AI and Ethics (2023) 3:303–314
https://doi.org/10.1007/s43681-022-00177-1

ORIGINAL RESEARCH

Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry

Georg Starke[1,2] · **Benedikt Schmidt**[1] · **Eva De Clercq**[1] · **Bernice Simone Elger**[1,3]

Received: 22 January 2022 / Accepted: 17 May 2022 / Published online: 7 June 2022

Abstract
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. 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. 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. 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.

Keywords Artificial intelligence · Machine learning · Bioethics · Explainability · Mental health

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 [1, 2], 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 [3]. 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 [4–7]. A broad debate about the ethical principles governing the development of ML-based psychiatric tools seems therefore more pressing then ever [8, 9].

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 [10]. For a debate within the context of European health care, the influential ethical framework of 'AI4people' seems particularly instructive [11]. It builds on the four principles of biomedical ethics by Beauchamp and Childress [12], i.e., respect for autonomy, beneficence, non-maleficence, and justice, supplementing them with an additional fifth principle of explicability.

✉ Georg Starke
georg.starke@unibas.ch

1 Institute for Biomedical Ethics, University of Basel, Bernoullistr. 28, 4056 Basel, Switzerland

2 College of Humanities, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

3 University Center of Legal Medicine, University of Geneva, Geneva, Switzerland

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 [13].

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 [14]. 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 brought up to speed, taking place in parallel to developments, or how ethical considerations could be embedded in research pipelines [15, 16]. Nevertheless, as of now, there are 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 [17]. 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 [18], whereas Blease et al. focused on the views of UK General Practitioners [19], and Tonekaboni et al. examined expectations towards explainability among ten Canadian acute care specialists [20].

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 [21]. 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 [22].

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 [23–25], however, with a unique focus on ethical challenges and on ML applications in psychiatry. Expert interviews are an established method to investigate attitudes and knowledge of people working in the field [26]. 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" [27]. 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 [28] and do not pay enough attention to ethically relevant, yet often neglected aspects such as environmental cost or exploitation of labor [29]. 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 [30] but also across Western European countries [31, 32]. 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 [28, 33, 34], as well as to tailor policy making for the introduction of ethically sound, trustworthy ML in the clinic [35–37].

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

Table 1 Relevant questions from the interview guide

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"?

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 min. The interviews were transcribed verbatim by the first and second author. Quotes used within this paper were translated by the first author 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 [38, 39]. 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 GS, BS, and EDC 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 [40]. In line with the previous findings, we did not find new codes after coding the 11th interview [41].

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, and (3) and we stored these 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.

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 can 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.

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 was reported to be at odds with methodological rigor, potentially already at the stage of collecting representative training data, including from 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 [42], 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 with ML you have a problem, because once you train on 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 the small number of experts in the field who 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 favors 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)

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 which they considered crucial to keep medical practice compatible with the 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)

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 eve-

> rything 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)

4 Discussion

Our findings offer 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 [9, 43] as are debates about the balancing of hierarchy between patients, physicians, and ML systems [44, 45]. 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 [22]. 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 [44]. Mirroring common tropes of the debate, our interviewees also called for considering ML systems as intelligent tools, not artificial colleagues [46] and did not foresee a step towards a full automation in the near future [47], 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 [48, 49]. However, with regard to trust, as an attitude that partially relinquishes the monitoring of algorithms [50–52], the interviewees represented a comprehensive spectrum of opinions. As in the ethical literature [53–55], 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 [28, 33, 52, 56]. 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 [11, 57], others have already noted that there is no uniform consensus among experts about the meaning of explainability [58, 59], and that expectations towards explainability vary across contexts [60]. This is also confirmed by our study, as are concerns about balancing explainability with accuracy [34], about the need of contestability [61], and about the importance of epistemological questions for an ethical use of ML systems [62].[1] 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 or, more precisely, of a principle of explicability, linking intelligibility and accountability [11, 13, 66]. 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 [67]. 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 favored models with low denunciatory power, i.e., they placed higher trust in "explainable" AI systems where unfair decisions were not perceived negatively [67].

Given these findings, further conceptual and empirical research should therefore critically investigate if, instead of providing a mediating principle enabling ethical scrutiny [11, 57], 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 [68], it may prove necessary to challenge the widely held belief that explainability is key to the acceptance of AI [69]. 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 [50]. In line with others, our finding also highlights themis need to refocus the view onto explainability and move towards more user-centered models of explainability that can provide meaningful understanding for physicians and patients [59, 60] and harness multiple levels of explanation [70].

