Poster presentation at the 2021 Annual International Neuroethics Society Meeting.
Awarded Best Contribution in Philosophical Neuroethics.

Participatory machine learning and social justice in the development of neurological interventions

Ashley Walton, PhD

1. Machine Learning & Social Justice

Recent work in computer science has focused on addressing issues of bias in machine learning (ML) algorithms in response to mounting evidence that using algorithms to make decisions such as providing a loan or to assign healthcare resources can exacerbate and proliferate structural inequalities, further oppressing individuals from marginalized groups[1,2]. Much of this work has focused on developing measures of algorithm "fairness" which aim to capture the extent to which a model's predictions are equivalent for individuals from marginalized and non-marginalized groups[3].

Jamelle Watson-Daniels from *Data for Black Lives*, demands for these efforts to go beyond the development of "fairness" metrics. She claims that for marginalized communities to truly be empowered, control over decisions made in the development of an algorithm must be shifted to individuals that are most impacted by their (mis)application[4]. Marginalized communities must be engaged as active contributors, participating in decisions including what data is collected, what data is used for training, how models are interpreted and assessed, and how models are deployed for decision-making[2].

Martin et al. (2020) argue that this engagement should include the process of causal theory formation used to motivate the structure of an ML model. They describe the example from Obermeyer et al. (2019) where an algorithm used to allocate health care resources to patients with a high risk of illness demonstrated significant racial bias because it used previous health care costs as a proxy for health. Unequal access to care results in less health care costs for Black patients, and thus less healthcare resources were allocated to Black individuals using this algorithm despite higher rates of illness and greater need[2]. Martin et al. (2020) explain that the problematic choice of healthcare spending as a proxy for healthcare need was rooted in causal theory formation that failed to incorporate the perspectives of marginalized communities. They proposed a participatory method, Community Based System Dynamics (CBSD), that uses visual tools and simulation to include communities vulnerable to algorithmic bias as part of the process of developing causal theories to motivate the structure of ML models[5].

2. Need for participatory methods in developing adaptive neurostimulation

Deep brain stimulation (DBS) methods that alter brain activity are FDA-approved for epilepsy, Parkinson's Disease, and obsessive-compulsive disorder with several other diseases under investigation. Closed-loop DBS delivers electrical stimulation to modulate neural circuits in response to recordings of an individual's brain activity in real-time. These closed-loop systems are being developed to be "adaptive", or algorithmically learn over time how to dynamically alter stimulation in order to predict and prevent hypothesized pathological brain states for a particular individual[6,7].

There is a particular need for participatory methods in the research and development of ML algorithms for adaptive neurostimulation, with the foundational aim of social justice and the empowerment of marginalized communities. First, to address the long-standing issue of a lack of representation of individuals from marginalized communities in clinical research, including research relevant to diseases targeted by neurostimulation treatments. For example, Parkinson's Disease (PD) has no differential effect on individuals of a particular race or ethnicity, however these groups make up a very small percentage of patients participating in PD research[8]. Participatory processes are essential to developing ML algorithms that account for the true diversity of disease experiences, as well as prevent further disparities in care.

Another characteristic of these interventions that require participatory development practices is how the algorithms delivering neurostimulation "learn", or change over time. Stimulation is automatically adjusted without conscious control of the patient, which could impact patient agency: the ability to act deliberately, autonomously, and authentically[9-11]. As the algorithm learns, participatory processes must be designed to continuously engage patients in order to understand when they would like to be able to provide input regarding how the intervention is adjusted in response to hypothesized pathological states (e.g. be kept "in the loop")[12,13].

Another relevant aspect of the development of adaptive neurostimulation is the role of personalization. As neurostimulation interventions are being developed to better target disease symptoms patients may experience undesired side effects. In some cases the effects of stimulation impact a patient's personality, or other aspects of their daily lives where different patients may have different preferences in regards to the optimal trade-off between how the intervention controls their disease symptoms versus affects other aspects of themselves and their behaviors. Because of these potential trade-offs, it is anticipated that there will be a need to personalize the intervention according to individual preferences for the different effects of neurostimulation[7]. Furthermore, because there will likely not be one optimal treatment effect for all individuals– patient input will be necessary to determining what effects of neurostimulation should be considered a personal choice.

