 2015). However, it remains open to debate whether trust constitutes a conceptually defensible attitude towards medical AI. As several authors have argued,

trust should be reserved exclusively to human agents, as opposed to any non-human systems, which supposedly lack the necessary motives to be recipients of trust (DeCamp & Tilburt, 2019; Hatherley, 2020).

In this paper, we will argue against this distinction, defending the notion of trust in medical AI by highlighting the motives enacted by potential applications. The crucial question we aim to answers is: if we are to trust AI systems, how could such trust be conceived? Despite the prominence "trustworthiness" enjoys in regulatory debates, fostered by the EU guidelines for trustworthy AI, substantive conceptualisations of trust in medical AI are still largely lacking (Gille, Jobin, & Ienca, 2020). We aim to address this gap by suggesting a three-dimensional framework of trust, distinguishing between the reliability, competence, and intentions of an AI system. In doing so, we hope to foster greater conceptual clarity while simultaneously defending trust as a meaningful attitude to deal with the inherent risks of medical AI.

3.2 Trust and trustworthiness

At the outset of any debate about trust, it seems vital to briefly revisit the distinction between trust and trustworthiness (Hardin, 2002). In general, trust is foremost a cognitive attitude of a trustor (an agent doing the trusting) towards a trustee (an agent who is a candidate for being trusted) to act or behave in a beneficial way toward the trustor (McLeod, 2015). Which characteristics a trustor takes to be important for assessing a trustee's trustworthiness may depend on the characteristics of the trustor (sometimes called trustfulness) (Tullberg, 2008), the situation in which the interaction takes place, the nature of the task, the type of agent the trustor is engaging with, and the propensity of the trustor to engage in trust relationships. For instance, if someone says "I don't trust Billy with that chainsaw", the meaning of trust differs dramatically

whether Billy is a curious four-year-old child or an outraged forty-year-old lumberjack. In the first case, the statement will likely refer to the fact that Billy is too young to know how to handle a chainsaw and might hurt someone (or himself) because of his incompetence. In the latter case, there is a good chance that Billy is very competent with a chainsaw, but might hurt someone because he has bad intentions. Trust is thus based on the subjective inferences of the trustor from the perceived features of the trustee and past experience with the trustee, reflecting its trustworthiness. Although trust in both situations is low, it can lead to quite different behavioural responses. In the first case, the response may well be to approach Billy and take away the chainsaw, whereas in the latter situation the appropriate response will arguably be to run away.

In either case, trust can be construed as a multi-part relationship in which the trustor (A) entrusts the trustee (B) with a specific task (T) or in matters (Y) in specific circumstances Z (Baier, 1986). Within this relation, A's trust in B is often limited to the specific task T or the field of expertise Y. A patient may very well trust his dermatologist to distinguish between nevi and melanoma but may be very reluctant to also trust her with the extraction of a carious tooth. The specific properties of the doctor leading to this judgement can in turn be summarized as trustworthiness, which concerns the prerequisites for a trust relationship. It refers to the qualities of a trustee, as observed and evaluated by the trustor, which give the trustor the confidence to engage in a trust relationship (Hardin, 2002).

While the exact nature of trust remains subject to much debate in philosophy and sociology (Misztal, 1996; Möllering, 2001; O'Neill, 2002b), two characteristics of this relation seem instructive here. First, trust requires a state of imperfect knowledge concerning the trustee B on the side of the trustor A: those with complete knowledge

need not trust while those without any knowledge cannot reasonably trust at all (Simmel, 1908 [1983]).[12] Second, the task T needs to include some risk for the trustor, putting her in a vulnerable relation towards the trustee (Baier, 1986). Trustworthiness on the other hand is a *property* of the trustee that renders A's trust in B with regard to T justifiable (McLeod, 2015). It is assembled of multiple different characteristics that are subject to much disagreement. However, some often-found characteristics seem rather uncontroversial, such as a trustee's capability to perform the task in question and a reliable performance in the past. One should note though that trust and trustworthiness are, on a factual level, not necessarily linked. Some credulous and gullible agents may place their trust in untrustworthy trustees, whereas others may be driven by caution to withhold trust even where it is warranted, foregoing a potentially useful interaction.

3.3 Trust in medical AI: conceptual nonsense?

The two mentioned conditions of trust, uncertainty and risk, generally characterise the practice of medicine and clearly apply in the context of medical AI (Grote & Berens, 2020). Few experts, if any, command the knowledge to understand medical AI systems, and even those may not be able to scrutinise the internals of a specific program, due to the much-discussed black-box nature of specific AI models (London, 2019). While this creates a multitude of ethical and regulatory challenges in medical contexts (Vayena, Blasimme, & Cohen, 2018), it certainly also fulfils the condition of an intermediate state of knowledge between complete knowledge and no knowledge at all. At the same time, the magnitude of risks created by AI systems for diagnostic, prognostic, therapeutic or

[12] «Der völlig Wissende braucht nicht zu vertrauen, der völlig Nichtwissende kann vernünftigerweise nicht einmal vertrauen.» (p. 263)

even predictive clinical decisions is evident. Patients and research participants as potential trustors are clearly vulnerable to physical or psychological harm throughout their exposure to the health care system. IBM Watson's incorrect and dangerous treatment suggestions for cancer patients that would have created severe harm if implemented are a common example (Ross, 2018), highlighting Luhmann's description of trust as a "risky investment" (Luhmann, 1979, p. 24). However, while the presence of risks and the absence of complete knowledge may constitute necessary conditions, they certainly do not render trust in medical AI plausible or justified by themselves. Rather, they create the conditions under which trust can play a meaningful and reasonable, albeit risky, role.

In fact, following the 2019 publication of the guidelines for trustworthy artificial intelligence by the European Commission's High Level Expert Group on Artificial Intelligence ("Ethics guidelines for trustworthy AI," 2019), there has been much renewed discussion whether artificial intelligence (AI) can be trustworthy at all. In a widely discussed newspaper comment, the philosopher Thomas Metzinger, himself a member of the EU's expert group, has posited that the very notion of trustworthy AI is "conceptual nonsense" (Metzinger, 2019). Fitting the brief format of his intervention and the different focus of his article, Metzinger claims: "Machines are not trustworthy; only humans can be trustworthy (or untrustworthy)" (ibid.). Other critics of trustworthy AI have since chimed in, e.g. arguing with a nod to Nietzsche that we can only trust beings that can make promises. However, since (current) AI does not have an inner emotional life nor feelings such as remorse or pride, it cannot make promises, and can thus not be worthy of our (direct) trust (Lauer, 2019).

With specific regard to medical AI, two recent articles have spelled out this criticism further (DeCamp & Tilburt, 2019; Hatherley, 2020). Both articles are important insofar as they highlight limitations of trust in medical AI, such as the difficulty of apportioning responsibility and liability for potential mistakes. Both articles rely on models of interpersonal trust that put a trustee's motives front and rightly point out that that a system's mere reliability or accuracy is too little to warrant trust. However, they also presume a categorical division between the realm of (potentially trustworthy) humans and a realm of artificial things, that cannot be worthy of trust: „Although well intentioned, applying trust to AI is a category error, mistakenly assuming that AI belongs to a category of things that can be trusted" (DeCamp & Tilburt, 2019). Supposedly, the reason for this lies in the lack of motives and good (or bad) will on side of the AI as the trustee. Key to this argument is the assumption that these systems do not have motives or intentions. Referring to theories of trust by Russell Hardin and Annette Baier, Hatherley puts his claim as follows: "AI systems lack the right kind of motivation for trust — either in the form of encapsulated interest or a sense of good will — since they lack motivation entirely" (Hatherley, 2020). Hence, „[t]o say that one can trust an AI system, or that the AI is trustworthy, is merely to say that one can rely on the AI system, or that the system is reliable" (ibid). In other words, the critics insist that proper trust presupposes some kind of (benevolent) motives, which only human agents possess.

We take it that there are at least three problems with this kind of argument. First, if we subscribe to a Wittgensteinian approach that (in most cases) the meaning of a word is its use in the language, trust factually describes a much broader phenomenon than mere interpersonal relations. From trust in local governments to trust in health care systems, trust is commonly used to denote an attitude towards non-human or non-living entities,

for instance towards bridges, cars or institutions (Gille, Smith, & Mays, 2017; Grimmelikhuijsen, 2010). Second, by focusing exclusively on motive-based models of interpersonal trust, the critics leave out many other models that would be better suited to encompass trust in AI. For example, it is far from uncontroversial that trust necessarily demands good will on part of the trustee (O'Neill, 2013). In the same vein, Ferrario et al. have recently suggested a multi-layered model of trust, distinguishing between incremental layers of simple, reflective, and paradigmatic trust, of which at least the first two are applicable to AI (Ferrario, Loi, & Viganò, 2020).[13] As they argue, one should therefore not model all forms of trust exclusively on "paradigmatic", interpersonal trust relations (ibid). Third, doing away with the notion of trust in AI also seemingly disregards decades of research from Human-Computer-Interaction studies that have embraced the notion of trust, if only in a very specific and narrow sense (Hancock et al., 2011; Lee & See, 2004; Sanders, Oleson, Billings, Chen, & Hancock, 2011). It thus seems worth revisiting the debate whether trust in AI is indeed conceptually flawed.

3.4 The intentions of machines

Let us for now assume that the motives of the trustee are in fact crucial for a trusting relation. Then, trust could, it would seem, come in two guises: a direct and an indirect form – a distinction which mirrors debates about trust in robots, distinguishing between direct and indirect trust in artefacts (Coeckelbergh, 2012). In its indirect, weaker sense, trust in AI does not require a fully independent agency of the program itself but rather ties trust to the intentions of its developers or those involved in its quality control,

[13] The authors have recently also used their framework to reply to the paper by Hatherley (Ferrario, Loi, & Viganò, 2021).

promoting "indirect trust in the humans related to the technology" (ibid, p. 54). For example, we may trust a system of medical AI because we trust the people who develop and regulate it. Even in this very limited sense, it may already be plausible to describe a potential attitude towards medical AI as "trusting" (Ferrario et al., 2021). However, it may indeed fall short of a proper concept of trust and could potentially be better described as "trust-as-reliance" (Coeckelbergh, 2012). A stronger, direct conception of trust in AI in turn requires defending the system as a somewhat independent agent, to which we can ascribe motives and intentions.

Defending the notion of trust in AI requires to look at the kind of agency one can find in inanimate objects. We build our argument on the rather strong assumption here that one can reasonably attribute agency to AIs. Since it would be beyond our scope here to cover the rich literature which discusses the status of AI as artificial agents,[14] we merely sketch one argument in its favour. Drawing on the work of Bruno Latour, Martin Hartmann has recently suggested such an approach towards AI in the context of trust that pays particular attention to the meaning constituted by an object itself through its relations in social contexts (Hartmann, 2020, pp. 230-232). Latour's famous example from *The Berlin key or how to do words with things* is a rather simple device, namely a door key (Latour, 2000). This key, common in Berlin tenant houses during the first half of the 20th century, is constructed in a way that it compels its user to re-lock the door of a building after entering: After unlocking a door, the key cannot be simply removed like a usual key but remains stuck in its position, unless it is pushed through the keyhole to

[14] A popular account is, for example, provided by Russell and Norvig (2021, pp. 34-59). See also Haselager (2005) and, for a discussion of agency in the specific context of medical AI, Braun, Hummel, Beck, and Dabrock (2020).

the other side of the door. Only after locking the door from the other side can it be removed. By its very design, the key thus contains a complex and specific action program that is born out of a set of specific motives, e.g. the proprietor's interest in a locked door. However, in Latour's view such objects are not mere intermediaries that simply transport or reflect the motives of the homeowner. Instead, by playing its part in a complex network of actors that would not be feasible without the material manifestation of the key, it contributes to the disciplinary relation itself: "Meaning does not antecede technological devices. […] From being a simple tool, the steel key assumes all the dignity of a mediator, a social actor, an agent, an active being" (Latour, 2000, p. 19).

If we follow this account and accept that even a simple key can be considered an agent in complex social relations, placing trust in AI seems no longer conceptually confused at all.[15] As Latour puts it: "To speak of "humans" and "non-humans" allows only a rough approximation that still borrows from modern philosophy the stupefying idea that there exist humans and non-humans, whereas there are only trajectories and dispatches, paths and trails" (Latour, 2000). Under this premise, the concern that we *cannot* trust medical AI simply due to it being non-human seems no longer convincing. However, this is still begging the arguably more important question if and under which conditions we *should* place trust in medical AI. In consequence, a substantial account of the determinants of trust in medical AI is urgently needed.

[15] Arguably, one could call such a form of trust "methodological trust" to distinguish it conceptually from e.g. interpersonal trust. Such a notion would highlight the primacy of methodology in the context of actor-network-theory, clarifying the notion of non-human agency involved here. (Sayes, 2014).

3.5 Dimensions of trust

We suggest that different aspects of the act of trusting itself can be distinguished, based on the features that the trustor utilizes to arrive at a trust decision. Specifically, we propose that the decision to trust an AI-based program depends on features of the trustor (e.g. overall willingness to trust), and the context (e.g. level of risk) in combination with the perceived reliability, competence, and intentions of the program. Given the broad range of potential applications of medical AI, these distinct aspects are likely to be weighed differently across different usages. Furthermore, such a model needs to be supported by empirical research into the factual ramifications of trust in medical AI, as Gille et al have stressed (Gille et al., 2020).

Fortunately, as a tentative approach, we can draw on the rich literature addressing human-machine interaction (HCI), which has explored the conditions of trust in machines in the past decades. Building on our earlier empirical work in human-robot interaction (Van den Brule, Bijlstra, Dotsch, Haselager, & Wigboldus, 2016; Van den Brule, Bijlstra, Dotsch, Wigboldus, & Haselager, 2013; Van den Brule, Dotsch, Bijlstra, Wigboldus, & Haselager, 2014), the work of Ososky et al. (2013), Hancock et al. (2011), and in parallel with Dennett's three levels of analysis (Dennett, 1989), we thus suggest that a trustor can take one of three stances towards the trustee. First, one can take a physical stance, considering a system's physical properties and the reliability of its functioning ("Will it break down?"). In many cases, this will require scrutinizing the hardware on which a specific system depends. Second, one can also take the design stance, focusing on a system's basic functions and performance ("What does it do, and what are the odds of its performance?"). Here, trust concerns its competence to achieve a task, evaluated by the likelihood of its success. Finally, one can take the intentional

stance, looking at the motives manifested in the system ("Why is it acting like this?"). Trust here concerns a system's "intentions", in the weak sense discussed above: not as something an AI develops on its own, but rather as something manifest in a system due to its embeddedness in a given social context with many different actors.[16]

3.5.1 Trusting reliability

One crucial aspect in determining an AI system's trustworthiness lies in assessing its basic reliability (i.e., whether it will work). A system can be considered reliable when it can perform its required function under stated conditions for an indicated amount of time. Evaluating reliability is distinct from assessing performance during normal functioning (i.e. competence) in that it addresses the robustness of an agent, i.e., whether deviations from normal functioning occur (usually labelled as "breakdowns" or "malfunctioning"). Importantly, reliability in this sense focuses on the avoidance of malfunctioning, not on a lower or higher degree of competence. For example, let us consider a hypothetical AI-aided detection system for adenoma, which assists physicians in their diagnostic process during colonoscopy. Let us further suppose that this system provided the most accurate diagnoses available to date, and would thus be trusted based on its competence. However, if the system was prone to breaking down during the procedure, whether due to hardware problems or software glitches, the same system could be distrusted with regard to its reliability.

[16] Also from Dennett's original "instrumentalist" use of the intentional stance, one may ascribe intentions to AI as they have been designed in a particular way or because humans attribute intentions to them automatically, especially if they are embodied in a robot, shaping our behaviour towards them (Dennett, 1978, 1991; Terada, Shamoto, & Ito, 2008).

3.5.2 Trusting competence

An important aspect of interpersonal evaluation, as studied by social psychology, is competence. In the context of AI systems, this translates into assessing features related to its performance as measured for instance by the validity and accuracy of its predictions.[17] To stay with a medical example, when using a diagnostic tool for the automated detection of melanoma or squamous cell carcinoma, developers, physicians and patients alike will closely scrutinize the rate of false-positive and false-negative decisions, and compare the system's results to the performance of trained physicians. Similarly, a system suggesting oncological treatment may foremost be evaluated based on its performance, i.e. whether it recommends a treatment regime in line with current guidelines, or whether patients treated according to its recommendations have a better outcome, e.g. measured in their 5-year survival rate. In contrast, recommendations that endanger a patient's life would drastically undermine the system's trustworthiness. From this perspective, the question is not whether the system occasionally breaks down in practice, but only whether, under ideal circumstances, it will perform well in a certain task.

3.5.3 Trusting intentions

Trust based on the intentions present in, or attributed to, a trustee has been studied especially in social psychology, particularly in the field of person perception (Fiske, Cuddy, & Glick, 2007), but has not yet received enough attention in the context of trust in AI. The prevailing view in person perception is that people are judged primarily based

[17] Of course, a particular challenge in the context of self-learning systems lies in the fact that the performance of the system may change over time, if its keeps being trained with new data. In these cases, evaluations of competence would need to be updated regularly.

on their intentions, which is conceived of as a "warmth" or valence (positivity and negativity) judgment. Already in 1946, Asch showed that a valence judgment can alter the impression of a person in a way that is relevant for our understanding of trust (Asch, 1946). Asch asked students to form an impression of two persons based on a list of traits (e.g., intelligent, skilful, practical, sincere), which also included either "warm" or "cold" depending on the experimental condition. Whereas participants described a warm intelligent person as wise, a cold intelligent person was seen more as sly. "Warmth" captures moral-social traits related to perceived intent, including friendliness, helpfulness, and sincerity.

In the context of medical AI, it is here, we believe, that most of the current debates about the ethical challenges posed by medical AI could be adequately reflected. Conflicts of interests, e.g. through financial ties to a company which would benefit from certain recommendations, a lack of transparency concerning the system's development as well as systematic biases against specific groups of patients would, for instance, all undermine trust in the (derived or indirect) intentions of such a system. A system's explainability or interpretability do, in turn, also seem crucial to foster this aspect of trust, insofar as they attempt to answer why a system behaves in a specific way (Ras, van Gerven, & Haselager, 2018; Shin, 2020).

3.5.4 Combining the aspects

For researchers striving to advance trustworthy medical AI, it seems crucial to distinguish between these three aspects of trust (i.e., competence, intention, and reliability). Of course, in any situation, more than one aspect of trust may influence the interaction; a system could be trusted well on its intentions, poorly on its reliability, and average on its competence in a certain situation. Although there are different ways to

visualize the possible combinations of the three aspects of trust, the most neutral approach is to consider them as dimensions of a three-dimensional space that encompasses all their logically possible combinations (see fig. 3.1). Thus, trust is not a single, uniform phenomenon, but can vary within that three-dimensional space, where there may, for instance, be no trust at all in competence of the system, complete trust in its good intentions, and only little trust in its reliability. Bringing together these different perspectives, a trustor may then decide whether to engage with a system in a specific situation.

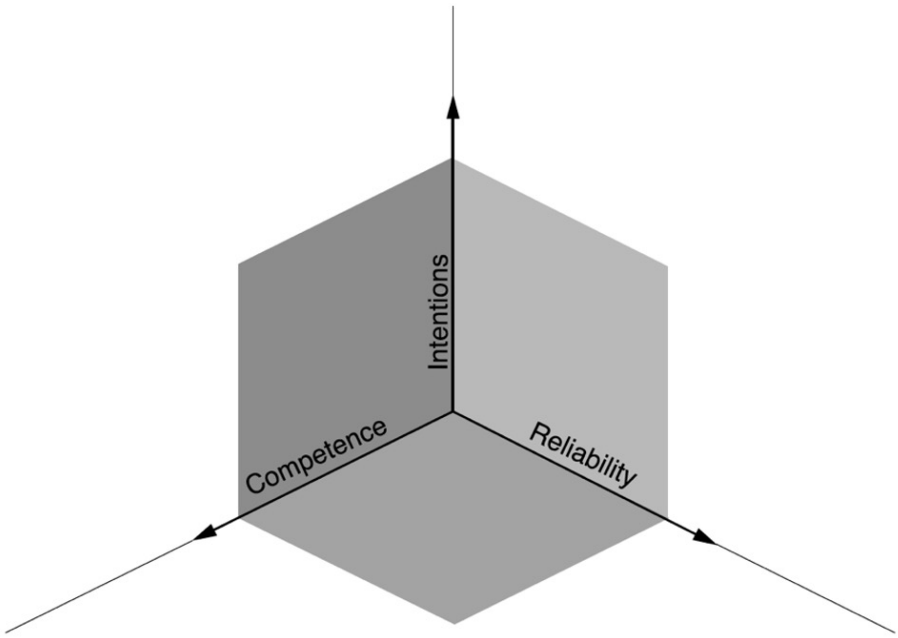

Fig 3.1: Three-dimensional space representing different aspects of trust: reliability, competence, and intentions

Whether all logical possibilities are psychologically meaningful is a topic of empirical investigation. Certain combinations may simply not be psychologically plausible whereas other combinations might be quite common. It is possible that dependencies between the aspects of trust are so strong that they can be viewed as having a

hierarchical relationship in which the different types build upon each other, similar to Maslow's famous hierarchy of needs (Maslow, 1943). For instance, the reason that intentions play such a large role in trust research in person perception may well be because in general a healthy person is assumed to meet some basic standards of reliability and competence. It is conceivable that also in the context of medical AI certain "precedence" relations exist between the various aspects. That is, it might be the case that trust in intentions is second to trust in statistically measurable degrees of reliability and competence in the context of medical AI, simply because we would otherwise not choose to interact with the system and may also not be allowed to do so, due to regulatory restraints.

In addition to features of the trustee, various trustor characteristics may play an important role in the eventual degree of human trust in a specific system: the propensity to trust may be quite influential in the intentional context, but of less relevance in assessing a systems reliability or performance. Likewise, background knowledge (e.g. a user's knowledge about and experience with medical procedures) may be most relevant when assessing competence, rather than intentions of a system. Such variations can influence a trustor's behaviour quite significantly, and it is therefore important that studies carefully consider the background of the trustors in empirical investigations of trust.

3.6 Towards trustworthy medical AI

In this article, we have defended the notion of trust in medical AI against recent charges that have criticised trust in non-human intelligences as conceptually confused. We did so by highlighting the kind of intentions that are manifest in inanimate objects. Having established the conceptual plausibility of trust in medical AI, we then suggested a novel

approach for understanding this kind of trust, distinguishing between the aspects of reliability, competence and intentions. In our view, such trust may not only appear as a possible attitude towards medical AI, but potentially even as a necessary condition for its successful implementation and acceptance at the bedside.

Importantly, in this sense, fostering justified trust does not diminish the importance of scrutiny and appropriate regulation of medical AI. On the contrary, to introduce a medical AI system for clinical use, it will be important to define acceptable thresholds of trustworthiness with regard to the three dimensions we have sketched. However, we believe that our model stresses the necessity to look beyond the mere reliability and performance of a system and take into account its wider ramifications, including ethical considerations that need be recognized under the aspect of intentions, relevant for instance during its conception, development and implementation (Char, Abràmoff, & Feudtner, 2020). Close attention to these challenges of medical AI systems seems thus as vital as ever, requiring public debates about an appropriate pipeline for their ethical evaluation (Wiens et al., 2019). In other words, it remains important to not trust blindly, but only place trust in agents, both human and non-human, that are worthy of it, and reconsider this judgement regularly (O'Neill, 2002b). Only then may we enable the different agents in healthcare systems to evaluate and weigh the risks and benefits that these systems entail, to make a justified judgement whether to trust ML integrated into clinical care.

In this sense, calls for trust in medical AI should not be mistaken for “ethics washing” Metzinger, 2019) or for a naïve embrace of authority, but rather as a well-grounded attitude towards systems that have been appropriately probed concerning their reliability, competence and intentions. The form of trust recommended here is not

opposed to searching for knowledge, but rather underlines its necessity while still acknowledging our epistemic limitations. In fact, as Steven Shapin has argued, such trust constitutes a fundamental necessity of science itself, as a collective enterprise where the overwhelming majority of propositions is accepted by trusting others, or as Shapin puts its:

> Science is a trusting institution. Trust is not an epistemic *problem* for science; it is - if one wants to engage in such evaluations - an evident epistemic virtue. (Shapin, 1995)

3.7 References

Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Ruschoff, J. H., & Claassen, M. (2018). Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific Reports, 8*(1), 12054. doi:10.1038/s41598-018-30535-1

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*(3), 258.

Baier, A. (1986). Trust and Antitrust. *Ethics, 96*(2), 231-260. doi:10.1086/292745

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics, 47*(e3). doi:10.1136/medethics-2019-105860

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *American Journal of Bioethics, 20*(11), 7-17. doi:10.1080/15265161.2020.1819469

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*(1), 53-60. doi:10.1007/s10676-011-9279-1

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *Lancet Digital Health, 1*(8), E390. doi:10.1016/S2589-7500(19)30197-9

Dennett, D. C. (1978). *Brainstorms* (Fortieth Anniversary Edition. ed.). Boston, MA: MIT Press.

Dennett, D. C. (1989). *The intentional stance*: MIT press.

Dennett, D. C. (1991). *Consciousness explained* (1st ed.). Boston: Little, Brown and Co.

Desai, D., & Kroll, J. (2017). Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law and Technology, 31*(1).

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, (2019). Retrieved from https://data.europa.eu/doi/10.2759/177365

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philosophy & Technology, 33*(3), 523-539.

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: it is possible to trust medical AI. *Journal of Medical Ethics, 47*(6), 437-438.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences, 11*(2), 77-83.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine, 1-2*, 100001. doi:10.1016/j.ibmed.2020.100001

Gille, F., Smith, S., & Mays, N. (2015). Why public trust in health care systems matters and deserves greater research attention. *Journal of Health Services Research & Policy, 20*(1), 62-64. doi:10.1177/1355819614543161

Gille, F., Smith, S., & Mays, N. (2017). Towards a broader conceptualisation of 'public trust' in the health care system. *Social Theory & Health, 15*(1), 25-43. doi:10.1057/s41285-016-0017-y

Grimmelikhuijsen, S. G. (2010). Transparency of Public Decision-Making: Towards Trust in Local Government? *Policy & Internet, 2*(1), 5-35.

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205-211. doi:10.1136/medethics-2019-105586

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527. doi:10.1177/0018720811417254

Hardin, R. (2002). *Trust and trustworthiness*. New York: Russell Sage Foundation.

Hartmann, M. (2020). *Vertrauen : Die unsichtbare Macht*. Frankfurt am Main: Fischer.

Haselager, W. F. G. (2005). Robotics, philosophy and the problems of autonomy. *Pragmatics & Cognition, 13*(3), 515-532. doi:10.1075/pc.13.3.07

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478-481. doi:10.1136/medethics-2019-105935

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine, 25*, 364-373.

Knoops, P. G. M., Papaioannou, A., Borghi, A., Breakey, R. W. F., Wilson, A. T., Jeelani, O., Zafeiriou, S., Steinbacher, D., Padwa, B. L., Dunaway, D. J., & Schievano, S. (2019). A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific Reports, 9*(1), 13597. doi:10.1038/s41598-019-49506-1

Latour, B. (2000). The Berlin key or how to do words with things. In P. M. Graves-Brown (Ed.), *Matter, materiality and modern culture* (pp. 10-21). London: Routledge.

Lauer, D. (2019). Nur liebende Roboter wären vertrauenswürdig. *Deutschlandfunk Kultur*. Retrieved from https://www.deutschlandfunkkultur.de/kommentar-zu-eu-ethik-leitlinien-fuer-kis-nur-liebende.2162.de.html?dram:article_id=446162

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum Factors, 46*(1), 50-80. doi:10.1518/hfes.46.1.50_30392

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report, 49*(1), 15-21. doi:10.1002/hast.973

Luhmann, N. (1968). *Vertrauen; ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart: F. Enke.

Luhmann, N. (1979). *Trust and power* (English ed.). Chichester: Wiley.

Maslow, A. H. (1943). A Theory of Human Motivation. *Psychological review, 50*(4), 370-396.

McLeod, C. (2015). Trust. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Misztal, B. A. (1996). *Trust in modern societies: the search for the bases of social order*. Cambridge: Polity Press.

Möllering, G. (2001). The nature of trust: From Georg Simmel to a theory of expectation, interpretation and suspension. *Sociology, 35*(2), 403-420.

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

O'Neill, O. (2002b). *A Question of Trust. The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

O'Neill, O. (2013). Trust before trustworthiness? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 237-238). Oxford: Routledge.

Ososky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). *Building appropriate trust in human-robot teams.* Paper presented at the 2013 AAAI Spring Symposium Series.

Ras, G., van Gerven, M., & Haselager, W. F. G. (2018). Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. . In H. Jair Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. A. J. van Gerven (Eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 19-36). Cham: Springer.

Ross, C., Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News.* Retrieved from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Hoboken: Pearson.

Sanders, T. L., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A Model of Human-Robot Trust: Theoretical Model Development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 1432–1462. doi:10.1177/1071181311551298

Sayes, E. (2014). Actor–Network Theory and methodology: Just what does it mean to say that nonhumans have agency? *Social Studies of Science, 44*(1), 134-149. doi:10.1177/0306312713511867

Shapin, S. (1995). Trust, Honesty, and the Authority of Science. In R. Bulger, Meyer Bobby, E., Fineberg, H. V. (Ed.), *Society's Choices: Social and Ethical Decision Making in Biomedicine* (pp. 388-408). Washington DC: National Academy Press.

Shin, D. (2020). The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. *International Journal of Human-Computer Studies*. doi:https://doi.org/10.1016/j.ijhcs.2020.102551

Simmel, G. (1908 [1983]). *Soziologie: Untersuchungen über die Formen der Vergesellschaftung*. Berlin: Duncker und Humblot.

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

Terada, K., Shamoto, T., & Ito, A. (2008). *Human goal attribution toward behavior of artifacts.* Paper presented at the RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56. doi:10.1038/s41591-018-0300-7

Tullberg, J. (2008). Trust—The importance of trustfulness versus trustworthiness. *Journal of Socio-Economics, 37*(5), 2059-2071.

Van den Brule, R., Bijlstra, G., Dotsch, R., Haselager, W. F. G., & Wigboldus, D. (2016). Warning signals for poor performance improve human-robot interaction. . *International Journal of Human-Robot Interaction, 5*(2), 69-89

Van den Brule, R., Bijlstra, G., Dotsch, R., Wigboldus, D. H. J., & Haselager, W. F. G. (2013). Signaling Robot Trustworthiness: Effects of Behavioral Cues as Warnings. In G. Herrmann, M. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *5th International Conference on Social Robotics, ICSR 2013* (pp. 583-584). Bristol, UK.

Van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, P. (2014). Do Robot Performance and Behavioral Style affect Human Trust? . *International Journal of Social Robotics, 6*(4), 519-531. doi:10.1007/s12369-014-0231-5

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. doi:10.1371/journal.pmed.1002689

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine, 25*(9), 1337-1340. doi:10.1038/s41591-019-0548-6

Chapter 4: Towards a Pragmatist Dealing with Algorithmic Bias in Medical Machine Learning

Towards a Pragmatist Dealing with Algorithmic Bias in Medical Machine Learning

Georg Starke[1], Eva de Clercq[1], Bernice S. Elger[1,2]

[1] Institute for Biomedical Ethics, University of Basel, Basel, Switzerland. [2] Center for Legal Medicine, University of Geneva, Geneva, Switzerland

Acknowledgement:

GS would like to thank the attending audience at the 2019 annual conference of the European Association of Centers for Medical Ethics at the University of Oxford for their critical and helpful input. The authors would also like to thank Christopher Poppe, Stuart McLennan, and Thomas Grote for their valuable input in personal discussions.

This is a post-peer-review, pre-copyedit version of an article published in *Medicine, Health Care and Philosophy*. The final authenticated version is available online at: https://doi.org/10.1007/s11019-021-10008-5.

Abstract

Machine Learning (ML) is on the rise in medicine, promising improved diagnostic, therapeutic and prognostic clinical tools. While these technological innovations are bound to transform health care, they also bring new ethical concerns to the forefront. One particularly elusive challenge regards discriminatory algorithmic judgements based on biases inherent in the training data. A common line of reasoning distinguishes between justified differential treatments that mirror true disparities between socially salient groups, and unjustified biases which do not, leading to misdiagnosis and erroneous treatment. In the curation of training data this strategy runs into severe problems though, since distinguishing between the two can be next to impossible. We thus plead for a pragmatist dealing with algorithmic bias in healthcare environments. By recurring to a recent reformulation of William James's pragmatist understanding of truth, we recommend that, instead of aiming at a supposedly objective truth, outcome-based therapeutic usefulness should serve as the guiding principle for assessing ML applications in medicine.

Keywords: Artificial Intelligence, Machine Learning, Pragmatism, Philosophy of Science, Algorithmic Bias, Fairness

4.1 Introduction

"Ethics and epistemology are always very closely related, and if we want to understand our ethics, we must look at our epistemology", the British philosopher and novelist Irish Murdoch noted in her early essay *Metaphysics and Ethics* (Murdoch, 1957, p. 113). Her statement rings eminently true with regard to ethical challenges posed by the integration of Artificial Intelligence (AI) into health care. Medical decisions are increasingly aided by recommender systems based on machine learning (ML) that support health care providers, e.g. in choosing an appropriate diagnosis or treatment for their patients. Particularly promising are programs using Deep Learning (DL) based on Artificial Neural Networks (ANN) (Esteva et al., 2019; Topol, 2019b). While much research has been devoted to ML-based diagnostic classifiers, ranging from oncology to psychiatry, recent advances also promise more robust predictive measures of immediate clinical utility. For example, it has been shown that ML-based systems can identify patients suffering from chronic lymphocytic leukaemia (CLL) for whom additional immunosuppression would constitute a major risk for infection (Agius et al., 2020). Another very recent application of ML promises early predictions of circulatory failure for patients in intensive care settings (Hyland et al., 2020) – without doubt of high interest during the Covid-19 pandemic –, and the list of such applications is ever increasing. For these reasons, many expect DL to revolutionize medicine and to constitute a major paradigm-shift in the practice of medicine towards an era of "Deep Medicine" (Topol, 2019a).

By enhancing treatment and freeing time for patient-physician interactions, these new developments have great potential to improve clinical care. Still, they also pose numerous ethical challenges that are narrowly tied to epistemological questions

concerning these programs. Of key concern are the replication and reinforcement of existing discriminatory practices by training ML programs on biased data. As is well documented, bias in medicine is pervasive, whether it is based on unconscious prejudices or rooted in systematically skewed data collection, e.g. through clinical trials carried out predominantly with male participants. In some instances, such biases can be easily detected and countered by appropriate data curation, for instance by assuring an appropriate balancing of male and female training cases. In other instances, such biases remain hidden and may prove impossible to trace, particularly if the target variable of interest, such as a diagnostic category, is based on medical convention.

Following the lead of others, we therefore turn our attention to questions of epistemology (Grote & Berens, 2020), and propose a different approach which takes inspiration from philosophy of science. Following a recent reformulation of William James' pragmatist theory of truth (Chang, 2017), we argue that for some medical contexts, the debate about bias can be improved by shifting the focus of attention beyond the mere correspondence of input and target variables. Instead of clinging to a supposedly objective truth of the training data, the outcome-based clinical utility of any medical ML program should be put to the forefront. The paper proceeds in three steps: first, we introduce the notion of algorithmic bias and provide some salient examples of bias in medicine. We then provide a critique of an understanding of ground truth based on the correspondence theory of truth and suggest an alternative pragmatist reading. Lastly, we show how such an alternative view could be applied to reshape the debate about biases of medical ML. Modifying Box's well-known maxim that all models are wrong, but some are useful (Box, 1976: 792), we propose what one may call James'

maxim: that some models are true precisely because they are useful (James, 1907 [1922]: 204).

