frontiers | Frontiers in Psychiatry

TYPE Conceptual Analysis
PUBLISHED 24 August 2023
DOI 10.3389/fpsyt.2023.1209862

OPEN ACCESS

EDITED BY
Giovanna Parmigiani,
Sapienza University of Rome, Italy

REVIEWED BY
Leda Tortora,
Sapienza University of Rome, Italy
Cristina Mazza,
University of Studies G. d'Annunzio Chieti and Pescara, Italy

*CORRESPONDENCE
Georg Starke
✉ georg.starke@epfl.ch

RECEIVED 21 April 2023
ACCEPTED 07 August 2023
PUBLISHED 24 August 2023

CITATION
Starke G, D'Imperio A and Ienca M (2023) Out of their minds? Externalist challenges for using AI in forensic psychiatry.
Front. Psychiatry 14:1209862.
doi: 10.3389/fpsyt.2023.1209862

Out of their minds? Externalist challenges for using AI in forensic psychiatry

Georg Starke[1,2,3]*, Ambra D'Imperio[1,4,5] and Marcello Ienca[1,2]

[1]Faculty of Medicine, Institute for History and Ethics of Medicine, Technical University of Munich, Munich, Germany, [2]École Polytechnique Fédérale de Lausanne, College of Humanities, Lausanne, Switzerland, [3]Munich School of Philosophy, Munich, Germany, [4]Department of Psychiatry, Hôpitaux Universitaires de Genève, Geneva, Switzerland, [5]Service of Forensic Psychiatry CURML, Geneva University Hospitals, Geneva, Switzerland

Harnessing the power of machine learning (ML) and other Artificial Intelligence (AI) techniques promises substantial improvements across forensic psychiatry, supposedly offering more objective evaluations and predictions. However, AI-based predictions about future violent behaviour and criminal recidivism pose ethical challenges that require careful deliberation due to their social and legal significance. In this paper, we shed light on these challenges by considering externalist accounts of psychiatric disorders which stress that the presentation and development of psychiatric disorders is intricately entangled with their outward environment and social circumstances. We argue that any use of predictive AI in forensic psychiatry should not be limited to neurobiology alone but must also consider social and environmental factors. This thesis has practical implications for the design of predictive AI systems, especially regarding the collection and processing of training data, the selection of ML methods, and the determination of their explainability requirements.

KEYWORDS

artificial intelligence, machine learning, ethics, forensic psychiatry, social determinants of health

The promises of AI-based precision psychiatry

Artificial Intelligence (AI) techniques, especially those based on machine learning (ML), are becoming integral part of procedures across medicine. Medical areas that can benefit from image classification enabled by computer vision, such as dermatology, radiology, pathology, or ophthalmology, provide ample examples for this development (1). Other domains of medicine are increasingly following suit though, and psychiatry is no exception. Here, ML-based models show a potential route to analyse complex multiscalar and multimodal data in a novel way, offering a way towards what has been called "precision psychiatry" (2).

As part of the broader move towards personalized medicine, precision psychiatry promises to enable healthcare strategies based on AI predictions and tailored more closely to individual patients. Clinical examples range from diagnostic and prognostic tools to improved opportunities for monitoring and treating psychiatric conditions (3). Identifying individual clinical phenotypes in bipolar disorders (4), predicting psychotic episodes in at-risk patients (5), managing mood disorders through using digital phenotyping (6) or selecting the most suitable psychopharmacological intervention in depression or schizophrenia (7–9) can seemingly all be improved by harnessing the computational power of ML for large-scale datasets. Ultimately,

even the very classification of psychiatric disorders may be overhauled or at least refined by drawing on results from AI-based research (10–12).

Despite these large promises, there are important ethical concerns with applying AI in psychiatry (13). As examples from other medical domains have shown, embedding AI in clinical care can jeopardize patients' safety if the AI has not been tested and validated rigorously in the correct context, resulting in potentially dangerous treatment recommendations (14). A study among US psychiatrists using different case vignettes showed that interacting with correct ML-based treatment recommendations did not improve physicians' accuracy while incorrect treatment recommendations paired with persuasive explanations even decreased physicians' accuracy of choosing a suitable psychopharmacological treatment (15). Such findings highlight the intricacies of involving AI in clinical decision-making processes and how overreliance on imperfect ML tools may adversely affect supposedly autonomous choices made by clinicians.

