Commentary

Biological Psychiatry: CNNI

Computational Psychiatry: From Mechanistic Insights to the Development of New Treatments

Quentin J.M. Huys, Tiago V. Maia, and Martin P. Paulus

Computational psychiatry is a young field that uses computational approaches to advance our understanding of mental health and to develop practical applications to improve treatment outcomes for patients (1). The use of computational tools is motivated by the recognition that mental health is hugely complex and requires sophisticated tools. A growing number of researchers from a wide range of disciplines related to psychiatry, including machine learning, computational neuroscience, neuroimaging, cognitive psychology, and others, are attracted by the challenge of applying sophisticated mathematical tools and the relevance of using these tools to improve patients' suffering.

However, computational psychiatry is not the first approach to raise hopes about improvements to mental health. Despite notable progress, it is difficult to ignore the frustrating inability of neuroscience to redefine disease and treatment categories, the slowing down of advances in psychopharmacology, and the glaring absence of genetic and neuroimaging methods from standard clinical practice after decades of intensive and expensive research. To earn its use of the term "psychiatry," the field will need to focus on improving patients' well-being: to understand and to treat are not always the same. The articles in this issue illustrate the breadth of computational psychiatry along the range from understanding to treatment.

In terms of promoting understanding, it is useful to remember Marr's (2) distinction between levels of description: what problem a system tries to solve; the algorithm it employs; and how it is implemented. Computational models can help us understand how these levels are linked (e.g., how modifying aspects of the circuit [the implementation] might affect task performance [the purpose]). Two classes of models, those of Bayesian inference and of reinforcement learning (RL), have been particularly prominent.

Bayesian theories view perception as an inferential process that combines sensory evidence with pre-existing prior expectations (3) that influence early sensory processing through top-down projections. Powers *et al.* (4) argue that hallucinations in schizophrenia might be due to overly strong top-down projections such that percepts are created even in the absence of stimuli. This could explain why contents of hallucinations often relate to the patient's inner life. On the other hand, the idea that schizophrenia involves overly strong top-down influences goes against evidence that it involves impairments in the formation or use of expectations (5) and in the prefrontal areas (6) that provide the highest level of top-down biases (7,8). A possible reconciliation of these seemingly contradictory ideas might be that patients with schizophrenia have difficulties forming appropriate expectations, even if they form inappropriate expectations. Indeed, patients with schizophrenia exhibit both reduced adaptive learning and increased aberrant learning, and the coexistence of these two disturbances can be explained by the specifics of the dopaminergic disturbances in schizophrenia (9). Notably, it is the increased aberrant learning—which could give rise to aberrant expectations—that relates to positive symptoms (10,11).

Harlé *et al.* (12) also use Bayesian approaches, but at a far more granular level, and applied to substance use. They examine the neural processes of trial-by-trial variation in impulsivity on the stop-signal task (13) in methamphetamine use disorder. Impulsivity seems related to a reduced ability to detect deviations from predictions. The regions showing this difference, the caudate and orbitofrontal cortices, are involved in goal-directed decision making. It raises the possibility that the inability to pursue longer-term goals in methamphetamine use disorder may involve an inability to adapt the relevant predictions.

Next, three articles examine RL approaches to psychiatric disorders. RL is orthogonal and complementary to Bayesian approaches and is concerned with inferring behavioral policies that maximize long-term rewards. Dopaminergic neurons report one of its key computational signals, the temporal prediction error (PE), in great detail (14,15), meaning that RL provides a particularly tight link between all three of Marr's levels. Such a link has proven particularly fruitful to shed light on the mechanistic substrates of psychiatric disorders with dopaminergic involvement, including addiction and schizophrenia (16).

Intuitively, RL provides a particularly compelling account of addiction: drugs of abuse affect dopamine, and clearly undermine adaptive choice. The link was made early (17) and has survived even though some specific predictions were not borne out (18). However, to what extent the very specific predictions really are borne out in particular disorders in humans is not clear. Huys *et al.* (19) review the neuroimaging evidence for PEs in alcohol addiction, finding it weak. This has important consequences for both translational and clinical studies. It points toward a more complex process that may involve homeostatic adjustments difficult to model in animal studies and questions the notion that alcohol use disorder (and possibly substance use disorders more generally) are learning disorders due to abnormal RL processes.

