Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Computational approaches to psychiatry

Klaas Enno Stephan[1,2,3] and Christoph Mathys[3]

A major reason for disappointing progress of psychiatric diagnostics and nosology is the lack of tests which enable mechanistic inference on disease processes within individual patients. The resulting inability to pursue formal differential diagnosis has forced the field to stick to symptom-based diagnostic schemes with limited predictive validity concerning treatment response and clinical outcome. A promising new approach is the use of computational modeling for inferring mechanisms which generate observed behavior and brain activity in psychiatric patients. However, while this computational approach to psychiatry is rapidly gaining attention, much work remains to be done to finesse existing computational models, making them 'fit for practice' in a clinical setting and proving their validity in longitudinal studies. This review outlines recent methodological advances and strategies in this regard, focusing on generative models which infer mechanistically interpretable parameters (of computational or physiological processes) from measured behavior and brain activity.

Addresses
[1] Translational Neuromodeling Unit (TNU), Institute of Biomedical Engineering, University of Zurich & Swiss Federal Institute of Technology (ETH Zurich), Switzerland
[2] Laboratory for Social and Neural Systems Research (SNS), University of Zurich, Switzerland
[3] Wellcome Trust Centre for Neuroimaging, University College London, UK

Corresponding author: Stephan, Klaas Enno (stephan@biomed.ee.ethz.ch)

Current Opinion in Neurobiology 2014, **25**:85–92

This review comes from a themed issue on **Theoretical and computational neuroscience**

Edited by **Adrienne Fairhall** and **Haim Sompolinsky**

For a complete overview see the Issue and the Editorial

Available online 29th December 2013

0959-4388/$ – see front matter,

http://dx.doi.org/10.1016/j.conb.2013.12.007

Why are computational approaches important for psychiatry?

The present diagnostic toolkit of psychiatry does not include diagnostic tests (other than those for excluding 'organic' causes of brain disease) which reveal precise mechanisms underlying a given behavioral symptom and predict clinical outcome or guide individual treatment [1••]. This is a major reason why psychiatry has been unable to move beyond descriptive categorizations (such as the Diagnostic and Statistical Manual of Mental Disorders, DSM) which define mental diseases phenomenologically as clusters of symptoms but have limited predictive validity [2••].

Some reasons for this absence of mechanistically grounded tests are easily named. Genetics and neuroimaging as key methods of biological psychiatry face considerable hurdles: genetics struggles with strong gene–environment interactions, which is a likely key reason why clinically relevant predictions based on genomic data alone have been unsuccessful so far; cf. [3]. In contrast, while neuroimaging has the advantage of providing readouts of the functional *status quo* of putatively symptom-producing circuits, its measurements are indirect and distal from the neuronal processes of interest, aggravating the formulation of mechanistic hypotheses. One important strategy for breaking this impasse rests on the use of 'computational' models [4•,5••,6,7,8••]. In this review, we consider two possible meanings of the broad term 'computational': first, modeling mechanisms of *information processing* and second, *inferring* physiological processes from measurements of brain activity.

Computational approaches to psychiatry are rapidly gaining attention, as demonstrated by transregional research programs (e.g., the joint initiative by University College London and the Max Planck Society on 'Computational Psychiatry and Ageing Research', [9]), the first conference dedicated to 'Computational Psychiatry' [10], and newly founded institutions specifically dedicated to translational neuromodeling [11]. Numerous encouraging proof-of-concept examples exist how computational modeling can be applied to patients, for example [12–15]. So far, however, so far none of these computational approaches has been evaluated using a prospective study design, which is essential for evaluating clinical utility.

Therefore, this review on recent advances in methods and strategies for unlocking the translational potential of the computational approach to psychiatry. We concentrate on so-called 'generative models' which specify a joint probability distribution over all variables (observations and parameters) and serve to infer on cognitive and physiological mechanisms from measured behavior or brain activity [4•] (see Box 1). By contrast, limited space prohibits us from discussing the rich modeling literature inspired by neuroeconomics, game theory, graph theory or machine learning applications to psychiatric neuroimaging; for comprehensive review on these topics, see [16–18].