Participatory methods for developing adaptive neurostimulation should be designed to consider issues of representation in research data, the autonomous adaptation of the intervention over time, and how the neurostimulation is personalized for individual patients. Below I provide two specific examples of opportunities for patient input: deciding when data collection is experienced as surveillance or support, and determining whether the way a model includes social categories like race and gender is a form of necessary abstraction, or oppression.

3. Support versus Surveillance

Recent developments in neurointerventions incorporate "passively" collected data about patients' state in addition to self-report and clinical assessments, usually with the aim of decreasing burden and improving algorithm predictions. This might include facial expressions of patients using video, continuous monitoring of movement and physiological measures using mobile sensors, or phone usage related to social communication.

While the incorporation of this type of data can help improve model predictions and decrease the burden of patient self-report, patient input should be used to identify issues of privacy and co-articulate justifications for recording this type of data and how it will be used to improve the intervention.

Example: ML algorithm for predicting pain severity

Take for example the development of an ML algorithm for predicting pain severity in patients with Sickle Cell Disease (SCD). SCD is a genetic disorder that results in the stiffening and distortion of red blood cells into a "sickled" shape that obstructs blood flow to different parts of the body, causing persistent chronic pain as well as intermittent acute pain crises that can occur quickly without warning and cause permanent organ damage or death[14]. There is significant variation with respect to mechanisms of pain, presentation of pain, and responses to different types of treatment both between patients and within a single patient across time[15,16]. This creates difficulties for an SCD patient experiencing an acute pain crisis, where they often need high doses of opioids for adequate relief. When seeking emergency medical support, patients' self-reports of pain intensity are often met with disbelief and suspicion from healthcare providers, where they are perceived as potential substance abusers[17,18]. Communicating their pain and obtaining relief becomes a stressful interpersonal negotiation, where patients will dress up or "act out" their suffering in a more visible way in order elicit the treatment they need from medical staff[17]. Research has documented how medical stereotyping results in significant disparities in the prescription of pain medication[19,20].

Evidence suggests that SCD pain might be predicted by heart rate, movement and even weather conditions. An adaptive intervention might be developed where mobile sensors are used to continuously monitor and record these bodily signals, so that it can be incorporated into an ML model for predicting acute pain in order to provide support outside the clinical setting[21].

Consider two different types of models that could be developed:

1) A model that predicts an SCD patient's pain, *measured using patients' self-reports of pain severity*, from changes in the recorded movement and heart rate.
2) A model that predicts an SCD patient's pain, *measured by opioid consumption tracked using an electronic pill bottle*, from changes in the recorded movement and heart rate.

In the first case, the model is taking as the ground truth of pain severity to be the self-reported pain experience of the patient. In the second case, pain severity is measured based upon a deviation from expected consumption of opioids prescribed by a physician.

The latter case might seem to be preferable because it decreases the burden that might incur from the patient self-reporting their pain experience. However, important questions to ask include:

- *To what extent is the recording of opioid consumption by an electronic pill bottle experienced as "surveillance" by the SCD patients?*
- *How does the use of opioid consumption versus reported pain severity change the types of predictions a model makes regarding what support a patient might need?*

In the case of SCD pain, researchers argue that because each SCD patient is a "unique human entity", interventions that support pain management should promote empowerment and recognize patients as the authority on their pain experiences[15]. Participatory methods for developing ML algorithms for neurostimulation should systematically include patient input regarding their experience of data collection, as well as tools for communicating and eliciting input for how data is used to improve the intervention.

4. Abstraction versus Oppression

Timothy Brown (2021) calls for engagement and participation of marginalized individuals within neuroscience research, where they are active researchers in the process of collecting, storing, analyzing neurological data. He advocates for them to have the authority to "interrogate what categories they [researchers] accept, propose or reject – to see if they exacerbate/create inequities"[22]. Brown introduces Robin Dembroff's term *ontological oppression* to describe the possible consequences of failing to engage marginalized communities in neuroscience research. Ontological oppression is the result of structures and practices within social contexts that either fail to recognize or construct social categories. Unwanted placement of an individual within a social category can unjustly constrain their behaviors, concepts or affect[23].