4.2 Bias in medical machine learning

Bias has been at the forefront of ethical debates both in ML and in medicine for decades. The word originates from the Old Provençal word *biais*, where it described the behaviour of balls with a greater weight on one-side (Oxford English Dictionary Online, 2020). In consequence, these balls tended to roll systematically in an oblique line into one particular direction and thus shifted the odds of a game. In the modern metaphoric sense, bias similarly describes such one-sided tendencies, usually with regard to decisions that systematically and erroneously favour or disadvantage particular decisions over others. In the context of ML, such biases take many different forms and can stem from various causes but are commonly summarized by the term *algorithmic bias.* Danks and London have suggested a useful taxonomy distinguishing between five different kinds of algorithmic bias, based on where in the design or use of a program the bias occurs (Danks & London, 2017). In the context of healthcare, Thomas Ploug and Søren Holm have recently distinguished between at least three different ways in which bias could lead to discrimination in ML-based diagnostics and treatment planning (Ploug & Holm, 2020). While our discussion here follows examples of algorithmic biases linked to training data, we believe that an outcome-oriented approach could equally address other forms of biases such as algorithmic processing bias, e.g. introduced through the choice of regularization or smoothing parameters (Danks & London, 2017: 4693).

In medicine, practical examples of biases are frequently based on gender or race. They shape a plethora of vital diagnostic and therapeutic decisions, leading for example to the

classic case of missed myocardial infarctions in women which do not show the supposedly typical symptoms of a heart attack prevalent in men (Hobson & Bakker, 2019). Another well-researched case concerns psychiatric decision making in black populations: black US-Americans are much more likely to be diagnosed with schizophrenia when presenting with affective symptoms than their Caucasian peers (Strakowski et al., 2003). White patients presenting with similar symptoms are in turn more likely to be diagnosed (arguably correctly) with mood disorders such as major depression. Partially, this persistent phenomenon of misdiagnosis is thought to arise from socially entrenched biases passed on by clinicians (Gara, Minsky, Silverstein, Miskimen, & Strakowski, 2019). Other examples include widespread misperceptions about pain management in black patients based on erroneous assumptions about physiological differences between black and white patients (Hoffman, Trawalter, Axt, & Oliver, 2016).

The rise of ML in medicine runs risk to exacerbate such biases, since structural racism is known to shape the collection and integration of data as well as the delivery of targeted therapeutic interventions (Genevieve, Martani, Shaw, Elger, & Wangmo, 2020). If, for example, one were to use the historical health records of black schizophrenic patients in the US to train a diagnostic ML program, it would arguably use race as a predictor for its calculations and continue the overdiagnosis of schizophrenia in its recommendations. However, one would not only risk purporting false clinical judgements from the past in the diagnostic ML program. More problematically, if such procedures would be dignified by the common belief in the objectivity of algorithms (Galison, 2019), discriminatory practices will become even more entrenched in medical practice and more difficult to address. Existing biases could become deeply hidden in

the hyperparameters of an ANN, beyond the grasp of human understanding and intervention. Such algorithmic bias would skew the recommendations systematically for one particular group resulting in unfair treatment.

What about instances where we actually *do* want to discern between different socially salient groups though? A different example, where ethnicity also plays a crucial role, may serve as a useful example here. Systemic lupus erythematosus (SLE), a severe autoimmune rheumatological disease which typically affects the skin, but also many other tissues and internal organs, is known to affect more women than men and have a significantly higher prevalence in people of African, Asian or Hispanic descent (Lewis & Jawad, 2017). Similar to schizophrenia, the exact underlying aetiology is, as of now, still unclear, rendering diagnosis rather difficult. With gender and ethnicity being crucial predictors for the occurrence of SLE, it would seem justified to include information on the ethnicity or gender of patients in the training data for a diagnostic program for this disease. In contrast to schizophrenia, such inclusion could be seen as warranted since it accurately mirrors true disparities between socially salient groups.[18]

Unfortunately, in most medical examples the relation between the predictor and the target variable, which shapes the so-called ground truth for an ML algorithm, is difficult to determine since the features of interest are based on medical convention. In such instances, the feature may prove to be somewhat of a shifting target, e.g. due to changing diagnostic classifications over time. Diagnostic categories in psychiatry, which have

[18] Of course, this is not to make any metaphysical claims or advocate a naturalistic understanding of diseases. We merely want to highlight that categories such as gender or ethnicity, intricately related to social and environmental factors, can often serve as useful predictors for diagnostic or therapeutic decisions.

shifted drastically over the past decades, as seen easily by the consecutive revisions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) over the past decades, may serve as a particularly salient example. Here, distinguishing between wrong biases that lead to misdiagnosis and erroneous treatment and justified differential treatment that mirrors true differences seems highly challenging – particularly for the many conditions and treatment options where underlying causal relations remain unclear (London, 2019). Yet, simply leaving out potentially discriminatory labels such as gender or ethnicity as input variables can apparently not solve the problem either. After all, to our best knowledge, ethnicity and gender seem to play a role for diagnostic, therapeutic and prognostic purposes in many diseases, as the example of SLE highlights. So, what may we do about these unclear instances of bias, in lieu of a clear standard against which to measure it?

4.3 Bias and the pragmatist theory of truth

A common strategy to address the problem of bias is to further the transparency of ML models (Mittelstadt, Russell, & Wachter, 2019; Vayena, Blasimme, & Cohen, 2018). The underlying assumption is that greater transparency will render algorithmic bias easier to detect and help understand a program's erroneous decisions, so that one can correct the algorithm's mistakes and avoid bias by curating the input variables accordingly. For many instances this solution can be sufficient, e.g. to identify so-called Clever Hans predictors that base a ML program's classification strategy on irrelevant correlations. A good example for such a misleading predictor is a program basing the classification of an image as "horse" on a source tag in the training images for horses (Lapuschkin et al., 2019). Based on such ill-curated input data, the program will erroneously assume that all future testing images displaying this source tag depict horses, largely independent

from the image's actual content. Increasing a program's transparency, one could identify the source tag as a decisive, yet meaningless factor for the decision-making process, enabling an ex-post correction. Transferred to the clinical example, if an explainable program allows seeing that a diagnosis of schizophrenia is at least partly based on a person's skin colour, anyone commanding trained judgement could notice this as erroneous and account for it.

In medicine, checking a program's decisions is not just a technical challenge though. Returning to the two clinical examples, both schizophrenia as well as SLE constitute heuristic constructs based on a number of diagnostic criteria, while the underlying aetiology remains subject to scientific debate. Put differently, there is no valid gold standard for establishing a ground truth – a wide-spread problem in medicine, that concerns all medical fields, even those with supposedly clear-cut pathological correlates such as oncology (Adamson & Welch, 2019). After all, most biological differences only become meaningful in medicine if they are correlated with symptoms and complaints – a process that is by definition highly conventional and ultimately also pragmatic. How may one distinguish in these instances between irrelevant correlations, shaped by human prejudice and convention, and causally relevant, yet currently unknown, predictors, e.g. based on genetic factors that are more prevalent in certain groups? One seemingly easy remedy to avoid discriminatory practices would be to forego the potentially problematic category altogether. For example, one could simply leave out ethnicity or gender as an input to achieve a non-discriminatory program.

Prima facie, this would safeguard the World Medical Association's Declaration of Geneva, prohibiting considerations of ethnic origin, gender or race to interfere with medical duties (Parsa-Parsi, 2017). However, this approach runs into two major

problems. First, it has been shown to be very challenging to implement, since the category which should not influence the training data may be inferred by other seemingly innocent input data, with ZIP-codes and socio-economic status being amongst the most obvious (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). Second, for specific instances, e.g. unequal distribution of genetic disease predisposing factors in humans from different races, some form of positive discrimination may seem warranted, justifying the inclusion of ethnicity in the training data. Again, the example of SLE can serve as a useful example here. As discussed above, SLE mainly affects populations which constitute minorities in most Western countries. If ethnicity was categorically excluded as a potential input in the training data, it would be more difficult to obtain a correct diagnosis for this vulnerable population. Rendering the diagnosis of a disease such as SLE less accurate in minority populations by disregarding race could easily be regarded as discriminatory.

One solution to this conundrum may lie in taking a step back and looking at the relation between input and output space anew, which as the terminology of the field already indicates, is supposed to be a truth relation. The way such truth is usually constructed assumes the classic understanding of truth, namely the correspondence theory of truth.[19] Commonly ascribed to Aristotle, this theory posits that a proposition "p" is true if and only if it corresponds to some fact. Put differently, according to this theory truth describes a relation between a truth bearer such as a proposition or a judgement and a

[19] Of course, this is not to make any claims concerning the factual epistemological beliefs of ML developers. Many may in fact embrace an instrumental understanding of truth, whether explicitly or implicitly. However, the term ground truth itself and its historical origins in geography and aerial reconnaissance, referring to the physical ground, seem to suggest a relation of correspondence (Gil-Fournier & Parikka, 2020).

truth-maker, such as an observable fact in the empirical world. In the context of ML, the ground truth relation can be similarly described as a mapping of different spaces onto each other.

In its simplest form, such mapping occurs between an observable input space and an intended output or decision space. In addition to these, some authors have proposed adding a so-called construct space in-between these two, capturing unobservable, yet meaningful predictors, as a third mediating space to formally address structural bias (Friedler, Scheidegger, & Venkatasubramanian, 2016). Seemingly, one could also apply their framework to the medical cases at hand here: the input space would contain clinical observations, e.g. symptoms or clinical findings, whereas the decision space would contain the recommended treatment. The construct space could be found in the agreed upon diagnostic criteria that are presumed to be of relevance by the medical community. Unfortunately, for the many and highly relevant cases in medicine where such causal relations between the different spaces continue to be unknown, such mapping remains highly spurious. As long as we do not know, for example, the causal link between brain-based pathology causing psychotic episodes, the presumed diagnostic construct of schizophrenia and the therapeutic mechanism of specific antipsychotic drugs, any such mapping will remain to some extent arbitrary and open to challenge.

However, there are also other ways to construe truth, that look at practices rather than at propositions and which may be better suited to the medical contexts at hand. One suggestive model is the pragmatic theory of truth, the best-known version of which was formulated by William James in 1907 (James, 1907 [1922]). James, who tellingly had received medical training himself, famously stressed the practical value of statements.

Turning against both rationalist and empiricist conceptions, James argued for defining their truth in terms of utility. In his lectures *Pragmatism: A New Name for Some Old Ways of Thinking*, James famously espoused this "instrumental view of truth", describing it as "any idea upon which we can ride, so to speak; any idea that will carry us prosperously from any one part of our experience to any other part, linking things satisfactorily, working securely, simplifying, saving labour" (James, 1907 [1922]: 58). His challenge to the correspondence theory finally culminates in his frequently cited statement that "you can say of it then either that "it is useful because it is true" or that "it is true because it is useful". Both these phrases mean exactly the same thing" (James, 1907 [1922]: 204).

Ever since their publication, these claims have subjected James to myriads of strong criticism, due to his supposedly antirealist stance (Capps, 2019). Notwithstanding this critique, we take it that a pragmatic approach may be worth reconsidering for construing truth in medical ML and, in particular, to address some of the ethical challenges posed by algorithmic bias. However, to do so, it may be more convenient to turn to a contemporary reading of James from the philosophy of science, which already addresses the criticism of James' account. The philosopher of science Hasok Chang, whose work has already been successfully employed to address other challenges in the context of nosology (Kendler & Parnas, 2012), prominently advocates a Jamesian pragmatist model of epistemology in the sciences. Chang reframes James' model to provide an understanding of truth based on operational coherence, rooted in action. While Chang explicitly rejects a correspondence theory of truth, his notion of coherence also "goes beyond consistency between propositions; rather, it consists in various actions coming together in an effective way towards the achievement of one's aims" (Chang, 2017: 109).

Applied to the medical context, such aims can entail simpler tasks such as immobilizing a broken bone with a plaster cast to promote its healing process, or highly complex aims requiring many different actors. The recent development of workable tracing apps to contain the spread of Covid-19 may serve as an example here. Within such given contexts, true statements are those necessary to achieve one's aims. As Chang puts it: "A statement is true in a given circumstance if (belief in) it is needed in a coherent activity" (Chang, 2017: 113). Based on the coherent system in question, different and possibly contradictory statements may have been adopted as true in the history of science insofar as they produced or improved certain kinds of knowledge for particular aims.

Given this historical contingency of science, one may be tempted to disregard the notion of truth in science altogether. While James's original approach may seem to support such a relativist stance, rendering the world dependent upon the interests of its describer (Putnam, 1994: 448), Chang's model of operational coherence does not severe the crucial connection between knowledge and reality in a similar fashion, precisely because it demands to be rooted in empirical facts: "operational coherence cannot be achieved in an arbitrary fashion by decree, wishful thinking, or mere mutual agreement. On the contrary, in order to do things successfully in the world, we need to have an understanding and mastery of our surroundings. It is operational coherence, not the mirage of correspondence, through which the mind-independent world is actually brought to bear on our knowledge" (Chang, 2017: 112).

Leaving more fundamental philosophical questions aside, this implies two crucial practical benefits for its application to medical ML. First, it does not undermine the powerful notion of scientific truth in the public sphere – a notion that seems to be intricately related to public trust in science (Shapin, 1995). Second, it supports retaining

the vocabulary of (ground) truth as a technical term for the necessary pairing of input and output variables (Gil-Fournier & Parikka, 2020), without making overly ambitious claims about medical truths – which as we have seen are frequently subject to contingent conventions. As Chang notes, his approach of "[c]hecking for pragmatic necessity may not live up to some overblown image of a philosophical test, but it is how we get on in science, and in the rest of life too" (Chang 2017: 115). In the following, we will show what this may mean practically in the context of medical ML.

4.4 Bias in medical ML: a pragmatist approach

We argue that a pragmatic understanding of "ground truth" can be highly informative for algorithmic bias in medical ML. Clearly, the overarching aim of the medical community needs to concur with the Ancient Hippocratic idea: the aim of medicine is to work for the benefit of the sick, to cure them or at least make them better. In our opinion, these general ambitions provide a rather clear purpose for our collective epistemic practices – even though the exact determination of its content will be subject to much debate for different applications in different diseases and diverse clinical contexts. To enable an open debate about the clinical utility of particular programs and their potential risks, it is worth trying to consider medical ML in terms of operational coherence guided by specific medical aims. Outcome-based therapeutic usefulness should serve as the guiding principle for their design, not a recourse to a supposedly objective truth based on a static correspondence theory.[20]

[20] With regard to diagnostic hypotheses, Stanley Donald and Rune Nyrup have recently made a similar point drawing on Charles Sanders Peirce, and suggested to conceptualise the diagnostic process as a form of strategic reasoning (2020).

Returning to the two clinical examples of schizophrenia and SLE, we can now apply this model to the context of bias in medical ML. In the case of schizophrenia, it seems clear that the diagnostic practice of US psychiatrists of readily diagnosing their black patients with schizophrenia did not further their well-being but may in fact have resulted in maltreatment and harmful medication and should hence be abolished. In comparison, the case of lupus provides quite a different picture. Here, a differentiation based on ethnicity could contribute to patients' well-being, if it increases diagnostic accuracy resulting in adequate treatment; it would thus be (to some degree) warranted to be included in the construction of ground truth.

There are at least three points of major concern that could be levelled against this position. First, one could argue that a utility-based account of ground truth is not adequate for all medical applications. And indeed, for some instances, mere data curation may be sufficient. When causal links between clinical observation, diagnostic construct and available treatment are clearly established, interpretable or explainable ML models can help to identify misleading or unnecessary input data. As a classical example one could think of diabetes mellitus type 1 (DM1), where the destruction of pancreatic beta-cells provides a clear aetiology that can be linked to clinical observations such as recurrent hyperglycaemia, the diagnostic construct of DM1 with certain predicted measurements under fasting, and the suggested treatment with insulin (Stegenga, 2018: 26). However, as we have shown, this is far from the rule in medicine – not only in specialties with a notoriously challenging nosology such as psychiatry but also in e.g. internal medicine or dermatology. As Alex London has argued, the unknown

aetiology of many medical conditions thus demands a primate of accurate diagnosis and treatment over explainability in the context of medical ML (London, 2019). [21]

A second and potentially more serious problem concerns the measurement and operationalisation of clinical utility.[22] After all, a pragmatist evaluation of medical ML based on clinical utility will need to be based on clear and operationalizable criteria to avoid an infinite circle and yield a useful guide for developers and regulating bodies. At the same time, one should also aim to avoid an overly prescriptive and potentially paternalistic definition of clinical utility, without patient involvement. The increasing use of short- and long-term Patient-Reported Outcome Measures (PROMs) aims to address this conundrum (McClimans, 2010), but relies on inherently subjective criteria (Alexandrova, 2017: 135-138). Of course, this is a problem that not only applies to medical ML, but evidence-based medicine more generally, and defies a simple and general answer. In consequence, heading Jacob Stegenga's advice that "(t)he instruments employed in clinical research should measure patient-relevant and disease-specific parameters" (Stegenga, 2015: 62), it may be fruitful to return to the concrete examples of SLE and schizophrenia.

In the case of schizophrenia, the very construct of the disorder, as laid out in ICD-10 and DSM5, largely relies on clinical observations by the attending psychiatrist, which are based on verbal self-reports from the patient. Outcome measures of schizophrenia are thus always multi-faceted attempts to grasp this complex reality – including neurobiological measures, drop-out from antipsychotic treatment, hospitalisations,

[21] While we agree with London (2019), who recommends prioritizing accuracy over explainability in the context of medical ML, we believe that a pragmatic focus on clinical utility may be better suited to stress the value-ladenness of ML systems as well as their embeddedness in a pragmatic context.
[22] We would like to thank two anonymous reviewers for their help in making this point more explicit.

structured symptom scales and patient-reported outcomes such as personal well-being – and have changed drastically since the disorder was first described by Kraepelin in 1896 (Burns, 2007). In addition, since the course of the disorder seems to be influenced heavily by social context (Leff, Sartorius, Jablensky, Korten, & Ernberg, 1992), outcome measures need to be adapted to specific contexts. In comparison, SLE seems to pose fewer problems, with standardized and congruent diseases activity scores, based on clinical observations such as seizures and objectifiable measures like proteinuria (American College of Rheumatology, 2004). Similarly, standardized PROMs for SLE have successfully been adapted for different cultural contexts (Bourré-Tessier et al., 2013; Kaya et al., 2014; Navarra et al., 2013), so the utility of an intervention for SLE could tentatively be evaluated based on a combination of these instruments. Still, as these examples highlight, choosing an appropriate outcome-measure will be context- and disease-specific, and always open for debate – in the context of ML as much as for other medicinal products. It is thus crucial that studies are explicit about their operationalization and measurement of clinical benefit, to allow patients and physicians to arrive at an informed choice regarding their individual use.

A third challenge relates directly to ethics. If we are to follow clinical utility as the single most important criterion for the evaluation of medical applications of ML, this could be misinterpreted as a call for a simplistic reading: that maximizing the benefit for the majority of patients justifies disregarding the needs of a potentially vulnerable minority. Our approach is different insofar as we deem it necessary to embrace explicit criteria for algorithmic fairness, derived from moral philosophy. We consider John Rawls' difference principle to be a potential contender, which prioritizes the well-being of those who are worst off (Rawls, 1999: 132-134).

Such a principle may be enacted by constrained optimization algorithms that maximize clinical utility but also need to satisfy other set conditions, based on evaluations calculated separately for various subgroups (Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017). There is extensive research from the area of fair ML that demonstrates how Rawls' theory of justice can be practically incorporated (Lundgard, 2020), e.g. as a constraint for classification (Jabbari, Joseph, Kearns, Morgenstern, & Roth, 2017; Joseph, Kearns, Morgenstern, Neel, & Roth, 2016) or as loss minimization (Hashimoto, Srivastava, Namkoong, & Liang, 2018). Of course, implementing a fairness constraint for ML algorithms requires intricate ethical judgements, e.g. concerning who counts as worse-off than others, a point which will often be contentious. In addition, there may be good reasons to implement fairness constraints that go beyond Rawls, "artificial intelligence's favorite philosopher" (Procaccia 2019, cit. in Lundgard, 2020: 3). For instance, in many ML applications ensuring the benefit of the patient will require a careful evaluation of different layers of vulnerabilities, as Paolo Corsico has recently argued with view to psychosis (2020). In the medical context, such approaches could translate to regulatory rules that demand tests whether an ML program performs worse in ethnic minorities, in terms of clearly defined outcome-measures, and denies approval to those which do.

4.5 Lessons for the evaluation of medical ML

For the design of medical ML programs, developers should thus focus on ex-post corrections of particular ML programs in medicine and evaluate a program's performance based on the relative treatment outcome within certain vulnerable

populations.[23] The examples of schizophrenia and SLE highlight this. Clearly, the pragmatic benchmark of a ML-based diagnostic program would be the treatment success that results from applying it to patients. Let us consider two options: (1) implementing a ML program designed to be blind to ethnicity and (2) designing the ML algorithm in a way that it explicitly or implicitly incorporates ethnicity as input variable in the training data. Embracing a pragmatic approach, the decision for using program 1 or 2 would focus on the clinical results which either program brings about. If, for example, algorithm 2 results in better outcomes in both black and white populations, than the differential treatment would be useful and hence, in the pragmatic sense, true. Based on our current knowledge, one would expect to find this result for the case of SLE. However, if algorithm 1 results in better treatment outcomes in both groups, than a differential treatment is apparently harmful, biased and should be disregarded, as may be the case in schizophrenia.

Still, based on the concept of operational coherence, these a priori assumptions require empirical testing. After all, one could similarly envision a contrary case, where an algorithm explicitly taking into account ethnicity performs better in terms of fairness for diagnosing schizophrenia. For example, depending on the design, ethnicity could be used as a correcting factor that counters the known overdiagnosis of schizophrenia among black patients. Here, transparency will be key for a critical reassessment of the assumptions underlying each particular program. Still, the choice of ML program will ultimately need to be adjudicated by its tangible clinical benefit.

[23] With regard to diagnostic decisions based on ML, we take it that these will also largely affect treatment outcomes since they determine the indication of therapeutic intervention.

While this may run counter to preferences in the machine learning community to focus on ex-ante mechanisms to ensure fairness, such an approach has been proven to be highly efficient in addressing discriminatory behaviour of algorithms, based for example on gender stereotypes (Zhao, Wang, Yatskar, Ordonez, & Chang, 2017). In medicine, some form of such ex-post tests on fairness could be integrated in clinical trials, conducted so that a specific program receives approval by regulatory bodies such as the US Food and Drug Administration (FDA) (He et al., 2019). This would also imply that both short- and long-term outcome of the ML system are tested and its safety and utility evaluated in different phases, transitioning from few healthy volunteers to large clinical trials in the target population (Paulus, Huys, & Maia, 2016).

We thus believe that a pragmatist approach focusing on a program's output would also constitute a viable and realistic way to address disparities for medical applications where ex-ante considerations are potentially impossible due to limited etiological knowledge and the often-conventional nature of medical practice. If we thereby move closer to accepting that also ML will replicate and not remove the shifty and often pragmatic ground of medicine, this could be a safeguard to avoid an overselling of the promises of medical ML. Such a viewpoint may further render us humbler and more willing to accept the epistemic limitations and historical contingency of much contemporary medical knowledge (Stegenga, 2018: 185-187). Thus, instead of focusing on potentially fruitless nosological speculations, we should instead try and privilege a focus on operational coherence, centred around the most crucial criterion in the medical domain: the betterment of the patient.

4.6 Conclusion

In this paper, we have argued for a pragmatic construction of truth in the context of supervised medical ML. Following two clinical examples with unknown etiological underpinnings, we have defended a position that stresses the importance of rigorous ex-post tests for medical ML programs to tackle harmful biases. Instead of aiming for a potentially unobtainable objective truth, developers, clinicians and regulators should pragmatically focus on clinical utility for specific socially-salient groups when evaluating the fairness of a ML system – as well as the many other ethical and value-laden considerations that Char, Abràmoff, and Feudtner (2020) have recently identified, such as: who devises these programs, based on which assumptions, and with which aims? If a pragmatist account of bias can help to clear the view for such questions, this may be all the more reason to embrace it.

4.7 References

Adamson, A. S., & Welch, H. G. (2019). Machine Learning and the Cancer-Diagnosis Problem - No Gold Standard. *New England Journal of Medicine, 381*(24), 2285-2287. doi:10.1056/NEJMp1907407

Agius, R., Brieghel, C., Andersen, M. A., Pearson, A. T., Ledergerber, B., Cozzi-Lepri, A., Louzoun, Y., Andersen, C. L., Bergstedt, J., & von Stemann, J. H. (2020). Machine learning can identify newly diagnosed patients with CLL at high risk of infection. *Nature communications, 11*(1), 1-17.

Alexandrova, A. (2017). *A philosophy for the science of well-being*. Oxford: Oxford University Press.

Bourré-Tessier, J., Clarke, A. E., Mikolaitis-Preuss, R. A., Kosinski, M., Bernatsky, S., Block, J. A., & Jolly, M. (2013). Cross-cultural validation of a disease-specific patient-reported outcome measure for systemic lupus erythematosus in Canada. *The Journal of rheumatology, 40*(8), 1327-1333.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association 71*(356), 791-799.

Burns, T. (2007). Evolution of outcome measures in schizophrenia. *British Journal of Psychiatry Supplement, 50*, s1-6. doi:10.1192/bjp.191.50.s1

Capps, J. (2019). The Pragmatic Theory of Truth. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 ed.). Retrieved from https://plato.stanford.edu/archives/sum2019/entries/truth-pragmatic

Chang, H. (2017). Operational Coherence as the Source of Truth. *Proceedings of the Aristotelian Society, 117*(2), 103-122.

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *American Journal of Bioethics, 20*(11), 7-17. doi:10.1080/15265161.2020.1819469

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806. doi:10.1145/3097983.3098095

Corsico, P. (2020). Psychosis, vulnerability, and the moral significance of biomedical innovation in psychiatry. Why ethicists should join efforts. *Medicine, Health Care and Philosophy, 23*(2), 269-279. doi:10.1007/s11019-019-09932-4

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4691-4697.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24-29. doi:10.1038/s41591-018-0316-z

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Galison, P. (2019). Algorists Dream of Objectivity In J. Brockman (Ed.), *Possible Minds: 25 Ways of Looking at AI* (pp. 231-239). New York: Penguin Press.

Gara, M. A., Minsky, S., Silverstein, S. M., Miskimen, T., & Strakowski, S. M. (2019). A Naturalistic Study of Racial Disparities in Diagnoses at an Outpatient Behavioral Health Clinic. *Psychiatric Services, 70*(2), 130-134. doi:10.1176/appi.ps.201800223

Genevieve, L. D., Martani, A., Shaw, D., Elger, B. S., & Wangmo, T. (2020). Structural racism in precision medicine: leaving no one behind. *BMC Medical Ethics, 21*(1), 17. doi:10.1186/s12910-020-0457-8

Gil-Fournier, A., & Parikka, J. (2020). Ground truth to fake geographies: machine vision and learning in visual practices. *AI & SOCIETY*. doi:10.1007/s00146-020-01062-3

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205-211. doi:10.1136/medethics-2019-105586

Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. *Proceedings of the 35th International Conference on Machine Learning, 80*, 1929-1938.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*(1), 30-36.

Hobson, P., & Bakker, J. (2019). How the heart attack gender gap is costing women's lives. *British Journal of Cardiac Nursing 14*(11), 1-3.

Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences of the United States of America, 113*(16), 4296-4301. doi:10.1073/pnas.1516047113

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine, 25*, 364-373.

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2017). Fairness in reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning,, 70*, 1617-1626.

James, W. (1907 [1922]). *Pragmatism: A New Name for Some Old Ways of Thinking*. New York: Longmans, Green & Co.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559, 1*(2).

Kaya, A., Goker, B., Cura, E. S., Tezcan, M. E., Tufan, A., Mercan, R., Bitik, B., Haznedaroglu, S., Ozturk, M. A., & Mikolaitis-Preuss, R. A. (2014). Turkish lupusPRO: cross-cultural validation study for lupus. *Clinical rheumatology, 33*(8), 1079-1084.

Kendler, K. S., & Parnas, J. (2012). Epistemic iteration as a historical model for psychiatric nosology: promises and limitations. In K. S. Kendler & J. Parnas (Eds.), *Philosophical issues in psychiatry II: Nosology* (pp. 303-322). Oxford: Oxford University Press.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications, 10*(1), 1096. doi:10.1038/s41467-019-08987-4

Leff, J., Sartorius, N., Jablensky, A., Korten, A., & Ernberg, G. (1992). The International Pilot Study of Schizophrenia: five-year follow-up findings. *Psychological Medicine, 22*(1), 131-145. doi:10.1017/s0033291700032797

Lewis, M. J., & Jawad, A. S. (2017). The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. *Rheumatology, 56*(suppl_1), i67-i77. doi:10.1093/rheumatology/kew399

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report, 49*(1), 15-21. doi:10.1002/hast.973

Lundgard, A. (2020). Measuring justice in machine learning. *arXiv preprint arXiv:2009.10050*.

McClimans, L. (2010). A theoretical framework for patient-reported outcome measures. *Theoretical Medicine and Bioethics, 31*(3), 225-240. doi:10.1007/s11017-010-9142-0

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *FAT* '19: Proceedings of the conference on fairness, accountability, and transparency*, 279-288. doi:10.1145/3287560.3287574

Murdoch, I. (1957). Metaphysics and Ethics. In D. Pears (Ed.), *The Nature of Metaphysics* (pp. 99-123). London: Macmillan.

Navarra, S., Tanangunan, R., Mikolaitis-Preuss, R., Kosinski, M., Block, J., & Jolly, M. (2013). Cross-cultural validation of a disease-specific patient-reported outcome measure for lupus in Philippines. *Lupus, 22*(3), 262-267.

Oxford English Dictionary Online. (2020). bias, adj., n., and adv. In *Oxford English Dictionary Online*. Oxford: Oxford University Press.

Parsa-Parsi, R. W. (2017). The Revised Declaration of Geneva: A Modern-Day Physician's Pledge. *JAMA, 318*(20), 1971-1972. doi:10.1001/jama.2017.16230

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A Roadmap for the Development of Applied Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 386-392. doi:10.1016/j.bpsc.2016.05.001

Ploug, T., & Holm, S. (2020). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care and Philosophy, 23*(1), 107-114. doi:10.1007/s11019-019-09912-8

Putnam, H. (1994). Sense, nonsense, and the senses: An inquiry into the powers of the human mind. *Journal of Philosophy, 91*(9), 445-517.

Rawls, J. (1999). *A Theory of Justice. Revised Edition*. Cambridge: Harvard University Press.

Rheumatology, A. C. o. (2004). The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials: measures of overall disease activity. *Arthritis & Rheumatology, 50*(11), 3418-3426. doi:10.1002/art.20628

Shapin, S. (1995). Trust, Honesty, and the Authority of Science. In R. Bulger, E. Meyer Bobby, & H. V. Fineberg (Eds.), *Society's Choices: Social and Ethical Decision Making in Biomedicine* (pp. 388-408). Washington, DC: National Academy Press.

Stanley, D. E., & Nyrup, R. (2020). Strategies in Abduction: Generating and Selecting Diagnostic Hypotheses. *Journal of Medicine and Philosophy, 45*(2), 159-178. doi:10.1093/jmp/jhz041

Stegenga, J. (2015). Measuring effectiveness. *Studies in History and Philosophy of Biological and Biomedical Sciences, 54*, 62-71. doi:10.1016/j.shpsc.2015.06.003

Stegenga, J. (2018). *Medical Nihilism*. Oxford: Oxford University Press.

Strakowski, S. M., Jr, P. E. K., Arnold, L. M., Collins, J., Wilson, R. M., Fleck, D. E., Corey, K. B., Amicone, J., & Adebimpe, V. R. (2003). Ethnicity and diagnosis in patients with affective disorders. *Journal of Clinical Psychiatry, 64*(7), 747-754. doi:10.4088/jcp.v64n0702

Topol, E. J. (2019a). *Deep medicine : how artificial intelligence can make healthcare human again*. New York: Basic Books.

Topol, E. J. (2019b). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56. doi:10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. doi:10.1371/journal.pmed.1002689

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Chapter 5: Karl Jaspers and Artificial Neural Nets: On the Relation of Explaining and Understanding Artificial Intelligence in Medicine

Karl Jaspers and Artificial Neural Nets: On the Relation of Explaining and Understanding Artificial Intelligence in Medicine

Georg Starke[1,2], Christopher Poppe[1]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland; [2] College of Humanities, École Polytechnique Fédérale de Lausanne, Switzerland

This version of the article has been accepted for publication in *Ethics and Information Technology*, after peer review, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s10676-022-09650-1

Abstract

Assistive systems based on Artificial Intelligence (AI) are bound to reshape decision-making in all areas of society. One of the most intricate challenges arising from their implementation in high-stakes environments such as medicine concerns their frequently unsatisfying levels of explainability, especially in the guise of the so-called black-box problem: highly successful models based on deep learning seem to be inherently opaque, resisting comprehensive explanations. This may explain why some scholars claim that research should focus on rendering AI systems understandable, rather than explainable. Yet, there is a grave lack of agreement concerning these terms in much of the literature on AI. We argue that the seminal distinction made by the philosopher and physician Karl Jaspers between different types of explaining and understanding in psychopathology can be used to promote greater conceptual clarity in the context of Machine Learning (ML). Following Jaspers, we claim that explaining and understanding constitute multi-faceted epistemic approaches that should not be seen as mutually exclusive, but rather as complementary ones as in and of themselves they are necessarily limited. Drawing on the famous example of Watson for Oncology we highlight how Jaspers' methodology translates to the case of medical AI. Classical considerations from the philosophy of psychiatry can therefore inform a debate at the centre of current AI ethics, which in turn may be crucial for a successful implementation of ethically and legally sound AI in medicine.

5.1 The promises of artificial intelligence for medicine

The integration of artificial intelligence (AI) seems bound to reshape the practice of medicine (Topol, 2019). Due to the convergence of Big Data, increased computational capacities and the rise of deep learning, a new generation of AI systems promises vast improvements, from new research approaches to their clinical implementation at the bedside. While for some authors the current hype of AI creates a danger of bringing about a new *AI winter*, i.e., a period of decreased interest and funding (Müller, 2020; Floridi 2020), the underlying technology may still usher in an age of *Deep Medicine,* given its tangible successes (Topol, 2019). After all, AI can provide tools that improve clinical outcomes across disparate medical specialties, from dermatology (Esteva et al., 2017) to pathology (Campanella et al., 2019), from intensive care (Hyland et al., 2020) to plastic surgery (Knoops et al., 2019) and psychiatry (Bzdok & Meyer-Lindenberg, 2019). Questions concerning the ethical and responsible design and use of medical AI are thus of high urgency and importance.

One major challenge to the implementation of AI in high-risk settings such as medicine lies in the lack of explainability of many current AI systems in healthcare (Vayena et al., 2018; Amann et al., 2020). This challenge results from the opacity of AI models, which in particular deep learning models exhibit (Burrell, 2018). Explainability seems of crucial instrumental value to foster trust in AI systems, to correct a model's errors and to enable vital ethical aspirations like informed consent. Accordingly, ethical guidelines for the implementation of AI have even granted explainability a place alongside the four influential principles of biomedical ethics by Beauchamp and Childress, complementing beneficence, non-maleficence, respect for autonomy and justice (Floridi et al, 2018;

Beauchamp & Childress, 2019). In addition, as the European General Data Protection Regulation (GDPR) highlights, explainability does not constitute a mere ethical recommendation but has become a legal requirement in some jurisdictions and is seen as a part of fundamental rights (Wachter et al., 2017).

