Schizophrenia Bulletin
https://doi.org/10.1093/schbul/sbab131

EDITORIAL

Racial and Ethnic Biases in Computational Approaches to Psychopathology

Kasia Hitczenko*,[1], Henry R. Cowan[2], Matthew Goldrick[1,2,3], and Vijay A. Mittal[2,3,4,5,6]

[1]Department of Linguistics, Northwestern University, Evanston, IL, USA; [2]Department of Psychology, Northwestern University, Evanston, IL, USA; [3]Institute for Innovations in Developmental Sciences, Northwestern University, Evanston/Chicago, IL, USA; [4]Department of Psychiatry, Northwestern University, Chicago, IL, USA; [5]Institute for Policy Research, Northwestern University, Evanston, IL, USA; [6]Medical Social Sciences, Northwestern University, Chicago, IL, USA

*To whom correspondence should be addressed; Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA; tel: 847-491-5831, fax: 847-491-3770, e-mail: kasia.hitczenko@northwestern.edu

Computational methods are a promising approach for the study, assessment, and treatment of mental illness.[1–4] Natural language processing, automatic speech recognition, and facial recognition technology have each been used to systematically and automatically identify abnormalities in speech,[5–9] language,[10–18] and facial expressivity[8,9,19–23] that characterize psychosis (see recent review[14] for additional background on these methods). Because computational methods automatically extract measures of language, behavior, and expressivity, they have the potential to save time on extensive expert training and eliminate costs on expensive apparatus often necessary for measurement in these domains.[24,25] Further, because the methods are amenable to naturalistic data sources, they can limit participant and patient burden, and can be used in contexts where the employment of cutting-edge assessment and treatment modalities have historically been severely limited or unavailable.[26] Taken together, automatic measures could serve as a resource multiplier for clinicians and researchers alike, allowing them to rigorously assess marginalized individuals who may otherwise have fallen through the cracks, helping to reduce existing disparities in mental health outcomes.[27] Indeed, half of the surveyed psychiatrists think that machine learning will significantly transform their jobs in the near future.[3,28]

However, it is not all good news. A common assumption is that computational methods avoid the harmful biases inherent in human raters—biases that can confound research findings and exacerbate structural inequity in clinical applications.[29] However, recent research suggests that in fact computational methods can reify and magnify, rather than reduce, existing health disparities.[3,30–39] Here, we review evidence that harmful racial biases may exist in computational methods already in use in studying psychosis and argue that a proactive approach to addressing these issues is urgently needed—especially as relates to racial identity.[1,3,40]

In initiating this discussion, we rely on macro-level sociodemographic groups (eg, "Black") that are currently widely used and discussed in the United States. This allows us to point out problems and discuss solutions within the current real-world research-policy-practice interface. However, this macro-level approach can also be problematic. It can reify existing biases, promote a false notion of a uniform construct, and ignore the significant heterogeneity that exists within racial and ethnic groupings both within[41,42] and across different cultures.[43] Macro-level analyses are therefore not the final word on the topic, but rather provide a foundation for more detailed analyses to engage with the full range of human experiences.

There are several potential issues with the automated methods currently used in the field. These methods have been shown to perform worse on racial minorities in other fields they have been applied in. In gender detection, the automated facial analysis systems that underlie facial emotional expressivity measures show error rates of only 0.9% on lighter-skinned men, but error rates of over 30% on darker-skinned women.[33] Similarly, automated technologies currently being used for vocal analysis make twice as many errors on speech from Black individuals than White.[44] It is highly likely that these higher error rates have

[1]It is important to note that similar biases could arise for other personal characteristics that social science has identified as bases for discrimination and bias (eg, age, education, gender, etc.). Given the clear societal links between race, ethnicity, inequality, and discrimination, we believe it is critical for us to attend specifically to these issues; that said, our conceptual arguments and methods could easily be extended to other types of bias.

been carried forward in studies detecting schizophrenia through automated vocal and facial emotion analyses.