In line with the wider literature on AI ethics, particular attention has also been devoted to questions of fairness and bias (16). AI systems are known to be susceptible to existing social biases, which they potentially entrench and amplify. For instance, commercial gender classification systems for facial analysis have been shown to systematically perform worse on images of female and darker-skinned persons, with the worst classificatory accuracy for the intersectional group of darker-skinned females (17). In clinical contexts, addressing biases is particularly intricate due to the manifold biological, social, psychological and cultural factors influencing health and their often unclear causal interactions (18). For instance, a recent study on an AI-based decision support system in the treatment of heart failure in the US highlighted how racial biases may not be apparent in a system's evaluation: the AI correctly predicted historical real-life treatment outcomes, yet these outcomes were themselves the result of a racially biased healthcare system (19). It may therefore sometimes be necessary to carefully curate training data and restrain optimization processes to achieve a less accurate but potentially more just model (19).

In light of this background, there are justified ethical concerns with expanding psychiatric predictions on a population level, as ML models with their well-documented propensity of reinforcing existing biases from the training data may provide many false positive predictions in specific disadvantaged communities. This would create self-fulfilling prophecies and further worsen discriminatory practices (20, 21). As we will see, such dangers become even more worrisome in the context of forensic psychiatry.

Using AI for predictions in forensic psychiatry

To our knowledge no AI-based tool has as of yet entered routine use in forensic psychiatry. However, several approaches have been suggested through proof-of-concept studies. These attempts of AI-based neuroprediction, i.e., prediction of health or behavioural outcomes based on neurobiological factors utilizing AI, can be seen as part of a long-standing search for neuro-markers that supposedly render risk assessment more objective and specific. Already in the early 2000s, researchers attempted to determine the value of specific variables to predict the restorability of criminal defendants using regression analysis (22), or drew on functional neuroimaging and multi-voxel pattern analysis (MVPA) to gain insights into defendants' thoughts (23). More recently, approaches within forensic psychiatry have integrated ML to assess psychopathy or make predictions about future aggressive behaviour. Training approaches as well as training data differ largely across such studies. A recent Danish study for instance predicted criminal offenses during or after psychiatric care using sociodemographic information, psychiatric history and criminal history as training data (24). In a similar approach, a Swiss team used machine learning to explore a large set of comprehensive information including forensic patients' psychiatric and criminal history, socio-demographic and prison data, social and sexual functioning, childhood experiences to identify variables that best predict aggression in patients with schizophrenia (25). Other research avenues have focused primarily on employing ML on neural data. A meta-analysis by Deming and Koenigs for instance analysed findings from 25 original studies employing functional MRI to identify functional neural correlates of psychopathy, which in turn is related to future criminal offenses (26).

Previous research has already raised ethical concerns about risk assessments of violence in forensic psychiatry, with and without the assistance of AI. For instance, in their ethical treatment of tools assessing risk of violence with structured questionnaires, Douglas and colleagues identified overreliance on the resulting scores, mismatches between applications and contexts, risks of discrimination and stigmatization, and the premature exclusion of contentious demographic variables as main concerns (27). These concerns are mirrored in the relatively scarce ethical literature dedicated specifically to AI in forensic psychiatry. Richard Cockerill for instance has drawn on the four principles of biomedical ethics by Beauchamp & Childress (28) to map and discuss ethical challenges posed by ML-based predictions of future violent behaviour with view to non-maleficence, beneficence, respect for autonomy and justice. Adding a neurolaw perspective to the debate about AI in forensic psychiatry, Tortora and colleagues have called for more research into the risks and benefits of neuroprediction as the technology matures (29).

In this paper, we approach the ethical debates surrounding the use of AI in forensic psychiatry from a complementary angle, expanding on the challenges that arise when employing AI in this field. We argue that, in addition to the many warranted ethical worries with AI in psychiatry and forensic psychiatry in particular, the very conceptualization of psychiatric disorders poses problems that have not yet received sufficient attention. In particular, we highlight that considering the external conditions that contribute to psychiatric disorders, rather than focusing exclusively on neural data, has practical implications for designing AI systems in forensic psychiatry. To do so, we first discuss the motivation for looking for AI-based tools by highlighting the unsatisfying status of current assessment practices. We then turn to the recent literature on the conceptualization of psychiatric disorders and highlight empirical and theoretical arguments supporting an externalist stance, i.e., the position that what goes on in a (disordered) mind cannot solely be explained by reference to individual bodily and neural processes (30–33). We then spell out the implications of these insights for ongoing research on AI-based tools in forensic psychiatry and provide four practical recommendations how to move forward.

A problematic status quo

A standard strategy to evaluate AI-based systems in medicine is to benchmark them against the current state of the art in clinical practice (34). When discussing potential ethical pitfalls of predictive AI in forensic psychiatry, it is therefore important to understand the current status quo of assessment practices in forensic psychiatry and their own potential ethical shortfalls, to have a clear point of comparison (27). In addition, being aware of existing problems in forensic evaluations may also foster a better understanding why many researchers are motivated to explore AI-based solution in forensic psychiatry, in the hopes of improved tools.

