VIEWPOINT

The Two Cultures of Computational Psychiatry

Daniel Bennett, PhD
Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey.

Steven M. Silverstein, PhD
Division of Schizophrenia Research, University Behavioral Health Care, Rutgers University, New Brunswick, New Jersey; and Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers University, New Brunswick, New Jersey.

Yael Niv, PhD
Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey; and Department of Psychology, Princeton University, Princeton, New Jersey.

Corresponding Author: Daniel Bennett, PhD, Princeton Neuroscience Institute, Princeton University, PNI 140, Princeton, NJ 08544 (daniel.bennett@princeton.edu).

Computational psychiatry is a rapidly growing field that uses tools from cognitive science, computational neuroscience, and machine learning to address difficult psychiatric questions. Its great promise is that these tools will improve psychiatric diagnosis and treatment while also helping to explain the causes of psychiatric illness.[1-3]

Within computational psychiatry, there are distinct research cultures with distinct computational tools and research goals: machine learning and explanatory modeling.[1] While each can potentially advance psychiatric research, important distinctions between the cultures sometimes go unappreciated in the broader psychiatric research community. We detail these distinctions, referring to Breiman's influential dichotomy between these cultures of statistical modeling[4] to identify limitations on the inferences that each culture can draw.

Breiman[4] defined the 2 cultures of statistical modeling in terms of a data-generating process that generates output data from input variables. His dichotomy distinguished "algorithmic modeling,"[4(p200)] which aims to predict what outputs a data-generating process will produce from a given set of inputs while treating the process itself as a black box,[2,3] from "data modeling,"[4(p199)] which uses the pattern of outputs and inputs to explain how the data-generating process works. In psychiatry, the data-generating processes are the psychological and neurobiological mechanisms that produce psychiatric illnesses. The output data produced by these processes are psychiatric outcomes (eg, symptoms, medication response) with input variables including family history, precipitating life events, and others. Breiman's distinction between prediction and explanation is also what separates machine-learning approaches to computational psychiatry, which aim to predict psychiatric outcomes, from explanatory modeling, which aims to explain the computational-biological mechanisms of psychiatric illnesses. While these approaches have also been termed *data-driven* and *theory-driven*,[1] we emphasize that the dual cultures of computational psychiatry share an overlapping set of statistical tools and practical methods but differ in whether the end goal is explanation or prediction. A deep neural network, for instance, can be either explanatory (as a biophysically realistic model of psychiatric dysfunction), or predictive (as a classifier used to predict a diagnosis), depending on context.

The culture of machine learning typically uses statistical techniques, such as support vector machines or deep neural networks, to predict psychiatric outcomes. These tools can be seen as lying on a continuum with classical statistics such as regression but with the addition of practices designed to reduce overfitting, such as parameter regularization and cross validation. For instance, a study by Webb et al[5] has used such tools to predict antidepressant response from a combination of variables, including demographic factors, symptom severity, and cognitive task performance. Despite good predictive performance, the study drew no conclusions about the mechanisms by which these variables were linked to antidepressant response. This is because in machine learning, the parameters of the models that are used to predict psychiatric outcomes are not assumed to correspond to any underlying psychological or neural process; consequently, these parameters cannot be interpreted mechanistically.

In comparison, the culture of explanatory modeling focuses on statistical models (expressed as equations) that define interacting processes with parameters that putatively correspond to neural computations. For instance, equations describing value updating in reinforcement-learning models are thought to correspond to corticostriatal synaptic modifications modulated by dopaminergic signaling of reward prediction errors. Consequently, explanatory model parameters fit to behavioral and/or neural data from patients with psychiatric diagnoses can directly inform inferences about dysfunctions in underlying neural computations, subject to several conditions being met. For instance, Huys et al[6] have shown that anhedonia is correlated across diagnoses with a model parameter corresponding to the blunting of experienced reward value but not with a parameter controlling the rate of learning from this experienced value, providing evidence against one dopaminergic explanation of depression.

Importantly, there are several conditions that must be met before an explanatory model can be used in this way. First, to support the model's correspondence to the true data-generating process and distinguish between different candidate models, the models must make sufficiently different predictions for the experimental data. Separately, to identify the model parameters accurately, the parameters' effects on model predictions should be relatively independent, and there must be sufficient data. One approach to testing these conditions is to simulate data from each candidate model and test the ability of a model-fitting routine to recover the true cognitive model and its parameters from these data. Because empirical data will not correspond as perfectly to any of the candidate models, this test is a necessary but not sufficient condition for reliable explanatory modeling. Indeed, a common error is to overinterpret results, forgetting that the best-fitting model is only better than models with which it was compared and parameter values are only estimates reliable to a level of statistical error.

