Schizophrenia Bulletin
https://doi.org/10.1093/schbul/sbac008

COMMENTARY

What's That Noise? Interpreting Algorithmic Interpretation of Human Speech as a Legal and Ethical Challenge

Mathias K. Hauglid*

Faculty of Law, University of Tromsø—The Arctic University of Norway, Tromsø, Norway

*To whom correspondence should be addressed; P.O. Box 6050 Langnes, 9037 Tromsø, Norway; tel: +47-970-75-055, e-mail: mathias.k.hauglid@uit.no

Introduction

The prospect of speech analysis by means of technologies based on natural language processing (NLP) lies in the anticipated ability of algorithms to hear what humans cannot. The premise is that even experienced psychiatrists dedicating their full attention to the patient cannot be expected to pick up on all the granular signals that might be present in the patient's speech or to utilize the complex relationships between those signals. Because of the limitations inherent in human data processing capacities, potentially useful information in patient speech might just be "noise" to the psychiatrist. As such, it might not be perceived as carrying meaningful information that can be used in a clinical assessment of the patient. NLP-based models can be implemented into clinical decision support systems (NLP-CDS) and give psychiatrists "hearing aid," thus improving assessments through automated analysis of acoustic as well as semantic features of the patient's speech.

However, NLP-based speech analysis in psychiatry invokes some of the most salient legal and ethical challenges that are known from the more general discourse around artificial intelligence (AI). Automated speech recognition systems have been suggested to perform disparately across ethnic groups,[1] and machine learning (ML) algorithms are likely to reflect historical biases when they are applied to natural language.[2,3] Moreover, currently available methods for interpretation of complex ML/NLP systems appear to be inadequate as a means of detecting potentially discriminatory behavior.[4]

Can Algorithmic Inferences From Speech Be Accounted for?

In the scholarship of law and moral philosophy, concerns about the impact of AI systems on privacy and equality have been framed not only as relating to the *outputs* produced by AI systems, but also as relating to the sometimes unverifiable and inappropriate *inferences* that might be drawn through the process of algorithmic learning and reasoning.[5] Such inferences can invoke a sense of privacy violation and lead to potential discrimination, if the implication is that the system relies on factors that would normally be seen as inappropriate (ie, "protected characteristics" under nondiscrimination law, eg, racial or ethnic origin, or sexual orientation).

In light of the legal/ethical discourse, one challenge for NLP in psychiatry is to ascertain whether a system makes potentially inappropriate inferences during training. In other words, how does one translate the "noises" that only algorithms can hear? Privacy interests suggest that the existence of sensitive inferences should be made transparent, regardless of whether they lead to discriminatory outcomes.[5] One important task for further research in the field should be to explore the extent to which NLP has similar capabilities as deep learning systems for medical image analysis—a recent study suggests that standard deep learning models can be trained to predict patients' self-identified ethnicity from medical images which radiologists deem as not containing any information about ethnicity.[6] The study arguably reinforces the concern about inappropriate inferences in AI-driven radiology, and the finding may be transferrable also to NLP. The implication is that it might be possible for a neural network to use ethnicity (or other protected characteristics) as a predictive factor in contexts where physicians would not consider it or even be aware of it. Especially if training data reflect a historic tendency to over-diagnose an ethnic minority,[7] it seems possible that the algorithm might use ethnicity as a shortcut to racially biased assessments. Due to current limitations in AI interpretability, it will probably be difficult to dissect algorithmic inferences in NLP-based speech analysis models and provide a complete account of the noises they hear. NLP developers

and researchers should nonetheless strive to understand the prevalence and implications of inappropriate inferences, for example, by experimenting to see which protected characteristics learning algorithms can be trained to predict in a dataset. Such efforts could enhance transparency and lay the groundwork for the consideration of safeguards against inappropriate algorithmic inferences.

Beyond Input- and Output-Oriented Approaches

Current efforts to evaluate and mitigate undesirable biases typically employ input-oriented measures, ie, measures that restrict the information in training data or rely on other data governance measures,[8] and output-oriented measures, eg, monitoring the output distribution for biases against vulnerable groups.[9] As explained by Palaniyappan, it is established that linguistic markers that are known to have predictive value in psychiatric assessments are correlated with both social and biological features of a person.[10] However, there is limited knowledge of how the speech patterns detected by complex and potentially opaque NLP algorithms correlate with protected characteristics such as ethnicity or sexual orientation. In the present issue, Cohen et al note that there has been little evaluation of systematic biases from factors such as demographic, cultural, linguistic, and other individual differences, which are often correlated with protected characteristics.[11] The authors are optimistic about the possibility of detecting and addressing such biases. To address biases, input- and output-oriented approaches should be encouraged. However, those approaches alone can only provide minimal understanding of why the distribution of outcomes is the way it is. If algorithmic inferences are not understood as such, transparency will remain limited, and tension will endure between privacy interests and the use of AI systems. Improved understanding of the inferences drawn by complex AI systems should, therefore, be a priority in further research.