Dowd *et al.* (20) investigated RL in schizophrenia and its relation to anhedonia and avolition. They replicated prior findings that patients are impaired at learning from positive but not negative feedback and that anhedonia/avolition correlates with reduced striatal activation to positive feedback (9,21–23). Contrary to previous studies, they failed to find group differences in PE signals in the ventral striatum, but

  http://dx.doi.org/10.1016/j.bpsc.2016.08.001
Biological Psychiatry: Cognitive Neuroscience and Neuroimaging September 2016; 1:382–385 www.sobp.org/BPCNNI ISSN: 2451-9022

interpreting null findings is always challenging; for robust conclusions, these results will have to be combined meta-analytically with those of similar studies (23). Their key finding, however, was that early in learning, patients engaged cognitive control regions less than controls did, which highlights the importance of considering executive function deficits when interpreting RL findings in schizophrenia (24,25). Such more general disturbances, however, do not explain the specificity of patients' deficits in learning from positive, but not negative, feedback, so they do not rule out a specific RL deficit (9).

Culbreth *et al.* (26) also investigated PE signals in two samples of medicated patients with schizophrenia, using a probabilistic reversal-learning task. As expected under the idea that negative symptoms relate to reduced phasic firing of dopamine neurons for relevant cues and outcomes (9), they replicated previous correlations between blunted PE signaling in the ventral striatum and negative symptoms (27). They did not, however, find an overall difference between patients and healthy controls, which is surprising given that several prior studies have reported such differences in both medicated and unmedicated patients, and even a meta-analysis has tentatively confirmed that such differences exist (23,24,27). Interpreting null results is fraught with difficulties, especially in complex multistep analyses such as those in model-based functional magnetic resonance imaging; nonetheless, the study by Culbreth *et al.* (26) is the largest investigating these issues so far and had 83% power to detect a medium effect size (0.5); a Bayesian analysis suggested that the null hypothesis was moderately more likely than a medium effect. They describe several possible explanations for the discrepancy between their findings and prior ones, including an explanation that seems particularly compelling: they did not analyze separately positive and negative PEs, and in reversal-learning tasks negative PEs are typically large and salient because they occur at block transitions. Their findings might therefore reflect the importance of negative PEs in their task. Indeed, as noted previously, patients with schizophrenia are not impaired at learning from negative feedback, and they have blunted striatal signaling of positive, but not negative, PEs (27,28). This interpretation dovetails nicely with ideas tying disrupted phasic bursting of dopamine neurons to schizophrenia (9,29), as such disturbances would affect positive, but not negative, PEs.

These RL studies, when considered together with the prior literature, demonstrate that the methods of computational psychiatry are sufficiently well developed, robust, and sensitive to uncover mechanistic insights that are commonly replicated across studies, while at the same time highlighting that other findings remain difficult to pin down conclusively. As in other areas of neuroscience, part of the problem is the use of excessively small samples, aggravated by the heterogeneity of methods used—both of which call for efforts for greater collaboration and standardization of methods in the field.

The second part of the issue illustrates how computational psychiatry may improve treatment outcomes. First, there is now an extensive literature on the use of electroencephalography to predict pharmacological treatment outcome in depression. Wade and Iosifescu's (30) detailed review highlights that due to a lack of replication it is still unclear which measures will be clinically useful. However, they also discuss the only existing phase II study (31) of computational techniques applied to practical clinical problems. They found that an algorithm that only has access to resting electroencephalography selected antidepressant treatments better than did clinicians following a state-of-the-art protocol. Machine learning may hence extract features from neuroimaging, behavioral, or other data that allow clinically relevant improvements in selecting medication for individuals. Jollans and Whelan (32) review advances in applying machine-learning techniques to these issues across several psychiatric disorders. Just as Wade and Iosifescu (30) did, they find that studies vary widely in sample size, selection of features, statistical approaches, and tools to avoid overfitting or improve generalization (33). Marquand *et al.* (34) finally take a step back and argue that clustering efforts have yielded, overall, highly variable results and have failed to yield stable subtypes, particularly when considering different data or algorithms. Their key suggestion is an alternative way forward: a type of normative modeling akin to growth curves, whereby pathological outliers are identified with respect to the distribution of values in both control and patient groups, rather than with respect to group distinctions that are difficult a priori. This addresses several key shortcomings of current approaches, particularly the need to assume homogeneity among clinical (sub)groups. As in *Anna Karenina*, happy families might be alike, but every unhappy family might be unhappy in its own way.