Current practice in forensic psychiatry is commonly supported by structured scales which are used to evaluate defendants and support professional recommendations in court. As there are large differences in the practice of forensic psychiatry worldwide, not least owing to different legal traditions (35), forensic practice in Switzerland may serve as an example here. Here, the prevailing practice involves subjecting a single defendant to evaluation by two distinct experts concurrently. These two experts are obligated to individually conduct interviews, each lasting approximately 60 min, during which they gather the defendant's comprehensive medical history and consider information provided by other medical professionals who may have been involved in the defendant's case. This dual assessment consitutes an important step to mitigate potential interpretational biases (36). To further foster impartiality, different psychometric scales are utilized. While the use of scales is not mandatory, they support a more objective assessment of the risk and responsibility associated with the defendant's actions, thereby facilitating a clearer presentation of evidence in the courtroom (36). Given that AI-based recommendations would likely play a role similar to such scales, it is important to understand their use and limitations. A prominent example, the Hare Psychopathy Checklist-Revised (PCL-R) (37) can serve as a useful point of comparison.

The PCL-R is used to distinguish between narcissistic and antisocial traits. It contains items such as shallow affects, superficial charm, and pathological lying, which are ranked on a scale from 0 to 2. Originally, the scale was based on a single psychiatric report, reflecting the degree of resemblance between the assessed individuum and a prototypical psychopath examined by Robert Hare in 1980 (38). The scale was later revised based on a larger study, yet drew exclusively on male prisoners in Northern America (39). Despite this origin, PCL-R is one of the most frequently used scales in forensic psychiatry, both in court and research (40, 41). Since the assessment relies on the judgment of individual assessors it can suffer from interpretability bias and is prone to influence by defendants unless the assessing expert is sensitive to a potential manipulative attitude of the assessed person. Accordingly, research suggests that the scale is unreliable, offers incorrect and harmful conclusions, and is prone to misuse in legal systems (42). A training course for the use of the PCL-R aimed at strengthening the skills of the forensic psychiatrist exists (43). However, the scales are to be considered only as a tool used at the discretion of the expert. Moreover, in a hypothetical view, we might also think a forensic expert may evaluate an NGIR (Not Guilty for Insanity Reason) condition because of its susceptibility to the manipulative and narcissistic defensiveness of the evaluee.

Problems with scales used in forensic psychiatry are not limited to the PCL-R though. A study investigating the precision of two so-called actuarial risk assessment instruments (ARAIs), namely the *Violence Risk Appraisal Guide* (VRAG; 44) and the *Static-99* (45), found that both instruments, designed to predict future violent behaviour, entailed so much statistical uncertainty on the individual level "as to render [their] risk estimates virtually meaningless (46)."

Despite all these shortfalls, forensic psychiatrists have to make judgements when called upon to assess the risk and dangerousness of pathological behaviour (e.g., determining the risk of violent recidivism in persons accused of murder). While the individual risk of recidivism remains shrouded in uncertainty, a medico-legal compromise must be reached in court. Many well-known cases confirm how delicate a balance must be struck in forensic evaluation, and how large the stakes are, for individual defendants as much as for society. A notorious example from Italy is the so-called "Circeo Massacre." In 1974, a year prior to the massacre, Mr. Angelo Izzo, one of the three perpetrators, was granted semi-release by a probation court after having been arrested for raping two women. This decision was made on the basis of his perceived "good behaviour" (47). Izzo served in jail only 10 months. Shortly after his sentence was suspended, he became one of the perpetrators of kidnapping and raping two young women, one of whom died. After serving approximately 25 years for the massacre (briefly interrupted by an escape to France in 1993), Izzo was granted in 2004 semi-freedom from Campobasso Prison based again on good behaviour in order to work at a cooperative called Città Futura (Future City). Nine months later, he murdered again two women (a 49 years old woman and her 14-year-old daughter). In opposition to such attention grabbing *false negative* cases, when defendants were falsely assessed to be unlikely to reoffend, there is also the danger of misclassifying defendants as high-risk, even though they do not pose a danger to society. Unfortunately, this may happen rather frequently. A systematic review and meta-analysis of 68 independent studies, including data from 24,847 persons from 13 countries, found that while the nine most frequently used assessment tools for risk of violence, sexual, and criminal behaviour had relatively high negative predictive value (median accuracy 91%), their positive predictive values were low to moderate (median accuracy 41%) (48). The authors therefore concluded that "even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence based" (48).

Given this unsatisfactory state of affairs, it is of little surprise that forensic psychiatry has turned towards machine learning to tackle complexity and provide better and more accurate predictive tools. Yet, also this approach is fraught with challenges and has to circumnavigate particular conceptual shallows if it is to move the debate forward. One key question researchers need to consider is what type of data should be included in the training of their models.