In our own work, we have found patterns consistent with racial bias in automated coherence models. These rely on machine learning to identify thought disorder in individuals with psychosis or at clinical high-risk for psychosis.[45] Given a patient's language sample, these models automatically assign a score intended to reflect the sample's semantic cohesion. These scores have been argued to differentiate individuals with formal psychotic disorders from healthy controls with accuracies reaching 100%.[10,12,13] However, past work has broadly compared psychosis/at risk for psychosis groups to healthy controls. In our work, we examined performance across racial groups. When analyzed by racial identity, these automated methods rated narrative speech samples from Black speakers as less coherent than those of White speakers—regardless of case vs healthy control status. Supposedly objective algorithms tended to rate entirely asymptomatic Black participants as having speech patterns consistent with thought disorder.[2] This is problematic, as the field already has a bias toward misdiagnosing Black participants with schizophrenia at disproportionately high rates.[46,47] Further, as this work was conducted in a clinical high-risk context, a next step translational or precision medicine version of the study (aiming to predict conversion and inform treatment) could yield false positives—which might lead to inappropriate treatment and additional stigma. This is not merely a theoretical possibility; researchers are already highlighting the limitations of screening instruments which predict psychosis-risk more accurately for White than for Black participants.[48] Finally, these effects are unlikely to be limited to language. Automated facial emotion detection systems are more likely to rate Black individuals as expressing negative emotions—regardless of reported emotions[30–32] and acoustic analysis systems make more errors on speech by Black individuals.[44] Consequently, clinical studies using these techniques may also misdiagnose Black participants at disproportionately high rates—due to erroneous measures of expressivity and vocal productions.

How can such issues arise? State-of-the-art machine learning algorithms are tuned based on a set of training data before being deployed. Unfortunately, these methods have often been primarily trained on samples from White individuals; this can distort algorithm performance on data from other social groups.[49] This is particularly salient in the context of language, which is pervasively used to communicate and reinforce systems of racial oppression.[50] For example, a recent study of personal narratives (a prototypical language sample for automated analysis) by one of our authors found that Black participants were more likely to focus on topics of danger and adversity, while White participants were more likely to focus on personal growth.[51] Another issue in machine learning is that algorithms are often trained to mimic human annotations (eg, emotion detection systems are trained on human annotations of whether a particular individual looks happy, angry, etc.). If human annotations are biased,[52] algorithms will detect and magnify these biases in their assessments, hard-wiring the very biases we hoped to avoid.

Given these issues, how should the field proceed? A key insight from previous work is that diagnosing and eliminating these biases should directly guide development of computational methods, rather than follow it. When ethical issues are instead treated as tangential, there can be severe negative consequences. For instance, Cambridge Analytica was able to exploit personality assessment methods to influence politics in the United States and the United Kingdom.[53] Similarly, Psychopathy Checklist-Revised (PCL-R) psychopathy scores are still widely used by prosecutors and parole boards, despite concerns about their reliability and validity.[54] In this regard, psychopathology research is in a strong position: as computational methods are still relatively new, concerted efforts to address these issues now could help avoid serious problems.

These critical challenges warrant large-scale and thorough investigations of the relationship between social identities—such as race—and automated algorithm performance in the study of mental illness. Collaborative, multidisciplinary studies applying a range of algorithms to large, diverse samples could better elucidate the presence and source of biases, fueling work aimed at improving upon existing methods. Papers that apply computational methods should treat generalizability across social groups as a central, explicit evaluation criterion. Work applying automated methods should always report relationships (or lack thereof) with key sociodemographic factors (ie, racial and gender identity, education, SES, etc.) and should diagnose why any identified relationships are observed. Such analyses should be guided by findings from previous social scientific studies of social differences or biases in human judgments. For example, if an automated system including language makes different predictions based on social group membership, social group differences in language use established by sociolinguistic research can inform the analysis of the algorithm as well as interventions applied to mitigate bias. Although these steps may delay implementation of computational methods, it will be most effective for psychosis researchers to address potential issues of bias early in the design process of computational methods before they are implemented in practice. By analogy, in complex systems design, early design defects are relatively simple to correct if they are addressed early. However, the cost dramatically increases if these defects are not addressed until after large-scale implementation.[55]

[2]Further analyses revealed that this was driven by the algorithm's sensitivity to sentence length. Black individuals tended to produce samples with shorter sentences than White individuals, and shorter sentences were assigned lower coherence scores than longer sentences.

Computational methods that automatically identify abnormalities in speech, language, facial expressivity, and other aspects of human behavior could be transformative for the field. However, by overfocusing on the successes of these models, we run a very real risk of worsening existing health disparities. As a result, it is critical that we prioritize generalizability across social groups, to ensure that psychiatric computational methods do not become another domain that perpetuates existing systemic biases.

Funding

This work was supported by the National Institutes of Health (grant numbers NIH R21 MH119677, T32 NS047987) and the Canadian Institutes of Health Research (grant number CIHR DFS-152268).