Further down the line, it might be feasible to implement inference-oriented safeguards alongside input- and output-oriented measures. At least, if an assessment is made of which protected characteristics it might be possible for an NLP model to infer, that assessment can be used to guide and optimize the use of input- and output-oriented measures. For example, if it is discovered that algorithms can predict patient ethnicity in a dataset where all information deemed by human data processors to reveal ethnicity has been removed, this could indicate that information-restriction measures should be abandoned for that dataset, while the use of output-oriented measures focusing on ethnic minorities should be intensified.

Emerging Legal Requirements

In recent years, global international organizations such as the OECD and the WHO have addressed known concerns relating to AI systems in their guidelines and policy recommendations, which lay down more or less common principles to promote "human-centric" and "trustworthy" AI.[12–14] In the EU, those principles are about to become binding legal requirements for developers and users, as the EU Commission has proposed the first comprehensive regulatory framework for AI (the AI Act). The AI Act subjects NLP-CDS systems to the requirement that their operation shall be "sufficiently transparent to enable users to interpret the system's output and use it appropriately" (Article 13).[15] The soft wording of this requirement leaves room for debate around what is sufficient and appropriate, but the proposed law does not seem to require that a comprehensive explanation must be provided of how the system "reasons" or of the logics it applies. Recital 47 to the AI Act's preamble provides that AI systems should be accompanied by documentation containing concise and clear information in relation to possible risks to fundamental rights and discrimination.[15] The WHO's guidance on AI states that AI developers should be aware of possible biases and the potential harms associated with them.[13] While it is an open question exactly what information NLP developers will need to disclose in the documentation, an account of suspected algorithmic inferences would contribute to the understanding of potential harms from biases and, consequently, possible risks to fundamental rights.

The EU's proposed AI Act further requires "human oversight" measures (Article 14) which shall enable natural persons to "fully understand the capacities and limitations" of the system. The demand for human oversight is reflected also in the WHO guidance, where it is stated that humans should remain in "full control" of medical decisions.[13] Similarly, the OECD stresses the need to ensure that AI systems have a "capacity for human determination," through the implementation of safeguards which shall be "appropriate to the context and consistent with the state of the art."[12] There is reason to expect that WHO and OECD guidance, as well as the EU AI Act, will influence future legislative processes globally. Outside of the EU, there is currently little AI-specific legislation (in the United States, a 2019 bill proposing a federal Algorithmic Accountability Act received media attention but has not moved forward).[16] In high-stakes medical decision making, the emerging legal requirements could mean that developers and users will be legally obligated to employ frameworks such as the "human-in-the-loop" methodologies which Chandler et al advocate for, in the present issue.[17] The approaches that are suggested therein appear to be particularly promising in terms of avoiding deployment of models that underperform when applied to minority groups, by combining input- and output-oriented measures. As a next step to address issues beyond those that are caused by unequal representation in training data, and to enhance understanding of the capacities and limitations of NLP, the feasibility of developing inference-oriented approaches to NLP should also be explored.

Funding

This work was funded by UiT the Arctic University of Norway's Strategic Project, "Data-Driven Health Technology" (project number 310230101)

Acknowledgments

The author declares that there are no competing interests.

References

1. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci USA*. 2020;117(14):7684–7689.
2. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017;356(6334):183–186.
3. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass*. 2021;15(8):e12432.
4. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* 2021;3(11):e745–e750.
5. Wachter S, Mittelstadt B. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum Bus L Rev.* 2019;2019:494.
6. Banerjee I, Bhimireddy AR, Burns JL, et al. Reading race: AI recognises patient's racial identity in medical images. *arXiv*, arXiv:210710356.2021, July 21, 2021, preprint: not peer reviewed.
7. Hitczenko K, Cowan HR, Goldrick M, Mittal VA. Racial and ethnic biases in computational approaches to psychopathology. *Schizophr Bull.* 2021. doi:10.1093/schbul/sbab131
8. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst*. 2012;33(1):1–33.
9. Žliobaitė I. Measuring discrimination in algorithmic decision making. *Data Min Knowl Disc*. 2017;31(4):1060–1089.
10. Palaniyappan L. More than a biomarker: could language be a biosocial marker of psychosis? *NPJ Schizophr*. 2021;7(1):42.
11. Cohen AS, Rodriguez Z, Warren K, et al. Psychometrics of NLP measures in psychosis research. *Schizophr Bull.* 2022;this issue.
12. OECD. *Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449*. OECD; 2019.
13. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*. Geneva: World Health Organization; 2021.
14. High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission; 2019.
15. European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts—COM(2021) 206 Final*. Brussels: European Commission; 2021.
16. Algorithmic Accountability Act of 2019. *S. 1108, H.R. 2231, 116th Cong*. House of Representatives; 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231/text
17. Chandler C, Foltz PW, Elvevåg B. Improving the applicability of AI for psychiatric applications through "human-in-the-loop" methodologies. *Schizophr Bull.* 2022;48(5):949–957.