Yet, despite its importance, exact, formal definitions of explainability are scarce and often differ across research domains (Adadi & Berrada, 2018). Mittelstadt and colleagues (2019) and Durán (2021) have examined the notion of explainability cautiously with regard to the philosophy of science, situating it in the broader context of scientific explanations. However, as Páez (2020) has convincingly argued, explanations resting on full model transparency which would allow to answer counterfactual questions run into severe and potentially insurmountable problems. Hence, the complexity of a model renders certain types of AI inherently opaque to causal explanations. While this may not preclude epistemically more modest explanations for specific, single decisions of an ML system, it still seems worth turning to a scientific tradition that has long struggled with the problem of explaining phenomena that defy full mechanistic explanation, namely philosophy of psychiatry.[24] In particular, we argue that Karl Jaspers' seminal framework of explaining and understanding in psychopathology provides a rich conceptual background that can be fruitfully adapted to address the challenges posed by current AI systems developed for medical purposes.

Our argument proceeds in five steps. First, we provide a short primer on current debates about the explainability of AI, highlighting its limits. Second, we turn to Jaspers, elaborating the elements of his theoretical framework for the debate at hand. In a third

[24] In the same vein, Páez also turns to a distinction derived from psychology between functional and mechanistic understanding to advance his argument (Lombrozo & Gwynne, 2014).

step, we argue why, psychopathology can serve as a model to develop a framework of explaining and understanding AI, and fourth, why applying a model from psychopathology to AI is warranted, despite the danger of anthropomorphism. Finally, bringing together these considerations, we suggest a framework for understanding and explaining medical AI inspired by Jaspers. We conclude by drawing on examples of medical AI to highlight the practical and ethical implications of our approach.

5.2 The challenge of explainable AI systems in medicine

Rendering AI systems explainable is commonly regarded as crucial for their successful implementation. Consequently, the development of explainable AI (XAI) takes centre stage in myriads of research efforts worldwide (Adadi & Berrada, 2018). Explainability has the instrumental value enabling crucial epistemic and ethical goals (Floridi et al., 2018). On the epistemic side, by allowing closer scrutiny of a system's decisions, XAI promises developers, regulators, and end users the possibility to spot systematic mistakes, correct erroneous decisions and improve the system's performance. In turn, these properties promote important ethical aims, such as fostering informed consent, accountability and avoiding discriminatory biases.

In clinical settings, the degree of a system's explainability may also have important consequences for the complex web of relations between software developers, regulatory bodies, physicians, and patients (Amann et al., 2020). For example, explainability is not only crucial for obtaining informed consent, which requires at least some minimal standards of knowledge, but is also a vital property for promoting trust in a specific system (Diprose et al. 2020). Furthermore, from the perspective of patients, some degree of explainability is required to be able to contest an AI's diagnostic decision – an

important ethical desideratum, rooted in the patients' right to defend themselves against harm (Ploug & Holm, 2020).

Unfortunately, the opacity of AI systems often resists simple explanations. Besides intentionally created secrecy measures within a program, opacity can come in the guise of technical illiteracy on the side of its users or as a system's property, necessarily following from its design and use (Burrell, 2016). Here, we are only interested in the latter. Such necessary opacity, commonly addressed as black-box problem in AI ethics, is particularly prevalent in deep learning models based on artificial neural nets (ANN). To some extent, this opacity may constitute a necessary characteristic of the program, following directly from an architecture with multiple hidden layers and a huge number of weights, optimized with vast and complex training data containing multiple features.

At the moment, approaches to increase an AI system's explainability often focus on visualizations, providing e.g. a heat or saliency map for a program's decision. As Mittelstadt and colleagues. (2019) have succinctly pointed out though, such approaches fall short of common human expectations towards a meaningful explanation, characterized by their contrastive, social, and selective nature. In the same vein, Páez (2020) has argued in favour of a pragmatic turn that cedes unrealistic attempts aiming at full causal explainability in favour of interpretative models that are easily accessible to the intended users. Within the specific context of medicine, Alex London has famously taken an even more provocative approach by arguing that we should prioritize the diagnostic or predictive accuracy of an AI system over its explainability (London, 2019). Similarly, we also agree with the view advocated by Durán and Jongsma (2021) that reliable, yet opaque black box algorithms can provide trustworthy tools for improving medical care.

Yet, given the ethical and epistemic importance of explainability, it would seem prudent to aim for a framework that retains the important aspirations ingrained in the project of rendering medical AI explainable wherever possible. As in other ML systems, explainability would comprise both ex-ante considerations, that focus on the input to a particular program, and ex-post evaluations, scrutinizing the output of a trained algorithm (Braun et al., 2020). Furthermore, in the specific context of medicine, explainability will also need to take into account the complex relation between physician, patient, and ML system, e.g., because physicians need to explain a decision to their patients (Braun et al., 2020). To enable successful forms of such communication and thereby establish the necessary preconditions for trust in a particular program, it will, as argued elsewhere, be crucial to not merely disclose information but render them intelligible, accessible, and assessable to the concerned parties (blinded for peer review).

These theoretical considerations are also supported empirically, e.g., by a recent survey among 170 physicians in New Zealand which confirmed that physicians' understanding of a ML model, their ability to explain the program's output to their patients and their trust in using it are indeed related to each other (Diprose et al., 2020). In light of these findings, it seems advisable to address the particular challenges of medical ML through a lens which not only discerns between different notions of explaining and understanding but relates them to each other in a systematic manner. As we will show in the following, Karl Jaspers' methodological groundworks in psychopathology offers this very kind of framework.

5.3 Karl Jaspers: explaining and understanding

In his seminal *Allgemeine Psychopathologie* (AP) from 1913 (cited in the 4th edition; Jaspers, 1946), Jaspers famously distinguished between different approaches to address

the epistemic difficulties of dealing with the inner life of his patients. Crucial to his writings is the distinction between explaining and understanding. This classic distinction drew on debates about methodological differences between humanities and natural sciences spearheaded by the German philosopher Wilhelm Dilthey in the late 19th century, who famously declared: "Nature we explain, but psychic life we understand" (1894, p. 144, quoted in Kumazaki, 2013). It also relates to Wilhelm Windelband's distinction between "nomothetic" and "idiographic" empirical sciences, with the former seeking "the general in the form of a law of nature", and the latter seeking "the particular in the form of the historically defined structure" (Windelband, 1980 [1894], p. 175).

Expanding on this framework, Jaspers developed a systematic approach encompassing a multi-faceted attempt to integrate subjective and objective phenomena and inferences, which comprised three consecutive steps. According to Jaspers, any attempt of explaining or understanding first needs to fully grasp the relevant facts (Jaspers, 1913, p. 22f.), that encompass both objective and subjective data. For an objective psychopathological assessment, the evaluation draws on outward observations and quantifiable data such as persons' interaction with their environment or their quantifiable performance in memory assessment (Jaspers, 1946, p. 130). Ideally, such objective assessment would imply that the clinician refrains from all theoretical and personal prejudices and presuppositions, relying for example on objective measures such as established psychometric scales, allowing for interindividual comparisons. In contrast, to take stock of the subjective facts of the inner life of a person such as their lived experience of a delusion, Jaspers suggests a 'phenomenological' approach, loosely based on Edmund Husserl's phenomenology, attempting to grasp an individual's own

perspective of their lived experience. As Jaspers describes the method with regard to patients in his psychopathology:

> "The task of phenomenology is to visualize the mental states that the sick really experience, to look at them according to their relationship, to limit them as sharply as possible, to distinguish between them and to assign them fixed terms." (Jaspers, 1946, p. 47) [25]

Jaspers himself calls this phenomenological realization and envisionment of a psychological state "static understanding" (Jaspers, 1946, p. 24). It should be noted though that this is not the kind of understanding in which we are interested here.

Having taken stock of the 'factual data', the psychopathologist then needs to make sense of these fragmentary data by investigating the relations between them (Jaspers, 1946, p. 23). Jaspers proposes two ways, and it is here that we finally encounter the distinction between understanding ("verstehende Psychologie") and explaining ("erklärende Psychologie") that is of interest to our argument.

> "We need to draw a distinction between these relations that is just as fundamental as the distinction between subjective psychopathology (phenomenology) and objective psychopathology. 1. By putting ourselves into the psychic situation, we *understand genetically* how one psychic event emerges from another. 2. By objectively linking several factual data into regularities based on repeated experiences, *we explain causally*." (Jaspers 1946, p. 250)

[25] Here as in the following, translation from the German original is provided by the authors.

For Jaspers, *explaining* therefore hinges on identifying a clear causal connection between cause and effect, and is commonly rooted in biology. According to Jaspers, establishing such an *explanatory* relation allows to formulate a rule that is valid for similar instances (Jaspers, 1946, p. 251). Such explanations therefore closely correspond to the methodological approach of the natural sciences, which according to Jaspers only investigate genuine causal relations (ibid.).[26]

In contrast, and going beyond the scope of natural science, subjective *understanding* concerns itself with comprehensible, meaningful relations that are related to personality and biography. It establishes meaningful connections by drawing on the psychopathologists' own inner experiences, resulting in a "direct evidence that we cannot trace back any further" (Jaspers 1946, p. 252). The evidence of these understandable relations is not based on genuine causal explanations but rather on psychological plausibility, and is achieved by contemplating mental life (Jaspers, 1946, p. 48). Jaspers calls such understanding "genetic", to distinguish it from the "static understanding" mentioned above (Jaspers, 1946, p. 252). Since we are only interested in this form of understanding, we will neglect the qualification as "genetic" in the following. In a nutshell, Jaspers proposes a model of psychopathology that offers a subjective and an objective approach both for the gathering of factual data and for establishing meaningful relations between them. The psychopathologist first needs to gather all relevant observations from their patient, including the patient's subjective mental state as well as their objective environment and biological state. Having brought both together in a full description, there are then two ways to establish meaningful relations

[26] It should be noted that Jaspers' original model from 1913 predates the vast philosophical debates concerning scientific explanations that take their cue from Carl Hempel's Deductive-Nomological Model from 1942 (Woodward 2021).

between them, either through subjective ("genetic") understanding or objective explanations. A schematic depiction of the complementary subjective and objective approaches is provided in Figure 1, to give a succinct overview over Jaspers' terminology.

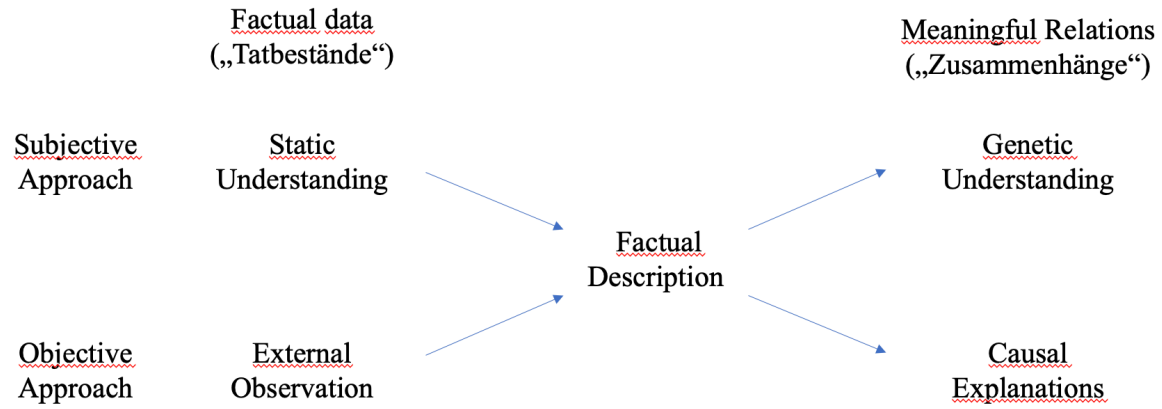

Fig. 5.1: Schematic representation of the subjective and objective evaluation in Jaspers' psychopathological approach.

Jaspers' model has been subject to fundamental criticism, including a recent call to give up the distinction between understanding and explaining in psychiatry altogether (Gough, 2021). More important to our argument, however, are attempts to disentangle the notion of causality in the context of Jaspers' distinction. For instance, many current scientific claims would possibly not fall under Jaspers' rigorous definition of explainability, as long as causal relations remain unclear.[27] Drawing on the writing of Elizabeth Anscombe, Christoph Hoerl (2013) has therefore suggested to describe both explaining and understanding in terms of causality, but with an important difference: Explaining provides "general causal claims linking types of events", whereas understanding "is concerned with singular causation […] – i.e. with the

[27] We would like to thank one of our anonymous reviewers for pointing this out.

particular way in which one psychic event emerges from or arises out of another on a particular occasion." (Hoerl 2013, p. 111)

This reading, distinguishing between general causal claims and singular causation, in fact mirrors Jaspers' own distinction between two different kinds of causality that seems in line with Husserl's distinction of volitional and natural causality (Spano, 2021; Husserl, 2020), yet sometimes renders Jaspers' arguments seemingly contradictory. While causal relations in the strict sense are, according to Jaspers, only to be found (Jaspers 1946, p. 250) in the objectifiable outward observations of the natural sciences, he sometimes also employs a notion of causality that grasps the understandable subjective phenomena:

> One has also called the intelligible connections of the mental *causality from within*, and thus denoted the unbridgeable abyss that exists between these merely parabolically causal connections and the genuine causal connections, the *causality from without*." (Jaspers 1946, p. 250)

If we follow this distinction by Jaspers and Hoerl's interpretation of it, we take it that there are important lessons to derive from his model for current debates about explaining and understanding AI.[28]

5.4 Dealing with the artificial black box: Explaining and Understanding AI

Models of explaining and understanding developed for dealing with human psychopathology may provide a promising approach to address the challenges of black-box AI systems and can elucidate how human users can attempt to make sense of an AI's behaviour in two different, yet complementary ways. Going back to Jaspers' framework,

[28] Jaspers' methodological convictions changed in the course of his life, and he moved away from his strict methodological dualism later in life (Schlimme et al 2012). We still rely on this early model here since it seems most instructive with regard to medical ML models.

we may take a new look at the problem of opacity. In accordance with Jaspers, we can distinguish two steps, the gathering of factual data and the establishment of relations between these data.

For the first step, we can distinguish between objective and subjective data. Objectively, we can observe the AI's *behaviour* by rigorous testing. Like with Jaspers, this objective stock taking should cover at least three different areas: (1) the AI's performance, measured e.g. by the accuracy of an AI's predictions, (2) its interaction with the world, measured e.g. by its behaviour in different settings, and (3), if applicable in instances such as the Deep Learning-based language model GPT3, the AI's *work*. On the subjective side based on phenomenology, our options for assembling factual data are necessarily limited: [29] We cannot grasp an ML models own perspective of their operation, unless we assume that the other mind is characterized to a large degree by human-likeness and has a similar capacity for consciousness (Shanahan, 2016). At least current AI models seem to lack both, barring us from a phenomenological *Vergegenwärtigung* of the machine mind. Here, Jaspers' model does therefore not offer any new insights.

However, we believe that Jaspers can contribute to a finer-grained analysis when it comes to the second step, aimed at establishing meaningful relations between factual data. It is here that we find room for Jaspers' distinction between understanding and explaining. The scope of explaining is in line with the many approaches of explainable AI that aim to establish general causal claims, in the sense of a "causality from without". Current approaches that e.g. use visualizations of weights given to specific factors to provide an "explanation interface" accessible to domain experts point in this direction

[29] To some extent, this is of course also true with view to the mind of other human beings, with the crucial difference that we are familiar with at least one human mind from an inward perspective: our own.

(Holzinger et al., 2019). On an even more fundamental level, ML attempts to provide causal models by learning causal mechanisms would satisfy Jaspers' model here (Schölkopf et al., 2012; Parascandolo et al., 2018). However, as outlined above, causal explanations are only available to a limited extend in current machine learning practice, especially when it comes to deep learning.

Like in psychopathology, we should therefore embrace a two-pronged strategy to make sense of opaque machine learning models, based on both understanding and explaining, on causality from within and from without. In this sense, understanding should be conceptualized as a valuable complementary route to explainability, allowing us to identify meaningful, comprehensible relations, that may become immediately evident to us. An example by Jaspers himself may highlight how understanding can provide epistemic evidence. When examining the evidence of understanding, Jaspers refers to Nietzsche's use of genealogy, especially his Genealogy of Morality: "When Nietzsche's shows us convincingly how being aware of our own frailty, wretchedness, and suffering gives rise to about moral demands and religions […] we experience an immediate evidence that we cannot trace back any further." We understand the relation Nietzsche construes *evidently*.

Similarly, we may understand certain observable behaviours of machine learning models by examining its *genealogy* and its training history. Emily Denton and colleagues have recently suggested this approach with view to the history of the ImageNet database (2021). Furthermore, if we engage in a form of intentional anthropomorphizing and follow the analogy of machine *learning*, we can also *understand* certain features by comparing the machine's learning to our own learning processes. For instance, we could infer from our own learning processes that an AI can only base its decisions and

recommendations on its past experiences – similarly to training medical staff receives, improving their clinical decision making through experience over time: a diagnostic tool trained to distinguish photographs of (malign) melanoma and (benign) naevi may perform very badly in Black patients if trained exclusively on white patients – just like a human dermatologist who only received training using examples of fairer skin. Here, we understand the program intuitively, based on inferences informed by introspection, in a sense which Jaspers calls "causality from within". Mathematically, such understanding could also be fostered by what Angelov and colleagues call a "cardinally different approach to explainability" (2021): By choosing actual training data samples based on local peaks of the data distribution which they call "typicality", Angelov and Soares provide "prototypes" that are easily understandable by human users (2020).

Importantly, just like in psychopathology, such understanding may be empirically falsified (Ebmeier, 1987). Nietzsche's account of the genealogy of morality may be historically false in the particular instance of Christianity despite being understandable, as Jaspers notes (Jasper 1946, p. 252). Similarly, looking at the genealogy of a training data set or prototypes among the training data could be misleading. It is therefore crucial to critically question the scope of understanding, as Jaspers repeatedly admonishes in in critical remarks against Freud, and not jump to general causal rules. Also in machine learning, understanding demands to closely observe the program, its design and behaviour, or as Jaspers puts it: "understanding […] needs to be grounded in actual facts" (Jaspers, 1946, p. 255). [30]

[30] It is in this factual grounding that we can also situate the difference between understanding and interpreting: If factual knowledge is lacking, one may still provide a general interpretation ("deuten"), which is lacking the properties of genuine understanding though (Jaspers, 1946, p. 252f.; cf. Hoerl, 2013). The distinction between the two may not always be clear though, especially in the context of incomplete knowledge.

Before we show how Jaspers model can inform debates about understanding and explaining medical AI in particular, it seems imperative though to address the potential objection that we misguidedly anthropomorphise AI despite its non-human characteristics.

5.5 Understanding AI as misguided anthropomorphism?

There is an obvious caveat to discussing the relation of explaining and understanding of AI with Jaspers. Jaspers originally discussed human psychopathology. We, however, want to draw on the relation of explaining and understanding with regard to AI. Indeed, the caveat is often brought up as a general objection to the use of human terms for artificial applications such as machine *learning* or artificial *intelligence*. This seems to be an instance of anthropomorphism which is defined as "the attribution of distinctively human-like feelings, mental states, and behavioural characteristics to inanimate objects, animals, and in general to natural phenomena and supernatural entities" (Salles, Evers, & Farisco, 2020, p. 89).

The alleged threat of anthropomorphism to our adequate understanding of AI has been widely discussed (Salles, Evers, & Farisco, 2020; Watson, 2019; DeCamp & Tilburt, 2019) and anthropomorphism has been accused of being ontologically and morally dubious (Salles, Evers, & Farisco, 2020). The issue has been most prominently raised in relation to moral ascriptions, such as responsibility and trustworthiness, of algorithms. DeCamp and Tilburt (2019) have argued that this has severe consequences: "Trust properly understood involves human thoughts, motives, and actions that lie beyond technical, mechanical characteristics. To sacrifice these elements of trust corrupts our thinking and values" (p. 390). Similarly pointing out the differences between humans and

algorithms, Watson (2019) writes: "Algorithms are not 'just like us' and the temptation to pretend they are can have profound ethical consequences" (p. 434). This finds expression in what Proudfoot (2011) calls the forensic problem of anthropomorphism, originally related to ascriptions of, say, intelligence to algorithms. As she writes:

> "But how can a researcher's effort to 'convince himself or anyone else' of intelligence in machines be trusted if the researcher readily succumbs to anthropomorphism and make-believe — ascribing joy to a robot vacuum cleaner, for example?" (p. 952).

Generally, Proudfoot (2011) calls this the forensic problem of anthropomorphism which describes the risk of introducing cognitive biases in favour of the algorithm's intelligence by anthropomorphizing it. Unless the risk is mitigated, such judgements are deemed suspect. Is our attempt to understand AI similarly based on make-believe? After all, some may argue that it is an obvious mistake to discuss algorithms with regard to Jaspers' human psychopathology.

However, it is similarly dubious that the abolition of anthropomorphism is something that can be easily done. Proudfoot (2011) points out that even the critics of anthropomorphism in AI describe algorithms as stupid at the same time — a clear anthropomorphism as being stupid is a human characteristic. Our answer is that the employment of anthropomorphism should be pragmatic: if anthropomorphism is useful, it should not be jettisoned.

In the case of AI, there is some indication that it is. Bos et al. (2019) argue that anthropomorphism is an effective strategy for human participants to predict whether a *high-performing image classifier* AI model would label an image correctly. The

participants of their study made reference to their own perception, either explicitly or implicitly, to predict the classifier's results. Interestingly, the researchers report that the mental model discussed "their own or general human abilities, indicating some cognitive separation of human and classifier abilities. The 'mental model' tag indicated awareness that participants were forming a mental model of the system as they did the task" (p. 954). This research is interesting for our context in at least two regards: first, it shows that anthropomorphism can be used for modelling (Cassini & Redmond, 2021) in the context of AI, making use of what we, as humans, know about our own abilities. Anthropomorphism in this sense seems also in line with a current human-centric approach to explainability in AI "which treats it as a human-centric (anthropomorphic) phenomena rather than reducing it to statistics" (Angelov et al., 2021, p. 8).

Second, the study by Bos and colleagues also helps to disentangle the question of modelling from the question under which circumstances such anthropomorphist fiction constitutes an empirically effective strategy. After all, an intentional cognitive effort to understand AI by comparison to similar human abilities may not always be useful. Since anthropomorphist modelling is irrespective of the model's veracity, it will be important to distinguish between contexts in which accurate representation is required (Nguyen, 2020) while other models may benefit from "felicitous falsehoods" (Elgin, 2017). Bos et al (2019) therefore rightly call for more empirical studies testing the factual effectiveness of anthropomorphist modelling in different contexts.

The human-centred distinction of explaining and understanding can therefore help to shed some light on explainability in AI. The discussion of understanding of AI should therefore not be hindered by general objections against anthropomorphism if it provides

a useful tool. However, this still demands a clear conception of what explaining and understanding in relation to AI means.

5.6 Explaining and understanding medical AI

So far, we have sketched how a model developed by Jaspers in the context of human psychopathology can help to augment debates about explainable AI. Based on his distinction of explaining, aimed at general causal claims, and understanding, elicited by plausible evidence in singular cases, we advocate for methodological pluralism, harnessing both routes to establish meaningful relations between the factual data of machine learning. While we therefore started with a theory derived for a clinical purpose and employed it in the context of machine learning, we return to the clinic in this section, highlighting what Jaspers' model may imply for explaining and understanding medical AI. To do so, we draw on the well-known and widely cited example of IBM Watson for Oncology (WFO) and its shortfalls here (Strickland, 2019).

As we have seen, the first step of assessing such an AI will require careful observation of the program. These will contain different kinds of evaluations, both ex-ante and ex-post, to establish a factual basis for understanding and explaining. For instance, one would need to determine how the model and its hyperparameters were chosen, how it was optimized, and on which data, as much as one would need to evaluate its performance in different validation samples and identify the factors that had the largest impact on the model's prediction. To stay with the example, one would e.g. need to look closely at the health records which IBM used to train WFO, relying heavily on input from oncologists at the Memorial Sloan Kettering Cancer Center in the US (Jie et al., 2021), and at the model itself. To enable this kind of scrutiny, the program's developers would

need to embrace open communication and share their "factual data" as openly as possible.

Having collected all this information, we would then have two routes to find meaningful relations in them. First, experts may aim for an *explanation* through an array of different methods (cf. Holzinger et al., 2019). Ideally, such an explanation would provide a general causal rule, which in turn may be used to improve the model. To stay with the example of WFO, it seems conceivable that by aiming for such a general causal rule, researchers may find a pattern in the program's decision that helps them to identify some novel (epi-)genetic causes underlying certain subtypes of cancer.

However, as a parallel, complementary approach, we should also aim at *understanding* the ML model. As we have shown at the beginning, to foster trust and enable important ethical goals such as informed consent, some grasp concerning the program's behaviour seems crucial for the end-users of a clinical ML application. As outlined, such an understanding can be based on plausible evidence, without establishing general causal claims – like we would, to use Jaspers' example, understand a connection between gloomy autumn weather and a tendency to commit suicide (Jaspers 1946, p. 252f.). In the case of WFO, such understanding may help us to make sense of observations that are immediately plausible to the lay person as well. A recent meta-analysis that compared WFO treatment recommendations with the recommendations of multidisciplinary teams of human experts found that concordance depended highly on regional differences and types of cancer. For instance, concordance of treatment recommendations was as low as 29.9% in gastric cancer, when comparing WFO with multidisciplinary teams from Asian countries (Jie et al., 2021). This observation becomes immediately plausible if one considers that WFO was trained and validated in the US

and may therefore not agree with experts from other regions. After all, there are "large difference between the surgical methods and guidelines for adjuvant treatment of gastric cancer in China and the United States" (ibid.), and "WFO recommended the use of agents that are considered outdated in Korea" (Choi et al., 2019).

In such cases, we can understand the program's behaviour considering its training history, drawing on a form of "causality from within". Such understanding will require some form of knowledge about the AI model that can be related to our own reasoning processes, e.g. on which data it has been trained, where, by whom, and with which intentions. Other, often more technical details may arguably not foster understanding, for instance, whether the underlying algorithm has been optimized using gradient descent, how many hidden layers were used in a deep learning architecture, or whether a sigmoid or a Rectified Linear Unit (ReLU) function has been used as activation function.

Like in psychopathology, it is important though to not mistake the evidence of understanding for the epistemic certainty granted by explaining (cf. Hoerl, 2013, p. 108). Jaspers notes this, when stressing that despite us understanding an autumnal death-wish, more people actually commit suicide in spring (Jaspers 1946, p. 253). Similarly, we may also find that the underlying reason for WFO's problematic treatment recommendations in gastric cancers was not attributable to differences in regional treatment guidelines but based on the prevalence of particular mutations as has been reported for lung cancer (Jie et al., 2021).

Put differently, understanding does not imply giving up on causal explanations, just like for Jaspers understanding based on causality from within and explaining based on

causality from without are not mutually exclusive. Yet, a complementary approach embracing both strategies to make sense of an AI model could prove fruitful in at least three ways. First, understanding meaningful correlations of an AI could be used to develop and test new hypotheses, thereby advancing genuinely causal explanations through the "encounter with the incomprehensible" (Jaspers, 1946, p. 254). Second, and particularly important in the context of medical AI, the differentiation between understanding and explaining could be seen as representing two different approaches tailored to different audiences. While explainability may continue to provide important technical tools for experts to improve and assess clinical AI, broader groups of end-users such as patients or physicians that do not command expertise in computer science may, at least partially, gain comprehension of an AI by means of understanding. Third, understanding and explaining could, in this sense, provide two complementary routes to increase an AI's trustworthiness: As recent research into the relation of explainability and trust has argued, the trustworthiness of an AI depends on both internal and external factors (Jacovi et al., 2021; Ferrario, & Loi, 2021). While the internal trustworthiness of a model depends on the questions whether the "reasoning process aligns with human reasoning" (Jacovi et al., 2021, p. 629) and may be promoted by a Jasperian understanding, the external path to trustworthiness relies on the observation from without and would therefore fall into the domain of what Jaspers calls explaining.

Jaspers' distinction between explaining and understanding, rooted in different accounts of causality, also connects well with recent philosophical contributions to the field such as Emily Sullivan's work on link uncertainty (Sullivan 2020), despite terminological differences. As she convincingly argues, when discussing the black box problem of (medical) AI, one should distinguish between uncertainty introduced by a particular

technical implementation – i.e., that we may not know how a particular deep learning model arrives at its predictions –, and link uncertainty, i.e. "the extent to which the model fails to be empirically supported and adequately linked to the target phenomena" (ibid.). Such link uncertainty can vary vastly in medical contexts. For instance, an opaque, deep learning-based program employed in pathology to diagnose cancer with rather clear aetiology and histological correlates acts on a fundamentally different link uncertainty than an algorithm employed in psychiatry to diagnose major depressive disorder. Given the many diverging levels of link uncertainty present in medical practice, is therefore crucial that the developers, users, and subjects of medical AI heed Jaspers' plead for methodological pluralism:

> "All categories and methods have their specific purpose. It makes no sense to play them off against each other. Each of them has its own pure and appropriate realization, which is necessarily limited. Each of them, through absolutization, results in empty demands, ineffective talk and in modes of behaviour through which the free view of the facts is destroyed." (Jaspers, 1946, p. 384)

5.7 Conclusion

In this article, we have argued that the distinction between explaining and understanding as developed by Karl Jaspers in the context of psychopathology can provide a fruitful framework for current debates about the explainability of medical AI. In line with Jaspers, we have argued that explaining and understanding should be conceptualized as complementary epistemic approaches that must not be pitted against each other. We have shown how these approaches relate to current positions in the ongoing philosophical debate about medical AI and provided a practical example of its implications, drawing on IBM's Watson for Oncology as case study. Recent

philosophical and ethical reflection on medical AI can therefore benefit from revisiting long-standing arguments from the philosophy of psychiatry to sketch a path towards ethically and legally sound, trustworthy AI in medicine.

5.8 References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 1-9.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, *130*, 185-194.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), e1424.

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.) Oxford: Oxford University Press.

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics, 47*(e3). doi:10.1136/medethics-2019-105860

Burr, C., Morley, J., Taddeo, M., & Floridi, L. (2020). Digital Psychiatry: Risks and Opportunities for Public Health and Wellbeing. *IEEE Transactions on Technology and Society*, *1*(1), 21-33.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Bos, N., Glasgow, K., Gersh, J., Harbison, I., & Lyn Paul, C. (2019, November). Mental models of AI-based systems: User predictions and explanations of image classification results. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 184-188). Sage CA: Los Angeles, CA: SAGE Publications.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223-230.

Cassini, A., & Redmond, J. *Models and Idealizations in Science*. Springer: Cham.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., ... & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, *25*(8), 1301-1309.

Choi, Y. I., Chung, J. W., Kim, K. O., Kwon, K. A., Kim, Y. J., Park, D. K., ... & Lee, U. (2019). Concordance rate between clinicians and watson for oncology among patients with advanced gastric cancer: early, real-world experience in Korea. *Canadian Journal of Gastroenterology and Hepatology*, *2019*.

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health*, 1(8), e390.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 20539517211035955.

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. Journal of the American Medical Informatics Association, 27(4), 592-600.

Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*, 103498.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

Elgin, C. Z. (2017). True enough. Cambridge, MA: MIT Press.

Ebmeier, K. P. (1987). Explaining and understanding in psychopathology. *The British Journal of Psychiatry*, *151*(6), 800-804.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118

Ferrario, A., & Loi, M. (2021). The meaning of "Explainability fosters trust in AI". Available at SSRN 3916396.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology*, 33(1), 1-3.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Gough, J. (2021). On the proper epistemology of the mental in psychiatry: what's the point of understanding and explaining? *The British Journal for the Philosophy of Science* (accepted). doi: 10.1086.715106

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Hoerl, C. (2013). Jaspers on explaining and understanding in psychiatry. In Stanghellini, G., & Fuchs, T. (Eds.). (2013). *One century of Karl Jaspers' general psychopathology*. Oxford: Oxford University Press.107-120.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Husserl, E. (2020). *Studien zur Struktur des Bewusstseins: Teilband III Wille und Handlung Texte aus dem Nachlass (1902-1934)*. Edited by U. Melle, & T. Vongehr. Cham: Springer.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).

Jaspers, K. (1946). *Allgemeine Psychopathologie* (4th ed). Berlin: Springer.

Jie, Z., Zhiying, Z., & Li, L. (2021). A meta-analysis of Watson for Oncology in clinical application. *Scientific reports*, *11*(1), 1-13.

Knoops, P. G., Papaioannou, A., Borghi, A., Breakey, R. W., Wilson, A. T., Jeelani, O., ... & Schievano, S. (2019). A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific reports*, *9*(1), 1-12.

Kumazaki, T. (2013). The theoretical root of Karl Jaspers' General Psychopathology. Part 1: Reconsidering the influence of phenomenology and hermeneutics. *History of Psychiatry*, *24*(2), 212-226.

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*, 700.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15-21.

Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*, 279-288.

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. (Winter 2020 ed.). Retrieved from https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>.

Nguyen, J. (2020). Do fictions explain? *Synthese*, 1-26.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441-459.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018, July). Learning independent causal mechanisms. *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, 4036 - 4044.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901.

Proudfoot, D. (2011). Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6), 950-957

Raket, L. L., Jaskolowski, J., Kinon, B. J., Brasen, J. C., Jönsson, L., Wehnert, A., & Fusar-Poli, P. (2020). Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *The Lancet Digital Health*.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, 11(2), 88-95.

Schlimme, J. E., Paprotny, T., & Brückner, B. (2012). Karl Jaspers. *Der Nervenarzt*, 83(1), 84-91.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Shanahan, M. (2016). Conscious exotica. *Aeon.* Retrieved from https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24-31.

Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axz035

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, *25*(1), 44-56.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, *15*(11), e1002689

Vogeley, K. (2013). A Social Cognitive Perspective on 'Understanding' and 'Explaining'. *Psychopathology*, 46(5), 295-300.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76-99.

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence, *Minds and Machines*. 29 (3) 417-440.

Windelband, W. (1980). Rectorial Address, Strasbourg, 1894. Translation by Guy Oakes. *History and Theory*, *19*(2), 169-185.

Woodward, J. (2021), Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.). Retrieved from https://plato.stanford.edu/archives/spr2021/entries/scientific-explanation/>.

Chapter 6: The Emperor's New Clothes? Transparency and Trust in Machine Learning for Clinical Neuroscience

The Emperor's New Clothes? Transparency and Trust in Machine Learning for Clinical Neuroscience

Georg Starke[1],

[1] Institute for Biomedical Ethics, University of Basel, Basel, Switzerland.

The following chapter is the accepted manuscript. Please cite the published version: Starke, G. (2021). The Emperor's New Clothes? Transparency and Trust in Machine Learning for Clinical Neuroscience. In O. Friedrich, A. Wolkenstein, C. Bublitz, R. J. Jox, & E. Racine (Eds.), *Clinical Neurotechnology Meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications* (pp. 183-196). Cham: Springer. https://doi.org/10.1007/978-3-030-64590-8_14

Abstract

Machine learning (ML) constitutes the backbone of many applications of Artificial Intelligence. In the field of clinical neuroscience, applying ML to neuroimaging data promises wide-ranging advancements. Yet, such potential diagnostic and predictive tools pose new challenges with regard to old problems of transparency and trust. After all, the very design of many ML applications can preclude comprehensive explanations of its inner workings and impede accurate predictions about its future behaviour, supposedly clashing with the ideal of transparency. It is often claimed that these shortcomings, inherent to many ML applications, are detrimental to their trustworthiness and thus hinder implementing new and potentially beneficial techniques. In this chapter, I will argue against beliefs that inextricably link transparency and trustworthiness. Drawing in particular on the framework of the British philosopher and bioethicist Onora O'Neill, I aim to show why, contrary to many intuitions, an obsession with transparency can be detrimental to tackling more fundamental ethical issues – and that hence transparency may not solve as many challenges for clinical ML applications as is usually assumed. I will conclude with a tentative suggestion on how to move forward from a practical point of view as to advance the trustworthiness of ML for clinical neuroscience.