The final contribution is ours (35). In it, we build on the insights of numerous prior investigations and attempt to chart a way forward for computational psychiatry. Our central suggestion is the adoption of a common pipeline inspired by the drug-development pipeline. We identify five phases in the development of computational tools for clinical purposes: 1) preclinical development, where potential tools are identified; 2) phase I, examining the robustness of the approach; 3) demonstration of clinical efficacy in phase II; 4) providing multisite evidence of clinical utility in phase III; and 5) extending the application to new target populations in phase IV. Key to realizing this pipeline is the establishment of an international community, which collaborates to rapidly develop promising tools and apply them to different psychiatric diseases. This in turn involves the sharing of tasks, analysis tools, computational resources, and trial protocols, all of which are becoming increasingly feasible.

Computational psychiatry is an emerging field encompassing a breadth of approaches. To bring these together and improve outcomes for patients, a number of challenges have to be met. First, we need a radically open data and model environment (i.e., an infrastructure and a community that are willing to share data and computational approaches among investigators). Open interaction and analysis of data should accelerate progress beyond what is possible within an individual laboratory. It will also further the expertise and understanding of academic psychiatrists who are not familiar with computational approaches. Second, we need to work with multilevel dimensional psychopathological approaches to redefine disease models. For example, computational treatment of the dynamical conceptualization of mental states may

help to develop prediction models that can be used to forecast the risk of an individual to experience a depressive or manic episode. Third, a mechanistic understanding of mental health and disease is the ultimate goal, but subject-specific prediction of future mental health or disease states may be the low-hanging fruit that will enable us to show the utility of computational psychiatry. A structured framework as proposed in Paulus *et al.* (35) may help to accomplish this.

Acknowledgments and Disclosures

This work was supported by grants from the National Institute of Mental Health (Grant No. R01 MH101453 to MPP), National Institute on Drug Abuse (Grant No. U01 DA041089 to MPP), and the Swiss National Science Foundation (Grant No. SNSF 320030L_153449 to QJMH), as well as funding from the Tourette Association of America and a Breakthrough Idea Grant from Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa (to TVM).

All authors report no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Psychiatry, Psychotherapy and Psychosomatics (QJMH), Hospital of Psychiatry, University of Zurich; and Translational Neuromodeling Unit (QJMH), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland; Institute for Molecular Medicine (TVM), School of Medicine, University of Lisbon, Lisbon, Portugal; Laureate Institute for Brain Research (MPP), Tulsa, Oklahoma; and the Department of Psychiatry (MPP), University of California, San Diego, San Diego, California.

Address correspondence to Martin P. Paulus M.D., Laureate Institute for Brain Research, 6655 South Yale Avenue, Tulsa, OK 74136-3326; E-mail: mpaulus@laureateinstitute.org.

Received Jul 29, 2016; accepted Aug 1, 2016.