Modeling computation

The majority of existing computational treatments of psychiatric diseases concern aberrant learning and

Box 1 Generative models

A generative model defines a joint probability distribution $p(y,\theta)$ over observations (measured data y) and parameters θ. It has two components, a likelihood function $p(y|\theta)$ and a prior density of the parameters $p(\theta)$. It is called 'generative' because one can generate synthetic data by sampling parameter values from the prior and plugging these into the likelihood. One can thus also regard a generative model as a 'forward model' from parameters to observed data. 'Model inversion' refers to the opposite process: estimating the posterior probability of the parameters, given some observed data.

Notably, by integrating out the dependency of the data on the parameters, one obtains the 'expected data', that is, the marginal likelihood or model evidence:

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta \quad (1)$$

The model evidence is a principled measure for the generalizability of a model (i.e., its trade-off between accuracy and complexity) and is widely used for model comparison; see [69,71].

decision-making as core components of maladaptive cognition. While many types of such models exist, two have found particularly widespread application to empirical data: models of reinforcement learning (RL) and Bayesian inference. While originating from different theoretical roots, the two frameworks share some conceptual links. Most importantly, as highlighted in a recent derivation of RL equations from a variational approximation to hierarchical Bayesian learning [19•], both frameworks posit a structurally similar driving force behind learning: prediction error (PE), weighted by learning rate (RL) or precision/uncertainty (Bayesian theories). In this review, we give particular emphasis to Bayesian approaches, given that several excellent recent reviews on developments of RL exist [20–24].

One research question of particular relevance for psychiatry concerns the difference between 'model-free' and 'model-based' systems which are supposed to mediate habitual and goal-directed learning, respectively [25]. Simply speaking, in the former case, the PE represents the difference between actual and expected outcomes (e.g., a reward PE); in the latter case, the model embodies explicit knowledge about the environment and updates its representations by 'state PEs' (the difference between implied and expected states).

This distinction has received much interest by RL approaches in recent years. This was motivated by ideas about potential competition between different learning systems, for example, counter-productive Pavlovian influences on goal-directed learning [26], or a disturbance in the balance between habitual and goal-directed learning in obsessive–compulsive disorder [27]. An initial fMRI study [28] found that healthy participants' learning behavior reflected both reward and state PEs, where the former were correlated with activity in the ventral striatum, consistent with many previous studies, while state PEs were encoded by activity in parietal and prefrontal areas. This was broadly compatible with subsequent fMRI results [29] of ventral striatal activations by reward PEs, while state PEs where reflected by activity in prefrontal areas. However, another study with a two-step task, designed to maximally distinguish model-free and model-based learning, showed that fMRI activity in the ventral striatum did not purely reflect model-free learning, but a mixture of both learning forms, with proportions identical to those which optimally explained behavior [30••]. According to the authors, 'these results challenge the notion of a separate model-free learner and suggest a more integrated computational architecture for high-level human decision-making.'

Moving from RL to Bayesian approaches, the 'Bayesian brain hypothesis' [31,32], which views the brain as constructing and continuously updating a generative model of its sensory inputs (cf. Box 1), has inspired recent modeling frameworks with considerable potential for applications to psychiatry. For example, the 'free-energy principle' [33••,34], posits that the continuous optimization of the brain's generative model depends on minimization of free energy, a principled and tractable approximation to surprise (see Box 2 for a formal definition). Simply speaking, this corresponds to minimization of net prediction error (across potentially many levels of inference) and can be achieved by either adjusting one's beliefs about the world (perception) or changing the way one samples the world through the sensorium (action).

This perspective has led to a series of recent theoretical treatments of (mal)adaptive cognition, particularly with regard to schizophrenia [4•,35,36••,37]. Moreover, it has inspired concrete strategies for analyzing empirical data. One such framework for practical applications is a meta-Bayesian approach which considers the Bayesian inference (by an experimenter or psychiatrist) on Bayesian inference processes (in the brain of a subject or patient) that underlie the observed behavioral responses [38,39]. In this framework one models how the subject's 'hidden' (internal) belief updating processes give rise to his/her overt responses which, in turn, are observed by the experimenter. The appeal of such a hierarchical approach is that the experimenter's beliefs (about the subjects' beliefs driving the observed behavior) can be estimated by inverting a single generative model and under the same assumption about how Bayesian inference is implemented in the brain (e.g., by free-energy minimization).

A particular implementation of such a meta-Bayesian approach is the Hierarchical Gaussian Filter (HGF; [19•]) which derives RL-like update equations from a variational approximation to ideal hierarchical Bayesian learning and contains parameters that represent the individual's approximation to Bayes-optimality. This