6.1 Introduction

> *Mehr an Information und Kommunikation allein erhellt die Welt nicht. Die Durchsichtigkeit macht auch nicht hellsichtig.*[31]
>
> Byung-Chul Han, *Transparenzgesellschaft* (2015b, p. 68)

Machine learning (ML) constitutes the backbone of many applications of Artificial Intelligence (AI). In the field of clinical neuroscience, applying ML to data from neuroimaging promises wide-ranging possibilities, from assisting clinicians in diagnostic and prognostic exams to enabling the selection of an optimal pharmacological intervention (Brodersen et al., 2014; Dwyer et al., 2018; Huys, Maia, & Frank, 2016; Janssen, Mourao-Miranda, & Schnack, 2018; Webb et al., 2018; Xiao et al., 2017). Despite the increasing body of bioethical literature on the subject of ML in medicine (Char, Shah, & Magnus, 2018; Darcy, Louie, & Roberts, 2016; Vayena, Blasimme, & Cohen, 2018), ethical discussions of ML with regard to neuroimaging data remain scarce (Bzdok & Meyer-Lindenberg, 2018; Martinez-Martin, Dunn, & Roberts, 2018). The aim of this chapter is to tackle this gap, focusing on the notion of transparency and its intricate relation to trust.

If we are to believe its proponents, transparency is key to solving the ethical challenges of clinically applied ML. Transparency is said to drive algorithmic fairness (Abdollahi & Nasraoui, 2018), guarantee patients' safety and enable informed consent (Turilli & Floridi, 2009). According to some, it constitutes "the first step towards ethical and fair

[31] Unfortunately, the English translation by Erik Butler cannot quite grasp the meaning of the German original: "More information and communication alone do not illuminate the world. Transparency also does not entail clairvoyance" (Han, 2015a).

ML models" (Zhou & Chen, 2018b). Unfortunately, the very design of many ML applications can preclude comprehensive explanations of its inner workings, thus posing particular challenges to an ideal of transparency (Kroll et al., 2017; Vayena et al., 2018). Black box algorithms may prevent accurate predictions about future behaviour, e.g. if the program continuously updates its inherent models based on newly available data. To many, such lack of scrutability of ML applications is of particular concern, as it can create gaps in responsibility for potential short fallings, which may also have legal implications (Bublitz, Wolkenstein, Jox, & Friedrich, 2018; Matthias, 2004). The use of poorly chosen or curated input data, for example, can result in errant or skewed output results, contributing to discriminatory or otherwise harmful practices (Cohen, Amarasingham, Shah, Xie, & Lo, 2014; Favaretto, De Clercq, & Elger, 2019). However, if the black-box program cannot be explained or understood, such errors may go unnoticed and evade remedy. Consequently, attributing responsibility to create clear, "transparent" patterns of accountability seems crucial for establishing a procedure's trustworthiness. In turn, trustworthiness may determine whether patients and physicians factually trust and thus embrace the clinical implementation of ML (Schnall, Higgins, Brown, Carballo-Dieguez, & Bakken, 2015). Hence, do we need to crack open the black boxes of ML applications in order to achieve transparency and render them trustworthy?

Placing so much burden on one scientific ideal certainly warrants scrutiny. Similar to the garments in Hans Christian Andersen's famous tale, which supposedly expose the viewers' own inadequacy, I will show why mere calls for transparency are too little to cloak the ethical challenges posed by applied ML in clinical neuroscience. Drawing on the writings of the philosopher and bioethicist Onora O'Neill, I will argue that

transparency is not an all-purpose remedy for fostering trustworthiness and that an obsession with transparency can be detrimental to tackling more fundamental issues. To do so, I will discuss the ideal of transparency and its relation to trust in clinical ML procedures using neuroimaging data, in order to give a more practical demonstration of an abstract debate. This example may prove particularly fruitful since both transparency and neuroimaging share a common aim: to make things visible. Nevertheless, similar points regarding trust and transparency could be raised with regard to other clinical ML applications as well.

The structure of this chapter will be as follows: I will first provide a brief definition of ML and offer some examples of potential applications for clinical neuroscience. I will then outline why transparency is commonly thought to be an ethically important epistemic ideal and why it is ascribed prudential value to foster public trust in such new technological developments. In a third section, I will draw on O'Neill's philosophy and argue against beliefs that inextricably link transparency and trustworthiness. Finally, I will conclude with a tentative suggestion on how to move forward from an O'Neillian point of view and advance trust in beneficial uses of ML in neurotechnology by moving toward a form of "intelligent openness" (O'Neill, 2018).

6.2 Opportunities for applied machine learning in clinical neuroscience

The notion of machine learning encompasses many different methods that constitute current state-of-the-art applications of artificial intelligence. An influential operational definition, which I will also use in this chapter, stems from the computer scientist Tom Mitchell, who characterised ML in terms of experience E, an area of tasks T and performance measure P. Following Mitchell, ML describes a program "if its performance at tasks in *T*, as measured by *P*, improves with experience *E*" (Mitchell, 1997). A classic

application would be recognising patterns, for instance across many different images. ML demarcates a narrower field then the broader term AI since the latter would also encompass the "holy grail" of AI research: generalised AI, capable of *any* intellectual task usually performed by humans, which for the moment remains in the domain of science fiction, such as the famous sentient computer HAL 9000 in Stanley Kubrick's *Space Odyssey*. At the same time, ML comprises a multitude of different computational approaches such as support vector machines (SVM) or artificial neural networks for deep learning (DL).

Neuroimaging, in turn, denotes as diverse an area as ML. Narrowly, it can be defined as "all techniques in which actual images of the brain are acquired" (Kellmeyer, 2017). Such processes rely on vastly different acquisition techniques, such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) or maps derived from electroencephalography (EEG) or transcranial magnetic stimulation (TMS). Within each of these, many further important distinctions can be made, e.g. in MRI between structural and (task- or resting state) functional MRI, between different acquisition sequences such as spin echo or gradient echo and so forth. Comprehensive overviews over the different techniques are offered elsewhere (Kellmeyer, 2017), but it is important to note that depending on the imaging modality it is already the case today that different degrees of computational efforts and human agency are required to generate images, rendering the techniques distinctly close to ideals of mechanical objectivity (Daston & Galison, 2007).

Of course, applications that combine the two very diverse areas of ML and neuroimaging are highly heterogeneous themselves, adding yet another layer of complexity. Nevertheless, some broad distinctions can be made based on the different purposes for

which ML-driven neuroimaging is employed. For the question at hand, the most fundamental distinction appears to be whether an application is intended (primarily) for research or for clinical purposes since this distinction shapes legal and ethical obligations involved in the process and may also determine the scope of the application. For example, let us suppose a project's only aim consists in contributing to a better understanding of pathologies underlying a certain disease or disorder. Let us further suppose, as a real-life example, that this project identifies different subtypes of schizophrenia based on distinct connectivity patterns by using ML on diffusion-weighted MRI scans. Critics may argue that such research could ultimately contribute to an unwarranted reification of psychiatric disorders (Hyman, 2010) by creating a dubious classificatory system of supposed "natural kinds" (Bzdok & Meyer-Lindenberg, 2018). But within such a research setting, patients would not be wronged by the tentative assignment to newly created, experimental subcategories of an already-diagnosed disorder.

However, where ML is applied to the clinic directly, stakes seem far higher. From identifying patients at risk of psychosis (Koutsouleris et al., 2015; Ramyead et al., 2016) to predicting the course of multiple sclerosis (Zhao et al., 2017), from prognostic tests for Alzheimer's dementia (Dallora, Eivazzadeh, Mendes, Berglund, & Anderberg, 2017) to algorithms suggesting ideal psychopharmacologic drugs for Major Depressive Disorder (Chekroud et al., 2016; Webb et al., 2018), future patients will certainly be treated based on recommendations by programs relying on ML and neuroimaging. If left unchecked, this could endanger the safety and well-being of patients, e.g. if the use of a diagnostic program provides the wrong diagnosis or misses a potentially life-threatening illness, leading in turn to wrong or delayed treatment. Potentially, recommendations

may also be skewed by biased input data, reinforcing direct or indirect discriminatory practices (Favaretto et al., 2019). Similar problems can arise with programs suggesting particular treatments. The shortcomings of IBM's Watson, recommending dangerous treatment strategies in oncology, may serve as a cautionary example here (Ross, 2018). Many authors have thus argued that the development of such applications warrants special scrutiny to safeguard against unjustified systematic biases, to ascertain the clinical safety of their deployment and to establish clear chains of responsibility. Yet, given the vast disparities in potential uses and methods of ML for clinical neuroscience, it seems surprising that transparency is often treated as an all-purpose remedy for potential challenges posed by its implementation. A recent review of translational ML for psychiatric neuroimaging for example suggested that "[f]or use in clinical support systems, transparency is both necessary and sufficient for legal certification and patient safety" (Walter et al., 2019). It thus seems worth shedding more light on this heavily burdened ideal.

6.3 The ideal of transparency

For a clear discussion of the ethical role of transparency, one firstly needs to note that the term itself encompasses two mutually exclusive meanings (Turilli & Floridi, 2009). On the one hand, in computer science transparency commonly refers to a process's property of being *invisible* to the user. For example, if a common text-editing program is updated to make it run faster, its external, visible user-interface may not change at all, even though the underlying computational processes may be quite different in the new version. Still, such invisible change in the program's internals would be called "transparent" (Turilli & Floridi, 2009). On the other hand, transparency also denotes a process's property of being *visible*, often in the combined term "algorithmic

transparency". As an ideal, it describes programs that enable the user to *see* and scrutinise its internals, i.e. the underlying computational processes taking place (Desai & Kroll, 2017). In the ethical and legal arguments that concern us here, transparency usually refers to the second meaning only, and I will adhere to it in the following.

Unfortunately, clear definitions of transparency are hard to come by, and quite frequently transparency serves as "an empty signifier that can be filled by very different interpretations" (Worthy, 2018). In a minimal definition, transparency merely describes "putting content in the public domain" (O'Neill & Bardrick, 2015). Such disclosure is thought to be commendable because it allows shedding light on something otherwise hidden, or as the former member of the US supreme court Louis Brandeis remarked: "Sunlight is said to be the best of disinfectants; electric light the most efficient policeman" (Brandeis, quoted in Hansen and Flyverbom (2015)). Related to this are positions which define transparency according to its opposition to concealment as "lifting the veil of secrecy" ((Davis, 1998, p. 121), quoted in (Meijer, 2009)). In political contexts, this renders it often interchangeable with notions of openness and even finds tangible expression in architecture. The glass dome of the Berlin Reichstag building by Sir Norman Foster, promising an open, transparent government, may serve as a prime example (Worthy, 2018). Where transparency and secrecy are pitched against each other, they are often portrayed rather simplistically as a clash between good and evil (Worthy, 2018).

Depending on the context, "transparency" is then adorned with a domain-specific epithet, from governmental transparency (Meijer, 2013) and information transparency (Turilli & Floridi, 2009) to algorithmic transparency (Desai & Kroll, 2017), to name but three prominent examples. Of course, the kind of transparency commonly invoked with

regard to ML aligns closely with the last kind. However, as Albert Meijer has argued, an understanding of transparency rooted in modern discourse about information technology has become so pervasive that the two have become almost synonymous: "Modern transparency *is* computer-mediated transparency" (Meijer, 2009, emphasis added). According to Meijer, three key components characterise this particular form of transparency. First, it is unidirectional in the sense that unlike in public assemblies, where information is mutually exchanged between various agents, information flows in one direction only. Second, computer-mediated transparency usually entails that the transferred information is taken out of its context, i.e. disconnected from its original theoretical and practical underpinnings. The reason for this lies in a third characteristic, namely that modern transparency is calculative and hence mostly requires quantifiable information: "computer-mediated forms of transparency reflect certain aspects of reality, namely those aspects that are being measured" (Meijer, 2009).

A crucial reason why many hold transparency, understood as "visibility contingent upon observation" (Brighenti (2007) quoted in Hansen and Flyverbom (2015)), to be key in resolving ethical challenges - such as the challenges of applied ML - lies in its supposed function of enabling other important ethical principles. Fittingly, Matteo Turilli and Luciano Floridi (Turilli & Floridi, 2009) have dubbed transparent access to certain kinds of information "pro-ethical", namely where disclosure of information has an impact on ethical principles. In the particular context of ML, transparency is thought to help establish accountability and promote fairness, by baring unjustified biases resulting in discriminatory differential treatment of salient social classes (Kroll et al., 2017). In doing so, transparency seems a vital condition for rendering clinically applied machine learning trustworthy.

6.4 Trust and trustworthiness

Interest in different forms of trust has begotten immense corpora of academic literature throughout the past decades. Philosophers and social scientists alike have tackled the topic from different angles, motivated by the fact that it constitutes a central, but long-neglected phenomenon (Baier, 1986, p. 232f.; Luhmann, 1968, p. FN1). The resulting definitions of trust differ largely, and so do its classifications, distinguishing between goodwill trust, competence trust, contractual trust, calculus-based trust, knowledge-based trust and identification-based trust, to name but a few (Bachmann, 2001). It suffices here to look at some of its general properties that are important to our inquiry and common to both interpersonal and institutional (or public) trust (Townley & Garfield, 2013).[32] Two overlapping perspectives shape the debate: sociologists such as Georg Simmel, Niklas Luhmann and Barbara Misztal are often primarily concerned with the *functions* of trust (Luhmann, 1968; Misztal, 1996; Möllering, 2001), while theorists who take a more philosophical approach such as Annette Baier, Russell Hardin or Onora O'Neill aim for *definitory* clarifications (Baier, 1991; Hardin, 2002; O'Neill, 2002b).

Most accounts agree, from a conceptual point of view, that trust takes the form of a three-part relation: A trusts B regarding X, where X might range from particular actions, including speech acts, to general assumptions about B's behaviour or character (Baier, 1986; Hardin, 2002, p. 9). Usually, A will only place her trust in B if she believes B to be *competent* with regard to X and also to be *committed* to X.[33] For example, I will only entrust my neighbour with taking care of my flowers while on vacation if I take her to

[32] In doing so, I will necessarily leave out many important facets of trust, especially conative or emotional components, which arguably are vital for full-fledged forms of trust. Some even hold that only non-cognitive trust should be regarded as trust in the fullest sense (Becker, 1996).
[33] However, exceptions exist, for example parents placing trust in their children as a means of education, even when they are convinced that the children will fall short of their trust (McGeer, 2008, p. 241).

be capable of treating plants appropriately and am also optimistic about her actually looking after them. At the same time, I might not trust her to take care of my pet, knowing that she is highly averse to dogs. Note that this does not presuppose that she bears some form of goodwill towards me; she might well find me to be the most annoying person and still live up to my expectations, e.g. because she cares about her view onto my balcony or is interested in friendly relations to her neighbours.[34] Analogous claims can be made about trust in institutions, where the parties may also not act out of emotional attachment but rather have an interest in future interactions with the trustor. As Russell Hardin put it: "I trust you because I think it is in your interest to take my interests in the relevant matter seriously" (Hardin, 2002, p. 1).

While exact conceptualisations of trust remain challenging, the phenomenon's central functions seem clearer. As an expectation in specific situations of uncertainty, trust serves as a means to reduce social complexity in absence of certainty (Luhmann, 1968) and thus constitutes "a solution for specific problems of risk" (Luhmann, 2000, p. 95). After all, trust is only necessary where one lacks comprehensive knowledge or the capacity to exercise full control and hence needs, to some extent, to rely on the actions of others. To stay with the previous example, I would not need to trust my neighbour if I could either water my flowers myself or had means to fully determine her actions. As a means to deal with risk, trusting itself remains a "risky engagement" (Luhmann, 2017) and renders us vulnerable to the actions of others. Annette Baier even uses the property of being open to betrayal as a core characteristic of trust, to distinguish it e.g. from reliance (Baier, 1986, p. 235).

[34] I am following O'Neill here (O'Neill, 2002a, p. 14), who disagrees on this point with Annette Baier (Baier, 1986, p. 234f).

Clinical applications of ML for decision-making in medicine doubtlessly entail risks and make patients vulnerable to failure on multiple levels. But is trust really the correct way of dealing with such risks? After all, it is crucial not to trust blindly and only place trust in trustworthy agents. “Not to trust rashly is the nerves and joints of wisdom,” Cicero advised his brother (Cicero, 1963, p. 39). For ethical debates with a view on practical implementation, evaluating the trustworthiness of an agent, institution or procedure may thus be more pressing than an abstract debate about the phenomenon of trust (O’Neill, 2013). While there is much debate about properties that warrant trustworthiness, a few points stand out as uncontroversially detrimental, as they follow from the previously discussed expectations of the trustees to be capable and committed to act in our interest. We do not want to trust people with tasks they cannot possibly achieve (Baier, 1991; Scanlon, 1990) and even less if they invite our trust by lying or withholding critical information in a deceiving manner (O'Neill, 2002a, chapter 6.4). Additionally, we expect trustees to act on our interests, so it can undermine agents’ trustworthiness if they have important competing interests that are opposed to our own. Importantly, anyone abusing their power against us would also be very untrustworthy, for we certainly do not want to increase this power further by placing unjustified trust in them.

What could this mean with regard to clinically applied ML? Many authors agree that as a remedy for achieving trustworthiness of such new methods we need to turn to transparency. The assumed mechanism seems to be that transparency increases knowledge about a procedure and thus renders it more trustworthy insofar as it decreases the uncertainty necessarily involved in trusting. In their recent guidelines, the EU Commission’s High Level Expert Group on Artificial Intelligence explicitly names

transparency as one of its seven key requirements for trustworthy AI. Similarly, and with particular regard to medical ML (MLm), Vayena et al state that a "lack of transparency can preclude the mechanistic interpretation of MLm-based assessments and, in turn, reduce their trustworthiness" (Vayena et al., 2018). Certainly, trustworthiness can be enhanced in certain instances by tangible forms of (algorithmic) transparency. However, it is far from clear that the relation between transparency and trust is as straightforward as is commonly assumed.

6.5 The paradox relation of trust and transparency

Does transparency beget trust - or are matters more complicated? The British philosopher and bioethicist Onora O'Neill has long addressed this question from several angles. To tackle the intricacies of transparency of clinically applied ML, her framework appears particularly well suited since she developed her account in the very context of biomedical ethics. Three of her works stand out as highly instructive: her seminal *Autonomy and Trust in Bioethics*, the BBC Reith Lectures *A Question of Trust* and *Rethinking Informed Consent in Bioethics,* co-authored with Neil Manson. From these three works, three key ideas can be distilled that have not yet been applied to clinical ML but can guide further discussions on its ethical dimension.

First, in her *Autonomy and Trust in Bioethics*, O'Neill highlights that factually, transparency does not simply beget trust. To do so, she extensively discusses trust in the so-called "risk society" (O'Neill, 2002a), referring to the work of the German sociologist Ulrich Beck (Beck, 1992). From a sociological perspective, risk societies are characterised by increased public fears and anxieties about hidden risks of increasingly complex social and technological practices. Frequently, these "focus particularly on hazards introduced (or supposedly introduced) by high-tech medicine" (O'Neill, 2002a, p. 8). Importantly

though, risk societies are *not necessarily* characterised by factually increased risks in all areas of society.[35] In fact, risk societies are often wealthier, healthier and less secretive than their historical precursors. Still, due to a changed *perception* of risks, people in these societies tend to be more reluctant to placing trust. The medical domain provides a particular striking example for such erosion of trust. In the United States, public trust in the medical profession declined sharply from 73% reported in 1966 to 34% in 2012 (Blendon, Benson, & Hero, 2014). Yet, provisions to increase transparency and foster the autonomous and informed decision making of patients have undoubtedly greatly increased since the 60s, when paternalist doctor-patient-relationships were still the norm.[36] How may one explain such observations, undermining a supposed close link between transparency and trust? One tentative explanation could be that if people are generally distrustful of technological advances, transparency may have limited influence on such a societal phenomenon since people may still remain distrustful against any transparently disclosed information. O'Neill calls such a public mood, potentially diminishing the impact of transparency of trust, a "culture of suspicion" (O'Neill, 2002b). In an imaginary society of trust though, things may be just the opposite. Due to higher generalised trust, people may be more prone to placing trust in transparently disclosed information about medical technologies. In other words, we find the paradox that effective disclosure, which could increase a procedure's trustworthiness, seemingly presupposes trust.[37]

[35] Of course, this is not to deny the grave risks created by modern societies in specific areas, e.g. newly introduced environmental risks (Beck, 1992).
[36] It may be worth recalling here that the principlist framework by Beauchamp and Childress, stressing respect for autonomy as a fundamental principle of medical practice, was only published in 1979.
[37] Annette Baier has made a similar point, stressing that "trust [is] a response to perceived trustworthiness, […] but it is equally true that trustworthiness is, to some degree, a response to trust."(Baier, 2013)

In her *Reith Lectures*, O'Neill discusses this entangled relation in more detail. In line with Meijer's previously discussed arguments about modern transparency being computer-mediated, also for O'Neill, transparency constitutes the "new ideal of the information age" (O'Neill, 2002b). In political contexts, increasing transparency often seems coextensive with improving structures of accountability, supposedly fostering trust in public and professional institutions. However, accountability does not aim at establishing relations of trust but rather at minimising risks, ideally improving a trustee's trustworthiness.[38] However, mere transparency in the sense of "putting content in the public domain" (O'Neill & Bardrick, 2015) may not increase trustworthiness, if it does not aim at increasing the trustor's knowledge, e.g. because it is accessible but incomprehensible. Thus, while increased transparency may often have beneficial effects, it is too little for establishing a trustee's trustworthiness (O'Neill & Bardrick, 2015). At the same time, focusing solely on transparency runs danger of marginalizing other, more basic obligations (O'Neill, 2002b). In particular, O'Neill argues, these comprise the true enemy of trust, which is neither secrecy nor a lack of information, but wilful deception. In fact, such deception can come in the very guise of supposed transparency, namely when agents hide behind a rallying cry of transparency to engage in information dumping, burying future trustors under a heap of unsorted or misleading data, confusing the addressee and obscuring the actually crucial information. Of course, this is not to say that some forms of transparency may not render a process trustworthy. Full and transparent disclosure of information can, in some instances and particularly after previous cases of deception, improve the trustworthiness of agents and institutions. Yet,

[38] Of course, this is not to say that trust is a mere matter of risk calculation.

O'Neill argues, an obsession with transparency can also undermine trust: like plants, trust does not flourish when constantly uprooted (O'Neill, 2002a).

In *Rethinking Informed Consent in Bioethics,* O'Neill and Manson provide a practical demonstration of these considerations and show how effective disclosure requires trust, with specific regard to informed consent (Manson & O'Neill, 2007). Following Willard Van Orman Quine's distinction of referential transparency and opacity, they stress that informed consent is referentially opaque (Quine, 1980). Their general idea is that if A consents to B doing p, she does thereby not consent to q, even if the propositions q and p are logically equivalent. Drawing on an example from Ruth Faden and Tom Beauchamp (Faden, Beauchamp, & King, 1986), O'Neill and Manson highlight that a person may consent to taking lysergic acid diethylamide as part of a study. Yet, the same person would possibly not consent to the very same procedure if she was asked to take LSD – even though, of course, the two are identical. On a more applied level, this abstract debate finds its expression in the requirements of *valid* informed consent, which entails that the patient or research subject actually understands the proposition to which she consents. Researchers or physicians asking for consent hence need to consider the level of knowledge and the beliefs of the consenters, who in turn may place their trust in them that the information provided is correct, comprehensive and adequately understandable. In other words, mere transparency is not sufficient since effective transference of information requires taking into account the audience's needs and interests to make sure that the information and its implications are properly understood. To stay with the example, putting LSD's technical name in a consent sheet instead of providing the commonly known abbreviation could thus very well amount to wilful

deception - as could the mere publication of an incomprehensible source code for a medical ML application.

6.6 Trust and transparency of applied ML for neuroimaging

If O'Neill is right and trust is required for effective communication, it will prove impossible to avoid. Certainly, this holds true within medicine, inherently shaped by unknown and unknowable risks. Accordingly, many authors have addressed issues of trust in the medical domain in past years and both interpersonal and public trust have received plenty of attention from medical ethicists (Gille, Smith, & Mays, 2015; O'Neill, 2002a). For example, Gille et al have proposed a model of public trust in medicine which draws on Habermas' Theory of Communicative Action and acknowledges the centrality of trust as a necessary condition between communicating parties (Gille, Smith, & Mays, 2017). However, the necessity of trust goes far beyond communication, or as Luhmann put it: "Without trust, everyday life would be impossible; indeed one would not even be able to get out of bed in the morning" (Luhmann, 2017, p. 20).

It should hence not come as a surprise if new advances in medical ML, whether for diagnostic, therapeutic or prognostic purposes, will require trust for their successful implementation. In fact, trust seems necessary for the acceptance of ML-assisted decision making not only by patients, but also by physicians, nurses and other health care professionals. However, as this chapter aimed to show, mere transparency cannot guarantee trustworthiness. So what could this mean practically if applications of ML drawing on neuroimaging data gain ground? As a first step, acknowledging that mere unidirectional disclosure is not a remedy for all potential problems may create space to make more fundamental notions central. Making the source code publicly available is certainly commendable in research contexts, but in itself too little to render a program

trustworthy for patients and physicians. Secondly, this opening would need to be filled by principles that are more substantial. In her most recent book, O'Neill sketches such an alternative model which she calls "intelligent openness": "Scientific communication [...] requires not mere transparency, but "intelligent openness" that ensures that communication is in principle accessible, intelligible and assessable for all others, so fully open to their check and challenge" (O'Neill, 2018, p. 51).

The recent EU guidelines for trustworthy AI seem to provide a step in this direction, as they link transparency explicitly to communication. With specific regard to medical ML, programs developed to be understandable by patients and physicians may be much better suited to build (on) trusting relationships, as they aim at successful communication. For example, so-called Influence Style Explanations, which estimate the impact of distinct inputs on a given output, already in-use for non-medical recommender systems, could be implemented for applications of medical ML as well (Abdollahi & Nasraoui, 2018). In clinical contexts, this could e.g. take the form of a weighted list of parameters taken into account for a program's suggestion. In neuroimaging, understandable ML programs could further provide visualizations, which have long been a focus for examining and explaining ML (Zhou & Chen, 2018a). Such applications could e.g. highlight visual aspects in the data taken to be salient, such as newly acquired white matter lesions for MS progress reports.

Two twists need to be noted in discussions of medical ML for neuroimaging. First, a debate regarding the transparency of ML may stress that already the input itself, i.e. neuroimaging visualisations, should not be considered belief-transparent, mechanically objective representations of the outward world. Given the underlying complex interactions between humans and machines, most imaging modalities cannot be

ascribed the same epistemic value as photographs. As Adina Roskies has noted with regard to fMRI: "We do not 'see through' the visual properties of neuroimages to the visual properties of their subjects; we do not understand the causal and counterfactual relationships between the images and the data they represent to the same extent that we understand them with photography" (Roskies, 2007). Reconsidering transparency of ML applications may hence inadvertently revitalise existing discourses about the epistemic status of neuroimaging modalities and draw further attention to the training required for their proper understanding (Racine, Bar-Ilan, & Illes, 2005).

Second, as Stephen John has argued with regard to science communication and climate change, expectations of transparency have become so engrained in our societal discourses that transparency can be seen as a fundamental principle of a "folk philosophy of science" (John, 2018). Hence, while transparency may in fact not beget trust, an open denial of transparency could damage trust by breaking (arguably unrealistic) expectations. Again, a stronger focus on a more intelligent form of openness oriented towards communication may offer a way out of this dilemma by providing a related, yet more substantial rallying cry.

Doubting the primacy of transparency will certainly prove challenging. The German-Korean philosopher Byung-Chul Han, a long-standing critic of the "transparency society", has warned against possible resistance: "The imperative of transparency is suspicious of everything that does not submit to visibility" (Han, 2015b, p. 24). Still, such change of focus may be vital to increasing public trust in clinically applied ML. As Annette Baier has noted, "trust comes in webs, not in single strands" (Baier, 1991). Weaving such webs anew to accommodate for new medical technologies thus seems to pose a particular challenge – but one worth pursuing.

6.7 References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning* (pp. 21-35). Basel: Springer International Publishing.

Bachmann, R. (2001). Trust, power and control in trans-organizational relations. *Organization Studies, 22*(2), 337-365. doi:Doi 10.1177/0170840601222007

Baier, A. (1986). Trust and Antitrust. *Ethics, 96*(2), 231-260. doi:10.1086/292745

Baier, A. (1991). *Trust. The Tanner Lectures on Human Values*. Salt Lake City: University of Utah Press.

Baier, A. (2013). What is trust? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill*. Oxford: Routledge.

Beck, U. (1992). *Risk society : towards a new modernity*. London ; Newbury Park, Calif.: Sage Publications.

Becker, L. C. (1996). Trust as noncognitive security about motives. *Ethics, 107*(1), 43-61.

Blendon, R. J., Benson, J. M., & Hero, J. O. (2014). Public trust in physicians--U.S. medicine in international perspective. *New England Journal of Medicine, 371*(17), 1570-1572. doi:10.1056/NEJMp1407373

Brighenti, A. (2007). Visibility - A category for the social sciences. *Current Sociology, 55*(3), 323-342. doi:10.1177/0011392107076079

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2014). Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clinical, 4*, 98-111. doi:10.1016/j.nicl.2013.11.002

Bublitz, C., Wolkenstein, A., Jox, R. J., & Friedrich, O. (2018). Legal liabilities of BCI-users: Responsibility gaps at the intersection of mind and machine? *International Journal of Law and Psychiatry*. doi:10.1016/j.ijlp.2018.10.002

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223-230. doi:10.1016/j.bpsc.2017.11.007

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *New England Journal of Medicine, 378*(11), 981-983. doi:10.1056/NEJMp1714229

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. *Lancet Psychiatry, 3*(3), 243-250. doi:10.1016/S2215-0366(15)00471-X

Cicero, M. (1963). Commentariolum Petitionis. In W. Watt (Ed.), *M. Tulli Ciceronis Epistulae.* (Vol. III). Oxford: Oxford University Press.

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs, 33*(7), 1139-1147. doi:10.1377/hlthaff.2014.0048

Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J., & Anderberg, P. (2017). Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. *PLoS One, 12*(6), e0179804. doi:10.1371/journal.pone.0179804

Darcy, A. M., Louie, A. K., & Roberts, L. W. (2016). Machine Learning and the Profession of Medicine. *JAMA, 315*(6), 551-552. doi:10.1001/jama.2015.18421

Daston, L., & Galison, P. (2007). *Objectivity*. New York: Zone Books.

Davis, J. (1998). Access to and Transmission of Information: Position of the Media. In V. Deckmyn & I. Thomson (Eds.), *Openness and transparency in the European Union* (pp. 121-126). Maastricht: European Institute of Public Administration.

Desai, D., & Kroll, J. (2017). Trust but verify: A guide to algorithms and the law. *Harvard Journal of Law and Technology, 31*(1).

Dwyer, D. B., Cabral, C., Kambeitz-Ilankovic, L., Sanfelici, R., Kambeitz, J., Calhoun, V., Falkai, P., Pantelis, C., Meisenzahl, E., & Koutsouleris, N. (2018). Brain Subtyping Enhances The Neuroanatomical Discrimination of Schizophrenia. *Schizophrenia Bulletin, 44*(5), 1060-1069. doi:10.1093/schbul/sby008

Faden, R. R., Beauchamp, T. L., & King, N. M. P. (1986). *A history and theory of informed consent*. Oxford: Oxford University Press.

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data, 6*(1), 12. doi:10.1186/s40537-019-0177-4

Gille, F., Smith, S., & Mays, N. (2015). Why public trust in health care systems matters and deserves greater research attention. *Journal of Health Services Research & Policy, 20*(1), 62-64. doi:10.1177/1355819614543161

Gille, F., Smith, S., & Mays, N. (2017). Towards a broader conceptualisation of 'public trust' in the health care system. *Social Theory & Health, 15*(1), 25-43. doi:10.1057/s41285-016-0017-y

Han, B.-C. (2015a). *The transparency society* (E. Butler, Trans.). Stanford, California: Stanford University Press.

Han, B.-C. (2015b). *Transparenzgesellschaft*. Berlin: Matthes & Seitz

Hansen, H. K., & Flyverbom, M. (2015). The politics of transparency and the calibration of knowledge in the digital age. *Organization, 22*(6), 872-889. doi:10.1177/1350508414522315

Hardin, R. (2002). *Trust and trustworthiness*. New York: Russell Sage Foundation.

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. doi:10.1038/nn.4238

Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual Review of Clinical Psychology, 6*, 155-179. doi:10.1146/annurev.clinpsy.3.022806.091532

Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(9), 798-808. doi:10.1016/j.bpsc.2018.04.004

John, S. (2018). Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology, 32*(2), 75-87. doi:10.1080/02691728.2017.1410864

Kellmeyer, P. (2017). Ethical and Legal Implications of the Methodological Crisis in Neuroimaging. *Cambridge Quarterly of Healthcare Ethics, 26*(4), 530-554. doi:10.1017/S096318011700007X

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., von Saldern, S., Cabral, C., Reiser, M., Falkai, P., & Borgwardt, S. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin, 41*(2), 471-482. doi:10.1093/schbul/sbu078

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. L. (2017). Accountable Algorithms. *University of Pennsylvania Law Review, 165*(3), 633-705. Retrieved from <Go to ISI>://WOS:000397048900003

Luhmann, N. (1968). *Vertrauen; ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart: F. Enke.

Luhmann, N. (2000). Familiarity, confidence, trust: Problems and alternatives. *Trust: Making and breaking cooperative relations, 6*, 94-107.

Luhmann, N. (2017). *Trust and power* (English edition. ed.). Malden, MA: Polity.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Martinez-Martin, N., Dunn, L. B., & Roberts, L. W. (2018). Is It Ethical to Use Prognostic Estimates from Machine Learning to Treat Psychosis? *AMA Journal of Ethics, 20*(9), E804-811. doi:10.1001/amajethics.2018.804

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183.

McGeer, V. (2008). Trust, hope and empowerment. *Australasian Journal of Philosophy, 86*(2), 237-254.

Meijer, A. (2009). Understanding modern transparency. *International Review of Administrative Sciences, 75*(2), 255-269.

Meijer, A. (2013). Understanding the Complex Dynamics of Transparency. *Public Administration Review, 73*(3), 429-439. doi:10.1111/puar.12032

Misztal, B. A. (1996). *Trust in modern societies : the search for the bases of social order*. Cambridge: Polity Press.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Möllering, G. (2001). The nature of trust: From Georg Simmel to a theory of expectation, interpretation and suspension. *Sociology, 35*(2), 403-420.

O'Neill, O. (2002a). *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

O'Neill, O. (2002b). *A Question of Trust. The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

O'Neill, O. (2018). *From Principles to Practice: Normativity and Judgement in Ethics and Politics*. Cambridge: Cambridge University Press.

O'Neill, O. (2013). Trust before trustworthiness? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 237-238). Oxford: Routledge.

O'Neill, O., & Bardrick, J. (2015). Trust, trustworthiness and transparency. *Brussels: European Foundation Centre*.

