npj | Schizophrenia

www.nature.com/npjschz

COMMENT OPEN

Check for updates

Neuroethics at the interface of machine learning and schizophrenia

Jacob McFarlane[1,2] and Judy Illes[2]✉

Ethical discourse around machine learning analysis of free speech for the detection of schizophrenia has largely focused on consent and personal privacy. We focus here on additional ethics concerns and principles that must be addressed to move the pendulum of risk over to benefit and propose solutions to achieve that shift.

npj Schizophrenia (2020)6:18 ; https://doi.org/10.1038/s41537-020-0108-6

Recent advances in biomedicine that utilize machine learning have shown promising results in identifying psychosis through automated analysis of speech and patterns of social media use[1–4]. Indeed, in a few years, such artificial intelligence (AI) methods will lead to the possibility of predicting psychosis before a human could ever reliably do so, as well as shedding light on the underlying mechanisms of the disorder[1,5,6]. Through earlier diagnosis and treatment, these advances could be instrumental in giving lives back to individuals at risk for schizophrenia[7,8]. For all modern neurotechnological innovation, however, risks invariably parallel benefits. Do the requisite ethics exist yet to optimally support the potential benefits of AI for schizophrenia?

For many in the public, AI involvement in health care is a daunting prospect[9]. Machine learning applied in some other areas has gone awry, and biased models run the risk of perpetuating and entrenching inequities in society[10]. Machine-learning algorithms are blind to which trends arise from bias, and which reflect real differences in the world[11]. Consequently, human bias and other issues with training data can lead to biased predictions when machine-learning models are utilized[12]. Models used in Google Photos to automatically identify images, for example, labeled African Americans as gorillas[13]. A model was used by the Florida justice department for predicting the likelihood that inmates up for parole would reoffend discriminated against African-American inmates[14]. Machine-learning software was developed by Amazon for hiring new employees which systematically penalized applicants for terms related to women[15].

Similar risks exist in using machine learning in the diagnosis of schizophrenia. Ethnic minorities and immigrants throughout the western world are diagnosed with schizophrenia more frequently than majority populations, but recent reviews suggest that these patterns do not reflect a real difference in prevalence of the disease[16]. Although the source of overdiagnosis in minorities is unclear, the consequences for machine-learning models are nonetheless troubling. Biased results from machine- learning models may further promulgate overdiagnosis and lead to misdirected treatment. Indeed, some prognostic models in development already use race as a predictor[17].

Given past issues with machine learning and the established bias in the diagnosis of psychotic disorders, current work in pursuit of machine learning for detecting psychosis—a kind of digital phenotyping—should be focused actively on ways to reduce bias as training data sets grow and become more representative of clinical populations. The development and testing of these models have also recently raised concerns about privacy for people who live with mental illness[18,19]. While previous work has led to great strides in creating an ethical framework for these applications of such technology[10,12,20], only limited attention has been given to how the release of models themselves or instructions on creating them might present a threat to privacy. For example, previous work has suggested that a possible future application of models that infers psychosis from prompted unstructured speech may be as part of publicly available online tools for self-assessment[21]. However, publicly accessible models will likely carry a higher risk of being misused. Unstructured speech and social media data can be easily obtained; a simple Google search grants access to twitter accounts full of possible samples. Nuanced decisions about model selection, typically thought to be technical decisions, have implications for privacy. The release of such a model or sufficiently detailed instructions for its creation threaten the privacy of not only those who choose to use it for themselves, but potentially for everyone with an online presence. These concerns must be balanced with the ethical imperatives of transparent and open research for which most frameworks for AI ethics advocate[19]. While detailed descriptions of development and testing as well as code sharing can support accountability and public trust, such practices may be harmful as these technologies begin to give fine-grained and more accurate insights about a person's mental state than ever before, and especially if it allows third parties to develop similar models. As a recent report on AI from the EU states, "In an age of ubiquitous and massive collection of data through digital communication technologies, the right to protection of personal information and the right to respect for privacy are crucially challenged"[22].