References

1. Huys QJ, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci 19: 404–413.
2. Marr D (1982): Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Cambridge, MA: MIT Press.
3. Weiss Y, Simoncelli EP, Adelson EH (2002): Motion illusions as optimal percepts. Nat Neurosci 5:598–604.
4. Powers AR, Kelley M, Corlett PR (2016): Hallucinations as top-down effects on perception. Biol Psychiatry Cogn Neurosci Neuroimaging 1: 393–400.
5. Adams RA, Huys QJ, Roiser JP (2016): Computational psychiatry: Towards a mathematically informed understanding of mental illness. J Neurol Neurosurg Psychiatry 87:53–63.
6. Hill K, *et al*. (2004): Hypofrontality in schizophrenia: A meta-analysis of functional imaging studies. Acta Psychiatr Scand 110:243–256.
7. Maia TV, Cleeremans A (2005): Consciousness: Converging insights from connectionist modeling and neuroscience. Trends Cogn Sci 9: 397–404.
8. Miller EK, Cohen JD (2001): An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.
9. Maia TV, Frank MJ (2016): An integrative perspective on the role of dopamine in schizophrenia [published online ahead of print June 1]. Biol Psychiatry.
10. Roiser JP, Howes OD, Chaddock CA, Joyce EM, McGuire P (2013): Neural and behavioral correlates of aberrant salience in individuals at risk for psychosis. Schizophr Bull 39:1328–1336.
11. Roiser JP, Stephan KE, den Ouden HEM, Barnes TRE, Friston KJ, Joyce EM (2009): Do patients with schizophrenia exhibit aberrant salience? Psychol Med 39:199–209.
12. Harlé KM, Zhang S, Ma N, Yu AJ, Paulus MP (2016): Reduced neural recruitment for Bayesian adjustment of inhibitory control in methamphetamine dependence. Biol Psychiatry Cogn Neurosci Neuroimaging 1:448–459.
13. Yu AJ, Dayan P, Cohen JD (2009): Dynamics of attentional selection under conflict: Toward a rational Bayesian account. J Exp Psychol Hum Percept Perform 35:700–717.
14. Huys QJ, Tobler PN, Hasler G, Flagel SB (2014): The role of learning-related dopamine signals in addiction vulnerability. Prog Brain Res 211:31–77.
15. Maia TV (2009): Reinforcement learning, conditioning, and the brain: Successes and challenges. Cogn Affect Behav Neurosci 9: 343–364.
16. Maia TV, Frank MJ (2011): From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154–162.
17. Redish AD (2004): Addiction as a computational process gone awry. Science 306:1944–1947.
18. Panlilio LV, Thorndike EB, Schindler CW (2007): Blocking of conditioning to a cocaine-paired stimulus: Testing the hypothesis that cocaine perpetually produces a signal of larger-than-expected reward. Pharmacol Biochem Behav 86:774–777.
19. Huys QJM, Deserno L, Obermayer K, Schlagenhauf F, Heinz A (2016): Model-free temporal-difference learning and dopamine in alcohol dependence: Examining concepts from theory and animals in human imaging. Biol Psychiatry Cogn Neurosci Neuroimaging 1: 401–410.
20. Dowd EC, Frank MJ, Collins A, Gold JM, Barch DM (2016): Probabilistic reinforcement learning in patients with schizophrenia: Relationships to anhedonia and avolition. Biol Psychiatry Cogn Neurosci Neuroimaging 1:460–473.
21. Deserno L, Boehme R, Heinz A, Schlagenhauf F (2013): Reinforcement learning and dopamine in schizophrenia: Dimensions of symptoms or specific features of a disease group? Front Psychiatry 4: 172.
22. Deserno L, Schlagenhauf F, Heinz A (2016): Striatal dopamine, reward, and decision making in schizophrenia. Dialogues Clin Neurosci 18: 77–89.
23. Radua J, Schmidt A, Borgwardt S, Heinz A, Schlagenhauf F, McGuire P, Fusar-Poli P (2015): Ventral striatal activation during reward processing in psychosis: A neurofunctional meta-analysis. JAMA Psychiatry 72:1243–1251.
24. Schlagenhauf F, Huys QJ, Deserno L, Rapp MA, Beck A, Heinze HJ, *et al*. (2014): Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. Neuroimage 89:171–180.
25. Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ (2014): Working memory contributions to reinforcement learning impairments in schizophrenia. J Neurosci 34:13747–13756.
26. Culbreth AJ, Westbrook A, Xu Z, Barch DM, Waltz JA (2016): Intact ventral striatal prediction error signaling in medicated schizophrenia patients. Biol Psychiatry Cogn Neurosci Neuroimaging 1:474–483.
27. Strauss GP, Waltz JA, Gold JM (2014): A review of reward processing and motivational impairment in schizophrenia. Schizophr Bull 40(suppl 2):S107–S116.
28. Koch K, Schachtzabel C, Wagner G, Schikora J, Schultz C, Reichenbach JR, *et al*. (2010): Altered activation in association with reward-related trial-and-error learning in patients with schizophrenia. Neuroimage 50:223–232.
29. Heinz A, Schlagenhauf F (2010): Dopaminergic dysfunction in schizophrenia: Salience attribution revisited. Schizophr Bull 36:472–485.
30. Wade EC, Iosifescu DV (2016): Using electroencephalography for treatment guidance in major depressive disorder. Biol Psychiatry Cogn Neurosci Neuroimaging 1:411–422.
31. DeBattista C, Kinrys G, Hoffman D, Goldstein C, Zajecka J, Kocsis J, *et al*. (2011): The use of referenced-EEG (rEEG) in assisting medication selection for the treatment of depression. J Psychiatr Res 45: 64–75.

32. Jollans L, Whelan R (2016): The clinical added value of imaging: A perspective from outcome prediction. Biol Psychiatry Cogn Neurosci Neuroimaging 1:423–432.
33. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF (2015): From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57:328–349.
34. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF (2016): Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. Biol Psychiatry Cogn Neurosci Neuroimaging 1:433–447.
35. Paulus MP, Huys QJM, Maia TV (2016): A roadmap for the development of applied computational psychiatry. Biol Psychiatry Cogn Neurosci Neuroimaging 1:386–392.