Locating mental disorders: the challenge of externalism

There is intense ongoing debate in the philosophy of psychiatry as to the nature of psychiatric disorders and how to properly conceptualize psychiatric diagnoses. Our aim here is not to weigh in on long-standing disagreements supported by a rich academic tradition (49–51) but to point towards specific implications of recent scholarship for employing AI in forensic psychiatry.

One widely held position among biologically oriented psychiatrists looks at mental disorders primarily as brain disorders (52). The development of the research domain criteria (RDoC), spearheaded by the US NIMH, aiming for a diagnostic classification based on biological differences instead of symptoms, constitutes a prominent example for such a line of reasoning (53). This approach, which is also rather common among proponents of computational psychiatry (54), therefore locates the psychiatric problem that requires evaluation and treatment *within* the patient or defendant themselves. This position is increasingly called into question though by theories of mental disorders that one may call externalist (49, 55–58). As Roberts and colleagues summarize, such positions "hold that a comprehensive understanding of mental disorder cannot be achieved unless we attend to factors that lie outside of the head: neural explanations alone will not fully capture the complex dependencies that exist between an individual's psychiatric condition and her social, cultural, and material environment (57)."

Embracing an externalist view does not entail rejecting the idea that psychiatric disorders are brain disorders. Rather, externalist theories emphasize the importance of looking beyond the brain in order to fully understand these disorders. In this regard, they are related to philosophical accounts that analyse mental processes as situated, embodied, embedded, enacted and extended within a specific extra-cranial environment (49, 59). For our argument, two points concerning the development and sustention of psychiatric disorders are particularly pertinent.

Ample empirical evidence highlights the etiological importance of environmental factors with view to the development of psychiatric disorders (60). Biological factors that are implicated in psychopathological aetiology are frequently linked to sociodemographic inequalities, such as a history of migration, living circumstances in urban areas, childhood adversity, or cannabis use (61). Schizophrenia with its many known individual genetic factors (62) is a case in point. While the heritability of schizophrenia is estimated to fall in the range of 41–87% (63), developmental factors heavily shaped by the respective environment play a key role for gene expression and co-determine whether an inherited genetic risk leads to schizophrenia in individual patients (64). Individual biological risks therefore constitute only one important factor in a complex, multifactorial aetiology.

At the same time, the individual expression and sustention of psychiatric disorders are similarly intertwined with an individual's social environment. Arguing for an ecological view of the human brain, Fuchs calls this circular causality (55, 65): social feedback loops contribute to eliciting and sustaining dysfunctional states, such as unrequited stress reaction, which in turn again influence the social environment. Empirically, such interactions can be traced in the rich field of social neuroscience, looking at brain processes during reciprocal social interactions (66). Given that social-cognitive skills are intricately intertwined with the ability to make moral decision (67), social external factors are especially relevant to consider in forensic psychiatry: the possible presence of a responsible third party could alleviate or aggravate the sentence of a convicted person, for whom a psychiatric expertise is required to determine his criminal responsibility.

A full account of a forensically relevant mental disorder therefore needs to look closely at social influences, for "what goes on inside the head cannot be isolated from an organism's interaction with the world (58)." This becomes especially clear when considering the expanding field of neuroscience that highlights the impact of poverty and social inequalities on cortical and subcortical brain structure as well as on brain function, affecting circuits that are implicated in language, emotion processing, memory, and executive functioning (68). Given that poverty seems to already affect brain function in infants (69) and has a lasting impact on the developing brains of children and adolescents (70), even a psychiatric diagnosis based purely on neurobiology may well reflect social inequalities. Adding a potentially opaque AI techniques to this complex causal mesh risks to further reify and amplify such existing inequalities.

Consequences for potential AI applications in forensic psychiatry

An externalist view of psychiatric disorders has important implications for using AI in forensic psychiatry. If mental illnesses are indeed "inseparable from the patient's lifeworld or social environment" (55), this should impact the selection of training data, the selection of appropriate models, the interplay between trained psychiatrists and AI models, and educative needs.

First, with view to selection of training data, researchers should always include social and environmental aspects in their data and go beyond, e.g., purely brain-based predictors of violent behaviour. Such data may include information about family and friendship networks, employment, income, place of residence, housing situation, and life events. Without controlling for such factors, there is a grave risk of turning social problems into supposedly psychiatric ones. At first, this may seem counterintuitive since an exclusive focus on biological data seems less prone to human bias. However, if the social and the biological dimension of the phenomenon cannot be disentangled, excluding environmental factors would not make AI less biased but rather render models blind to important mediating factors. Instead, developers should include such environmental and social determinants of (mental) health and actively scrutinize their data and models with view to potential sources of biases (71).