There are several points in the development of an ML algorithm where categories are defined that serve as quantitative abstractions of patient attributes and their disease experiences. This includes the process of causal theory formation, defining input variables and how they are measured, and defining the target outcome the model predicts. Furthermore, these categories may be continuously redefined as the algorithm is trained, evaluated, and deployed over time. In the case of adaptive neurostimulation, an ML algorithm is often trained using data from a small sample of patients, or as part of an n-of-1 study. Because there is significant heterogeneity with respect to individual neuroanatomy, neurophysiological instantiation of disease, and symptom experience, the algorithm will need to be developed over time to maintain predictive performance as it is exposed to data from new patients.

There are also many different ways an algorithm might be changed over time as it is exposed to new data. For example, features for predicting the target outcome may be added or subtracted, or changes might be made to the model architecture, sampling methods and how it's optimized (i.e. the objective function used to adjust weights as the model learns from new data). Therefore there is a need to develop procedures for *model governance*, or approaches to "auditing" and changing ML algorithms for neurostimulation as they are deployed over time, to ensure that they continue to embody ethically desirable values[24]. Particularly when personalizing neurostimulation, the categories used to change an ML algorithm in response to variability related to patients from marginalized identities need to be thoughtfully "interrogated"[22] through the participation of individuals from these communities.

Example: Genetics differences and Race

Take for example scientific studies that quantify genetic differences with the aim of characterizing aspects of an individual's race. Dorothy Roberts in *Fatal Invention: How Science, Politics, and Big Business Re-create Race in the Twenty-First Century*, investigates the way statistics is applied to describe race as a genetic category, providing a detailed account of how each step of data analysis from defining the data sample to deciding how the results apply to our every day lives– is dependent on and driven by preconceived notions of race. She points out: "Science is the most effective tool for giving claims about human difference the stamp of legitimacy"[20]. She interviews epidemiologists that explain "…differences between racial groups are usually too small to warrant using this variable as a predictive tool or as a factor in clinical decision making. The practice risks 'stereotyping and the tendency to misapply quantitative differences between groups as though they were categorical differences'"[20]. For example, Roberts gives an account of research focused on finding a genetic characteristic of black and Puerto Rican children that account for their higher rates and severity of asthma. She also describes research studies focused on environmental allergens that trigger asthma, where elements in dust particles collected from inner-city homes were found to cause asthmatic symptoms in mice. She points out that while it is widely accepted that genetic and environmental contributions to health cannot be separated, genes are frequently described as the "cause" of disease while environmental contributions are merely "triggers". She asks:

- *Do black children have more severe asthma because they are genetically susceptible to triggers?*
- *Or could it be because they are more likely to live in neighborhoods where these triggers are concentrated?*

It is necessary to be mindful of the use of categories related to marginalized identities, particularly how their use can imply causal relationships that neglect the role of environmental factors and the impact of structural inequalities that lead to quantitative differences between groups. Participatory methods must engage marginalized individuals to determine when an ML algorithm is utilizing a necessary and acceptable abstraction to account for variability associated with social categories such as race and gender, or if its use results in a form of ontological oppression.

5. Goals of participatory ML for adaptive neurointerventions

As patients are engaged in processes that contribute towards the development of adaptive neurostimulation, three key goals of this work should be:

1. The development of long-term partnerships that recognize participation as labor, and compensating individuals and communities for their contributions[25].
2. Ongoing efforts to develop communication tools for sharing "technical" knowledge in order to elicit meaningful input from patients.

3. The integration of neuroscience research with concerns of population health, where understanding the impact of structural inequalities is essential to and inseparable from the successful generation of scientific knowledge that drives neurotechnological progress.