Quine, W. V. O. (1980). *From a logical point of view: 9 logico-philosophical essays* (Vol. 9). Cambridge, MA: Harvard University Press.

Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews Neuroscience, 6*(2), 159-164. doi:10.1038/nrn1539

Ramyead, A., Studerus, E., Kometer, M., Uttinger, M., Gschwandtner, U., Fuhr, P., & Riecher-Rössler, A. (2016). Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naive at-risk patients. *World Journal of Biological Psychiatry, 17*(4), 285-295. doi:10.3109/15622975.2015.1083614

Roskies, A. L. (2007). Are neuroimages like photographs of the brain? *Philosophy of Science, 74*(5), 860-872. doi:Doi 10.1086/525627

Ross, C., Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News* Retrieved from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Scanlon, T. (1990). Promises and practices. *Philosophy & Public Affairs*, 199-226.

Schnall, R., Higgins, T., Brown, W., Carballo-Dieguez, A., & Bakken, S. (2015). Trust, Perceived Risk, Perceived Ease of Use and Perceived Usefulness as Factors Related to mHealth Technology Use. *Medinfo 2015: Ehealth-Enabled Health, 216*, 467-471. doi:10.3233/978-1-61499-564-7-467

Townley, C., & Garfield, J. L. (2013). Public trust. *Trust: Analytic and Applied Perspectives*, 95-108.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105-112. doi:10.1007/s10676-009-9187-9

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. doi:10.1371/journal.pmed.1002689

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. doi:10.1016/j.pnpbp.2018.09.014

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Adams, P., Trombello, J. M., Cooper, C., Deldin, P., Oquendo, M. A., McInnis, M. G., Huys, Q., Bruder, G., Kurian, B. T., Jha, M., DeRubeis, R. J., & Pizzagalli, D. A. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological Medicine, 49*(7), 1118-1127. doi:10.1017/S0033291718001708

Worthy, B. (2018). Transparency. In B. H. Nerlich, Sarah; Raman, Sujatha; Smith, Alexander (Ed.), *Science and the politics of openness. Here be monsters* (pp. 23-32). Manchester: Manchester University Press.

Xiao, Y., Yan, Z., Zhao, Y., Tao, B., Sun, H., Li, F., Yao, L., Zhang, W., Chandan, S., Liu, J., Gong, Q., Sweeney, J. A., & Lui, S. (2017). Support vector machine-based classification of first episode drug-naive schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*. doi:10.1016/j.schres.2017.11.037

Zhao, Y., Healy, B. C., Rotstein, D., Guttmann, C. R., Bakshi, R., Weiner, H. L., Brodley, C. E., & Chitnis, T. (2017). Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One, 12*(4), e0174866. doi:10.1371/journal.pone.0174866

Zhou, J., & Chen, F. (2018a). 2D Transparency Space—Bring Domain Users and Machine Learning Experts Together. In *Human and Machine Learning* (pp. 3-19). Basel: Springer International Publishing.

Zhou, J., & Chen, F. (2018b). *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Basel: Springer International Publishing.

Chapter 7: Computing Schizophrenia: Ethical Challenges for Machine Learning in Psychiatry

Computing Schizophrenia: Ethical Challenges for Machine Learning in Psychiatry

Georg Starke[1], Eva De Clercq[1], Stefan Borgwardt[2,3], Bernice Simone Elger[1,4]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland, [2] Department of Psychiatry, University of Basel, Switzerland, [3] Department of Psychiatry and Psychotherapy, University of Lübeck, Germany, [4] University Center of Legal Medicine, University of Geneva, Switzerland.

Acknowledgements:

The authors would like to thank Andrea Martani and Christopher Poppe for their helpful comments as well as the three anonymous reviewers for substantial improvements on a previous draft.

The following chapter is the accepted manuscript. Please cite the final published version: Starke, G., De Clercq, E., Borgwardt, S., & Elger, B. (2021). Computing schizophrenia: Ethical challenges for machine learning in psychiatry. *Psychological Medicine, 51*(15), 2515-2521. https://doi.org/10.1017/S0033291720001683. Reproduced with permission.

Abstract

Recent advances in machine learning (ML) promise far-reaching improvements across medical care, not least within psychiatry. While to date no psychiatric application of ML constitutes standard clinical practice, it seems crucial to get ahead of these developments and address their ethical challenges early on. Following a short general introduction concerning ML in psychiatry, we do so by focusing on schizophrenia as a paradigmatic case. Based on recent research employing ML to further the diagnosis, treatment, and prediction of schizophrenia, we discuss three hypothetical case studies of ML applications with view to their ethical dimensions. Throughout this discussion, we follow the principlist framework by Tom Beauchamp and James Childress to analyse potential problems in detail. In particular, we structure our analysis around their principles of beneficence, non-maleficence, respect for autonomy, and justice. We conclude with a call for cautious optimism concerning the implementation of ML in psychiatry if close attention is paid to the particular intricacies of psychiatric disorders and its success evaluated based on tangible clinical benefit for patients.

7.1 Introduction

The quest for objective measures of mental disorders has been a long-standing ambition of psychiatry (Kapur, Phillips, & Insel, 2012; Singh & Rose, 2009). Given the notorious difficulties of classifying mental disorders and the challenge of establishing psychiatric biomarkers, many recent advances put their hope in approaches using machine learning (ML) as a paradigm-shifting way forward (Bzdok & Meyer-Lindenberg, 2018; Janssen, Mourao-Miranda, & Schnack, 2018; Shatte, Hutchinson, & Teague, 2019). By applying ML on large-scale datasets, it seems feasible to distinguish between healthy controls and patients diagnosed with major depressive disorder or schizophrenia on an individual level – although reported diagnostic accuracies differ largely across studies (Ebdrup et al., 2018; Gao, Calhoun, & Sui, 2018; Kambeitz et al., 2015). Furthermore, ML techniques can differentiate successfully between subgroups within psychiatric categories (Drysdale et al., 2017; Dwyer et al., 2018) and predict the success of specific psychopharmacological interventions for single subjects (Chekroud et al., 2016; Webb et al., 2018). Of high clinical interest are ML applications that provide robust probabilistic estimates regarding future onset of psychosis (Borgwardt et al., 2013; Chung et al., 2018; Koutsouleris et al., 2018) or the risk of suicide (Franklin et al., 2017; Just et al., 2017; Walsh, Ribeiro, & Franklin, 2017). However, to allow translation to current clinical practice, further multicentre imaging studies, integrating clinical measures and multivariate imaging data, are needed to replicate promising initial findings (Giordano & Borgwardt, 2019).

Currently, there is no established ML application in psychiatric clinical practice. The drastic increase of FDA approvals for medical applications of artificial intelligence (AI) in the past two years (Topol, 2019) suggests that some ML programs could soon be

integrated into standard clinical care, improving prediction and early detection, diagnostic certainty and individual treatment outcome in the sense of personalized psychiatry (Perna, Grassi, Caldirola, & Nemeroff, 2018). Unfortunately, the majority of ML applications in psychiatry still lack in-depth ethical analysis. With few exceptions discussing specific case studies (Martinez-Martin, Dunn, & Roberts, 2018), ethical concerns are often voiced in a general form (Char, Shah, & Magnus, 2018; Topol, 2019; Vayena, Blasimme, & Cohen, 2018), thus necessarily neglecting the particular intricacies of potential psychiatric applications.

ML is an extremely broad term, covering many distinct computational approaches for even more heterogeneous real-world problems. We aim to demonstrate that any categorical rejection of the use of ML in psychiatry would be ethically wrong given its potential benefits but that careful evaluation is needed whether a particular procedure improves clinical care or merely constitutes a nifty computational exercise. Using schizophrenia as a paradigmatic case, we will first sketch some fundamental distinctions of different ML methods, before turning to three (hypothetical) case studies. To support our main claim, we will discuss these cases following the principlist framework of Beauchamp and Childress (2013), which has recently been embraced as providing suitable principles for the ethical use of AI as well ("Ethics guidelines for trustworthy AI," 2019; Floridi et al., 2018).

7.2 Machine learning in psychiatry

The meaning of the term 'machine learning' is often ambiguous. In the present paper, we use ML to describe learning algorithms which improve their performance in a certain task based on prior computation (Iniesta, Stahl, & McGuffin, 2016; Mitchell, 1997). ML in this sense comprises a narrower field than AI, which includes generalized AI and

incidentally describes ”whatever hasn't been done yet” (Hofstadter, 1980). At the same time, ML itself entails many specific computational approaches, from deep learning (DL) using artificial neural networks to algorithms relying on support vector machines (SVM). Across the many different methods of ML, a common distinction is drawn between three types: supervised, unsupervised and reinforced learning.

Typical tasks performed by supervised learning are problems of discriminative classification where the ML algorithm assigns a probability of belonging to a certain category Y based on feature X. To do so, supervised learning requires labelled training data, matching the training instances to labels such as “diseased” – “healthy”, “developed psychosis” - “did not develop psychosis” or “positive treatment outcome” – “negative treatment outcome”. After training, the ML algorithm can then assign these labels correctly to new data. Unsupervised learning, on the other hand, does not require labelled training data. Instead, it can make use of often more readily available, unlabelled data, such as whole-genome sequences or cell phone metadata, to find clusters within these data points. In real-life settings, applications may fall between these two approaches and are described as “semi-supervised” or, as recently suggested by Yann LeCun, as “self-supervised” (LeCun, 2018), complementing labelled training data with large bits of unlabelled data (Chapelle, Schölkopf, & Zien, 2010). Finally, reinforcement learning denotes ML programs that optimize their interaction with an environment by trying to maximize reward over time (Mnih et al., 2015). While this approach, inspired by neuroscientific accounts of learning, does not require fully labelled data, it needs some formalization of rewards, e.g. winning an ATARI game.

The schematic distinction of these three general ML types can also be instructive for ethical debate of applied ML in psychiatry. For as we will show, differences in

methodology do not only have a big impact on feasibility since labelling of data often requires cost- and labour-intensive efforts but may also account for important ethical implications.

ML type	Required data	Typical Problem	Exemplary application in schizophrenia
Unsupervised	Unlabelled training data	Clustering	Refine diagnostic criteria **(case 1)**
Supervised	Labelled training data	Classification and regression	Improve diagnostic accuracy **(case 2)**
Reinforced	Labelled and unlabelled data	Dynamic decision-making	Suggest optimal treatment regime **(case 3)**

Table 7.1: Supervised, unsupervised and reinforced ML

Before turning to the potential of ML techniques to improve clinical care, some methodological limitations of psychiatric ML need to be mentioned, recently stressed by Vieira et al. (2019). Some of these concerns, such as small sample size or publication bias, are pervasive across different research areas and neuroscientific research in particular (Button et al., 2013; Kellmeyer, 2017; Schnack & Kahn, 2016). Other methodological issues arise with specific regard to ML, for example regarding failure to rigorously employ nested cross-validation, testing the predictions of a ML program on a fully independent sample (Stahl & Pickles, 2018). In addition, psychiatry's high-dimensional and often noisy data demand particular consideration and may hinder adopting computational strategies popular in other medical areas. While DL is frequently considered the method of choice for medical image analysis (Shen, Wu, & Suk, 2017), some recent results suggest that for imaging-based predictions of cognitive

and behavioural measures classical kernel regression is at least as successful as DL (He et al., 2019; Mihalik et al., 2019), rendering a linear and more interpretable approach (Heinrichs & Eickhoff, 2020) potentially preferable. These methodological challenges may partially account for inconsistent results across different studies, e.g. reporting largely variable accuracies for potential biomarkers of schizophrenia based on ML and neuroimaging (Kambeitz et al., 2015).

The potentially deepest challenge for implementing ML in psychiatry lies in its long-embattled nosology though (Kendler, 2016; Kendler, Zachar, & Craver, 2011; Zachar, 2015), calling into question the choice of appropriate data for training. Given that psychiatry arguably still lacks a successful diagnostic scheme that is valid and reliable (Barron, 2019), establishing psychiatric ML programs relies on a shaky ground truth. This problem is exacerbated by fundamental concerns whether a reductionist framework, considering psychiatric disorders as mere brain diseases to be investigated with neuroimaging and genetics, is convincing (Borsboom, Cramer, & Kalis, 2018). While we largely focus on neuroimaging studies in our examples for the sake of simplicity, research should thus be careful to not restrain their input a priori to biological data but also include social and idiosyncratic information on individual patients. Using natural language processing (NLP) on narrative electronic health records could provide a starting point for such an endeavour (Rumshisky et al., 2016).

7.3 Applications of ML for schizophrenia

Future ML applications for patients with schizophrenia may differ largely. For research purposes, using unsupervised learning to identify altered brain structures in patients with schizophrenia is common. In some of these possible approaches, which have been described as data- or discovery-oriented (Huys, Maia, & Frank, 2016; Krystal et al., 2017),

the algorithm is provided with neuroimaging data of patients with schizophrenia and left to find clusters (Dwyer et al., 2018; Schnack, 2017). Hence, apart from sample choice, little human labelling determines the data. Instead, the algorithm is left to find clusters that may or may not map onto a given hypothesis and can, in some cases, correlate with clinical data. Indeed, given the manifold disputes over psychiatric categorizations, some authors hope that embracing such a data-driven ML approach may provide new insights into neurobiological mechanisms of psychiatric diseases (Adams, Huys, & Roiser, 2016; Huys et al., 2016; Madsen, Krohne, Cai, Wang, & Chan, 2018; Skatun et al., 2017). A recent study that associated neuroanatomically distinct subtypes of schizophrenia with different illness duration and degrees of negative symptoms may serve as an example for this aspiration (Dwyer et al., 2018).

Also for diagnostic purposes, ML presents new opportunities for psychiatry. Based on specific changes in brain volume, several groups have shown that ML can distinguish non-medicated, first-episode patients with schizophrenia from healthy controls using volumetric MRI data (Chin, You, Meng, Zhou, & Sim, 2018; Gould et al., 2014; Haijma et al., 2013; Lee et al., 2018; Rozycki et al., 2018; Xiao et al., 2017). As noted, findings so far have been rather inconsistent and one should avoid overoptimistic interpretations of these results (Kambeitz et al., 2015; Vieira et al., 2019). Still, it seems reasonable to assume that in the future some ML techniques could assist physicians in their diagnostic process. Such applications could provide probabilistic estimates regarding one or several diagnostic labels such as schizophrenia, based on overlap with previously diagnosed patients. Arguably, most such methods would fall under the label of supervised learning since the training data need to be labelled, consisting of a vector of individual data such as brain data assigned to a category of “diseased” vs “healthy” respectively.

Finally, recent psychiatric advances employing ML have seen a turn towards predicting certain quantifiable events beyond diagnostic labels, e.g. providing probabilities for the likelihood of an onset of psychosis (Koutsouleris et al., 2018; Koutsouleris et al., 2015) or for the treatment success of one certain drug (Chekroud et al., 2016; Webb et al., 2018). While the majority of these approaches draw on supervised or unsupervised ML, some also use reinforcement learning to derive recommendations for optimal dynamic treatment regimes, using e.g. longitudinal data from so-called *Sequential Multiple Assignment Randomized Trials* (SMARTs). For example, by considering the treatment success of specific antipsychotics from the CATIE study (Stroup et al., 2003), Ertefaie et al. have constructed a Q-learning approach which optimizes treatment outcome based on a patient's characteristics (2016). Even more to the point, Koutsouleris et al. have shown that a cross-validated ML tool trained on diverse data from 334 patients could identify individuals which were more likely to benefit from treatment with amisulpride or olanzapine than with haloperidol, quetiapine or ziprasidone (2016). Such studies should be taken with a grain of salt though, given that there is no agreement what constitutes useful measures of treatment outcomes in psychiatry (Zimmerman & Mattia, 1999; Zimmerman, Morgan, & Stanton, 2018) – a conundrum the introduction of ML seems unlikely to solve.

7.4 Three cases and four principles

To highlight the dissimilarities between different usages, we provide three schematic cases that fall within the range of possible applications, from research to diagnosis and choice of treatment (Tbl. 2). All three cases, we hold it, touch upon important ethical concerns that can be discussed in accordance with the four principles put forth by

Beauchamp and Childress: beneficence, non-maleficence, respect for autonomy, and justice (Beauchamp & Childress, 2013).

Three potential applications for ML in schizophrenia
Case 1: *R is presenting with newly developed negative and positive symptoms at a university psychiatry department. Based on a clinical interview, R is diagnosed with schizophrenia by a psychiatrist. As part of a research program that aims to distinguish amongst schizophrenia subtypes, Z undergoes structural cranial magnetic resonance imaging (MRI) scanning which is analysed by a ML algorithm trained to find commonalities and differences of brain volume in specific cortical areas across all brain scans acquired from first-episode patients with schizophrenia presenting to the university hospital. Based on his brain scan, R is assigned to a subtype of schizophrenia with a typical pattern of superior-temporal grey matter loss.*
Case 2: *D is presenting at a psychiatric day-clinic with mild psychotic symptoms and is diagnosed with schizophrenia after a clinical interview. Given her markedly depressed mood and further reported symptoms such as insomnia, psychomotor retardation and strong headache, the attending psychiatrist also considers differential diagnoses such as a major depressive episode or a space-consuming intracerebral process. To exclude the latter, the attending psychiatrist refers her antipsychotic-naïve patient to a neuroradiologist to obtain a structural MRI. After segmentation of white- and grey-matter, the radiological data are fed to a machine learning algorithm which, based on previous training data in a comparable population, classifies the patient as suffering from schizophrenia with a probability of 70%. The psychiatrist sees her diagnosis confirmed and commences psychopharmacological treatment.*

Case 3: *T is diagnosed with a first episode of schizophrenia based on a clinical interview. To choose the most effective drug for his individual situation, his psychiatrist recommends a newly approved routine employing functional MRI during a reward-learning task. Based on T's brain activity and a plethora of other available information, from demographic data to his clinical records, the ML algorithm suggests one specific anti-psychotic drug as ideal for T's specific situation. Following the automated recommendation, the psychiatrist prescribes the drug to her patient.*

Table 2: Case Vignettes

7.3.1 Beneficence

The principle of beneficence expresses an aspiration to further the welfare and interests of others, potentially implying particular obligations of acting (Beauchamp & Childress, 2013, pp. 165-176). As our previous points and cases indicate, patients may benefit from applied ML in many different ways, both directly and indirectly.

Direct

Firstly, ML-supported diagnostic tools aim at improving diagnostic certainty. Techniques such as in the case of D (case 2) may serve as an automated second opinion, confirm a psychiatrist's judgement and help with unclear cases. In fact, if the algorithm is trained on data of the highest quality, which are e.g. labelled independently by several internationally leading and experienced psychiatrists, it could provide patients with a reliable diagnosis. Considering the difficulty of establishing whether schizophrenia is accurately diagnosed and given the considerable inter-rater disagreement among experts (Mokros, Habermeyer, & Kuchenhoff, 2018), a diagnostic algorithm supporting psychiatrists in their decision making could increase the likelihood of patients receiving

a correct diagnosis and hence of receiving an adequate treatment. By providing prognostic estimates concerning the future course of a disorder, such as the occurrence of psychotic episodes, or the success of specific treatments, ML applications may also help to reduce extraneous psychopharmacological interventions (Martinez-Martin et al., 2018) and track the progression of the disorder. This is the case for T (case 3), who may be spared an arduous trial-and-error regime of medication by an algorithm suggesting one potentially ideal medication early on. Of course, the benefits of a correct diagnosis might be infringed dramatically by additional risks, to which we turn later, if these diagnostic or predictive processes were to be left unchecked. However, at least for now, such a development seems rather unlikely, both technically and socially, in most medical specialties (Topol, 2019).

Indirect

Beyond these immediate clinical uses, patients may also benefit from research projects similar to our first case, leading to more accurate diagnostic categories. After all, most current psychiatric diagnoses as enshrined in the DSM or ICD are purely descriptive, optimized primarily for validity and inter-rater reliability, not for underlying pathophysiology – but this lack of concern for etiological underpinnings has long been of concern to many in the field (Hyman, 2011). In contrast, computational approaches based on ML aspire "to automatically segregate brain disorders into natural kinds" (Bzdok & Meyer-Lindenberg, 2018). Notwithstanding conceptual questions regarding the nature of psychiatric disorders (Kendler, 2016; Zachar, 2015), ML may be eminently suited to develop biologically more plausible diagnostic categories, allowing for more specific treatment options. After all, concerns of insufficiently grasping psychiatric complexity has long accompanied the development of psychiatric biomarkers (Singh &

Rose, 2009). ML drawing on rich data, from detailed biological information such as (f)MRI scans or whole genome sequences to demographic data and electronic health records, could arguably accommodate such complexity. Still, the concern remains that ML applications drawing on ML may overtly reify diagnostic categories designed as heuristic constructs (Hyman, 2010) – and thus end up harming patients.

7.3.2 Non-maleficence

Abstaining from harm is a bedrock of clinical practice (Smith, 2005). How does ML in psychiatry fare with regard to this crucial principle? Firstly, privacy concerns may come to mind here (Vayena et al., 2018). How is sensitive medical information disclosed to an algorithm and how can data created by the algorithm be protected appropriately? These are essential questions but only concern ML techniques indirectly, via the data used and produced by its applications. Since privacy issues of big data have been addressed extensively elsewhere (Price & Cohen, 2019), we will leave them aside here to focus on harm potentially caused by ML in psychiatry. As in the case of benefits, there are both direct and indirect ways in which its use may harm patients.

Direct

First, using an algorithm may bring about harm directly, e.g. when the diagnosis or predictions made by the ML application are erroneous. Previous shortcomings of health-related ML can be instructive here. IBM's ML-based computer system Watson, advertised as a revolutionary tool for cancer care, has been shown to recommend unsafe treatments endangering patients' safety and health (Ross, 2018). Such errors are particularly worrying if recommendations of algorithms are readily accepted by medical staff, as in T's case, or if the process would become fully automated. Although an erroneous algorithm is likely to affect more patients compared to an individual mistake

made by a physician, errors are far from exclusive to algorithms (McLennan et al., 2013) and these concerns could be tackled by a model of shared responsibility in which competent human agents check the ML-based suggestions (Topol, 2019). However, as opposed to human physicians, a trained ML algorithm may not be flexible enough to account for contextual changes such as the swift rise of smartphone usage or altered eating habits. Given the dependency of psychiatric conditions on contingent societal contexts, even a tested and approved program may thus require regular overhauling and retraining to avoid systematic misjudgements.

Indirect

The more intricate questions seem to arise from indirect effects of using ML in patients with schizophrenia. By potentially modifying the expectations of doctors, the result of a computationally assigned risk-category will most likely influence downstream diagnostic and therapeutic decision-making. For example, in mammography screening risk stratification affects the detection performance of radiologists: a known BRCA mutation strongly decreases the number of missed visible breast cancer lesions in MRI scans (Vreemann et al., 2018). Timing the disclosure of ML-based computations to the physician is thus crucial: should she have to decide on one diagnosis first before being confronted with the results of ML diagnostics? Furthermore, the impact of incorporating ML in the clinical setting will require additional scrutiny regarding its effects on the therapeutic relationship. How do patients perceive the use of ML by their physicians to arrive at diagnostic judgements or prognostic estimates? Does it impair their trust in health care professionals and if so, could it harm their compliance and the therapeutic outcome? These questions are of particular importance in the case of psychiatric patients who are particularly vulnerable to so-called “diagnostic overshadowing”, i.e.

health care professionals falsely attributing somatic symptoms to known mental health issues (Callard, Bracken, David, & Sartorius, 2013; Jones, Howard, & Thornicroft, 2008; Shefer, Henderson, Howard, Murray, & Thornicroft, 2014). These challenges merit ongoing attention and require accompanying efforts of clinical ML implementation with corresponding empirical bioethical research to explore potential negative impact.

7.3.3 Patients' autonomy and clinicians' judgement

Respect for autonomy demands conveying sufficiently detailed and understandable information to patients about planned medical procedures and asking for their consent (Manson & O'Neill, 2007). Such disclosure may be particularly challenging in cases of applied ML, used by medical practitioners who may themselves not fully understand the mathematical underpinnings of an algorithm. Does the, to some extent, unavoidable opacity of ML, commonly discussed as "black box"-problem, clash with the requirement to appropriately inform patients? And should one ask patients for their explicit consent when using (existing) data before providing it to the algorithm at all? After all, obtaining informed consent for the use of predictive analytics is not legally mandatory at the moment (Cohen, Amarasingham, Shah, Xie, & Lo, 2014). One could wonder whether discussing ML algorithms with a group as vulnerable as patients at risk of psychosis or paranoid symptoms might not exacerbate their situation and cause severe additional psychological stress (Martinez-Martin et al., 2018).

Questions of autonomy also stretch to the domain of medical doctors' discernment and respecting clinicians' judgement is vital in the context of modern health care systems (Faden et al., 2013). Much depends on the conceptualization of the relation between human expert and ML algorithm. One analogy, recently proposed by Eric Topol (Topol, 2019), suggests that we conceptualize the relation of clinician and algorithm similarly to

assisted driving and increasingly autonomous cars. While the machine may take over some tasks, the drivers or physicians need to remain in charge as a backup, checking the machine's output by comparing it to their own judgements. This would facilitate attributing degrees of responsibility to health care personnel, clarifying important issues of accountability and liability. It implies that human agents need to remain able to weigh ML recommendations and potentially decide against them. Ideally, as a safeguard against bad judgements by single individuals one could envision provisions in which disagreements between physicians and ML application lead to consultations with other clinicians, e.g. during departmental meetings, providing an opportunity to sharpen the clinical skills of everyone involved. Furthermore, an institutional framework may be needed to test and approve ML applications in a similar fashion as pharmaceutical products (Paulus, Huys, & Maia, 2016).

7.3.4 Fair allocation and systematic biases

Finally, using ML in psychiatry also raises important issues concerning justice, from financial aspects to systematic biases. Does increased diagnostic certainty justify the allocation of scarce financial means to additional computational efforts and vindicate even highly expensive exams such as (f)MRI? Integrating the data from examinations such as MRI into psychiatric routines may pose additional serious challenges for equal treatment if certain patients cannot undergo scanning due to limited availability or contraindications such as claustrophobia. Arguably, any new technique needs to establish a measurable clinical benefit over a conventional psychiatric assessment to vindicate its cost (Iwabuchi, Liddle, & Palaniyappan, 2013), or show that it can avoid costs elsewhere. With regard to discerning different diagnostic entities, research based on ML could also lead to issues commonly known as salami slicing: even without

understanding the underlying pathophysiological mechanisms, lobbying by pharmaceutical companies might have an interest to split psychiatric disorders into many distinct categories to gain advantages in the approval of new drugs. On the other hand we should not forget that in many countries only a very limited amount of the overall healthcare budget is allocated to mental health (World Health Organization, 2018). More precise diagnoses and better treatments might convince policymakers to overcome this health disparity, ultimately empowering psychiatric patients.

Of further concern are systematic biases, easily induced by poor training data and particularly worrisome in diagnostic contexts (Vayena et al., 2018). The example of schizophrenia is a case in point, with its long-standing disproportionate number of diagnoses in African-Americans and Latin-Americans, arguably influenced by stereotypes, the clinician's own ethnicity or the under-diagnosis of other psychiatric diseases (Schwartz & Blankenship, 2014). ML trained on data with these or other biases could further purport and reify misconceptions (Tandon & Tandon, 2018). If training data are less than carefully curated, ML applications might hence not constitute an independent diagnostic tool for enhancing diagnostic accuracy, undermining the endeavour's very aim. To avoid perpetuating pathophysiologically misleading biases, developing appropriate supervision strategies for the ML algorithm thus seems key to a successful clinical implementation. Such supervision should (1) track which parameters are taken into account by the algorithm to arrive at its recommendations and (2) compare the results of algorithms trained on different databases. Such strategies would also help to foster explicability which the initially mentioned AI4people initiative rightly suggests as a fifth principle for ethical AI use, enabling the other four (Floridi et al., 2018). The implementation of such safety measures will be critical for minimizing biases

in decision making but it is not yet clear how ML algorithms will nonetheless capitalize on existing biases in the data.

7.5 Conclusion

A plethora of context-specific ethical issues might arise in applied ML in psychiatry and the treatment of schizophrenia. For now, ML remains in the domain of research and should be accompanied by exploring its ethical aspects as there is no standard rule to determine when an application is ethically permissible given the complexity of each singular case. Further, empowering psychiatric patients can only happen with the help of important support systems such as family, peer and community members. Still, if some of the vast potential benefits of psychiatric ML can indeed lead to tangible improvements for patients, we believe it is not only permissible but it may in fact be a moral obligation to pursue them further and aim at their successful clinical implementation.

7.6 References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry, 87*(1), 53-63. doi:10.1136/jnnp-2015-310737

Barron, D. (2019). Should Mental Disorders Have Names? [Scientific American]. *Scientific American*. Retrieved from https://blogs.scientificamerican.com/observations/should-mental-disorders-have-names

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Borgwardt, S., Koutsouleris, N., Aston, J., Studerus, E., Smieskova, R., Riecher-Rössler, A., & Meisenzahl, E. M. (2013). Distinguishing prodromal from first-episode psychosis using neuroanatomical single-subject pattern recognition. *Schizophrenia Bulletin, 39*(5), 1105-1114. doi:10.1093/schbul/sbs095

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. doi:10.1017/S0140525X17002266

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi:10.1038/nrn3475

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(3), 223-230. doi: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. doi:10.1136/bmj.f4312

Chapelle, O., Schölkopf, B., & Zien, A. (2010). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *New England Journal of Medicine, 378*(11), 981-983. doi:10.1056/NEJMp1714229

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. *Lancet Psychiatry, 3*(3), 243-250. doi:10.1016/S2215-0366(15)00471-X

Chin, R., You, A. X., Meng, F., Zhou, J., & Sim, K. (2018). Recognition of Schizophrenia with Regularized Support Vector Machine and Sequential Region of Interest Selection using Structural Magnetic Resonance Imaging. *Scientific Reports, 8*(1), 13858. doi:10.1038/s41598-018-32290-9

Chung, Y., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., McGlashan, T., Perkins, D., Seidman, L. J., Tsuang, M., Walker, E., Woods, S. W., McEwen, S., van Erp, T. G. M., Cannon, T. D., North American Prodrome Longitudinal Study, C., the Pediatric Imaging, N., & Genetics Study, C. (2018). Use of Machine Learning to Determine Deviance in Neuroanatomical Maturity Associated With Future Psychosis in Youths at Clinically High Risk. *JAMA Psychiatry, 75*(9), 960-968. doi:10.1001/jamapsychiatry.2018.1543

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs, 33*(7), 1139-1147. doi:10.1377/hlthaff.2014.0048

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., Schatzberg, A. F., Sudheimer, K., Keller, J., Mayberg, H. S., Gunning, F. M., Alexopoulos, G. S., Fox, M. D., Pascual-Leone, A., Voss, H. U., Casey, B. J., Dubin, M. J., & Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine, 23*(1), 28-38. doi:10.1038/nm.4246

Dwyer, D. B., Cabral, C., Kambeitz-Ilankovic, L., Sanfelici, R., Kambeitz, J., Calhoun, V., Falkai, P., Pantelis, C., Meisenzahl, E., & Koutsouleris, N. (2018). Brain Subtyping Enhances

The Neuroanatomical Discrimination of Schizophrenia. *Schizophrenia Bulletin, 44*(5), 1060-1069. doi:10.1093/schbul/sby008

Ebdrup, B. H., Axelsen, M. C., Bak, N., Fagerlund, B., Oranje, B., Raghava, J. M., Nielsen, M. O., Rostrup, E., Hansen, L. K., & Glenthoj, B. Y. (2018). Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naive schizophrenia patients. *Psychological Medicine*, 1-10. doi:10.1017/s0033291718003781

Ertefaie, A., Shortreed, S., & Chakraborty, B. (2016). Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in Medicine, 35*(13), 2221-2234. doi:10.1002/sim.6859

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, (2019). Retrieved from https://data.europa.eu/doi/10.2759/177365

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. *Hastings Center Report, 43*(1), S16-27. doi:10.1002/hast.134

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143*(2), 187-232. doi:10.1037/bul0000084

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics, 24*(11), 1037-1052. doi:10.1111/cns.13048

Giordano, G. M., & Borgwardt, S. (2019). Current goals of neuroimaging for mental disorders: a report by the WPA Section on Neuroimaging in Psychiatry. *World Psychiatry, 18*(2), 241-242. doi:10.1002/wps.20652

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. doi:10.1016/j.nicl.2014.09.009

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: a meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin, 39*(5), 1129-1138. doi:10.1093/schbul/sbs118

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. T. T. (2019). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*, 116276. doi:10.1016/j.neuroimage.2019.116276

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*(6), 1435-1444. doi:10.1002/hbm.24886

Hofstadter, D. R. (1980). *Gödel, Escher, Bach: an eternal golden braid*. New York: Vintage Books.