Second, researchers should prefer dimensional and dynamic models over categorical and static assessments. It has rightly been argued that, as clinical utility of AI models in psychiatry increases, so does their complexity (72). Nevertheless, research should still aim for simplicity were feasible and for conceptualizing a standardized and high-quality system, to avoid creating a complex algorithm that only calculates the error of the weighting error and moves away from the goal of the research. A dynamic application of variables, in which items are intended to evolve over time, could aim at predicting the treatability of the defendant, including environmental protective factors in the risk assessment. In this sense, AI could mirror existing scales such as the SAPROF (Structured Assessment of Protective Factors for violence risk), which considers potential reintegration as a distinctive feature (73).

A third point concerns the interaction between human practitioners and predictive AI models. One of the most important goals of forensic psychiatry is not only to assess a particular diagnosis of the evaluee, but also to evaluate the pre-existing dialectic between a psychiatric diagnosis, if any, and the crimes charged. It follows from this that the forensic psychiatric evaluation is aimed at a deep psychopathology investigation, often considering details that may

TABLE 1 Overview of points to consider and their associated normative implications.

Points to consider	Normative implications
Selection of training data	Environmental aspects should be included in training data to avoid turning social problems into psychiatric ones. Excluding environmental factors may render researchers blind to important mediating factors.
Models to be preferred	Dimensional and dynamic models should be preferred over categorical and static assessments. A dynamic application of variables could aim at predicting the treatability of the defendant, including environmental protective factors in the risk assessment.
Interaction between human practitioners and AI	AI systems should not replace physicians but assist them in their practice. Explicability and contestability are important in AI systems, especially in forensic psychiatry where freedom is restricted. A personalized rehabilitation strategy should be developed that considers all the multifactorial factors involved in the defendant's unique lifeworld.
Education of all parties	Potential knowledge gaps need to be addressed through extensive education of all parties who rely on AI recommendations. Additional education remains crucial to avoid the pitfalls of genetic essentialism and to make sure that AI is employed in a beneficial manner.

be overlooked in simpler psychiatric assessments. To preserve this benefit, AI systems in medicine should therefore not replace physicians, as recent ethical guidelines have stressed again (74), but merely assist them in their practice. In addition, minimal demands of explicability and contestability, which are of general importance in medical AI, need to be respected, *especially* in a context such as forensic psychiatry where freedom is restricted. Consequently, one important goal for AI in forensic psychiatry would be the development of a personalized rehabilitation strategy that considers not only the diagnosis of the person being evaluated but also all the multifactorial factors, including cultural ones, involved in their unique lifeworld.

Finally, using AI in forensic psychiatry will require extensive education of all parties who rely on its recommendations, both from the medical as well as from the legal field. The earlier example of a complex interplay between genetics and environment in psychiatric disorders can be seen as paradigmatic here, raising similar critical issues with view to prediction (75). Ethical concerns have been raised that psychiatrists and genetic counselors may at times not fully understand the procedure and implication of psychiatric genetic testing, and require further training before using them in a beneficial manner (76). Potential knowledge gaps are even more concerning when it comes to the responsibility of predicting future criminal behaviour. As has been suggested, genetics can take on a dual role here: it can either serve to exculpate a defendant, who is subject to the unstoppable force of their genes, or it can be used to fuel essentialist intuitions that a supposedly objective test tell us something fundamental about a person's very core (77). Employing AI in psychiatry should avoid both pitfalls, for which additional education remains crucial (78). Table 1 provides an overview of these points to consider and their associated normative implications.

Conclusion

In conclusion, we argue that any potential predictive AI system in forensic psychiatry must take into account the influence of social and environmental factors on the presentation and development of psychiatric disorders. Adopting such an externalist perspective on mental disorders has critical implications for the design and implementation of AI systems in forensic psychiatry. By emphasizing the need to consider the external environment, such as social and environmental factors, in the selection of training data and machine learning models, AI systems can avoid the risk of turning social problems into psychiatric ones and better account for important mediating factors. Additionally, the use of dimensional and dynamic models, human-machine interaction, and personalized rehabilitation strategies can help to improve the precision and humaneness of forensic psychiatry practices. However, these developments should be accompanied by extensive education for all parties involved to address potential knowledge gaps and ethical concerns, especially when it comes to predicting future criminal behaviour. Overall, our paper emphasizes the importance of responsible and ethical development of AI systems in forensic psychiatry that aim for better assessment and treatment. Yet, until these points have been addressed, doing justice to the complex interaction of social, mental and biological factors, forensic psychiatrists should not rely uncritically on predictive AI techniques, to avoid unintended consequences and negative societal impact.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work has been supported by the ERA-NET NEURON project HYBRIDMIND (Swiss National Science Foundation 32NE30_199436).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Esteva A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med*. (2021) 4:1–9. doi: 10.1038/s41746-020-00376-2

2. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging*. (2018) 3:223–30. doi: 10.1016/j.bpsc.2017.11.007