While general ethics frameworks may be sufficient to guide discussions about the utility and benefits of AI for psychotic disorders, academic, industry, and public cooperation will further advance discourse about its consequences, and the understanding, awareness, and solutions to the breadth of the associated technical and ethical trade-offs. Reducing the potential harms of machine learning in medicine, not the censure of innovation, should be a key part of the conversation[12,23]. Concerns for privacy —with perhaps a modernized definition of privacy for this digital age that foregrounds some form of bounded sharing over information protections and confidentiality—should inform decisions about the selection of predictor variables and the

[1]Cognitive Systems, Faculty of Arts, University of British Columbia, Vancouver, BC, Canada. [2]Neuroethics Canada, Division of Neurology, Department of Medicine, University of British Columbia, Vancouver, BC, Canada. ✉email: jilles@mail.ubc.ca

Published in partnership with the Schizophrenia International Research Society

npj nature partner journals

dissemination of methods. While unstructured free speech is a tool for communication that would be difficult to keep private, approaches such as federated learning may help maintain user control over their data[10]. Models need only be less biased than physicians to be an improvement over the *status quo*. Regulatory response may ultimately prove to also be a necessary element of a potential solution.

Scientists, engineers, and bioinformaticians who are pursuing machine-learning approaches for diagnosing psychosis should begin to acknowledge and engage with these ethical questions in a public setting now to preempt future issues. Ultimately, trust of patients in advances in biomedicine and the willingness to try new interventions to mitigate the burden of mental illness is key to the success of any new innovation. Dedicated conversations about the ethical issues for psychoses that so profoundly impede brain and human health will forestall harms and promote the benefits AI is intended to achieve.

Received: 24 March 2020; Accepted: 29 May 2020;
Published online: 17 July 2020

REFERENCES

1. Bedi, G. et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* **1**, 15030 (2015).
2. Corcoran, C. M. et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17**, 67–75 (2018).
3. Birnbaum, M. L. et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *npj Schizophr.* **5**, 1–9 (2019).
4. Rezaii, N., Walker, E. & Wolff, P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* **5**, 1–12 (2019).
5. Mota, N. B., Furtado, R., Maia, P. P. C., Copelli, M. & Ribeiro, S. Graph analysis of dream reports is especially informative about psychosis. *Sci. Rep.* **4**, 3691 (2014).
6. de Boer, J. N. et al. Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts. *npj Schizophr.* **6**, 1–10 (2020).
7. Haas, G. L. & Sweeney, J. A. Premorbid and onset features of first-episode schizophrenia. *Schizophr. Bull.* **18**, 373–386 (1992).
8. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* **30**, 67–76 (2008).
9. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: addressing ethical challenges. *PLoS Med.* **15**, e1002689 (2018).
10. Yuste, R. et al. Four ethical priorities for neurotechnologies and AI. *Nature* **551**, 159–163 (2017).
11. Bishop, C. M. Pattern recognition and machine learning. *CERN Document Server*. https://cds.cern.ch/record/998831 (2006).
12. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866 (2018).
13. Garcia, M. Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J.* **33**, 111–117 (2016).
14. Julia Angwin, J. L. Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
15. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (2018).
16. Schwartz, R. C. & Blankenship, D. M. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. *World J. Psychiatry* **4**, 133–140 (2014).
17. Fusar-Poli, P. et al. Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry* **74**, 493–499 (2017).
18. Insel, T. R. Digital phenotyping: technology for a new science of behavior. *J. Am. Med. Asoc.* **318**, 1215–1216 (2017).
19. Jobin, A., Ienca, M. & Vayena, E. Artificial intelligence: the global landscape of ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).
20. Martinez-Martin, N., Insel, T. R., Dagum, P., Greely, H. T. & Cho, M. K. Data mining for health: staking out the ethical territory of digital phenotyping. *Npj Digit. Med.* **1**, 1–5 (2018).
21. Ben-Zeev, D., Buck, B., Kopelovich, S. & Meller, S. A technology-assisted life of recovery from psychosis. *npj Schizophr.* **5**, 1–4 (2019).
22. European group on ethics in science and new technologies. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*. (2018).
23. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).

ACKNOWLEDGEMENTS

This work was supported in part by the Canada Research Chairs Program (J.I.).

AUTHOR CONTRIBUTIONS

J.M. generated the content for the paper, wrote the first draft, and worked closely with supervisor Dr. Illes who provided oversight on all aspects of this work through to the final product.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to J.I.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.