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. doi:10.1038/nn.4238

Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual Review of Clinical Psychology, 6*, 155-179. doi:10.1146/annurev.clinpsy.3.022806.091532

Hyman, S. E. (2011). Diagnosing the DSM: Diagnostic Classification Needs Fundamental Reform. *Cerebrum, 2011*, 6. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23447775

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine, 46*(12), 2455-2465. doi:10.1017/S0033291716001367

Iwabuchi, S. J., Liddle, P. F., & Palaniyappan, L. (2013). Clinical utility of machine-learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. *Frontiers in Psychiatry, 4*, 95. doi:10.3389/fpsyt.2013.00095

Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3*(9), 798-808. doi:10.1016/j.bpsc.2018.04.004

Jones, S., Howard, L., & Thornicroft, G. (2008). 'Diagnostic overshadowing': worse physical health care for people with mental illness. *Acta Psychiatrica Scandinavica, 118*(3), 169-171. doi:10.1111/j.1600-0447.2008.01211.x

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour, 1*, 911-919. doi:10.1038/s41562-017-0234-y

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., & Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: a meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology, 40*(7), 1742-1751. doi:10.1038/npp.2015.22

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry, 17*(12), 1174-1179. doi:10.1038/mp.2012.105

Kellmeyer, P. (2017). Ethical and Legal Implications of the Methodological Crisis in Neuroimaging. *Cambridge Quarterly of Healthcare Ethics, 26*(4), 530-554. doi:10.1017/S096318011700007X

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry, 15*(1), 5-12. doi: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. doi:10.1017/S0033291710001844

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry, 3*(10), 935-946. doi:10.1016/S2215-0366(16)30171-7

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R. K. R., Pantelis, C., Meisenzahl, E., Wood, S. J., Beque, D., Brambilla, P., Borgwardt, S., & Consortium, P. (2018). Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry, 75*(11), 1156-1172. doi:10.1001/jamapsychiatry.2018.2165

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., von Saldern, S., Cabral, C., Reiser, M., Falkai, P., & Borgwardt, S. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin, 41*(2), 471-482. doi:10.1093/schbul/sbu078

Krystal, J. H., Murray, J. D., Chekroud, A. M., Corlett, P. R., Yang, G., Wang, X. J., & Anticevic, A. (2017). Computational Psychiatry and the Challenge of Schizophrenia. *Schizophrenia Bulletin, 43*(3), 473-475. doi:10.1093/schbul/sbx025

LeCun, Y. (2018). The Power and Limits of Deep Learning. *Research-Technology Management, 61*(6), 22-27. doi:10.1080/08956308.2018.1516928

Lee, J., Chon, M. W., Kim, H., Rathi, Y., Bouix, S., Shenton, M. E., & Kubicki, M. (2018). Diagnostic value of structural and diffusion imaging measures in schizophrenia. *Neuroimage-Clinical, 18*, 467-474. doi:10.1016/j.nicl.2018.02.007

Madsen, K. H., Krohne, L. G., Cai, X. L., Wang, Y., & Chan, R. C. K. (2018). Perspectives on Machine Learning for Classification of Schizotypy Using fMRI Data. *Schizophrenia Bulletin, 44*(suppl_2), S480-S490. doi:10.1093/schbul/sby026

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Martinez-Martin, N., Dunn, L. B., & Roberts, L. W. (2018). Is It Ethical to Use Prognostic Estimates from Machine Learning to Treat Psychosis? *AMA Journal of Ethics, 20*(9), E804-811. doi:10.1001/amajethics.2018.804

McLennan, S., Engel, S., Ruhe, K., Leu, A., Schwappach, D., & Elger, B. (2013). Implementation status of error disclosure standards reported by Swiss hospitals. *Swiss Medical Weekly, 143*, w13820. doi:10.4414/smw.2013.13820

Mihalik, A., Brudfors, M., Robu, M., Ferreira, F. S., Lin, H., Rau, A., Wu, T., Blumberg, S. B., Kanber, B., Tariq, M., Del Mar Estarellas Garcia, M., Zor, C., Nikitichev, D. I., Mourao-Miranda, J., & Oxtoby, N. P. (2019). ABCD Neurocognitive Prediction Challenge 2019: Predicting individual fluid intelligence scores from structural MRI using probabilistic segmentation and kernel ridge regression. *arXiv e-prints*. Retrieved from https://ui.adsabs.harvard.edu/abs/2019arXiv190510831M

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529-533. doi:10.1038/nature14236

Mokros, A., Habermeyer, E., & Kuchenhoff, H. (2018). The uncertainty of psychological and psychiatric diagnoses. *Psychological Assessment, 30*(4), 556-560. doi:10.1037/pas0000524

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A Roadmap for the Development of Applied Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 386-392. doi:10.1016/j.bpsc.2016.05.001

Perna, G., Grassi, M., Caldirola, D., & Nemeroff, C. B. (2018). The revolution of personalized psychiatry: will technology make it happen sooner? *Psychological Medicine, 48*(5), 705-713. doi:10.1017/S0033291717002859

Price, W. N., 2nd, & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine, 25*(1), 37-43. doi:10.1038/s41591-018-0272-7

Ross, C., Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News.* Retrieved from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., Fan, Y., Gur, R. E., Gur, R. C., Meisenzahl, E. M., Zhuo, C., Yin, H., Yan, H., Yue, W., Zhang, D., & Davatzikos, C. (2018). Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. *Schizophrenia Bulletin, 44*(5), 1035-1044. doi:10.1093/schbul/sbx137

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry, 6*(10), e921. doi:10.1038/tp.2015.182

Schnack, H. G. (2017). Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia Research*. doi:10.1016/j.schres.2017.10.023

Schnack, H. G., & Kahn, R. S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry, 7*, 50. doi:10.3389/fpsyt.2016.00050

Schwartz, R. C., & Blankenship, D. M. (2014). Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World Journal of Psychiatry, 4*(4), 133-140. doi:10.5498/wjp.v4.i4.133

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426-1448. doi:10.1017/s0033291719000151

Shefer, G., Henderson, C., Howard, L. M., Murray, J., & Thornicroft, G. (2014). Diagnostic overshadowing and other challenges involved in the diagnostic process of patients with mental illness who present in emergency departments with physical symptoms-a qualitative study. *PLoS One, 9*(11), e111682. doi:10.1371/journal.pone.0111682

Shen, D., Wu, G., & Suk, H. I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering, 19*, 221-248. doi:10.1146/annurev-bioeng-071516-044442

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature, 460*(7252), 202-207. doi:10.1038/460202a

Skatun, K. C., Kaufmann, T., Doan, N. T., Alnaes, D., Cordova-Palomera, A., Jonsson, E. G., Fatouros-Bergman, H., Flyckt, L., KaSp, Melle, I., Andreassen, O. A., Agartz, I., & Westlye, L. T. (2017). Consistent Functional Connectivity Alterations in Schizophrenia Spectrum Disorder: A Multisite Study. *Schizophrenia Bulletin, 43*(4), 914-924. doi:10.1093/schbul/sbw145

Smith, C. M. (2005). Origin and uses of primum non nocere--above all, do no harm! *Journal of Clinical Pharmacology, 45*(4), 371-377. doi:10.1177/0091270004273680

Stahl, D., & Pickles, A. (2018). Fact or fiction: reducing the proportion and impact of false positives. *Psychological Medicine, 48*(7), 1084-1091. doi:10.1017/S003329171700294X

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., McGee, M. F., Simpson, G. M., Stevens, M. C., & Lieberman, J. A. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. *Schizophrenia Bulletin, 29*(1), 15-31. doi:10.1093/oxfordjournals.schbul.a006986

Tandon, N., & Tandon, R. (2018). Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia. *Schizophrenia Bulletin, 44*(5), 939-941. doi:10.1093/schbul/sby101

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56. doi:10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689. doi:10.1371/journal.pmed.1002689

Vieira, S., Gong, Q. Y., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., Tordesillas-Gutierrez, D., Ortiz-Garcia, V., Setien-Suero, E., Scheepers, F. E., Van Haren, N. E. M., Marques, T. R., Murray, R. M., David, A., Dazzan, P., McGuire, P., & Mechelli, A. (2019). Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophrenia Bulletin*. doi:10.1093/schbul/sby189

Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., & Mann, R. M. (2018). The frequency of missed breast cancers in women participating in a high-risk MRI screening program. *Breast Cancer Research and Treatment, 169*(2), 323-331. doi:10.1007/s10549-018-4688-z

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science, 5*(3), 457-469. doi:10.1177/2167702617691560

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., Fava, M., McGrath, P. J., Weissman, M., Parsey, R., Adams, P., Trombello, J. M., Cooper, C., Deldin, P., Oquendo, M. A., McInnis, M. G., Huys, Q., Bruder, G., Kurian, B. T., Jha, M., DeRubeis, R. J., & Pizzagalli, D. A. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological Medicine, 49*(7), 1118-1127. doi:10.1017/S0033291718001708

World Health Organization, W. (2018). *Mental health atlas 2017*. Geneva, Switzerland: WHO.

Xiao, Y., Yan, Z., Zhao, Y., Tao, B., Sun, H., Li, F., Yao, L., Zhang, W., Chandan, S., Liu, J., Gong, Q., Sweeney, J. A., & Lui, S. (2017). Support vector machine-based classification of first episode drug-naive schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*. doi:10.1016/j.schres.2017.11.037

Zachar, P. (2015). Psychiatric disorders: natural kinds made by the world or practical kinds made by us? *World Psychiatry, 14*(3), 288-290. doi:10.1002/wps.20240

Zimmerman, M., & Mattia, J. I. (1999). Psychiatric diagnosis in clinical practice: is comorbidity being missed? *Comprehensive Psychiatry, 40*(3), 182-191. doi:10.1016/s0010-440x(99)90001-9

Zimmerman, M., Morgan, T. A., & Stanton, K. (2018). The severity of psychiatric disorders. *World Psychiatry, 17*(3), 258-275. doi:10.1002/wps.20569

Chapter 8: Explainability as Fig Leaf? An Exploration of Researchers' Ethical Expectations Towards Machine Learning in Psychiatry

Explainability as Fig Leaf? An Exploration of Researchers' Ethical Expectations Towards Machine Learning in Psychiatry

Georg Starke[1], Benedikt Schmidt[1], Eva De Clercq[1], Bernice Simone Elger[1,2]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland, [2] University Center of Legal Medicine, University of Geneva, Switzerland.

Acknowledgements: First and foremost, we would like to thank all our interviewees for their time and willingness to participate in our study, despite their multiple obligations during the pandemic. GS would further like to thank the Fondation Brocher, Hermance, Switzerland, and its staff for their generous support during a 1-month fellowship that allowed the completion of this paper.

This version of the article has been accepted for publication in *AI and Ethics*, after peer review, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s43681-022-00177-1

Abstract

Background: The increasing implementation of programs supported by machine learning in medical contexts will affect psychiatry. It is crucial to accompany this development with careful ethical considerations informed by empirical research involving experts from the field, to identify existing problems and to address them with fine-grained ethical reflection.

Methods: We conducted semi-structured qualitative interviews with 15 experts from Germany and Switzerland with training in medicine and neuroscience on the assistive use of machine learning in psychiatry. We used reflexive thematic analysis to identify key ethical expectations and attitudes towards machine learning systems.

Results: Experts' ethical expectations towards machine learning in psychiatry partially challenge orthodoxies from the field. We relate these challenges to three themes, namely (1) ethical challenges of machine learning research, (2) the role of explainability in research and clinical application, and (3) the relation of patients, physicians, and machine learning system. Participants were divided regarding the value of explainability, as promoted by recent guidelines for ethical artificial intelligence, and highlighted that explainability may be used as an ethical fig leaf to cover shortfalls in data acquisition. Experts recommended increased attention to machine learning methodology, and the education of physicians as first steps towards a potential use of machine learning systems in psychiatry.

Conclusion: Our findings stress the need for domain-specific ethical research, scrutinizing the use of machine learning in different medical specialties. Critical ethical research should further examine the value of explainability for an ethical development of machine learning systems and strive towards an appropriate framework to communicate ML-based medical predictions.

8.1 Introduction

The integration of diagnostic, predictive, and therapeutic tools based on machine learning (ML) into clinical care is accelerating – a development also apparent in psychiatry. Beyond increasingly popular direct-to-consumer apps, offering for instance digital psychotherapy (Lui, Marcus, & Barry, 2017; Martinez-Martin & Kreitmair, 2018), the US Food and Drug Administration (FDA) recently approved the first ML-based psychiatric tool, providing diagnostic aid based on joint inputs from caregivers and attending physicians (Dattaro, 2021). Many further attempts to employ ML in psychiatry are under way, covering a multitude of psychiatric disorders and ranging from diagnostic and prognostic tools to the prediction of treatment outcomes (Chekroud et al., 2021; Chivilgina, Elger, & Jotterand, 2021; Chivilgina, Wangmo, Elger, Heinrich, & Jotterand, 2020; Salazar de Pablo et al., 2021). A broad debate about the ethical principles governing the development of ML-based psychiatric tools seems therefore more pressing then ever (Jacobson et al., 2020; Starke, De Clercq, Borgwardt, & Elger, 2021).

With view to artificial intelligence (AI) in general, many recent guidelines have attempted to spell out specific ethical principles that researchers and regulators should respect. While different guidelines around the globe stress different ethical aspirations, there is a substantive convergence with regard to a handful of fundamental principles, such as transparency, fairness, and non-maleficence (Jobin, Ienca, & Vayena, 2019). For a debate within the context of European health care, the influential ethical framework of 'AI4people' seems particularly instructive (Floridi et al., 2018). It builds on the four principles of biomedical ethics by Beauchamp and Childress (Beauchamp & Childress, 2013), i.e. respect for autonomy, beneficence, non-maleficence, and justice, supplementing them with an additional fifth principle of explicability. Within the EU,

this framework has exerted particular influence as it served as a blue-print for the EU commission's ethical guidelines for trustworthy AI ("Ethics guidelines for trustworthy AI," 2019).

Yet, despite international attempts to provide ethical guidelines for the development of responsible or trustworthy AI and develop a suitable regulatory framework, there remains large uncertainty whether and how such principles translate into practice (Floridi, 2019). Since regulation and ethical debates typically lag behind the newest technological developments, several models have recently been suggested how ethical research using social science methods could be brough up to speed, taking place in parallel to developments, or how ethical considerations could be embedded in research pipelines (Jongsma & Bredenoord, 2020; McLennan et al., 2020). Nevertheless, as of now, there is little empirical data on how physicians and researchers perceive current guidelines, and whether their own ethical expectations towards ML systems are in alignment with recommended general principles. Yet research involving people working in the field is crucial to improve bioethical theory and develop appropriate policy suggestions, as the 'empirical turn' in bioethics has stressed (Wangmo et al., 2018). Notable exceptions that extend to multiple medical specialties include, for instance, Nichol et al. who have investigated experts' ethical perspectives on using ML to predict HIV risk in sub-Saharan Africa (Nichol, Bendavid, Mutenherwa, Patel, & Cho, 2021), whereas Blease et al. focused on the views of UK General Practitioners (Blease et al., 2019), and Tonekaboni et al. examined expectations towards explainability among 10 Canadian acute care specialists (Tonekaboni, Joshi, McCradden, & Goldenberg, 2019).

Findings with specific view to psychiatric practice are even scarcer and current research does only provide qualitative reasons of limited depth, due to being based on online surveys with comment boxed, as opposed to (semi-)structured interviews. A recent evaluation of an online survey among psychiatrists in 22 countries found a surprising lack of engagement with AI ethics, reporting that only 9 out of a sample of 791 participants mentioned ethical considerations when asked about the impact of ML and AI on future psychiatric practice (Blease, Locher, Leon-Carlyle, & Doraiswamy, 2020). An online survey among Swiss postgraduate students in clinical psychology, some of which were intending to pursue a psychotherapeutic career, reported greater concern with ethical questions (Blease, Kharko, Annoni, Gaab, & Locher, 2021).

Our study contributes to this emerging field of research. It provides a first insight into the attitudes of academic experts whose work is concerned with the use of ML systems in psychiatry by eliciting their explicit and implicit knowledge of ethical challenges posed by such systems. It thereby adds to recent qualitative research interviewing experts on the implementation of ML in healthcare (Cai, Winter, Steiner, Wilcox, & Terry, 2019; Morgenstern et al., 2021; Pumplun, Fecho, Wahl, Peters, & Buxmann, 2021), however, with a unique focus on ethical challenges and on ML applications in psychiatry. Expert interviews are an established method to elicit both explicit and implicit knowledge from people working in the field (Döringer, 2021). In the context of ethics, they provide a tool to better understand the actual challenges in the field, enabling ethical reflection that pays close attention to its context and thereby fill "blind spots in AI ethics" (Hagendorff, 2021). In current debates about medical ML, such close attention seems all the more necessary since prominent scholars have criticized forms of AI ethics that merely provide formulaic checklists (Braun, Bleher, & Hummel, 2021) and do not

pay enough attention to ethically relevant, yet often neglected aspects such as environmental cost or exploitation of labor (Crawford, 2021). Investigating potentially problematic conditions of academic knowledge and ML production therefore demands qualitative research examining the actual ramifications of academic research in the field.

In our study, we focused on scholars affiliated to psychiatric departments in Switzerland and Germany. Given the interconnected regulatory frameworks and the high number of German physicians and researchers in Switzerland, our sample offers a relatively homogeneous sample, providing insights into the attitudes of experts from the largest Western European language community. Such homogeneity seemed crucial to gathering context-sensitive information since large cultural differences regarding technology acceptance have been reported not only between Europe and the US or China (Bröhl, Nelles, Brandl, Mertens, & Nitsch, 2019) but also across Western European countries (Conti, Cattani, Di Nuovo, & Di Nuovo, 2015; Van den Berg, 2012). Here, we focus on what experts on psychiatric ML consider the most pressing ethical challenges for their field if asked under the condition of anonymity, and how they suggest solving them.

To our knowledge, our paper reports the first findings from qualitative expert interviews on the ethical challenges posed by ML in the context of psychiatry. Besides the clinical and research community from this specific field, our findings are also of interest to researchers working on the ethics of medical AI, informing the lively debate about opaque ML in medicine more generally (Braun et al., 2021; Durán & Jongsma, 2021; London, 2019), as well as to tailor policy making for the introduction of ethically sound, trustworthy ML in the clinic (Char, Abramoff, & Feudtner, 2020; Paulus, Huys, & Maia, 2016; Walter et al., 2019).

8.2 Methods

Our study included Swiss and German experts on the use of ML in psychiatry. Our recruiting strategy was two-pronged. 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 stage and gender. Potential candidates were invited to participate in our study via e-mail and received a reminder after a week in case they did not reply. We only invited experts who held at least a doctorate in a relevant field

Interviews were conducted between April 2020 and July 2021 by the first author, a physician (MD) with additional degrees 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. The first three interviews served as pilot interviews, after which a critical revision of the interview guide by all authors resulted in minor changes. Owing to the constraints of the pandemic, interviews took place exclusively via phone (10) or online video call (5), were conducted in German (13) or English (2), depending on the experts' preferences, and lasted 25 to 66 minutes. The interviews were transcribed verbatim by the first and second author. Quotes used within this paper were translated by GS and checked by BS and EDC. The interviewer knew three of the participants through prior research activities.

To identify important ethical themes within the interviews, we analyzed our data by conducting a reflexive thematic analysis (Braun & Clarke, 2006, 2019). We assigned individual codes to each segment of the transcripts of our interviews, with one segment representing a unit of meaning, consisting of one or more sentences. The coding was conducted jointly by all authors for four interviews. Having agreed upon a coding tree structure, comprising themes and subthemes, the remaining transcripts were coded by the first author, using MaxQDA software. To monitor data saturation, conceptualized as thematic redundancy indicated by recurrent coding, data analysis took place in parallel to data collection (Given, 2015). In line with previous findings, we did not find new codes after coding the 11th interview (Guest, Bunce, & Johnson, 2006).

Prior to the pilot interviews, we submitted a description of our study design including the consent sheet and the interview guide for review to the cantonal research ethics committee (Ethikkommission Nordwest- und Zentralschweiz, EKNZ). Within the Swiss legal framework, the ethics committee judged that the project did not fall under restrictions imposed on research with human subjects, as stated in a certificate of non-objection (Req-2019-00920). Nevertheless, to ensure high ethical standards of our bioethical project, we adhered to the following procedures (1) we asked participants for their written informed consent prior to their participation in our study and again orally at the beginning of the interview, (2) we omitted identifying information such as names and places in the transcripts, (3) and stored this de-identified data separately on our secure university servers.

To allow for a more detailed analysis of our findings, we divided our data into two separate manuscripts. Here, we focus on ethical concerns that relate to the use of AI in the clinic more generally, whereas the second manuscript covers themes that are

particular to the practice of psychiatry, such as the definition of psychiatric disorders. Questions from the interview guide that are relevant to the current manuscript are provided in Table 1.

What would you consider the biggest ethical challenge for successfully implementing ML in clinical contexts? - What do you think is the best way to address this issue? Do you have an example?
What specific expectations would you have for the transparency of such programs? Which technical strategies for making machine learning more transparent do you think are most promising? Could you give an example?
Should black box programs be used for clinical purposes? Why/why not?
Do you think trust is a justifiable way of dealing with the risks of medical AI? Why / why not? What expectations would you have for a program to be considered "trustworthy"?

Table 8.1. Relevant questions from the interview guide

8.3 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. Having achieved data saturation, we stopped recruiting additional participants. All participants held at least a doctorate and considered themselves experts on the use of ML in psychiatry (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 diseases. Ten participants were licensed physicians and five had degrees in psychology or neuroscience. Reflecting the multidisciplinary nature of the research field, eight participants reported additional formal education in mathematics, physics, engineering, and philosophy. Given the lack

of established ML routines in psychiatry and our recruitment strategy that focused on research outputs, the interviewed experts should rather be considered to be involved in the development of ML systems but also reflect the views of potential users, as indicated by their involvement in clinical contexts.

Analysis of the interviews resulted in three major themes, namely 1) ethical challenges of machine learning research, (2) the role of explainability in research and clinical application, and (3) the relation of patients, physicians, and machine learning system.

8.3.1. Ethical challenges of machine learning research

While only one interviewee was familiar with current ethics guidelines such as the EU guidelines for trustworthy AI, the experts exhibited great awareness of the ethical problems they encounter in their work, and in the development of new ML models. Many of these challenges concern the ramifications of academic research itself. Continuous pressure to produce promising results and publish frequently in high-ranking journals were reported to be at odds with methodological rigor, potentially already at the stage of collecting representative training data, including non-Western contexts because, as one participant put it, "everyone wants to get their paper out and not be told: go to Malaysia and collect data from 500 more people. That's difficult, expensive, and complicated, and that's why nobody does it." (P11) Yet, as several participants stressed, such shortfalls could lead to systematic bias if there is no incentive to acquire training data that fully mirror a phenomenon's complexity. Another respondent argued:

> There are these examples that algorithms are partly racist or so, simply because of their experiences - their lack of experiences - that they have collected. Just like a human being who lives in a small white village and has reservations about

> foreigners – that's just how a machine works as well. If it's fed the same information over and over again and never sees certain things (P2).

In consequence, all participants were concerned with questions of justice and algorithmic fairness resulting from training data that lacked diversity in the recruited cohort. Several interviewees named discrimination based on ethnicity, gender, or socio-economic status as major ethical concern for using ML in clinical contexts; a problem that mirrored existing bias in current medical practice.

> Of course, it is a methodological and ethical challenge to avoid such unintentional bias or at least make it visible. I believe that this has the potential to cause real damage. Of course, it is also the case that in the current medical system we already have a fairly high degree of bias and probably also systematic bias for the majority population and against minorities. But due to the learning aspect of AI algorithms, this is a real problem that one must not fall prey to. It has to be addressed. (P7)

Recommended strategies to control for systematic bias often focused on proper and independent external validation, i.e. the testing of a model in an independent sample. Yet, some experts were skeptical of current practices of external validation, namely if performed by the same experts who ran the original experiment.

> It really has to be a clean external validation. And I just have the feeling that often external validation studies [...] have not really been carried out independently. Most of the time, they may have been done in the same paper, or some predictive model has been developed, and part of the data has been omitted to test this predictive model. But the people who did the statistics of course already had this external data set when they developed the model, and that's why I ask myself whether they really only tested the model at the end or whether they didn't look a bit beforehand to see how it worked, and then maybe, if it didn't work, improved the model a bit more. And then it's not really an independent external validation. (P13)

As a result, studies reporting ML-based results may be biased and not tailored to broader clinical practice, but only to the specific contexts from which the training data were obtained. Drawing on the example of IBM Watson Oncology that was famously accused of suggesting erroneous cancer treatments (Ross, 2018), one participant highlighted that

such attention to context is crucial if a program is supposed to be incorporated into clinical routines.

> The task of the machine is to minimize its cost function. That's it. And the users have to understand that the machine does not have the context, or if we need it, if we want to use it clinically at some point, then we need machines that have been trained in the correct context or can switch between sub-models for different concepts for different contexts. And that is actually totally simple and all machine learners know that, but there is a relatively big temptation to say 'I now have a machine that can predict therapy response for schizophrenia, and that it might work quite differently in Spain, I'll ignore for now'. (P11)

In the view of several interviewees, this problem could be addressed through more extensive and international data sharing between different research groups. Yet again, interviewees reported that this demand seemed at odds with pressure to turn your research group's data into high-ranking publications first, before sharing them with anyone else, and that it also contradicted intuitions concerning privacy protections.

> I don't like my data to be shared with anybody if I don't want it to be, and definitely not (…) in a way that can come back to me. And you know wit ML you have a problem, because once you train data, naturally you probably can't go back and say: ok, this part is based on X's data. But at some point, if you pool the data together, it could come back to you. (P12)

In consequence, several experts were skeptical concerning current research outputs, because a small number of experts in the field that are competent to scrutinize results in peer review processes, and the complexity of the used models could render reported findings questionable in terms of generalizability.

> It's not as rosy as things seem. And I think that will change as the field matures, but at the moment - because there are more parameters, because its more complex, because people don't understand it, it opens the door to a lot of ambiguity in a lot of things. And it won't be solved by putting code online or something because (…) the problem is happening earlier on in the pipeline. It's that classical thing of running a few thousand models and then, when you are reporting: two. (…) The same sort of thing is happening, and it is happening even with external validation. So - don't believe everything that people say. (P14)

The reasons for this may partially lie in the current hype around Artificial Intelligence that favours publications with a focus on machine learning techniques, as one interviewee remarked:

> And it always sounds so great, doesn't it? You just throw around terms like gradient boosting machine and support vector machine, and people are then somehow totally impressed, but that's a bit of a danger. (…) It's easy to publish a paper when you've used such a method because it's trendy and because it sounds so sophisticated and so modern, so whatever, and everyone is trying to get a piece of the pie for themselves. But for me, to a large extent, I have the feeling that it's old wine in new bottles. (P13)

Being more optimistic about the promises of ML, one interviewee expressed frustration that at the moment, psychiatry is often left out of large ML initiatives, despite the high burden of disease and a potentially large benefit, both for the individual patient and for the healthcare system.

> Why does so little take place? (…) When I look at the large medical technology or data initiatives, (…) they all leave out psychiatry. And the reasons are always the same: it's too complicated, we have fuzzy diagnoses in psychiatry, imaging is difficult to handle anyway, and on the other hand, I would say that psychiatric diseases are actually the ones that cause the greatest financial and health economic and subjective burden. (…) In fact, one has to say that the added value, the gain in psychiatry would be particularly high. But obviously the least research in this direction is currently taking place there. I find that interesting when you think about: why not? Are our drugs too cheap, are the surgical techniques that depend on them too simple? I don't think it's just because of the academic complexity of the concept of psychiatric diagnosis, I think there are certainly other reasons as well. (P5)

8.3.2. The role of explainability in research and clinical application

Questions concerning explaining and understanding ML systems in research and clinic appeared to be a topic of particular relevance throughout the interviews. Some participants were very vocal in their support for explainability and considered it crucial to keep medical practice compatible with current ethical standards of medical practice.

> If I have a black box prediction, the inside of which is unknown to me, then I can only accept that and have to trust that everything went well, regarding the intentions and the execution of the validation. If that happens, then we are moving into a whole new kind of medicine, which in my view is not compatible with the idea of the patient's right to self-determination. Within such a medicine, we become objects who can no longer understand where certain recommendations come from. And that is, from my point of view, completely contrary to the developments in medicine in the last decades and something that I personally do not strive for. (P4)

As minimal requirement for such scientific scrutiny and understanding, many mentioned transparent disclosure of both training data and of the used code.

> I am absolutely in favor of publishing data, and also of publishing the scripts used for analysis. Even if probably no one takes the trouble to exactly understand the script afterwards. (P13)

Some interview partners went further though, demanding a form of contestability:

> [The program] must allow itself to be questioned, it must be able to give answers, and it must be able to say what it cannot. (…) So, let's say metaphorically: it must be capable of dialog. For the doctor anyway, that's clear, but also for the patient. (P3)

At the same time, some interviewees hinted at the necessity of weighing accuracy and explainability against each other, and countered calls for explainability with recourse to utilitarian thought:

> I think we will come down to more like an accuracy trade-off. If something is 90% [accurate] and it is not interpretable, and then you get an interpretable model, and it's like 70%, then you have got to think about what to use. So I don't really have a big problem with it. (P14)

Positions that doubted the necessity of high degrees of explainability often drew comparisons between the lack of explainability of an ML system and current medical practice that also often involves incomplete knowledge on the side of practitioners and patients, for instance concerning clinical chemistry and pharmacy.

> Maybe it's not such a new thing at all compared to now. I'm pretty sure that clinical chemists understand clinical chemistry, but a lot of people in clinical

> practice don't understand it. They might understand the meaning, but not how the values come about (…). So maybe it is really not that different from what we already do in medicine. (P1)

> In the end, I would say it's like pharmacology. I mean, we've all learnt something about the way drugs work. I probably can't recite most of them to you now, but you have a rough idea of where the problems are and how it works and can therefore classify it well. But in the end, you rely on your experience, your clinical experience and see what helps the patient: If they come to me with symptom X, I prescribe drug Y, and then I have experience of how that works. (P7)

Yet, as argued by several participants, a crucial difference between these examples and ML, is that physicians have received training in these subjects, and thus have, in principle, at least a rough idea of potential pitfalls. Accordingly, many experts recommended to include education on the fundamentals of ML in medical curricula to better deal with the uncertainty associated with ML systems, as we highlight in the following section.

In this debate about explainable AI, several aspects came up that were specific to the context of psychiatry. Notable were repeated remarks that the mechanisms underlying current psychotropic drugs are also black boxes, and that we may impose double standards by demanding a higher degree of explainability from ML systems.

> I come from psychiatry. We have no idea how drugs work in psychiatry. So: why not? You know, they are both black boxes, we trust those. (P14)

This aspect seemed even more decisive in the views of many since, due to these existing therapeutic black boxes, there may be a particularly large benefit of using ML-based treatment recommendations when it comes to psychotropic drugs.

> If you consider how uncertain a method is compared to how much you can gain with it, then the possible gain in information in the area of therapy response for antidepressants is so great that even the marginal increase in prediction accuracy is already relevant, because antidepressants have to be taken for at least two to three weeks and many patients say after 10 days, well, it hasn't worked yet, I just have this dry mouth and beads of sweat on my forehead and have sexual side

> effects - should I really continue taking it? And the adherence falls in the critical phase where we are still waiting for the response, in this - this is currently a therapeutic black box! The patient has to wait 3-4 weeks to see if it has worked. In this phase, of course, an ML algorithm can help us a lot and say: yes, the patient should take the trouble and definitely take the medication for another week, and if you think about how many depressive patients there are, how many of them are treated (...), I would say that the additional expense (...) is justifiable given the probability of success and the expected benefit. (P5)

Finally, one interviewee applied the idea of a black box also to their own decision-making process, drawing on a metaphorical comparison between themselves and an artificial neural net:

> When I make a decision, I am a neural network too, and I may be able to explain to you 50% of my logical decisions, why I make a decision, but then a lot is also unconscious and I decide based on experience, even if it is not accessible to me or if I am not conscious of it myself. (P2)

Some also questioned the role of explainability as an ethical principle with view to its utility for end users.

> Explainability is a tool for machine learning developers to find out whether their model works or not. We should not give this to a user so that they have to find out whether some weights are as we imagine them to be. It's actually simply a measuring instrument for technically oriented machine learning developers to find out whether it works. (P14)

Instead, there was worry that recourse to explainability may at times serve as a smoke screen, to cover shortcomings in methodology:

> "What I mean is not this stupid short-circuited 'then we have to open the black box' talk that you hear again and again. That's a substitute for 'I don't have a proper solution, and it's too much effort on my part. Then I'll just map some weights out somewhere.' That's just gross nonsense. What I need to know as a user, or even as a patient, is how did they make this thing – probably - work well. And here the question is: what did I train it on, so what are the properties of the data, not of the algorithm or my weights or something. That's not relevant to it at all. The relevant point is: what does my training data look like? (...) And that's my problem - you use explainability as a fig leaf because you don't want to do the hard, difficult, expensive task of measuring proper populations and testing on those." (P11)

8.3.3. The relation of patients, physicians, and machine learning systems

As a third theme, the interviewed experts articulated ethical expectations concerning the relationship between patients, physicians and ML systems – i.e. problems that need to be addressed even if challenges concerning development and explainability were to be solved in the future.

As with any interpersonal relationship, communication was considered key for interactions between physicians and patients. In particular, there was tangible worry that in the absence of an established framework to communicate statistical findings appropriately, patients and physicians may find their perceived scope of possible actions narrowed by ML-based predictions.

> Generally, most patients but also many physicians run danger of interpreting predictors too little in terms of statistics, and therefore severely limit possibilities of how something can develop. And that would be a big problem. Because self-fulfilling prophecies are a big problem, they limit the scope of action, the possibilities of action enormously, both on the part of the physician and on the part of the patient. There is actually no real framework, no conceptual framework how this information can be used to generate more possibilities. (P6)

Similar concerns for self-determined actions also found their expression with explicit regard to patients' autonomy. The dreaded impact on the relation between algorithm, physician, and patient, as a mere shift in hierarchy, was succinctly expressed by one interviewee:

> It is crucial that the patient does not end up in a position of powerlessness as a result of any therapeutic intervention, be it conversation, medication or algorithm. This is a basic law in psychotherapy. Because if that happens, then the therapy has already failed. And I see the risk in these giant programs (…) that the power imbalance is no longer between psychiatrist and patient, but between algorithm and patient, and that is no better. So autonomy, the central word in psychiatry is autonomy, and that also applies in this context. (P3)

At the same time, the interviewed experts agreed that algorithms could play a useful role for clinical treatments, and some even argued that it may be ethically questionable to reserve specific tasks for humans even if an algorithm outperforms clinicians in this regard. All interviewees agreed that ML would play an assistive role, not replacing physicians, and that the last say should remain with physicians, also for legal reasons:

> Ultimately, the physician has to sign, and that will remain the case for a long time. It will not be the algorithm that prescribes the medication or admits the patient but the physician. (P8)

Such attribution of responsibility was taken to be particularly important in light of potentially erroneous ML-based decisions, whether resulting from a systematically biased model or an adversarial attack with purposively manipulated inputs for one particular patient. Concerning psychiatric diagnoses, such errors may for instance lead to harmful stigmatization that is not open to recourse:

> When I make an unfavorable diagnosis, there is of course always the problem in psychiatry that we give labels, that we stigmatize in some way. I think that is a general problem of psychiatry, perhaps less of AI, but (...) if we can then not even justify on what basis we have made a decision.... And we are not doing that at the moment either, that needs to be said quite clearly. But let's assume that you use (ML) for diagnostic purposes, and you can't even justify it in any way, then of course it could be stigmatizing. (P2)

As crucial necessity to address these problems interviewees unanimously suggested that more education on computer science needed to be integrated into medical curricula. While several interviewees acknowledged the problems of further burdening medical education, conveying some basic knowledge was considered crucial.

> Doctors ought to gain an understanding, and I believe that this would be possible without any problems, to address the mathematical dimensions. This could be integrated into medical training without any problems. Therefore, I assume that in 10 years we ought to have ensured that doctors are roughly informed about the dimensions and the significance of machine learning and its susceptibility to errors. (P5)

However, there were perceived limits of what to expect from additional training, as highlighted by the comparisons with training in clinical chemistry and pharmacology, that are merely meant to convey basic knowledge of the underlying techniques. A certain level of trust, supported by thorough regulatory oversight and certification, may therefore remain inevitable:

> I believe that also up to now, people have trusted certain methods and not understood them in detail. I think the basic approach is right, i.e. to say, ok, there is a certain committee or certain experts who look at everything in detail and understand it and then make a recommendation. And all the other "half-experts" or users, they trust in that. Basically, I think this is the right approach, or the only feasible approach, because it won't be possible, if you want to apply it, for every doctor to become a medical informatician. That's unrealistic. The alternative would be to say, no, it's too complex, we can't apply it. (P1)

This was also mirrored in comments that stressed the necessity for specialization, due to the rapidly evolving landscape of ML:

> We currently have some colleagues in medicine working on the applications of ML who have immersed themselves heroically and very far into the subject and the current medical debate, the research in this area, is carried out by colleagues who have a relatively good overview of the state of the art of both medicine and in this area [ML]. […] I believe that this will increasingly fade into the background because the development in machine learning is so rapid and outside of medicine that in a few years even doctors with an affinity for technology won't be able to follow the topic and just like now, when you use medical devices, i.e. products from companies, you will no longer be thinking about the functioning of the device or the algorithm, […] and doctors will rather remain experts on a higher level of abstraction. (P5)

In a nutshell, interviewees who brought up the topic of doctor-patient-relationship pleaded for a more conscious communication and a careful balancing of power, hierarchy, and responsibility, with no single side taking general precedent over the other, so that the room of possible action is increased by the introduction of clinical ML systems.

> If someone only has a hammer, then everything becomes a nail. And that must not happen with artificial intelligence. If I have a great computer, then this computer isn't everything, but there is still the patient who sits in front of me crying and says: everything is shit, I'm going to kill myself now. That must not be played off against each other. (P3)

8.4 Discussion

Our findings provide a first glimpse on the ethical reasoning of experts on ML in psychiatry in Germany and Switzerland, to the best of our knowledge. With view to the existing theoretical literature from ethics, they provide three crucial additions. First, they highlight that even within our small sample, both agreements and disagreements concerning fundamental ethical principles ran along the line of debates that enjoy prominence in the ethical literature. This demonstrates that current ethical debates are not merely placed in the infamous philosophical armchair, but mirror actual concerns of people in the field. Second, our findings lend support to critical voices that have denounced AI ethics for being too focused on principles and not being attentive enough to the conditions of AI production. Third, the often sceptical attitudes of our experts can be read as a warning to reflect critically on overly optimistic statements in the literature and provide an exhortation to spend more attention to methodological scrutiny. In the following, we discuss all three points with view to our interviews.