3. Chen ZS, Galatzer-Levy IR, Bigio B, Nasca C, Zhang Y. Modern views of machine learning for precision psychiatry. *Patterns*. (2022) 3:100602. doi: 10.1016/j.patter.2022.100602

4. Wu M-J, Mwangi B, Bauer IE, Passos IC, Sanches M, Zunta-Soares GB, et al. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. *NeuroImage*. (2017) 145:254–64. doi: 10.1016/j.neuroimage.2016.02.016

5. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiat*. (2021) 78:195–209. doi: 10.1001/jamapsychiatry.2020.3604

6. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *NPJ Digit Med*. (2019) 2:1–11. doi: 10.1038/s41746-019-0166-1

7. Chekroud AM, 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*. (2021) 20:154–70. doi: 10.1002/wps.20882

8. Del Fabro L, Bondi E, Serio F, Maggioni E, D'Agostino A, Brambilla P. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. *Transl Psychiatry*. (2023) 13:75. doi: 10.1038/s41398-023-02371-z

9. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychol Med*. (2021) 51:2742–51. doi: 10.1017/S0033291721003871

10. Starke G, Elger BS, De Clercq E. Machine learning and its impact on psychiatric nosology: findings from a qualitative study among German and Swiss experts. *Philos Mind Sci*. (2023) 4. doi: 10.33735/phimisci.2023.9435

11. Ghosh CC, McVicar D, Davidson G, Shannon C, Armour C. What can we learn about the psychiatric diagnostic categories by analysing patients' lived experiences with machine-learning? *BMC Psychiatry*. (2022) 22:1–17. doi: 10.1186/s12888-022-03984-2

12. Chen J, Patil KR, Yeo BT, Eickhoff SB. Leveraging machine learning for gaining neurobiological and nosological insights in psychiatric research. *Biol Psychiatry*. (2022) 93:18–28. doi: 10.1016/j.biopsych.2022.07.025

13. Starke G, De Clercq E, Borgwardt S, Elger BS. Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychol Med*. (2021) 51:2515–21. doi: 10.1017/S0033291720001683

14. Ross C, Swetlitz I. (2018). IBM's Watson supercomputer recommended "unsafe and incorrect" cancer treatments, internal documents show. Available at: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.

15. Jacobs M, Pradier MF, McCoy TH Jr, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry*. (2021) 11:108. doi: 10.1038/s41398-021-01224-x

16. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. (2019) 1:389–99. doi: 10.1038/s42256-019-0088-2

17. Buolamwini J, Gebru T, Editors. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency. PMLR. 81:77–91. Available from https://proceedings.mlr.press/v81/buolamwini18a.html

18. Starke G, De Clercq E, Elger BS. Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Med Health Care Philos*. (2021) 24:341–9. doi: 10.1007/s11019-021-10008-5

19. Kostick-Quenet KM, Cohen IG, Gerke S, Lo B, Antaki J, Movahedi F, et al. Mitigating racial bias in machine learning. *J Law Med Ethics*. (2022) 50:92–100. doi: 10.1017/jme.2022.13

20. Martinez-Martin N, Dunn LB, Roberts LW. Is it ethical to use prognostic estimates from machine learning to treat psychosis? *AMA J Ethics*. (2018) 20:E804–11. doi: 10.1001/amajethics.2018.804

21. Lawrie SM, Fletcher-Watson S, Whalley HC, McIntosh AM. Predicting major mental illness: ethical and practical considerations. *BJPsych Open*. (2019) 5:e30. doi: 10.1192/bjo.2019.11

22. Mossman D. Predicting restorability of incompetent criminal defendants. *J Am Acad Psychiatry Law*. (2007) 35:34–43.

23. Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*. (2003) 19:261–70. doi: 10.1016/S1053-8119(03)00049-1

24. Trinhammer M, Merrild AH, Lotz JF, Makransky G. Predicting crime during or after psychiatric care: evaluating machine learning for risk assessment using the Danish patient registries. *J Psychiatr Res*. (2022) 152:194–200. doi: 10.1016/j.jpsychires.2022.06.009

25. Hofmann LA, Lau S, Kirchebner J. Advantages of machine learning in forensic psychiatric research—uncovering the complexities of aggressive behavior in schizophrenia. *Appl Sci*. (2022) 12:819. doi: 10.3390/app12020819

26. Deming P, Koenigs M. Functional neural correlates of psychopathy: a meta-analysis of MRI data. *Transl Psychiatry*. (2020) 10:133. doi: 10.1038/s41398-020-0816-8

27. Douglas T, Pugh J, Singh I, Savulescu J, Fazel S. Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *Eur Psychiatry*. (2017) 42:134–7. doi: 10.1016/j.eurpsy.2016.12.009