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 [[71]: 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 rigor [72]. Methodological concern was also tangible in calls for proper external validation to ensure the generalizability of ML systems across different demographics [65], and with view to the increasing importance placed on the diversity of cohort and data in clinical research [73]. Other much-discussed aspects of fairness, e.g., the problem of competing fairness standards [74, 75], 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 [76, 77]. 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

[1] 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 [63]. Similarly, some form of contestability is increasingly implemented in ML by virtue of counterfactual reasoning [64]. 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 [65].

of psychiatry.[2] 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 clinical ML employed in psychiatry, more empirically informed research will be crucial, accompanying the implementation of psychiatric ML [15]. 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 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.

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 an ML model gets involved. Early involvement of ethical considerations in the development pipeline [16] seems therefore as crucial as stratified basic education on computer science both of physicians and the public, in line with the detailed recommendations of others [80]. 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.

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.

Funding Open access funding provided by University of Basel.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Martinez-Martin, N., Kreitmair, K.: Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. JMIR Ment. Health. **5**(2), e32 (2018). https://doi.org/10.2196/mental.9423
2. Lui, J.H., Marcus, D.K., Barry, C.T.: Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. Prof. Psychol. Res. Pract. **48**(3), 199–210 (2017). https://doi.org/10.1037/pro0000122
3. Dattaro, L.: Green light for diagnostic autism app raises questions concerns. Spectrum (2021). https://doi.org/10.53053/IZWC9259
4. Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., et al.: The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry **20**(2), 154–170 (2021). https://doi.org/10.1002/wps.20882
5. der Salazar Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., et al.: Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. Schizophr. Bull. **47**(2), 284–297 (2021). https://doi.org/10.1093/schbul/sbaa120
6. Chivilgina, O., Elger, B.S., Jotterand, F.: Digital technologies for schizophrenia management: a descriptive review. Sci. Eng. Ethics **27**(2), 1–22 (2021)
7. Chivilgina, O., Wangmo, T., Elger, B.S., Heinrich, T., Jotterand, F.: mHealth for schizophrenia spectrum disorders management: a systematic review. Int. J. Soc. Psychiatry **66**(7), 642–665 (2020)
8. Jacobson, N.C., Bentley, K.H., Walton, A., Wang, S.B., Fortgang, R.G., Millner, A.J., et al.: Ethical dilemmas posed by mobile health and machine learning in psychiatry research. Bull. World

[2] It should be noted that there has been much progress in the development of context-sensitive ML recently [78, 79]. We are indebted to an anonymous reviewer for pointing this out.