First, the attitudes of the interviewed experts mirrored current debates on the ethics of medical ML. This was present in both their agreements and disagreements. While the majority of interviewees was not aware of ethical guidelines such as the EU guidelines for trustworthy AI, many of the experts' attitudes reflect common principles of medical ethics and AI ethics, such as concerns about systematic biases, privacy violations or respect for autonomy. Concerns regarding respect for autonomy, algorithmic fairness,

and breaches of privacy are largely commensurate with conceptual research in this domain (Morley et al., 2020; Starke, De Clercq, et al., 2021) as are debates about the balancing of hierarchy between patients, physicians, and ML systems (Braun, Hummel, Beck, & Dabrock, 2020; Grote & Berens, 2022). They also fit the few empirical studies from the field which reported infringement of privacy, undue exploitation of patient data, and worries about autonomy as main ethical concerns Swiss psychology students had with the use of ML (Blease et al., 2021). Finding an appropriate balance between physicians, patients, and ML systems was widely seen by our participants as a way to foster the acceptance of specific ML systems at the bedside (Braun et al., 2020). Mirroring common tropes of the debate, our interviewees also called for considering ML systems as intelligent tools, not artificial colleagues (Dennett & Chalmers, 2019) and did not foresee a step towards a full automation in the near future (Topol, 2019), yet considered the use of ML as potentially valuable assistance. Similarly, we found shared concern with view to responsibility and legal liability, two dimensions that have long enjoyed great prominence in the field (Bublitz, Wolkenstein, Jox, & Friedrich, 2018; Matthias, 2004). However, with regard to trust, as an attitude that partially relinquishes the monitoring of algorithms (Ferrario & Loi, 2021; Ferrario, Loi, & Viganò, 2020, 2021), the interviewees represented a comprehensive spectrum of opinions. As in the ethical literature (DeCamp & Tilburt, 2019; Hatherley, 2020; Metzinger, 2019), some voices were entirely opposed to the notion of trust and considered it "completely contrary to the developments in medicine in the last decades" (P4, see above). Others strongly endorsed it as "the only feasible approach" (P1, see above), similar to proponents of trust in medical AI (Braun et al., 2021; Durán & Jongsma, 2021; Ferrario et al., 2021; Starke, van

den Brule, Elger, & Haselager, 2021). Our study therefore supports the relevance of current theoretical debates on trust, also from the view of experts working in the field.

Second, our findings call attention to ethical questions that seem to be underdeveloped in the ethical discourse so far. In particular, these relate to questions of explainability and of self-fulfilling promises. While much current ethical debate is concerned with explainability of ML models, treating it as a mediating principle enabling other ethical principles (Floridi et al., 2018; Turilli & Floridi, 2009), others have already noted that there is no uniform consensus among experts about the meaning of explainability (Adadi & Berrada, 2018; Arbelaez Ossa et al., 2022), and that expectations towards explainability vary across contexts (Mittelstadt, Russell, & Wachter, 2019). This is also confirmed by our study, as are concerns about balancing explainability with accuracy (London, 2019), about the need of contestability (Ploug & Holm, 2020) and about the importance of epistemological questions for an ethical use of ML systems (Grote & Berens, 2020).[39] Yet, there has not yet been sufficient debate whether the ethical focus of explainability could potentially yield ethically detrimental results. The concern reported here that explainability could be used by technical experts as an ethical fig leaf, covering methodological shortfalls by providing end-users with a false sense of understanding, has to our knowledge not yet been discussed elsewhere. Yet, it seems paramount to reflect in depth on this problem since both ethical literature and ethical guidelines, including the EU guidelines for trustworthy AI, stress the importance of explainability

[39] Many of the interviewees' responses seemed informed by the assumption of a trade-off between accuracy and explainability in ML models. This assumption, prevalent early in the current wave of explainable AI, is increasingly challenged and considered a fallacy (Rudin & Radin, 2019). Similarly, some form of contestability is increasingly implemented in ML by virtue of counterfactual reasoning (Verma, Dickerson, & Hines, 2020). These findings therefore further highlight the need of continued education on recent developments in the field that seem to move increasingly away from the "black boxes" which dominate the bioethical literature (Cearns, Hahn, & Baune, 2019).

or, more precisely, of a principle of explicability, linking intelligibility and accountability ("Ethics guidelines for trustworthy AI," 2019; Floridi et al., 2018; Herzog, 2021). Our finding is also in line with those of a very recent experimental study that has shown how certain forms of explainability can convey the illusion that an algorithm is attentive to context and ethical questions whereas in reality it is blind to ethical incidents (John-Mathews, 2022). Simulating a sexist decision of an AI that denies a loan to a woman based on her gender, the randomized study showed that 800 participants favoured models with low denunciatory power, i.e., they placed higher trust in "explainable" AI systems where unfair decisions were not perceived negatively (John-Mathews, 2022).

Given these findings, further conceptual and empirical research should therefore critically investigate if, instead of providing a mediating principle enabling ethical scrutiny (Floridi et al., 2018; Turilli & Floridi, 2009), explainability is indeed misused as "fig leaf" that brings about ethically undesirable results. While efforts based on explainable AI will remain crucial to developers and could potentially even contribute to better deal with the complexity of diagnosing and treating mental disorders (Roessner et al., 2021), it may prove necessary to challenge the widely held belief that explainability is key to the acceptance of AI (Chandler, Foltz, & Elvevåg, 2020). As Ferrario and Loi have recently highlighted, explainability does not necessarily foster acceptance and trust in medical AI, and can in fact only do so in a narrowly limited number of cases (Ferrario & Loi, 2021). In line with others, our finding also highlights themis need to refocus the view onto explainability and move towards more user-centred models of explainability that can provide meaningful understanding for physicians and patients (Arbelaez Ossa et al., 2022; Mittelstadt et al., 2019) and harness multiple levels of explanation(Vu et al., 2018).

Beyond issues with explainability, our findings also stress the concern that ML-based predictors could function as self-fulfilling prophecies, particularly in psychiatric contexts. From a sociological point of view, this could be interpreted as a classic instance of the influential Thomas theorem, postulating that situations which are defined as real, are real in their consequences (Thomas & Thomas, 1928: 572). Tellingly, William Thomas and Dorothy Swain Thomas developed this thought in the very context of psychiatry, where paranoid delusions may bring about very real consequences. Statistical outputs from ML models should similarly be treated cautiously, so that they do not bring about the very events they predict by limiting the scope of interventions that is perceived as possible by physicians and patients. Education about the principles of modern information-based diagnostic theories will be key to avoid such developments.

Third, our findings call for increased attention to methodological debates that also impact ethical considerations. Our interviewees pointed to the broader ramifications of how ML models are trained in academic research to highlight ethical shortfalls. Many reflected critically on the current climate of hype and the danger of a new AI winter, brought about by overly optimistic promises and a lack of methodological rigour (Floridi, 2020). Methodological concern was also tangible in calls for proper external validation to ensure the generalizability of ML systems across different demographics (Cearns et al., 2019), and with view to the increasing importance placed on the diversity of cohort and data in clinical research ("Striving for Diversity in Research Studies," 2021). Other much-discussed aspects of fairness, e.g. the problem of competing fairness standards (Barocas, Hardt, & Narayanan, 2017; Friedler, Scheidegger, & Venkatasubramanian, 2016), were not raised. These findings suggest that more empirical research is needed on how closely current studies of ML in psychiatry adhere to

established reporting guidelines such as SPIRIT or CONSORT (Liu, Rivera, Moher, Calvert, & Denniston, 2020; Rivera, Liu, Chan, Denniston, & Calvert, 2020). Debates on policy should also further address whether additional incentives are needed, as suggested by the experts, to foster the collection of representative and context-sensitive training data and to encourage multi-centered collaborations in the particular context of psychiatry.[40] Such policy debates should also address the issue of sharing not only data but also the models itself, for which clear theoretical foundations need to be established.

There are several limitations to our study. As with any qualitative research, our findings are not generalizable and only reflect the attitudes and opinions within a limited sample of experts in Germany and Switzerland. Due to our highly targeted sampling, our participants were not representative of society, as highlighted for instance by the small number of female participants, reflecting the underrepresentation of women in the field. In addition, our interviews do not reflect the views and attitudes of potentially larger groups of stakeholders that will be affected by the introduction of ML into psychiatry, first and foremost the affected patients. While ethical research interviewing experts on psychiatric ML seemed most promising at the moment, given the nascent stage of the clinical ML employed in psychiatry, more empirically informed research will be crucial, accompanying the implementation of psychiatric ML(Jongsma & Bredenoord, 2020). Furthermore, the direct involvement of the interviewer in the research field may have shaped his interaction with participants, while in turn social desirability, e.g., being critical of ML when talking to a colleague from ethics, may have shaped answers to our

[40] It should be noted that there has been much progress in the development of context-sensitive ML recently (Elayan, Aloqaily, & Guizani, 2021; Nascimento, Alencar, Lucena, & Cowan, 2018). We are indebted to an anonymous reviewer for pointing this out.

open questions. However, since the aim of our qualitative study was exploratory rather than striving for a representative depiction, we do believe that these limitations do not draw away from the novelty of our insights.

8.5 Conclusion

Our study adds to the emerging corpus of empirical literature on the ethics of using ML in psychiatric settings. It highlights the need for further ethical reflection concerning the ramifications of developing and using ML models for mental health to avoid that predictions become self-fulfilling prophecies, and to ascertain that promises of explainability do not serve as ethical fig leaf. We have pointed out that the conditions of academic research in the field may require further incentives for rigorous methodology, that current attempts of explainability should be questioned concerning their utility for end-users, and that a careful balance needs to be found to safeguard important features of doctor-patient relationships once a ML model gets involved. Early involvement of ethical considerations in the development pipeline (McLennan et al., 2020) seem therefore as crucial as stratified basic education on computer science both of physicians and the public, in line with the detailed recommendations of others (Gauld, Micoulaud-Franchi, & Dumas, 2021). This may in turn also facilitate to not overstate the promises of ML and safeguard the importance of the interpersonal interactions fundamental to medical practice.

8.6 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access, 6*, 52138-52160.

Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J., Shaw, D., & Elger, B. S. (2022). Re-focusing explainability in medicine. *Digital health, (in print)*.

Barocas, S., Hardt, M., & Narayanan, A. (2017). *Fairness in machine learning*. Retrieved from http://www.fairmlbook.org

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., & DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *Journal of medical Internet research, 21*(3), e12802. doi:10.2196/12802

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. doi: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. doi:10.1177/2055207620968355

Braun, M., Bleher, H., & Hummel, P. (2021). A Leap of Faith: Is There a Formula for "Trustworthy" AI? *Hastings Center Report, 51*(3), 17-22. doi:10.1002/hast.1207

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics, 47*(e3). doi:10.1136/medethics-2019-105860

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101. doi:10.1191/1478088706qp0630a

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

Bröhl, C., Nelles, J., Brandl, C., Mertens, A., & Nitsch, V. (2019). Human–robot collaboration acceptance model: development and comparison for Germany, Japan, China and the USA. *International Journal of Social Robotics, 11*(5), 709-726.

Bublitz, C., Wolkenstein, A., Jox, R. J., & Friedrich, O. (2018). Legal liabilities of BCI-users: Responsibility gaps at the intersection of mind and machine? *International Journal of Law and Psychiatry*. doi:10.1016/j.ijlp.2018.10.002

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction, 3*(CSCW), 1-24.

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry, 9*(1), 1-12.

Chandler, C., Foltz, P. W., & Elvevåg, B. (2020). Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophrenia Bulletin, 46*(1), 11-14.

Char, D. S., Abramoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *American Journal of Bioethics, 20*(11), 7-17. doi:10.1080/15265161.2020.1819469

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. doi:10.1002/wps.20882

Chivilgina, O., Elger, B. S., & Jotterand, F. (2021). Digital Technologies for Schizophrenia Management: A Descriptive Review. *Science and Engineering Ethics, 27*(2), 1-22.

Chivilgina, O., Wangmo, T., Elger, B. S., Heinrich, T., & Jotterand, F. (2020). mHealth for schizophrenia spectrum disorders management: A systematic review. *International Journal of Social Psychiatry, 66*(7), 642-665.

Conti, D., Cattani, A., Di Nuovo, S., & Di Nuovo, A. (2015). A cross-cultural study of acceptance and use of robotics by future psychology practitioners. *24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 555-560. doi:10.1109/ROMAN.2015.7333601

Crawford, K. (2021). *Atlas of AI : power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

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

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *Lancet Digital Health, 1*(8), E390. doi:10.1016/S2589-7500(19)30197-9

Dennett, D., & Chalmers, D. (2019). Is Superintelligence Impossible? On Possible Minds: Philosophy and AI. *Edge*. Retrieved from https://www.edge.org/conversation/david_chalmers-daniel_c_dennett-is-superintelligence-impossible

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. doi:10.1080/13645579.2020.1766777

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329-335. doi:10.1136/medethics-2020-106820

Elayan, H., Aloqaily, M., & Guizani, M. (2021). Digital twin for intelligent context-aware iot healthcare systems. *IEEE Internet of Things Journal, 8*(23), 16749-16757.

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, (2019). Retrieved from https://data.europa.eu/doi/10.2759/177365

Ferrario, A., & Loi, M. (2021). The Meaning of "Explainability Fosters Trust in AI". *Available at SSRN*. doi:10.2139/ssrn.3916396

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philosophy & Technology, 33*(3), 523-539.

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: it is possible to trust medical AI. *Journal of Medical Ethics, 47*(6), 437-438.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*(2), 185-193.

Floridi, L. (2020). AI and its new winter: from myths to realities. *Philosophy & Technology, 33*(1), 1-3.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*.

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. doi:10.1017/S0033291720003906

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

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205-211. doi:10.1136/medethics-2019-105586

Grote, T., & Berens, P. (2022). How competitors become collaborators—Bridging the gap (s) between machine learning algorithms and clinicians. *Bioethics, 36*(2), 134-142. doi:10.1111/bioe.12957

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods, 18*(1), 59-82.

Hagendorff, T. (2021). Blind spots in AI ethics. *AI and Ethics*, 1-17. doi:10.1007/s43681-021-00122-8

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478-481. doi:10.1136/medethics-2019-105935

Herzog, C. (2021). On the risk of confusing interpretability with explicability. *AI and Ethics*, 1-7. doi:10.1007/s43681-021-00121-9

Jacobson, N. C., Bentley, K. H., Walton, A., Wang, S. B., Fortgang, R. G., Millner, A. J., Coombs III, G., Rodman, A. M., & Coppersmith, D. D. (2020). Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bulletin of the World Health Organization, 98*(4), 270-276. doi:10.2471/BLT.19.237107

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399.

John-Mathews, J.-M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change, 174*, 121209.

Jongsma, K. R., & Bredenoord, A. L. (2020). Ethics parallel research: an approach for (early) ethical guidance of biomedical innovation. *BMC Medical Ethics, 21*(1), 1-9.

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine, 26*, 1364–1374. doi:10.1038/s41591-020-1034-x

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report, 49*(1), 15-21. doi:10.1002/hast.973

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. doi: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. doi:10.2196/mental.9423

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175-183.

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., Haddadin, S., & Buyx, A. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence, 2*(9), 488-490.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *FAT* '19: Proceedings of the conference on fairness, accountability, and transparency*, 279-288. doi:10.1145/3287560.3287574

Morgenstern, J. D., Rosella, L. C., Daley, M. J., Goel, V., Schünemann, H. J., & Piggott, T. (2021). "AI's gonna have an impact on everything in society, so it has to have an impact on public health": a fundamental qualitative descriptive study of the implications of artificial intelligence for public health. *BMC Public Health, 21*(1), 1-14.

Morley, J., Machado, C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The Debate on the Ethics of AI in Health Care: a Reconstruction and Critical Review. *SSRN*. doi:10.2139/ssrn.3486518

Nascimento, N., Alencar, P., Lucena, C., & Cowan, D. (2018). *A context-aware machine learning-based approach.* Paper presented at the Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering.

Nichol, A. A., Bendavid, E., Mutenherwa, F., Patel, C., & Cho, M. K. (2021). Diverse experts' perspectives on ethical issues of using machine learning to predict HIV/AIDS risk in sub-Saharan Africa: a modified Delphi study. *BMJ open, 11*(7), e052287. doi:10.1136/bmjopen-2021-052287.

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A Roadmap for the Development of Applied Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 386-392. doi:10.1016/j.bpsc.2016.05.001

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine, 107*, 101901.

Pumplun, L., Fecho, M., Wahl, N., Peters, F., & Buxmann, P. (2021). Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study. *Journal of medical Internet research, 23*(10), e29301.

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine, 26*, 1351–1363. doi:10.1038/s41591-020-1037-7

Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., & Beste, C. (2021). Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research. In (Vol. 30, pp. 1143-1146): Springer.

Ross, C., Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News.* Retrieved from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition.

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. doi:10.1093/schbul/sbaa120

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

Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2021). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics*. doi:10.1111/bioe.12891

Striving for Diversity in Research Studies. (2021). *New England Journal of Medicine, 385*(15), 1429-1430. doi:10.1056/NEJMe2114651

Thomas, W., & Thomas, D. (1928). *The child in America*. New York: Knopf.

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). *What clinicians want: contextualizing explainable machine learning for clinical end use.* Paper presented at the Machine learning for healthcare conference.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44-56. doi:10.1038/s41591-018-0300-7

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105-112. doi:10.1007/s10676-009-9187-9

Van den Berg, B. (2012). *Differences between Germans and Dutch people in perception of social robots and the tasks robots perform.* Paper presented at the 16th Twente Student Conference on IT.

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Vu, M.-A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V. S., & Widge, A. S. (2018). A shared vision for machine learning in neuroscience. *Journal of Neuroscience, 38*(7), 1601-1607.

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. doi:10.1016/j.pnpbp.2018.09.014

Wangmo, T., Hauri, S., Gennet, E., Anane-Sarpong, E., Provoost, V., & Elger, B. S. (2018). An update on the "empirical turn" in bioethics: analysis of empirical research in nine bioethics journals. *BMC Medical Ethics, 19*(1), 1-9.

Chapter 9: Machine Learning and its Impact on Psychiatric Nosology: Findings from a Qualitative Study Among German and Swiss Experts

Machine Learning and its Impact on Psychiatric Nosology: Findings from a Qualitative Study Among German and Swiss Experts

Georg Starke[1, 2], Bernice Simone Elger[1,3], Eva De Clercq[1]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland, [2] College of Humanities, École Polytechnique Fédérale de Lausanne, Switzerland, [3] University Center of Legal Medicine, University of Geneva, Switzerland.

Acknowledgements: 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.

The following chapter is the accepted manuscript, forthcoming in *Philosophy and the Mind Sciences*. Please cite the published version: Starke, G., Elger, B., & De Clercq, E. (2022). Machine learning and its impact on psychiatric nosology. Findings from a qualitative study among German and Swiss experts.

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.

9.1 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, Koppe, & Meyer-Lindenberg, 2019; Jacobson & Bhattacharya, 2022; Quaak, van de Mortel, Thomas, & van Wingen, 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, Marcus, & Barry, 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, Diefenbacher, Courtet, & Genieys, 2020; Eitel, Schulz, Seiler, Walter, & Ritter, 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, Chapman-Rounds, Verma, Bzdok, & Georgatzis, 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, Maia, & Frank, 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. 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, Cramer, & Kalis, 2018) and is, as highlighted by a recent review, also supported by a large corpus of empirical results (Robinaugh, Hoekstra, Toner, & Borsboom, 2020). Similarly, Peter Zachar's description of psychiatric disorders 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 best, for 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, Correia, Esparza, Saxena, & Maj, 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, Kuhl, & Kupfer, 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, Kharko, Annoni, Gaab, & Locher, 2021; Blease, Locher, Leon-Carlyle, & Doraiswamy, 2020), 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.

9.2 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, 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, Schmidt, De Clercq, & Elger, 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 9.1.

- 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 9.1: Relevant questions from the interview guide

9.3 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.

9.3.1 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)

9.3.2 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 a 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 a 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 him 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 a 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)

9.4 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, Stephan, Montague, & Dolan, 2014; Montague, Dolan, Friston, & Dayan, 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, Bracken, David, & Sartorius, 2013) and with positions that favour more pragmatic definitions of mental disorders (Kendler, Zachar, & Craver, 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, Ryan, Robinaugh, Waldorp, & Borsboom, 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 my 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, Kilbertus, Rojas-Carulla, & Schölkopf, 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, 2008; 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, 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.

9.5 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.

9.6 References

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

Bird, A., & Tobin, E. (2008). Natural kinds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.).

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. doi: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. doi:10.1177/2055207620968355

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5-13. doi: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. doi:10.1017/S0140525X17002266

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

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

Brinkmann, S. (2017). Perspectives on diagnosed suffering. *Nordic Psychology, 69*(1), 1-4. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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). Cham: Springer.

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. doi: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. doi:10.1186/1741-7015-11-126

Dattaro, L. (2021). Green light for diagnostic autism app raises questions, concerns. *Spectrum*. Retrieved from 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. doi: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. doi:10.1080/13645579.2020.1766777

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry, 24*(11), 1583-1598. doi: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. doi:10.1016/j.expneurol.2021.113608

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

Fleck, L. (1986). Some Specific Features of the Medical Way of Thinking [1927]. In *Cognition and Fact: Materials on Ludwik Fleck* (pp. 39-46). Dordrecht: Springer Netherlands.

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. doi:10.1016/S2215-0366(14)70275-5

Given, L. M. (2015). *100 questions (and answers) about qualitative research*. London: 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. doi:10.1007/BF00385836

Hacking, I. (1996). The looping effects of human kinds. In D. Sperber, Premack, D., James Premack, A. (Ed.), *Causal cognition: A multi-disciplinary debate* (pp. 351-383). Oxford: Oxford Academic.

Hacking, I. (1999). *The Social Construction of What?* Cambridge, MA: 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*. doi:10.1037/met0000303

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

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. doi:10.1038/nn.4238

Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. *Science, 348*(6234), 499-500. doi: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. doi: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. doi: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. doi: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. doi:10.1016/j.shpsa.2018.01.011

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry, 15*(1), 5-12. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1016/j.tics.2011.11.018

Nagel, T. (1986). *The View from Nowhere*. Oxford: 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. doi:10.1177/0959354320921464

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., & Schölkopf, B. (2018). *Learning Independent Causal Mechanisms*. Paper presented at the Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v80/parascandolo18a.html

Quaak, M., van de Mortel, L., Thomas, R. M., & van Wingen, G. (2021). Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *NeuroImage: Clinical, 30*, 102584. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi: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. doi:10.1038/s41598-020-68858-7

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

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology, 25*, 85-92. doi: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. doi: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. doi:10.1080/02698590701589601

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

Walter, H. (2013). The third wave of biological psychiatry. *Frontiers in psychology, 4*, 582. doi: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. doi: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. doi: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. doi:10.1038/s41380-021-01116-y

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

Chapter 10: Why Educating for Clinical Machine Learning Still Requires Attention to History

Why Educating for Clinical Machine Learning Still Requires Attention to History: a Rejoinder to Gauld et al.

Georg Starke[1], Eva De Clercq[1], Stefan Borgwardt[2,3], Bernice Simone Elger[1,4]

[1] Institute for Biomedical Ethics, University of Basel, Switzerland, [2] Department of Psychiatry, University of Basel, Switzerland, [3] Department of Psychiatry and Psychotherapy, University of Lübeck, Germany, [4] University Center of Legal Medicine, University of Geneva, Switzerland.

The following chapter is the accepted manuscript. Please cite the published version: Starke, G., De Clercq, E., Borgwardt, S., & Elger, B. S. (2021). Why educating for clinical machine learning still requires attention to history: a rejoinder to Gauld et al. *Psychological Medicine*, 51 (14), 2512 – 25131-2. https://doi.org/10.1017/S0033291720004766. Reproduced with permission.

10.1 Introduction

We are very grateful that Christophe Gauld, Jean-Arthur Micoulaud-Franchi and Guillaume Dumas have added their valuable comment to our article (Starke, De Clercq, Borgwardt, & Elger, 2021). We fully agree with their response, highlighting the importance of an appropriate framework for educating young psychiatrists (Gauld, Micoulaud-Franchi, & Dumas, 2021). Indeed, basic knowledge about the fundamentals of computer science, cognitive neuroscience, computational psychiatry, clinical practice as well as ethics seems crucial for a successful and responsible implementation of machine learning (ML) in psychiatry. Similarly, we fully concur with them and others (Grote & Berens, 2020) that developing an appropriate epistemological framework will be crucial to advance the ethical debates surrounding AI in healthcare.

Still, expanding on the useful practical guide Dr Gauld and his colleagues have provided to develop a curriculum fit for educational purposes, we would like to draw further attention to the persistent importance of teaching history of psychiatry. While this is no new demand (Shorter, 2008), it may not have received enough attention in the context of psychiatric ML yet. Of course, we are aware that curricula run danger of being overburdened in the context of ML, and agree with Gauld et al. (2021) and McCoy et al. (2020) that training should focus on fundamental concepts. However, education about the historical development and employment of psychiatric classifications should be considered part of these fundamental issues and will remain crucial to counter potential ethical, clinical and conceptual pitfalls of ML in psychiatry. Once more, the example of schizophrenia seems particularly well suited to highlight these challenges.

10.2 From the history of schizophrenia to machine learning

With view to ethical questions, education about the historical ramifications surrounding the development of particular classificatory concepts helps to elucidate the fact that they are human made. When developing and using diagnostic ML tools in psychiatry, this may help to stress their historical contingence as heuristic concepts, countering tendencies to reify the categories which a particular system has been trained to classify (Hyman, 2010). Furthermore, attention to historical atrocities and gross abuse of power in psychiatry, e.g. during the Nazi era, can serve as a cautionary tale in educative settings, raising awareness for ethical pitfalls today (Strous, 2007). In fact, some old ethical problems of psychiatry may return under new guise with ML-based systems. For example, it has been argued that in the US during the 1960s and 1970s, the diagnosis of schizophrenia was disproportionately applied to African-Americans connected to the civil rights movement, on account of their alleged aggressive behaviour (Metzl, 2009). Given that even today there remain significant disparities between ethnic groups with regard to the diagnosis of schizophrenia (Gara, Minsky, Silverstein, Miskimen, & Strakowski, 2019), educative curricula should draw attention to such historical injustices, fostering particular attention to discrimination and biases potentially ingrained in ML-based systems.

For the current clinical practice of psychiatry, obtaining an historically informed view seems highly beneficial as well. In particular, historical education may promote clinical qualities that critics fear could fade into the background with the introduction of ML systems. For example, looking closely at the original conditions under which a specific concept was introduced may inspire close attention to clinical context. Again, the case of schizophrenia can serve to illustrate this. The term “schizophrenia” was famously

coined by the Swiss psychiatrist Eugen Bleuler in 1908, arguably in rejection of a Kraepelinian nosology based on prognosis (Maatz & Hoff, 2014). In turn, Bleuler has been read as an early proponent of a bio-psycho-social model of disease, aiming for an understanding of the disorder that integrates the underlying neurobiology with individual psychological and social aspects (Maatz, Hoff, & Angst, 2015). In a similar vein, recent research has highlighted the irreducible and subjective psychological nature of Bleuler's so-called first-rank symptoms, stressing the importance of the individual, lived experiences of patients for his psychopathology (Moscarelli, 2020). With regard to ML systems, teaching about the historical origins of the concept of schizophrenia may thus serve to avoid an overly simplified view of the disorder and stress their cumulative nature. In other words, recent advances in ML notwithstanding, psychiatry will need to keep paying close attention to the social conditions of disorders as well as the individual phenomenological perspectives of patients.

Finally, with view to conceptual questions, attention to the history of psychiatric theory will also remain fundamental to the development and improvement of diagnostic categories. We fully agree with Dr Gauld and his colleagues that an appropriate framework of medical epistemology requires a "to-ing and fro-ing" between philosophy and science. However, in line with contemporary philosophy of science, we also hold that this process needs to retain attention to historical detail, in the sense of an integrated history and philosophy of science (Chang 2008). Kenneth Kendler has sketched the consequences of such an historical approach with regard to the classification of schizophrenia, driven by a process of "epistemic iterations" (Kendler, 2009). Attempts to redefine psychiatric classification based on ML may thus need to reflect upon their own historically contingent role in this evolutive process, so that

psychiatric nosology may mature "historically from top-down essentialist views of our categories to bottom-up empirically defined entities that reflect with increasingly accuracy the world as we can best understand it." (Kendler, 2009)

10.3 References

Gara, M. A., Minsky, S., Silverstein, S. M., Miskimen, T., & Strakowski, S. M. (2019). A Naturalistic Study of Racial Disparities in Diagnoses at an Outpatient Behavioral Health Clinic. *Psychiatric Services, 70*(2), 130-134. doi:10.1176/appi.ps.201800223

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. doi:10.1017/S0033291720003906

Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205-211. doi:10.1136/medethics-2019-105586

Hyman, S. E. (2010). The diagnosis of mental disorders: the problem of reification. *Annual Review of Clinical Psychology, 6*, 155-179. doi:10.1146/annurev.clinpsy.3.022806.091532

Kendler, K. S. (2009). An historical framework for psychiatric nosology. *Psychological Medicine, 39*(12), 1935-1941. doi:10.1017/S0033291709005753

Maatz, A., & Hoff, P. (2014). The birth of schizophrenia or a very modern Bleuler: a close reading of Eugen Bleuler's 'Die Prognose der Dementia praecox' and a re-consideration of his contribution to psychiatry. *History of Psychiatry, 25*(4), 431-440. doi:10.1177/0957154X14546606

Maatz, A., Hoff, P., & Angst, J. (2015). Eugen Bleuler's schizophrenia--a modern perspective. *Dialogues in clinical neuroscience, 17*(1), 43-49. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25987862

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), 86. doi:10.1038/s41746-020-0294-7

Metzl, J. (2009). *The protest psychosis : how schizophrenia became a Black disease*. Boston: Beacon Press.

Moscarelli, M. (2020). A major flaw in the diagnosis of schizophrenia: what happened to the Schneider's first rank symptoms. *Psychological Medicine, 50*(9), 1409-1417. doi:10.1017/S0033291720001816

Shorter, E. (2008). History of psychiatry. *Current Opinion in Psychiatry, 21*(6), 593-597. doi:10.1097/YCO.0b013e32830aba12

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

Strous, R. D. (2007). Psychiatry during the Nazi era: ethical lessons for the modern professional. *Annals of General Psychiatry, 6*, 8. doi:10.1186/1744-859X-6-8

Chapter 11: Discussion

11.1 Filling the gaps: what this thesis adds to current debates

This thesis set out to investigate whether we can trust medical ML, and if so, under which conditions. While its cumulative structure implies that the individual chapters necessarily form independent scholarly contributions, taken together, their main claims and findings form a mosaic answer to the research question.

This thesis argues why trust can serve as a meaningful concept in bioethical debates about medical ML and introduces a multidimensional model of trust (**chapter 3**). Highlighting the frequent problem of unknown causality and missing gold standards when modelling medical phenomena, the thesis suggests approaches to algorithmic fairness and understanding medical ML that aim to circumvent these underlying challenges, supporting two key conditions of trustworthiness (**chapters 4 & 5)**. It also embraces a concept of intelligent openness, taking communicative conditions into account to foster trust in medical ML (**chapter 6**). Turning to ML applications in psychiatry, the findings of our interview study point to the necessity of reflecting critically on explainability as a means to achieve trustworthy ML and call for closer attention to ML methodology (**chapter 8)**. In addition, analysis of our qualitative findings with regard to the definition of mental disorders stresses why philosophical reflection is crucial to harness the potential of ML in psychiatry (**chapter 9**). Looking at the specific example of schizophrenia, the thesis finally spells out key ethical challenges which trustworthy ML needs to address in the context of clinical psychiatry and advocates for the continued importance of historical education (**chapters 7 & 10**).

This final chapter of the thesis brings these different strains of thought together and shows what they add to current bioethical debates about medical AI. Following the research question, I first examine the implication of our findings for the debate about

trust in medical ML (10.2), before turning to questions of *trustworthiness* (10.3). I complement these arguments with a methodological plea for a greater integration of history into bioethical inquiry (10.4), and sketch limitations of my work as well as implications for future research (10.5). The chapter concludes with a call for improved training of all neural nets involved, both artificial and human (10.6).

11.2 Towards a new model of trust in medical ML

I started my investigation with the question whether trust constitutes a meaningful concept to address ethical challenges posed by black-box medical ML. In my answer, I diverged from more widely held positions that argue against trust, on conceptual and normative grounds (Bryson, 2018; DeCamp & Tilburt, 2019; Hatherley, 2020; Metzinger, 2019; Ryan, 2020). Instead, I have shown in the more conceptual papers how trust does indeed seem a defensible stance if (a) one respects the term's factual use in ordinary language and (b) does not construe a notion of trust that presupposes an insurmountable difference between non-human and human actors (Latour, 1994, p. 46).

Since our paper on the topic was submitted and published (Starke, van den Brule, Elger, & Haselager, 2021), various voices in bioethics have defended trust in medical ML, from slightly different angles (Braun, Bleher, & Hummel, 2021; Durán & Jongsma, 2021; Ferrario, Loi, & Viganò, 2021). It therefore seems imperative to distinguish our own approach from these three groups of authors. First, Braun et al. (2021) have advocated an understanding of trust as a "leap of faith" and criticised "formulaic approaches" based on their position "that a rigid set of principles and regulation will suffice to govern AI threatens to be an oversimplification" (ibid.). Such criticism is certainly warranted, especially if ethical regulatory frameworks only appeal to abstract general ethical demands such as fairness or transparency, that need to be spelled out in detail for any

particular application. Yet, while the paper also provides many other important suggestions, e.g., the systematic involvement of relevant stakeholders, it largely skips over the technical details of opaque ML, robustness, or accuracy. In addition, while also our own model requires agents to take a decision after evaluating its different dimensions, rephrasing trust as a binary leap of faith may inhibit a finer grained ethical analysis of our interaction with opaque ML systems.