28. Beauchamp T, Childress J. Principles of biomedical ethics: marking its fortieth anniversary. *Am J Bioeth*. (2019) 19:9–12. doi: 10.1080/15265161.2019.1665402

29. Tortora L, Meynen G, Bijlsma J, Tronci E, Ferracuti S. Neuroprediction and ai in forensic psychiatry and criminal justice: a neurolaw perspective. *Front Psychol*. (2020) 11:220. doi: 10.3389/fpsyg.2020.00220

30. Rowlands M. *Externalism: putting mind and world back together again*. Montreal: McGill-Queen's Press-MQUP (2003).

31. Gallagher S. Philosophical antecedents of situated cognition In: M Aydede and P Robbins, editors. *The Cambridge handbook of situated cognition*. Cambridge: Cambridge University Press (2009). 35–53.

32. Bateson G. *Steps to an ecology of mind: collected essays in anthropology, psychiatry, evolution, and epistemology*. Chicago: University of Chicago Press (2000).

33. Noë A. *Out of our heads: why you are not your brain, and other lessons from the biology of consciousness*. New York: Hill and Wang (2009).

34. Starke G, Ienca M. Misplaced trust and distrust: how not to engage with medical artificial intelligence. *Camb Q Healthc Ethics*. (2022):1–10. doi: 10.1017/S0963180122000445

35. Beis P, Graf M, Hachtel H. Impact of legal traditions on forensic mental health treatment worldwide. *Front Psych*. (2022) 13:876619. doi: 10.3389/fpsyt.2022.876619

36. Fonjallaz J, Gasser J. (2017). *Le juge et le psychiatre: une tension nécessaire*. Chêne-Bourg: RMS éditions.

37. Hare RD. Psychopathy checklist—revised. *Psychol Assess*. (1991).

38. Hare RD. "Psychopathy," in *Research in Psychophysiology*. eds. P. Venables, M. Christie, New York: Wiley (1975):325–48.

39. Hare RD, Harpur TJ, Hakstian AR, Forth AE, Hart SD, Newman JP. The revised psychopathy checklist: reliability and factor structure. *Psychol Assess J Consult Clin Psych*. (1990) 2:338–41. doi: 10.1037/1040-3590.2.3.338

40. Vien A, Beech AR. Psychopathy: theory, measurement, and treatment. *Trauma Violence Abuse*. (2006) 7:155–74. doi: 10.1177/1524838006288929

41. Higgs T, Tully RJ, Browne KD. Psychometric properties in forensic application of the screening version of the psychopathy checklist. *Int J Offender Ther Comp Criminol*. (2018) 62:1869–87. doi: 10.1177/0306624X17719289

42. Martens WH. The problem with Robert Hare's psychopathy checklist: incorrect conclusions, high risk of misuse, and lack of reliability. *Med Law*. (2008) 27:449.

43. Hare.Org. (2023). The Hare PCL-R training program. Available at: http://www.hare.org/training/.

44. Quinsey VL, Harris GT, Rice ME, Cormier CA. Violent offenders: Appraising and managing risk. American Psychological Association (1998).

45. Hanson RK, Thornton D. Static 99: Improving actuarial risk assessments for sex offenders. Ottawa, Ontario: Solicitor General Canada (1999).

46. Hart SD, Michie C, Cooke DJ. Precision of actuarial risk assessment instruments: evaluating the "margins of error" of group v. individual predictions of violence. *Br J Psychiatry*. (2007) 190:s60–5. doi: 10.1192/bjp.190.5.s60

47. Albinati E. *The catholic school*. London: Picador (2019).

48. Fazel S, Singh JP, Doll H, Grann M. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *BMJ*. (2012) 345:345. doi: 10.1136/bmj.e4692

49. Cooper R. Where's the problem? Considering Laing and Esterson's account of schizophrenia, social models of disability, and extended mental disorder. *Theor Med Bioeth*. (2017) 38:295–305. doi: 10.1007/s11017-017-9413-0

50. Kendler KS. The nature of psychiatric disorders. *World Psychiatry*. (2016) 15:5–12. doi: 10.1002/wps.20292

51. Zachar P. *A metaphysics of psychopathology*. Cambridge, MA: MIT Press (2014).

52. Insel TR, Cuthbert BN. Brain disorders? Precisely. *Science*. (2015) 348:499–500. doi: 10.1126/science.aab2358

53. Cuthbert BN. Research domain criteria: toward future psychiatric nosologies. *Dialogues Clin Neurosci*. (2022) 7:89–97. doi: 10.31887/DCNS.2015.17.1/bcuthbert

54. Wiese W, Friston KJ. AI ethics in computational psychiatry: from the neuroscience of consciousness to the ethics of consciousness. *Behav Brain Res*. (2021) 420:113704. doi: 10.1016/j.bbr.2021.113704

55. Fuchs T. Are mental illnesses diseases of the brain In: *Critical neuroscience: a handbook of the social and cultural contexts of neuroscience*. eds. S Choudhury, J Slaby, Chichester: Blackwell (2012). 331–44.