Health Organ. **98**(4), 270–276 (2020). https://doi.org/10.2471/BLT.19.237107
9. Starke, G., De Clercq, E., Borgwardt, S., Elger, B.S.: Computing schizophrenia: ethical challenges for machine learning in psychiatry. Psychol. Med. **51**(15), 2515–2521 (2021). https://doi.org/10.1017/S0033291720001683
10. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
11. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al.: AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
12. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics, 7th edn. Oxford University Press, New York (2013)
13. Intelligence H-LEGoA.: Ethics guidelines for trustworthy AI. In: High-Level Expert Group on Artificial Intelligence. Brussels: EU Commission (2019)
14. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019)
15. Jongsma, K.R., Bredenoord, A.L.: Ethics parallel research: an approach for (early) ethical guidance of biomedical innovation. BMC Med. Ethics **21**(1), 1–9 (2020)
16. McLennan, S., Fiske, A., Celi, L.A., Müller, R., Harder, J., Ritt, K., et al.: An embedded ethics approach for AI development. Nat. Mach. Intell. **2**(9), 488–490 (2020)
17. Wangmo, T., Hauri, S., Gennet, E., Anane-Sarpong, E., Provoost, V., Elger, B.S.: An update on the "empirical turn" in bioethics: analysis of empirical research in nine bioethics journals. BMC Med. Ethics **19**(1), 1–9 (2018)
18. Nichol, A.A., Bendavid, E., Mutenherwa, F., Patel, C., Cho, M.K.: 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 (2021). https://doi.org/10.1136/bmjopen-2021-052287
19. Blease, C., Kaptchuk, T.J., Bernstein, M.H., Mandl, K.D., Halamka, J.D., DesRoches, C.M.: Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J. Med. Internet Res. **21**(3), e12802 (2019). https://doi.org/10.2196/12802
20. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. Machine learning for healthcare conference: PMLR 359–80 (2019)
21. Blease, C., Locher, C., Leon-Carlyle, M., Doraiswamy, M.: Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. Digital Health. **6**, 2055207620968355 (2020). https://doi.org/10.1177/2055207620968355
22. Blease, C., Kharko, A., Annoni, M., Gaab, J., Locher, C.: Machine learning in clinical psychology and psychotherapy education: a mixed methods pilot survey of postgraduate students at a Swiss University. Front. Public Health **9**, 623088 (2021). https://doi.org/10.3389/fpubh.2021.623088
23. Pumplun, L., Fecho, M., Wahl, N., Peters, F., Buxmann, P.: Adoption of machine learning systems for medical diagnostics in clinics: qualitative interview study. J. Med. Internet Res. **23**(10), e29301 (2021)
24. Morgenstern, J.D., Rosella, L.C., Daley, M.J., Goel, V., Schünemann, H.J., Piggott, T.: "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 (2021)
25. Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M.: "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 (2019)
26. Döringer, S.: The problem-centred expert interview'. Combining qualitative interviewing approaches for investigating implicit expert knowledge. Int. J. Soc. Res. Methodol. **24**(3), 265–278 (2021)
27. Hagendorff, T.: Blind spots in AI ethics. AI Ethics. (2021). https://doi.org/10.1007/s43681-021-00122-8
28. Braun, M., Bleher, H., Hummel, P.: A leap of faith: is there a formula for "trustworthy" AI? Hastings Cent. Rep. **51**(3), 17–22 (2021). https://doi.org/10.1002/hast.1207
29. Crawford, K.: Atlas of AI : power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven (2021)
30. Bröhl, C., Nelles, J., Brandl, C., Mertens, A., Nitsch, V.: Human–robot collaboration acceptance model: development and comparison for Germany, Japan, China and the USA. Int. J. Soc. Robot. **11**(5), 709–726 (2019)
31. Van den Berg B.: Differences between Germans and Dutch people in perception of social robots and the tasks robots perform. 16th Twente Student Conference on IT p. 1–6 (2012)
32. Conti, D., Cattani, A., Di Nuovo, S., Di Nuovo, A.: 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). 2015:555–60. doi: https://doi.org/10.1109/ROMAN.2015.7333601
33. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. J. Med. Ethics **47**(5), 329–335 (2021). https://doi.org/10.1136/medethics-2020-106820
34. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**(1), 15–21 (2019). https://doi.org/10.1002/hast.973
35. Walter, M., Alizadeh, S., Jamalabadi, H., Lueken, U., Dannlowski, U., Walter, H., et al.: Translational machine learning for psychiatric neuroimaging. Progress Neuropsychopharmacol. Biol. Psychiatry. **91**, 113–121 (2019). https://doi.org/10.1016/j.pnpbp.2018.09.014
36. Paulus, M.P., Huys, Q.J., Maia, T.V.: A roadmap for the development of applied computational psychiatry. Biol. Psychiatry. **1**(5), 386–392 (2016). https://doi.org/10.1016/j.bpsc.2016.05.001
37. Char, D.S., Abramoff, M.D., Feudtner, C.: Identifying ethical considerations for machine learning healthcare applications. Am. J. Bioeth. **20**(11), 7–17 (2020). https://doi.org/10.1080/15265161.2020.1819469
38. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)
39. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qual. Res. Sport Exerc. Health. **11**(4), 589–597 (2019)
40. Given, L.M.: 100 questions (and answers) about qualitative research. SAGE publications, London (2015)
41. Guest, G., Bunce, A., Johnson, L.: How many interviews are enough? An experiment with data saturation and variability. Field Methods **18**(1), 59–82 (2006)
42. Ross, C., Swetlitz, I.: IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ (2018). Accessed 31 July 2019
43. Morley, J., Machado, C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., et al.: The debate on the ethics of AI in health care: a reconstruction and critical review. SSRN. (2020). https://doi.org/10.2139/ssrn.3486518