Acknowledgments

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Insel TR. Digital phenotyping: technology for a new science of behavior. *JAMA.* 2017;318(13):1215–1216.
2. Corcoran CM, Cecchi GA. Computational approaches to behavior analysis in psychiatry. *Neuropsychopharmacology.* 2018;43(1):225–226.
3. Grzenda A, Kraguljac NV, McDonald WM, et al. Evaluating the machine learning literature: a primer and user's guide for psychiatrists. *Am J Psychiatry.* 2021;178(8):715–729. doi:10.1176/appi.ajp.2020.20030250.
4. Redish AD, Gordon JA. *Computational Psychiatry: New Perspectives on Mental Illness*. Cambridge, MA: MIT Press; 2016.
5. Compton MT, Lunden A, Cleary SD, et al. The aprosody of schizophrenia: computationally derived acoustic phonetic underpinnings of monotone speech. *Schizophr Res.* 2018;197:392–399.
6. Cohn JF, Kruez TS, Matthews I, et al. Detecting depression from facial actions and vocal prosody. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops; September 10–12, 2009; Amsterdam, Netherlands; 2009:1–7.
7. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. *Curr Opin Psychiatry.* 2014;27(3):203–209.
8. Cohen AS, Cowan T, Le TP, et al. Ambulatory digital phenotyping of blunted affect and alogia using objective facial and vocal analysis: proof of concept. *Schizophr Res.* 2020;220:141–146.
9. Abbas A, Sauder C, Yadav V, et al. Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study. *Front Digit Health.* 2021;3:28. doi:10.3389/fdgth.2021.610006.
10. Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr Res.* 2007;93(1–3):304–316.
11. Elvevåg B, Foltz PW, Rosenstein M, Delisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *J Neurolinguistics.* 2010;23(3):270–284.
12. Bedi G, Carrillo F, Cecchi GA, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* 2015;1(1):1–7.
13. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry.* 2018;17(1):67–75.
14. Hitczenko K, Mittal VA, Goldrick M. Understanding language abnormalities and associated clinical markers in psychosis: the promise of computational methods. *Schizophr Bull.* 2021;47(2):344–362.
15. Mota NB, Vasconcelos NA, Lemos N, et al. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One.* 2012;7(4):e34928.
16. Iter D, Yoon J, Jurafsky D. Automatic detection of incoherent speech for diagnosing Schizophrenia. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; June 5, 2018; CLPsych@NAACL-HTL, New Orleans, LA, USA. Stroudsburg, PA: Association for Computational Linguistics; 2018:136–146.
17. Just S, Haegert E, Kořánová N, et al. Coherence models in schizophrenia. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; June 6, 2019; Minneapolis, MN, USA. Stroudsburg, PA: Association for Computational Linguistics; 2019:126–136.
18. Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: a natural language processing approach. *Schizophr Res.* 2020;226:158–166.
19. Wang P, Barrett F, Martin E, et al. Automated video-based facial expression analysis of neuropsychiatric disorders. *J Neurosci Methods.* 2008;168(1):224–238.
20. Pampouchidou A, Pediaditis M, Kazantzaki E, et al. Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation. *Mach Vis Appl.* 2020;31(4):30. doi:10.1007/s00138-020-01080-7.
21. Kupper Z, Ramseyer F, Hoffmann H, Kalbermatten S, Tschacher W. Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. *Schizophr Res.* 2010;121(1-3):90–100.
22. Gupta T, Haase CM, Strauss GP, Cohen AS, Mittal VA. Alterations in facial expressivity in youth at clinical high-risk for psychosis. *J Abnorm Psychol.* 2019;128(4):341–351.
23. Cohen AS, Morrison SC, Callaway DA. Computerized facial analysis for understanding constricted/blunted affect: initial feasibility, reliability, and validity data. *Schizophr Res.* 2013;148(1–3):111–116.
24. Andreasen NC. Scale for the assessment of thought, language, and communication (TLC). *Schizophr Bull.* 1986;12(3):473–482.
25. Ekman P, Rosenberg EL, eds. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. New York, NY: Oxford University Press; 1997.
26. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology.* 2016;41(7):1691–1696.