56. Maung HH. Externalist argument against medical assistance in dying for psychiatric illness. *J Med Ethics*. (2022) 49:553–7. doi: 10.1136/jme-2022-108431

57. Roberts T, Krueger J, Glackin S. Psychiatry beyond the brain: externalism, mental health, and autistic spectrum disorder. *Philos Psychiatry Psychol*. (2019) 26:E-51–68. doi: 10.1353/ppp.2019.0030

58. Zachar P, Kendler KS. Psychiatric disorders: a conceptual taxonomy. *Am J Psychiatr*. (2007) 164:557–65. doi: 10.1176/ajp.2007.164.4.557

59. Newen A, De Bruin L, Gallagher S. *The Oxford handbook of 4E cognition*. Oxford: Oxford University Press (2018).

60. Schmitt A, Malchow B, Hasan A, Falkai P. The impact of environmental factors in severe psychiatric disorders. *Front Neurosci*. (2014) 8:19. doi: 10.3389/fnins.2014.00019

61. Robinson N, Bergen SE. Environmental risk factors for schizophrenia and bipolar disorder and their relationship to genetic risk: current knowledge and future directions. *Front Genet*. (2021) 12:686666. doi: 10.3389/fgene.2021.686666

62. Smeland OB, Frei O, Dale AM, Andreassen OA. The polygenic architecture of schizophrenia—rethinking pathogenesis and nosology. *Nat Rev Neurol*. (2020) 16:366–79. doi: 10.1038/s41582-020-0364-0

63. Chou I-J, Kuo C-F, Huang Y-S, Grainge MJ, Valdes AM, See L-C, et al. Familial aggregation and heritability of schizophrenia and co-aggregation of psychiatric illnesses in affected families. *Schizophr Bull*. (2017) 43:1070–8. doi: 10.1093/schbul/sbw159

64. Birnbaum R, Weinberger DR. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat Rev Neurosci*. (2017) 18:727–40. doi: 10.1038/nrn.2017.125

65. Fuchs T. *Ecology of the brain: the phenomenology and biology of the embodied mind*. Oxford: Oxford University Press (2017).

66. Redcay E, Schilbach L. Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat Rev Neurosci*. (2019) 20:495–505. doi: 10.1038/s41583-019-0179-4

67. Bzdok D, Schilbach L, Vogeley K, Schneider K, Laird AR, Langner R, et al. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Struct Funct*. (2012) 217:783–96. doi: 10.1007/s00429-012-0380-y

68. Noble KG, Giebler MA. The neuroscience of socioeconomic inequality. *Curr Opin Behav Sci*. (2020) 36:23–8. doi: 10.1016/j.cobeha.2020.05.007

69. Troller-Renfree SV, Costanzo MA, Duncan GJ, Magnuson K, Gennetian LA, Yoshikawa H, et al. The impact of a poverty reduction intervention on infant brain activity. *Proc Natl Acad Sci*. (2022) 119:e2115649119. doi: 10.1073/pnas.2115649119

70. Rakesh D, Whittle S. Socioeconomic status and the developing brain–a systematic review of neuroimaging findings in youth. *Neurosci Biobehav Rev*. (2021) 130:379–407. doi: 10.1016/j.neubiorev.2021.08.027

71. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. *Nat Mach Intell*. (2021) 3:659–66. doi: 10.1038/s42256-021-00373-4

72. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*. (2017) 22:37–43. doi: 10.1038/mp.2016.201

73. Burghart M, de Ruiter C, Hynes SE, Krishnan N, Levtova Y, Uyar A. The Structured Assessment of Protective Factors for violence risk (SAPROF): A meta-analysis of its predictive and incremental validity. *Psychological Assessment*. (2023) 35:56–67. doi: 10.1037/pas0001184

74. Deutscher Ethikrat. *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. Berlin: (2023).

75. Corsico P, Singh I. "The ethics of identifying and treating psychosis risk," in *Risk factors for psychosis* Paradigms, Mechanisms, and Prevention. eds. Thompson A, Broome M, London: Academic Press (2020). 335–50.

76. Appelbaum PS, Benston S. Anticipating the ethical challenges of psychiatric genetic testing. *Curr Psychiatry Rep*. (2017) 19:39. doi: 10.1007/s11920-017-0790-x

77. Tabb K, Lebowitz MS, Appelbaum PS. Behavioral genetics and attributions of moral responsibility. *Behav Genet*. (2019) 49:128–35. doi: 10.1007/s10519-018-9916-0

78. 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*. (2021) 51:2509–11. doi: 10.1017/S0033291720003906