44. Braun, M., Hummel, P., Beck, S., Dabrock, P.: Primer on an ethics of AI-based decision support systems in the clinic. J Med. Ethics. (2020). https://doi.org/10.1136/medethics-2019-105860
45. Grote, T., Berens, P.: How competitors become collaborators—bridging the gap (s) between machine learning algorithms and clinicians. Bioethics **36**(2), 134–142 (2022). https://doi.org/10.1111/bioe.12957
46. Dennett, D.: What can we do? In: Brockman, J. (ed.) Possible Minds. Twenty-Five Ways of Looking at AI, pp. 41-53. Penguin, New York (2019)
47. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7
48. Matthias, A.: The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf Technol. **6**(3), 175–183 (2004)
49. Bublitz, C., Wolkenstein, A., Jox, R.J., Friedrich, O.: Legal liabilities of BCI-users: responsibility gaps at the intersection of mind and machine? Int. J. Law Psychiatry (2018). https://doi.org/10.1016/j.ijlp.2018.10.002
50. Ferrario, A., Loi, M.: The meaning of "explainability fosters trust in AI." Available SSRN. (2021). https://doi.org/10.2139/ssrn.3916396
51. Ferrario, A., Loi, M., Viganò, E.: In AI we trust incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. Philos. Technol. **33**(3), 523–539 (2020)
52. Ferrario, A., Loi, M., Viganò, E.: Trust does not need to be human: it is possible to trust medical AI. J. Med. Ethics **47**(6), 437–438 (2021)
53. Metzinger, T.: Ethics washing made in Europe. Der Tagesspiegel (2019)
54. Hatherley, J.J.: Limits of trust in medical AI. J. Med. Ethics **46**(7), 478–481 (2020). https://doi.org/10.1136/medethics-2019-105935
55. DeCamp, M., Tilburt, J.C.: Why we cannot trust artificial intelligence in medicine. Lancet Digital Health. **1**(8), E390 (2019). https://doi.org/10.1016/S2589-7500(19)30197-9
56. Starke, G., van den Brule, R., Elger, B.S., Haselager, P.: Intentional machines: a defence of trust in medical artificial intelligence. Bioethics (2021). https://doi.org/10.1111/bioe.12891
57. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics Inf Technol. **11**(2), 105–112 (2009). https://doi.org/10.1007/s10676-009-9187-9
58. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. **6**, 52138–52160 (2018)
59. Arbelaez Ossa, L., Starke, G., Lorenzini, G., Vogt, J., Shaw, D., Elger, B.S.: Re-focusing explainability in medicine. Digital Health (2022). https://doi.org/10.1177/20552076221074488
60. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. FAT* '19: Proceedings of the conference on fairness, accountability, and transparency. 2019:279–88. doi: https://doi.org/10.1145/3287560.3287574
61. Ploug, T., Holm, S.: The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. Artif. Intell. Med. **107**, 101901 (2020)
62. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**(3), 205–211 (2020). https://doi.org/10.1136/medethics-2019-105586
63. Rudin, C., Radin, J.: Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Sci. Rev. (2019). https://doi.org/10.1162/99608f92.5a8a3a3d
64. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: a review. arXiv preprint. arXiv:201010596. (2020)
65. Cearns, M., Hahn, T., Baune, B.T.: Recommendations and future directions for supervised machine learning in psychiatry. Transl. Psychiatry **9**(1), 1–12 (2019)
66. Herzog, C.: On the risk of confusing interpretability with explicability. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00121-9
67. John-Mathews, J.-M.: Some critical and ethical perspectives on the empirical turn of AI interpretability. Technol. Forecast. Soc. Chang. **174**, 121209 (2022)
68. Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., Beste, C.: Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research, pp. 1143–1146. Springer (2021)
69. Chandler, C., Foltz, P.W., Elvevåg, B.: Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. Schizophr. Bull. **46**(1), 11–14 (2020)
70. Vu, M.-A.T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., et al.: A shared vision for machine learning in neuroscience. J. Neurosci. **38**(7), 1601–1607 (2018)
71. Thomas, W., Thomas, D.: The child in America. Knopf, New York (1928)
72. Floridi, L.: AI and its new winter: from myths to realities. Philos. Technol. **33**(1), 1–3 (2020)
73. Rubin, E.: Striving for diversity in research studies. N. Engl. J. Med. **385**(15), 1429–1430 (2021). https://doi.org/10.1056/NEJMe2114651
74. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im)possibility of fairness. arXiv preprint. arXiv:160907236. (2016)
75. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning. Limitations and opportunities (2019)
76. Rivera, S.C., Liu, X., Chan, A.-W., Denniston, A.K., Calvert, M.J.: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat. Med. **26**, 1351–1363 (2020). https://doi.org/10.1038/s41591-020-1037-7
77. Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., Denniston, A.K.: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat. Med. **26**, 1364–1374 (2020). https://doi.org/10.1038/s41591-020-1034-x
78. Nascimento, N., Alencar, P., Lucena, C., Cowan, D.: A context-aware machine learning-based approach. Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering. p. 40–7 (2018)
79. Elayan, H., Aloqaily, M., Guizani, M.: Digital twin for intelligent context-aware iot healthcare systems. IEEE Internet Things J. **8**(23), 16749–16757 (2021)
80. Gauld, C., Micoulaud-Franchi, J.-A., Dumas, G.: Comment on Starke et al. 'Computing schizophrenia: ethical challenges for machine learning in psychiatry': from machine learning to student learning: pedagogical challenges for psychiatry. Psychol. Med. **51**(14), 2509–2511 (2021). https://doi.org/10.1017/S0033291720003906