27. Gara MA, Vega WA, Arndt S, et al. Influence of patient race and ethnicity on clinical assessment in patients with affective disorders. *Arch Gen Psychiatry.* 2012;69(6):593–600.
28. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med.* 2020;102:101753.
29. Jago AS, Laurin K. Assumptions about algorithms' capacity for discrimination. *Pers Soc Psychol Bull.* 2021. doi:10.1177/01461672211016187.
30. Xu T, White J, Kalkan S, Gunes H. Investigating bias and fairness in facial expression recognition. In: Bartoli A, Fusiello A, eds. *Computer Vision – ECCV 2020 Workshops*; Glasgow, UK. Rochester, NY: Springer International Publishing; 2020:506–523.
31. Rhue L. *Racial Influence on Automated Perceptions of Emotions*. Social Science Research Network; 2018.
32. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK. Face recognition performance: role of demographic information. *IEEE Trans Inf Forensics Security.* 2012;7(6):1789–1801. doi:10.1109/TIFS.2012.2214212.
33. Buolamwini J, Gebru T. Gender Shades: intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency 2018*; February 23–24, 2018, New York, NY; 77–91.
34. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science.* 2017;356(6334):183–186.
35. Bolukbasi T, Chang K-W, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, eds. *Advances in Neural Information Processing Systems 29*. Barcelona, Spain: Curran Associates, Inc.; 2016:4349–4357. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf. Accessed August 4, 2020.
36. Tatman R, Kasten C. Effects of talker dialect, gender and race on accuracy of Bing speech and YouTube automatic captions. In: Proceedings of the Annual Conference of the International Speech Communication Association – INTERSPEECH 2017; August 20–24, 2017; Stockholm, Sweden; 2017:934–938.
37. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–453.
38. Yuste R, Goering S, Arcas BAY, et al. Four ethical priorities for neurotechnologies and AI. *Nature.* 2017;551(7679):159–163.
39. Palaniyappan L. More than a biomarker: could language be a biosocial marker of psychosis? *npj Schizophr.* 2021;7(1):42.
40. Cockerill RG. Ethics implications of the use of artificial intelligence in violence risk assessment. *J Am Acad Psychiatry Law.* 2020;48(3):345–349.
41. King S. From African American vernacular English to African American language: rethinking the study of race and language in African Americans' speech. *Annu Rev Linguist.* 2020;6(1):285–300. doi:10.1146/annurev-linguistics-011619-030556.
42. Rosenfield PJ, Pauselli L, Jiang D, Malaspina D. Letter to the Editor regarding concept of race. *Schizophr Bull.* 2021;47(4):884–885.
43. Chen JM, de Paula Couto MCP, Sacco AM, Dunham Y. To be or not to be (black or multiracial or white): cultural variation in racial boundaries. *Soc Psychol Person Sci.* 2018;9(7):763–772. doi:10.1177/1948550617725149.
44. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci U S A.* 2020;117(14):7684–7689.
45. Hitczenko K, Cowan HR, Mittal V, Goldrick M. Automated coherence measures fail to index thought disorder in individuals at risk for psychosis. In: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access; June 11, 2021; Online. Stroudsburg, PA: Association for Computational Linguistics; 2021:129–150.
46. Baker FM, Bell CC. Issues in the psychiatric treatment of African Americans. *Psychiatr Serv.* 1999;50(3):362–368.
47. Anglin DM, Ereshefsky S, Klaunig MJ, et al. From womb to neighborhood: a racial analysis of social determinants of psychosis in the United States. *Am J Psychiatry.* 2021;178(7):599–610.
48. Millman ZB, Rakhshan Rouhakhtar PJ, DeVylder JE, et al. Evidence for differential predictive performance of the prime screen between Black and White help-seeking Youths. *Psychiatr Serv.* 2019;70(10):907–914.
49. Schulz E, Dayan P. Computational psychiatry for computers. *iScience.* 2020;23(12):101772.
50. Rosa J, Flores N. Unsettling race and language: toward a raciolinguistic perspective. *Lang Soc.* 2017;46(5):621–647. doi:10.1017/S0047404517000562.
51. Turner A, Couch N, Otto-Meyer R, et al. Narrative identity in Black and White: stories of life's high and low points told by African American and White adults. Published online under review.
52. Elfenbein HA, Ambady N. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol Bull.* 2002;128(2):203–235.
53. Hinds J, Joinson A. Human and computer personality prediction from digital footprints. *Curr Dir Psychol Sci.* 2019;28(2):204–211. doi:10.1177/0963721419827849.
54. DeMatteo D, Edens JF, Galloway M, et al. Investigating the role of the Psychopathy Checklist–Revised in United States case law. *Psychol Public Policy Law.* 2014;20(1):96–107. doi:10.1037/a0035452.
55. Tan JJY, Otto KN, Wood KL. Relative impact of early versus late design decisions in systems development. *Des Sci.* 2017;3:e12. doi:10.1017/dsj.2017.13.