Works Cited

1. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. *Fairness through awareness*. In ITCS, pages 214–226, 2012.
2. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. Science, 366(6464), 447-453.
3. Kasy, M., & Abebe, R. (2021, March). *Fairness, equality, and power in algorithmic decision-making*. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 576-586).
4. Watson-Daniels, J. (2020, July). *Beyond Fairness and Ethics: Towards Agency and Shifting Power*. Paper presented at the meeting of International Conference on Machine Learning. https://slideslive.com/38930952/beyond-fairness-and-ethics-towards-agency-and-shifting-power
5. Martin Jr, D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Isaac, W. S. (2020). *Participatory problem formulation for fairer machine learning through community based system dynamics*. arXiv preprint arXiv:2005.07572.
6. Arlotti, M., Rosa, M., Marceglia, S., Barbieri, S., & Priori, A. (2016). *The adaptive deep brain stimulation challenge*. Parkinsonism & related disorders, 28, 12-17.
7. Klein, E. (2015). *Models of the patient-machine-clinician relationship in closed-loop machine neuromodulation*. In Machine medical ethics (pp. 273-290). Springer, Cham.
8. Siddiqi, B., & Koemeter-Cox, A. (2021). *A Call to Action: Promoting Diversity, Equity, and Inclusion in Parkinson's Research and Care*. Journal of Parkinson's Disease, (Preprint), 1-4.
9. Roskies, A. L. (2015). *Agency and intervention*. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1677), 20140215.
10. Goering, S., Klein, E., Dougherty, D. D., & Widge, A. S. (2017). *Staying in the loop: Relational agency and identity in next-generation DBS for psychiatry*. AJOB Neuroscience, 8(2), 59-70.
11. Zuk, P., & Lázaro-Muñoz, G. (2019). *DBS and autonomy: Clarifying the role of theoretical neuroethics*. Neuroethics, 1-11.
12. Gilbert, F., O'brien, T., & Cook, M. (2018). *The effects of closed-loop brain implants on autonomy and deliberation: what are the risks of being kept in the loop?* Cambridge Quarterly of Healthcare Ethics, 27(2), 316-325.
13. Klein, E., Goering, S., Gagne, J., Shea, C. V., Franklin, R., Zorowitz, S., ... & Widge, A. S. (2016). *Brain-computer interface-based control of closed-loop brain stimulation: attitudes and ethical considerations*. Brain-Computer Interfaces, 3(3), 140-148.
14. Ballas, S. K. (2015). *Pathophysiology and principles of management of the many faces of the acute vaso-occlusive crisis in patients with sickle cell disease*. European journal of haematology, 95(2), 113-123.
15. Ballas, S. K. (2011). *Update on pain management in sickle cell disease*. Hemoglobin, 35(5-6), 520-529.
16. Dampier, C., Palermo, T. M., Darbari, D. S., Hassell, K., Smith, W., & Zempsky, W. (2017). *AAPT Diagnostic Criteria for Chronic Sickle Cell Disease Pain*. The Journal of Pain.
17. Ciribassi, R. M., & Patil, C. L. (2016). *"We don't wear it on our sleeve": Sickle cell disease and the (in) visible body in parts*. Social Science & Medicine, 148, 131-138.
18. Jenerette, C. M., & Brewer, C. (2010). *Health-related stigma in young adults with sickle cell disease*. Journal of the National Medical Association, 102(11), 1050.
19. Todd, K. H., Samaroo, N., & Hoffman, J. R. (1993). *Ethnicity as a risk factor for inadequate emergency department analgesia*. Jama, 269(12), 1537-1539.

20. Roberts, D. (2011). *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. The New Press.
21. Walton, A., Crosby, L., Kiefer, A., Chemero, A., Murphy, S., Richardson, M. (2018). *Multi-scaled assessment for predicting pain experience in adolescents with Sickle Cell Disease*. (doctoral dissertation). University of Cincinnati, Cincinnati, OH.
22. Brown, T. (2021, June). Symposium: *Achieving Diversity, Equity and Inclusion in Neuroscience Research in Under-Served, Under-Resourced and Remote Settings*. Paper presented at the National Institutes of Health BRIAN Investigators Annual Meeting, Virtual.
23. Dembroff, R. (2018). *Real talk on the metaphysics of gender*. philosophical topics, 46(2), 21-50.
24. Papernot, N. (2020, July). *What Does it Mean for ML to be Trustworthy?* Paper presented at the meeting of International Conference on Machine Learning. https://www.youtube.com/watch?v=UpGgIqLhaqo&ab_channel=NicolasPapernot
25. Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). *Participation is not a design fix for machine learning*. arXiv preprint arXiv:2007.02423.