A potential limitation of explanatory modeling in computational psychiatry is that theories (ie, models) may be ill-matched to available data, because data collected for other purposes may not distinguish between subtly (but importantly) different hypotheses regard-

ing the mechanisms underlying psychiatric dysfunction. It is therefore crucial that explanatory modeling studies be carefully designed to ensure they provide data that can be used to accurately identify model parameters and discriminate between models. Another pitfall is the ubiquity of generalized performance deficits in individuals with mental illness, owing to factors such as low motivation, poor understanding of task instructions, and medication-induced sedation. Computational modeling can address this directly by developing specific predictions regarding what data would look like if generalized deficits are present and then determining whether other models better account for the data.

Although these limitations mean that explanatory modeling can be a challenging enterprise, its potential benefits are also great. One exciting possibility is that parameters from explanatory models can be used as computational markers of psychiatric illness.[1] Using such markers, it may be possible to (1) distinguish diagnoses that might initially have similar symptom profiles, such as major depression and bipolar disorder; (2) characterize within-diagnosis heterogeneity (and potentially generate new diagnostic categories) with reference to the disordered computational mechanism; or even (3) predict relapse and/or treatment responses based on shifts in computational markers. Explanatory models may also help associate psychiatric dysfunction with failures of canonical neural computations (eg, predictive coding, divisive normalization, contextual modulation), and therefore lead to a greater appreciation of shared mechanisms across cognitive impairments, symptom domains and disorders, consistent with the National Institutes of Mental Health Research Domain Criteria initiative.

Conclusions

Applying Breiman's dichotomy[4] between the cultures of statistical modeling to computational psychiatry helps to parse the promises of this growing field. Crucially, it suggests that the 2 cultures of computational psychiatry are fundamentally suited for drawing different kinds of inferences from psychiatric data. This marks a point of difference between our dichotomy and Breiman's dichotomy.[4] Whereas Breiman espoused the virtues of prediction over explanation, we wish to emphasize the value of both cultures in asking distinct research questions and the importance of ongoing crosstalk between cultures. Although we have treated these cultures as separate, hybrid approaches[1,2] have already proven powerful: generative embedding approaches incorporate parameter estimates from explanatory models as variables in machine-learning algorithms,[7] and clusters of symptoms identified using machine-learning approaches can prompt explanatory modeling to determine the mechanisms underlying each specific cluster.[2] Indeed, as long as we remain far from understanding the provenance of mental illness, it behooves us to use all appropriate methods to their full extent.

ARTICLE INFORMATION

Published Online: April 24, 2019. doi:10.1001/jamapsychiatry.2019.0231

Conflict of Interest Disclosures: Dr Niv reports grants from the Army Research Office during the conduct of the study. No other disclosures were reported.

Funding/Support: This work was funded by grant R01DA042065 from the National Institute on Drug Abuse (Drs Niv and Bennett), and grant R61MH115119 from the National Institute of Mental Health (Dr Silverstein).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the article; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci*. 2016;19(3):404-413. doi:10.1038/nn.4238

2. Gillan CM, Whelan R. What big data can do for treatment in psychiatry. *Curr Opin Behav Sci*. 2017;18:34-42. doi:10.1016/j.cobeha.2017.07.003

3. Paulus MP. Pragmatism instead of mechanism: a call for impactful biological psychiatry. *JAMA Psychiatry*. 2015;72(7):631-632. doi:10.1001/jamapsychiatry.2015.0497

4. Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001;16(3):199-231. doi:10.1214/ss/1009213726

5. Webb CA, Trivedi MH, Cohen ZD, et al. Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychol Med*. 2018:1-10. doi:10.1017/S0033291718001708

6. Huys QJ, Pizzagalli DA, Bogdan R, Dayan P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biol Mood Anxiety Disord*. 2013;3(1):12. doi:10.1186/2045-5380-3-12

7. Brodersen KH, Deserno L, Schlagenhauf F, et al. Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin*. 2013;4:98-111. doi:10.1016/j.nicl.2013.11.002