In comparison, a detailed ethical analysis is exactly one of the greatest strengths the model of trust by Andrea Ferrario and colleagues has to offer (Ferrario, Loi, & Viganò, 2020; Ferrario et al., 2021). In their understanding, trust comes in three incremental layers, namely *simple*, *reflective*, and *paradigmatic* trust (ibid.). They delineate that simple trust describes the non-cognitive attitude of a trustor to rely on a trustee to perform a specific action, "without intentionally generating and/or processing further information about Y's capabilities to achieve G" (Ferrario et al., 2020, p. 530). In consequence, any properties of the trustee that justify trust, i.e., properties of trustworthiness, only play a role at the two higher levels of trust. Both forms, reflective and paradigmatic trust, require that the trustor holds beliefs about the trustee that vindicate a trusting relationship. The difference between the two is merely the degree of certainty with which such beliefs are held, and that the trustor is willing to forego control of the trustee in the case of paradigmatic trust (Ferrario et al., 2020, p. 532). The account of Ferrario and colleagues seems a compelling description to describe human-AI-interactions. It also shares with our model a notion of degrees, mirroring the complexity and variability of actual medical ML applications. However, it may provide only limited guidance for ethical analysis since its inclusion of a trust form that is defined

independently of trustworthiness may move the latter to the background.[41] In a response to Annette Baier, Onora O'Neill has stressed the ethical danger of such approaches:

> Attitudes of trust can indeed diverge from, or disregard questions about trustworthiness, often at great cost to those who place their trust poorly. I am concerned with the practical demands of trust, so think that it matters that it be placed in the trustworthy and denied to the untrustworthy, and that we need therefore to grasp the importance of placing trust in the trustworthy. (O'Neill, 2013, p. 238)

The stance by Juan Durán and Karin Jongsma avoids this problem by grounding trustworthiness in a notion of computational reliabilism (Durán & Jongsma, 2021). Drawing on the prior work of one of the authors (Durán & Formanek, 2018), they provide a list of reliability indicators that, in their opinion, render black-box algorithms trustworthy, namely "verification and validation methods, robustness analysis, a history of (un)successful implementations, and expert knowledge" (Durán & Jongsma, 2021, p. 4). In many ways, the approach of Durán and Jongsma is close to the position defended in this thesis. Similar to the arguments advanced in the sixth chapter of this thesis, the authors hold that "transparency is a methodology that does not offer sufficient reasons to believe that we can reliably trust black box algorithms. At best, transparency *contributes* to building trust in the algorithms and their outcomes, but it would be a mistake to consider it as a solution to overcome opacity altogether" (Durán & Jongsma, 2021, p. 2; emphasis in original). At the same time, computational reliabilism runs into the very problems criticised by Braun et al. (2021), insofar as it also only provides a rather formulaic list of items that can hardly ever do full justice to the complexity of assessing

[41] In a different paper, the authors have recently embraced a warning by Jacovi, Marasović, Miller, and Goldberg (2021) though, stressing that unwarranted trust is ethically unacceptable and should be avoided (Ferrario & Loi, 2021).

a system's trustworthiness. In consequence, Thomas Grote has already pointed out that the model lacks information that is crucial for individual decision making such as uncertainty estimates (Grote, 2021).

In comparison, our model accommodates the benefits of all three rivalling attempts. It offers a multi-dimensional space of analysis that allows for a fine-grained ethical investigation of trustworthiness along its three axes.[42] At the same time, these dimensions are sufficiently accommodating to not provide a mere formulaic approach that lends itself to simplification. It further includes technical measures in its dimensions of reliability and competence that can go beyond the list of computational reliabilism and accommodate, for instance, probability distributions of a model's uncertainty. Finally, similarly to the model by Ferrario and colleagues, it can also help to conceptualise different degrees of trust, depending on the different aspects of trustworthiness. This seems all the more important in light of O'Neill's admonishment that trustworthiness, not trust should be at the centre of bioethical inquiry.

11.3 Fostering the trustworthiness of medical ML[43]

While I have refuted conceptual arguments raised by the sceptics of trust based on a dichotomy of human and non-human agents, there is much merit in their normative warnings: employing the complex concept of trust in the context of ML should not be misused to encourage users into accepting ML-based appliances that are not trustworthy (DeCamp & Tilburt, 2019; Hatherley, 2020). Given that trustworthiness

[42] Similarly to the argument advanced in chapter 5, there is a remarkable parallel here between ML and psychopathology, where the last decade has also seen a gradual move to dimensional instead of binary systems for classifying psychiatric disorders (Appelbaum, 2017).

[43] For a largely expanded discussion of these thoughts please refer to: Starke, G., & Ienca, M. (2022): Misplaced trust and ill-placed distrust: How not to engage with medical AI. *Cambridge Quarterly of Healthcare Ethics* (in print).

seems key to bioethical deliberations, it comes as little surprise that the larger part of this dissertation addressed the second part of its research question, namely, under which conditions we can trust opaque medical ML. So how do the different suggestions of this thesis contribute towards a bigger picture of trustworthy medical ML?

The EU guidelines for trustworthy AI provide an excellent starting point to this question. As mentioned in the introduction, similarly to Beauchamp and Childress (2019), the EU Commission's High-Level Expert Group suggested four ethical principles, namely the established principles of respect for human autonomy, prevention of harm, and fairness, complemented by a principle of explicability ("Ethics guidelines for trustworthy AI," 2019). Since this thesis investigates opaque ML models, it is little wonder that of these four, explicability – a principle combining the epistemic question of intelligibility with the normative question of accountability – takes centre stage in several chapters. Like other defenders of trust (Braun et al., 2021; Durán & Jongsma, 2021), I take it for given that ML opacity can, at least for now, not simply be circumvented by technical methods from XAI but that some form of understanding remains an important desideratum of trustworthy ML models.[44] To gain a systematic view on trustworthy AI that goes beyond ethics checklists, it therefore seems crucial to relate other ethical desiderata such as fairness to the question of explainability in a systematic way.[45]

[44] For discussions whether there is also a legal right to explainability under the EU General Data Protection Regulation (GDPR), see Crabtree, Urquhart, and Chen (2019); Edwards and Veale (2017); Wachter, Mittelstadt, and Floridi (2017).

[45] On the disambiguities of the term "explainability" in the context of ML see for instance Adadi and Berrada (2018) and Vilone and Longo (2020, 2021) as well as our own work on the question (Arbelaez Ossa et al., 2022).

A recent model of trust in AI suggested by Alon Jacovi and colleagues (2021) seems instructive for this task, and complements our own dimensional model, distinguishing between reliability, competence and intentions, with a further distinction. In contrast to the three models discussed above, the authors do not focus on the context of medicine but make a helpful distinction between two different ways of bringing about *trust* in AI, namely an intrinsic and extrinsic way (ibid).[46] Reformulating the considerations of Jacovi et al. with view to *trustworthiness* allows to sort the topics discussed in this thesis more systematically (see fig. 11.1).

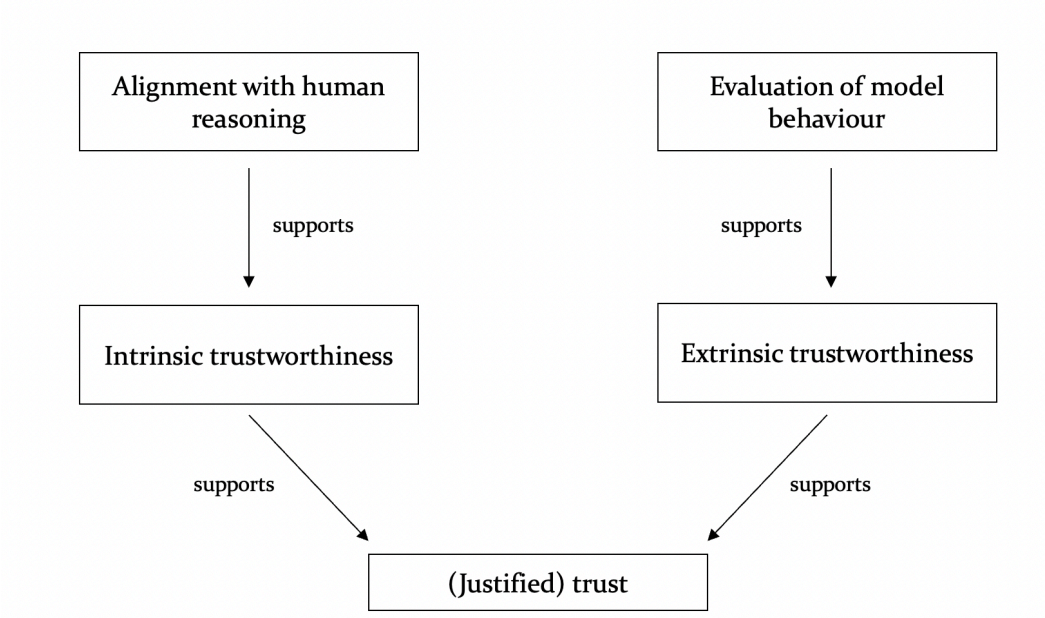

Fig. 11.1: Two kinds of ML trustworthiness, based on the model by Jacovi et al. (2021)

On the one hand, trustworthiness can rely on *internal* factors, namely when the reasoning process of the ML model is explainable, whether ex-ante or ex-post, and aligns with human reasoning (Jacovi et al., 2021). Here, the insights of chapter five and six can complement their view. If indeed a model gains trustworthiness through its alignment

[46] While the authors do not reference the paper, their account seems closely related to a distinction Mark Coeckelbergh has proposed in the context of robotics, between direct and indirect trust (2012). For a critical discussion of the model by Jacovi et al. see also Ferrario and Loi (2021).

with human reasoning, as seems plausible following recent simulation-based surveys (Alam & Mueller, 2021; Diprose et al., 2020), a model's perceived internal trustworthiness may not only be fostered by technical explainability but could also be supported by other approaches that focus on understanding, as argued in chapter five. In addition, as argued in chapter six, attempts to increase a model's trustworthiness require not mere disclosure but successful communication – which in turn, as I have stressed with an example from Onora O'Neill, presupposes a trusting relationship itself (Manson & O'Neill, 2007). The call to focus on communication was also shared by the interviewed experts, as presented in chapters eight and nine, who presented the need to carefully use and confer information in psychiatric contexts that may turn into self-fulfilling prophecies and could affect the symptoms a patient exhibits. By heeding O'Neill's advice, we may also avoid dangers of using explainability as a cover for poorly trained models since an explainable model would need to be designed to foster understanding on the side of its user and could not provide a mere "fig leaf".

On the other hand, *extrinsic* trustworthiness relies on observations not of the inner workings of an opaque algorithm, but of its behaviour in a specific context (Jacovi et al., 2021). Accepting the black-box nature of the model, this type of trustworthiness is supported by methodological rigour in the evaluation process, and potentially by proper regulation. Our plea for an outcome-oriented evaluation of algorithmic fairness in chapter four as well the practical demands from respecting the principles of biomedical ethics in the context of ML for diagnosing, predicting, and treating schizophrenia in chapter seven do equally fall into the domain of such *external* trustworthiness.

As has become clear throughout this thesis, the medical domain, and psychiatry in particular, provide particular challenges to ML modelling insofar as they are

characterised by a *dual black box*: The much-discussed opacity of ML is conjoined with potentially equally opaque black boxes, such as historically and socially contingent definitions of disorders and diseases, or unknown mechanism underlying medical interventions (Adamson & Welch, 2019; Holzinger, Langs, Denk, Zatloukal, & Müller, 2019). These problems relate to topics that can only be addressed appropriately by also drawing on empirical investigation as well as on the history and philosophy of science and medicine.

11.4 Integrating history into integrated empirical bioethics

From its outset, the thesis has espoused a methodological approach of integrated empirical bioethics, as described in detail in chapter two. This approach is particularly visible in the chapters that draw on qualitative interviews with experts commanding specific domain knowledge on applications of ML in psychiatry. The empirical research presented here is integrated, beyond a strict dichotomy of values and facts, in at least three ways. The first concerns the position of the interviewees, who had often received interdisciplinary training, including philosophy in five instances. Given that experts were thus informed and influenced by debates in ethics, the attitudes and opinions reported should not be considered as mere facts but are inherently value laden. Second, reflecting on my own positionality as researcher with dual education and knowledge from both fields, my interaction with the interviewees was also influenced by my own ethical attitudes and opinions. Third, such interaction between the empirical and normative was already present, to some extent, in the design of the project itself, providing a qualitative empirical methodology, but one that was normatively laden due to the questions and concepts used in the interview guide. The thesis therefore adhered

to its intended methodology. However, the question remains in which ways the empirical findings presented here can inform bioethical debate.

Kon (2009) distinguishes between four different roles of empirical research in bioethics, namely (1) to provide a "lay of the land", investigating the status quo and its practices and beliefs, (2) to compare such status quo with ideal (normative) theory, (3) to improve existing practices, and (4) to change and improve ethical theory. In its empirically driven chapters eight and nine, the thesis addresses aspects of all four roles. Investigating the opinions of experts on ML in psychiatry, it provides *the* first "lay of the land", reporting the views and attitudes of researchers' on questions of ethics (chapter 8) and psychiatric nosology (chapter 9). Chapter eight also allows for a comparison of the status quo in research with an ideal normative theory, in particular with view to the principle of explicability that is commonly accepted as fundamental in AI ethics, yet may be misused in practice. Pointing to this reported discrepancy, this chapter further aims to improve (research) practices, demanding more attention to methodological rigour, as do our empirical findings on ML's impact on nosology, demanding closer attention to the conceptualization of mental disorders when pursuing ML-based research in the field. Finally, while the empirical chapters themselves do not propose specific amendments to normative theories, read in the greater context of this thesis, they do support the arguments advanced in chapters five and six, calling for a focus on effective communication, aimed at understanding and trustworthiness.

As argued with view to schizophrenia in chapter ten, critical reflection informed by history remains crucial to both education and research in ethics. An integrated approach to bioethics will therefore need to go beyond the empirical social sciences and integrate historical reflection as well to achieve ethical reflection that is attentive to context.

While our contribution has already highlighted why this is the case with regard to the history of psychiatry, there are many further pertinent examples from the field of AI. The scholar Kate Crawford has recently highlighted the importance of history in the context of algorithmic bias (Crawford, 2021). The database ImageNet, spearheaded by Stanford professor Fei-Fei Li since 2009 and widely used for the training of ANNs, provides millions of annotated images that have been manually labelled by workers on platforms such as Amazon Mechanical Turk (ibid.). While criticism has often focused on the fact that ImageNet includes predominantly images of White people (Zou & Schiebinger, 2018), there are also less apparent problems hidden in its taxonomy of 21 841 hierarchical categories (Crawford & Paglen, 2021). To understand why an AI trained with ImageNet classifies certain people as "crazy", "ape-man", or "hooker" (Crawford, 2021, p. 109), one needs to look at its history. ImageNet derived its classificatory taxonomy from the WordNet database from the 1980s, which in turn draws on sources from the 1960s, providing a hierarchical list of classificatory terms (ibid., p. 98) – some of which are highly problematic in nature. For instance, with regard to sexual orientation, part of the classification is derived, through various stages, from the way in which books on LGBTQ themes were sorted in the US Library of Congress until 1972, falling under the category "Abnormal Sexual Relations, Including Sexual Crimes" (Crawford & Paglen, 2021).

As this example highlights, understanding the history of a program and its training data is therefore vital to understanding its current problems – and to fixing them. However, as stressed in this thesis, historical understanding alone is also not sufficient. Critical epistemological and normative reasoning from philosophy, empirical insights from the social science, e.g., into a novel technology's real-life effects, as well as the necessary

domain knowledge concerning the object of investigation, such as medicine and computer science in the context of this thesis, are equally required to provide a fully integrated approach to empirical bioethics (see fig. 11.2).[47]

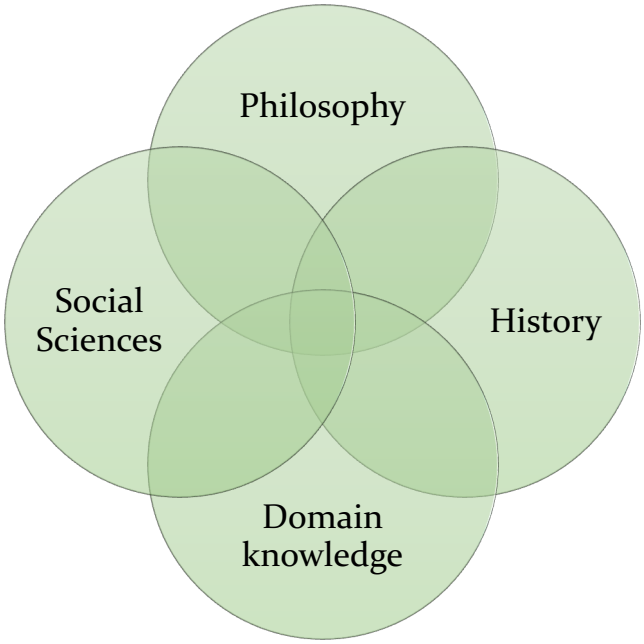

Fig. 11.2 Integrating integrated empirical bioethics

11.5 Limitations and implications for future research

There are several limitations to this thesis, comprising both its theoretical angle and its empirical methods. First, the dissertation is limited insofar as its conceptual prong reflects a Western-European context. While critics have long raised concerns about the eurocentrism of Kantian philosophy (Mignolo, 2002; Zanotti, 2021) and of the principlist framework of Beauchamp and Childress (Bach, 2021; Behrens, 2017), some authors have also asked whether explicability as the new principle of AI ethics is applicable in non-Western contexts (Carman & Rosman, 2021). Further research should therefore discuss

[47] Notably, similar methodological debates as in empirical bioethics, about to how to relate the abstract and the concrete to each other, have long riddled history and philosophy of science as well. See for instance (Chang, 2011; Herring, Jones, Kiprijanov, & Sellers, 2019; Mauskopf & Schmaltz, 2011; Sauer & Scholl, 2016).

whether the models and modes of trustworthiness presented here require adaptation in non-Western contexts.

Second, the majority of medical applications of ML discussed in this dissertation are not yet implemented at the bedside, let alone form part of established clinical routines. For example, Canvas Dx, the first ML-based program for psychiatric use, received FDA-approval only in June 2021 (Schuman, 2021). This limits the results of the thesis in two directions: Conceptually, my discussions of medical ML applications, e.g., for schizophrenia, necessarily drew on examples that are still in the state of research, while empirically, not one of the interviewees was able to report about practical experience with implementing ML at the bedside. It is thus crucial that future research examines and accompanies the introduction of clinical ML applications with critical bioethical investigation – e.g., if it assigns, as Canvas Dx does, the potentially stigmatising label of diagnosis of autism spectrum disorder to toddlers and young children. At the same time, explorative bioethical inquiry seems crucial even at this early stage of developments, to promote what has been called "ethics parallel research", anticipating and proactively guiding technological developments (Jongsma & Bredenoord, 2020), and to enable the embedding of ethics in the very development of clinical ML models (McLennan et al., 2020).

Third, an apparent limitation of the empirical prong of this thesis is its lack of generalisability, based on its small sample comprising only Swiss and German experts on psychiatric ML. Here, the research may have benefitted from including other professions as well, and originally, this is also what I set out to do, comparing psychiatrists with computer scientists. However, this strategy ran into two major obstacles: First, my recruitment efforts coincided exactly with the COVID-19 pandemic,

providing a major obstacle to recruit experts caught between the challenges of lockdown such as home schooling and additional clinical duties to carve out time for a qualitative interview by phone. Second, I soon discovered that the majority of people working in the field commanded an interdisciplinary training, as described in chapters eight and nine, so that no clear-cut distinction between different backgrounds would have been possible. Going beyond the rather small subfield of psychiatric ML, it may prove useful though if future research would explore the different educational worlds of computer scientists and physicians in greater depths.

Fourth, due to the focus of this thesis on ML applications in psychiatry, its findings should not be transferred uncritically to other medical fields. Explainability, for instance, may have different requirements in the context of a neurobiological ML model in psychiatry than explainable machine learning employed for image analysis in radiology, and successfully establishing relationships of warranted trust in the interaction of physicians, patients, and ML systems may also have very different demands in psychiatry than in, e.g., ophthalmology. While such differences may also be of relevance elsewhere, psychiatry still occupies a special place compared to other medical field since nowhere else, the very status as medical discipline is contested, or the question whether it treats illness and disease at all (Double, 2019). Further research should therefore dedicate particular attention to the particularities of individual medical specialties and not draw hastily on our findings from psychiatry.

Fifth, and finally, the centrality with which explainability featured as a topic in this dissertation was initially unintended and may have therefore been insufficiently covered in the interview guide (see appendix). Building on the findings of this dissertation as well as on other mostly conceptual research (Adadi & Berrada, 2018; Markus, Kors, &

Rijnbeek, 2021; McCoy et al., 2020; Shin, 2020; Tonekaboni, Joshi, McCradden, & Goldenberg, 2019; Vilone & Longo, 2021), further empirical research is therefore urgently needed to examine the bioethical demands of explainability in medical ML. Only with such an empirically grounded understanding of the communicative needs of physicians and patients may we hope to achieve ML-based systems that are deserving of trust.

11.6 Conclusion

This thesis has argued that we can trust medical ML models, employed for instance in psychiatry, if they fulfil certain conditions of extrinsic and intrinsic trustworthiness. However, the actual impact of opaque ML algorithms on clinical practice will not only depend on their technical realisation but also on their acceptance by medical professionals – i.e., whether these systems are treated as competitors or as collaborators (Grote & Berens, 2022). As this thesis has highlighted, both paths are lined with ethical pitfalls.

Simply shunning ML from clinical use would not sufficiently harness its potential for the improvement of current clinical practices and would ignore a valuable opportunity to advance research in which AI conspires with human creativity and intuition, as has been the case in mathematics (Davies et al., 2021; Stump, 2021). Doing so without good reason, e.g., based on an unfounded general scepticism towards novel technologies, would therefore contradict the principle of beneficence. At the same time, overreliance on medical ML could prove harmful to patients, contradicting the fundamental Hippocratic duty of physicians to do no harm. Such harm seems particularly likely if unwarranted trust is expedited to what the artist and author Hito Steyerl has called "artificial stupidity": automated processes that reflect and reinforce moral and intellectual shortfalls of society (Steyerl & Crawford, 2017).

As I have argued, to avoid falling prey to either danger demands extended teaching efforts: in clinical contexts to medical students, in academic contexts to engineers, computer scientists, and ethicists, and the fostering of collaborative practices across disciplines to promote joint reflection (Gauld, Micoulaud-Franchi, & Dumas, 2021). In sum, this thesis should therefore be read as an exhortation for proper training of all neural nets involved, both artificial and human: the most trustworthy medical ML system will be of little benefit to patients unless it rests in the hands of an equally competent and trustworthy user.

11.7 References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access, 6*, 52138-52160.

Adamson, A. S., & Welch, H. G. (2019). Machine Learning and the Cancer-Diagnosis Problem - No Gold Standard. *New England Journal of Medicine, 381*(24), 2285-2287. doi:10.1056/NEJMp1907407

Alam, L., & Mueller, S. (2021). The Myth of Diagnosis as Classification: Examining the Effect of Explanation on Patient Satisfaction and Trust in AI Diagnostic Systems. *BMC Medical Informatics and Decision Making, 21*, 178. doi:10.1186/s12911-021-01542-6

Appelbaum, P. S. (2017). Moving toward the future in the diagnosis of mental disorders. *Psychological Science in the Public Interest, 18*(2), 67-71.

Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J., Shaw, D., & Elger, B. S. (2022). Re-focusing explainability in medicine. *Digital health, (in print)*.

Bach, M. C. (2021). When the universal is particular: a re-examination of the common morality using the work of Charles Taylor. *Medicine, Health Care and Philosophy*, 1-11. doi:10.1007/s11019-021-10059-8

Beauchamp, T., & Childress, J. (2019). Principles of biomedical ethics: marking its fortieth anniversary. *American Journal of Bioethics, 19*(11), 9-12.

Behrens, K. G. (2017). Hearing sub-Saharan African voices in bioethics. *Theoretical Medicine and Bioethics, 38*(2), 95-99.

Braun, M., Bleher, H., & Hummel, P. (2021). A Leap of Faith: Is There a Formula for "Trustworthy" AI? *Hastings Center Report, 51*(3), 17-22. doi:10.1002/hast.1207

Bryson, J. (2018). No One Should Trust AI. *AI & Global Governance*. Retrieved from https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Carman, M., & Rosman, B. (2021). Applying a principle of explicability to AI research in Africa: should we do it? *Ethics and Information Technology, 23*(2), 107-117.

Chang, H. (2011). Beyond case-studies: History as philosophy. In *Integrating history and philosophy of science* (pp. 109-124): Springer.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*(1), 53-60. doi:10.1007/s10676-011-9279-1

Crabtree, A., Urquhart, L., & Chen, J. (2019). Right to an explanation considered harmful. *Edinburgh School of Law Research Paper, available at SSRN*. doi:10.2139/ssrn.3384790

Crawford, K. (2021). *Atlas of AI : power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY*, 1-12.

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. doi:10.1038/s41586-021-04086-x

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *Lancet Digital Health, 1*(8), E390. doi:10.1016/S2589-7500(19)30197-9

Diprose, W. K., Buist, N., Hua, N., Thurier, Q., Shand, G., & Robinson, R. (2020). Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association, 27*(4), 592-600.

Double, D. B. (2019). Twenty years of the critical psychiatry network. *The British Journal of Psychiatry, 214*(2), 61-62.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645-666. doi:10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329-335. doi:10.1136/medethics-2020-106820

Edwards, L., & Veale, M. (2017). Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review, 16*, 18. doi:10.2139/ssrn.2972855

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, (2019). Retrieved from https://data.europa.eu/doi/10.2759/177365

Ferrario, A., & Loi, M. (2021). The Meaning of "Explainability Fosters Trust in AI". *Available at SSRN*. doi:10.2139/ssrn.3916396

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philosophy & Technology, 33*(3), 523-539.

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: it is possible to trust medical AI. *Journal of Medical Ethics, 47*(6), 437-438.

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. doi:10.1017/S0033291720003906

Grote, T. (2021). Trustworthy medical AI systems need to know when they don't know. *Journal of Medical Ethics, 47*(5), 337-338.

Grote, T., & Berens, P. (2022). How competitors become collaborators—Bridging the gap (s) between machine learning algorithms and clinicians. *Bioethics, 36*(2), 134-142. doi:10.1111/bioe.12957

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478-481. doi:10.1136/medethics-2019-105935

Herring, E., Jones, K. M., Kiprijanov, K. S., & Sellers, L. M. (2019). *The Past, Present, and Future of Integrated History and Philosophy of Science*. London: Routledge.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). *Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai.* Paper presented at the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Jongsma, K. R., & Bredenoord, A. L. (2020). Ethics parallel research: an approach for (early) ethical guidance of biomedical innovation. *BMC Medical Ethics, 21*(1), 1-9.

Kon, A. A. (2009). The role of empirical research in bioethics. *The American Journal of Bioethics, 9*(6-7), 59-65.

Latour, B. (1994). On technical mediation. *Common knowledge, 3*(2), 29-64.

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge: Cambridge University Press.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655.

Mauskopf, S., & Schmaltz, T. (2011). *Integrating history and philosophy of science: Problems and prospects*. Dordrecht: Springer

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), 86. doi:10.1038/s41746-020-0294-7

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., Haddadin, S., & Buyx, A. (2020). An embedded ethics approach for AI development. *Nature Machine Intelligence, 2*(9), 488-490.

Metzinger, T. (2019). Ethics washing made in Europe. *Der Tagesspiegel*. Retrieved from https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html

Mignolo, W. (2002). *The many faces of cosmo-polis: Border thinking and critical cosmopolitanism*. Durham, NC: Duke University Press.

O'Neill, O. (2013). Trust before trustworthiness? In D. Archard, M. Deveaux, N. C. Manson, & D. Weinstock (Eds.), *Reading Onora O'Neill* (pp. 237-238). Oxford: Routledge.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749-2767.

Sauer, T., & Scholl, R. (2016). Towards a methodology for integrated history and philosophy of science. In R. Scholl & T. Räz (Eds.), *The philosophy of historical case studies* (pp. 69-91). Cham: Springer.

Schuman, A. J. (2021). Facilitating Autism Diagnosis. *Psychiatric Times, 38*(10), 22-23.

Shin, D. (2020). The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. *International Journal of Human-Computer Studies*. doi:https://doi.org/10.1016/j.ijhcs.2020.102551

Starke, G., van den Brule, R., Elger, B. S., & Haselager, P. (2021). Intentional machines: A defence of trust in medical artificial intelligence. *Bioethics*. doi:10.1111/bioe.12891

Steyerl, H., & Crawford, K. (2017). Data streams. *The New Inquiry, 23*. Retrieved from https://thenewinquiry.com/data-streams/

Stump, C. (2021). Artificial intelligence aids intuition in mathematical discovery. *600*. doi:10.1038/d41586-021-03512-4

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). *What clinicians want: contextualizing explainable machine learning for clinical end use.* Paper presented at the Machine learning for healthcare conference.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89-106. doi:10.1016/j.inffus.2021.05.009

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76-99. doi:10.1093/idpl/ipx005

Zanotti, L. (2021). De-colonizing the political ontology of Kantian ethics: A quantum perspective. *Journal of International Political Theory, 17*(3), 448-467.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature, 559*(7714), 324-326. doi:10.1038/d41586-018-05707-8

Appendix

Appendix 1: COREQ Checklist

Based on the SRQR guidelines

			Page No
Title			
	1.	Concise description of the nature and topic of the study identifying the study as qualitative or indicating the approach (e.g. ethnography, grounded theory) or data collection methods (e.g. interview, focus group) is recommended.	176
Abstract			
	2.	Summary of the key elements of the study using the abstract format of the intended publication; typically includes background, purpose, methods, results and conclusions.	177
Introduction			
Problem formulation	3.	Description and significance of the problem / phenomenon studied: review of relevant theory and empirical work; problem statement.	178-180
Purpose or research question	4.	Purpose of the study and specific objectives or questions.	180-181
Methods			
Qualitative approach and research paradigm	5.	Qualitative approach (e.g. ethnography, grounded theory, case study, phenomenology, narrative research) and guiding theory if appropriate; identifying the research paradigm (e.g. postpositivist, constructivist / interpretivist) is also recommended; rationale. The rationale should briefly discuss the justification for choosing that theory, approach, method or technique rather than other options available; the assumptions and limitations implicit in those choices and how those choices influence study conclusions and transferability. As appropriate the rationale for several items might be discussed together.	181-182
Researcher characteristics and reflexivity	6.	Researchers' characteristics that may influence the research, including personal attributes, qualifications / experience, relationship with participants, assumptions and / or presuppositions; potential or actual interaction between researchers' characteristics and the research questions, approach, methods, results and / or transferability.	181-182
Context	7.	Setting / site and salient contextual factors; rationale.	181-182
Sampling strategy	8.	How and why research participants, documents, or events were selected; criteria for deciding when no further sampling was necessary (e.g. sampling saturation); rationale.	180-181
Ethical issues pertaining to human subjects	9.	Documentation of approval by an appropriate ethics review board and participant consent, or	182

		explanation for lack thereof; other confidentiality and data security issues.	
Data collection methods	10.	Types of data collected; details of data collection procedures including (as appropriate) start and stop dates of data collection and analysis, iterative process, triangulation of sources / methods, and modification of procedures in response to evolving study findings; rationale.	181
Data collection instruments and technologies	11.	Description of instruments (e.g. interview guides, questionnaires) and devices (e.g. audio recorders) used for data collection; if / how the instruments(s) changed over the course of the study.	181, 183
Units of study	12.	Number and relevant characteristics of participants, documents, or events included in the study; level of participation (could be reported in results).	183-184
Data processing	13.	Methods for processing data prior to and during analysis, including transcription, data entry, data management and security, verification of data integrity, data coding, and anonymisation / deidentification of excerpts.	182
Data analysis	14.	Process by which inferences, themes, etc. were identified and developed, including the researchers involved in data analysis; usually references a specific paradigm or approach; rationale.	182
Techniques to enhance trustworthiness	15.	Techniques to enhance trustworthiness and credibility of data analysis (e.g. member checking, audit trail, triangulation); rationale.	182-183
Results/findings			
Syntheses and interpretation	16.	Main findings (e.g. interpretations, inferences, and themes); might include development of a theory or model, or integration with prior research or theory.	183-193
Links to empirical data	17.	Evidence (e.g. quotes, field notes, text excerpts, photographs) to substantiate analytic findings.	185-193
Discussion			
Integration with prior work, implications, transferability and contribution(s) to the field	18.	Short summary of main findings; explanation of how findings and conclusions connect to, support, elaborate on, or challenge conclusions of earlier scholarship; discussion of scope of application / generalizability; identification of unique contributions(s) to scholarship in a discipline or field.	193-198
Limitations	19.	Trustworthiness and limitations of findings.	197-198
Other			
Conflicts of interest	20.	Potential sources of influence of perceived influence on study conduct and conclusions; how these were managed.	182, 198
Funding	21.	Sources of funding and other support; role of funders in data collection, interpretation and reporting.	176

Appendix 2: Interview guide

Part 1: Expert's work on ML and its potential for clinical use

1. Which role do AI and machine learning play for your work?
    - In the clinic? In research?
    - Which methods and data do you use?
    - Are any of these methods already used in clinical practice?

2. For which applications of machine learning do you see the greatest potential in future psychiatry?
    - For which clinical objective (diagnostic, prognostic, therapeutic/response prediction)?
    - For which psychiatric disorders or symptoms?
    - In what time frame could you imagine that these applications are ready to be implemented in patient care?

Part 2: Ethics of medical ML

3. What do you consider the biggest ethical challenge for successfully implementing ML in clinical contexts?
    - Why?
    - How would you address this issue?
    - Do you have an example?

4. A frequently discussed problem is that of so-called black-box programs, for example in the form of ANNs, so programs that may be inaccessible to human understanding in principle. In your opinion, should such black-box programs be used for clinical purposes?
    - Why/why not?
    - If yes: What should doctors and patients know about such programs?
    - Should the information be stratified for different groups of users? Why/why not?
    - Do you see specific barriers to effective communication between the groups involved, and if so, how do you think these could be addressed?

5. The EU's expert group on artificial intelligence famously put its ethical guidelines for AI under the heading of "trustworthy AI". Do you think trust is a justifiable way of dealing with the risks of medical AI?
    - Why / why not?
    - In a clinical context, when would you consider a program "trustworthy"?
    - Do you think the EU Commission guidelines are useful?
    - Do you feel that the guidelines have had or are having an impact on your research area? If so, in what way?

6. As part of trustworthiness, one commonly finds calls for transparency. What specific expectations would you have for the transparency of such programmes?
    - To whom should this information be disclosed?
    - Which technical strategies for making machine learning more transparent do you think are the most promising? Could you give an example?

Part 3: Specific questions of psychiatry

7. Are there, in your opinion, any *particular* ethical problems for using ML in psychiatry?
    - (If no answer:) How would you suggest to deal with cases of impaired judgement? E.g., a ML program that can recommend the most suitable antipsychotic medication during a psychotic episode?

8. As you know, some authors argue that machine learning, and Deep Learning in particular, promise a way to finally define psychiatric disorders as natural kinds and solve the old problems of psychiatric nosology. Where would you stand on this?

9. Is there any other topic that seems central to you which we have not yet covered?

Appendix 3: Jurisdictional inquiry

EKNZ

Eth kommission
Nordwest- und
Zentralschweiz

Präsident
Prof. Christoph Beglinger
Vizepräsidenten
Dr. Angela Frotzler
Dr. Marco Schärer

Georg Starke
Institute for Biomedical Ethics
Bernoullistr. 28
4056 Basel

Basel, 14th October 2019 / nj

Statement of the Ethics Committee Northwest and Central Switzerland (EKNZ) according to HRA Art.51

Project-ID	Req-2019-00920
Project title	Trusting Black-Box Algorithms? Ethical Challenges for Biomedical Machine Learning
Submission Date	30/09/2019
Applicant	Georg Starke

Decision

The research project doesn't fall under the remit of the cantonal or federal law HRA (Human Research Act), because your project is not defined as a research project as per HRA Art. 2. Therefore, the EKNZ cannot officially approve your project.

The EKNZ has reviewed the submitted documents and can confirm that the research project fulfils the general ethical and scientific standards for research with humans (see Art. 51 Abs. 2 HRA).

Fees

Tariff code: 6.0 **Amount: 300** CHF
In accordance with the current swissethisc fee schedule.

With the Committee's best wishes for the success of this project.

Yours sincerely,

i.V. mfjane.

Prof. Christoph Beglinger
President of the Ethics Committee
Northwest and Central Switzerland / EKNZ

Attachment 1. Liste of documents, submitted on 30. September 2019

Geschäftsführerin Irene Oberli | Hebelstrasse 53 | 4056 Basel | Tel 061 268 13 50 | Fax 061 268 13 51 | eknz@bs.ch | www.eknz.ch