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**

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

Box 2 Free energy

The free energy F represents an upper bound on the surprise (negative log probability) of encountering the data y, given a generative model m. The difference is given by the Kullback-Leibler divergence (KL; a measure of the dissimilarity of two probability densities) between an approximate posterior density $q(\theta)$ and the true but unknown posterior density $p(\theta|y,m)$:

$$F = -\log\, p(y|m) + KL[q(\theta),\, p(\theta|y,m)] \qquad (2)$$

Eqn (2) reveals two important things. First, instead of computing the posterior density directly (which can be intractable or computationally expensive), one can minimize free energy; this will minimize the KL divergence term and thus optimize the approximate posterior. Second, free energy represents a lower bound on the log model evidence (negative surprise) and can thus be used for model comparison [69,71].

framework has been used by several recent studies to adjudicate between competing hypotheses of learning and decision-making, using pathophysiologically relevant paradigms, such as perceptual learning [40] or cued eye movements [41]. It has also served as the basis for theoretical work on 'emotional valence' (in terms of the negative rate of change of free-energy) [42•].

Hierarchical Bayesian approaches are particularly useful for paradigms where uncertainty plays a crucial role, for example, induced by stimulus-bound (sensory noise) or environmental factors (volatility). In addition to the HGF, several other Bayesian models have been introduced recently, for example [43,44]. In particular, these have contributed to studies of neuromodulatory transmitters (e.g., dopamine, DA; acetylcholine, ACh; norepinephrine, NE), an application domain of particular relevance for psychiatry and a 'classical' target of computational modeling [45,46]. In the past two years, RL and Bayesian modeling of behavioral and neuroimaging data has yielded new insights into the roles of different neuromodulators, in particular in the context of the proposal by Yu and Dayan [47] that ACh and NE release encodes levels of 'expected uncertainty' and 'unexpected uncertainty', respectively. The proposed involvement of NE in signaling unexpected uncertainty has received support by studies of pupil size changes [48,49] and fMRI [50]. An outstanding issue is that fMRI shows a decrease of locus coeruleus activity with increase in unexpected uncertainty [50]; a relation with the opposite sign to that predicted [47].

An important aspect of neuromodulatory function concerns the adaptive scaling of PE signals [51]. For example, the impact of a PE on learning depends on its precision (inverse uncertainty) [19•,34]. While midbrain neuron activity has been found to reflect precision-weighting for rewards [52,53], it has been unclear whether such precision-weighting extends beyond rewards and the dopaminergic system. A recent fMRI study on sensory learning found that precision-weighted PEs about visual outcome activated the midbrain (unrelated to reward or novelty) [40]. By contrast, the precision-weighted PE on conditional probability (of the visual outcome given an auditory cue) — a quantity conceptually related to expected uncertainty — was encoded by activity in the cholinergic basal forebrain [40].

Modeling neurophysiology

While attempts to understand brain pathophysiology through mathematical models date back many decades, the interest in mathematical modeling of physiological processes relevant to psychiatric diseases has grown considerably in recent years. Two general approaches can be distinguished. Whereas one is based on inverting generative models of brain activity (discussed below), the more classical strategy rests on biophysically detailed dynamic system models describing either local microcircuits or ensembles thereof, linked by long-range connections. One salient example to which this approach has been successfully applied in recent work is the role of dopamine and NMDA receptors (NMDARs) for the occurrence of a frequent cognitive dysfunction in schizophrenia: impaired working memory, for example [54,55].

The complexity of these models prohibits parameter estimation; however, augmented with suitable forward models, they can predict changes in measured fMRI or EEG data which arise from changes in neuronal parameters susceptible to experimental manipulation. A nice example of this strategy [56••] used a biophysical model of prefrontal cortex to predict that blocking NMDARs would lead to less segregated representations of working memory contents by pyramidal cell activity and, as a result, a specific behavioral pattern of errors. This prediction was confirmed in a group of healthy volunteers who received the NMDAR antagonist ketamine versus placebo.

While the above approach is useful to generate testable predictions about the average pathophysiology in a conventionally (DSM) defined group of patients, it is not suited to address what is perhaps the most critical challenge for psychiatric diagnostics: differential diagnosis, that is, to infer, from observed behavior and brain physiology, on the most likely disease mechanism in a given individual patient. In other domains of medicine, such differential diagnosis is often supported by (biochemical) assays which allow for inference on 'hidden' disease mechanisms from peripherally accessible tissue (e.g., blood). An attractive idea is to use computational models for establishing equivalent procedures in psychiatry, using non-invasive functional read-outs instead of tissue samples. These 'computational assays' have been suggested in the form of generative models that can be fitted to measurements of brain activity and behavior [4•].

The hope is that such assays could detect the expression of (unknown) pathophysiological processes in individuals and help demarcating subgroups in heterogeneous disorders. Ideally, such assays would map onto processes that are directly amenable to existing therapeutic approaches (pharmacological or cognitive); this would allow for differential treatment predictions which could (and would have to) be evaluated in longitudinal studies.

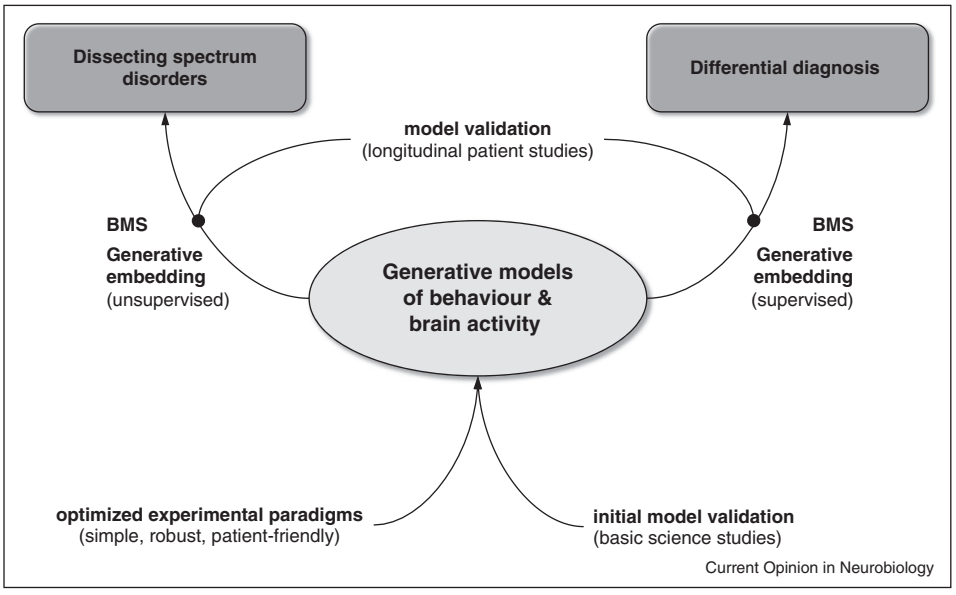
While not a trivial undertaking and still far from any major successes, in the physiological domain some important initial steps have been made in the last years. These typically rested on models of neuronal population dynamics which are sufficiently simplified to enable parameter estimation (model inversion) from fMRI or EEG data, yet sufficiently detailed that they retain a meaningful summary of physiological processes [57–59]. An established Bayesian system identification framework of this sort is dynamic causal modeling (DCM; [59,60]). For fMRI, DCM rests on a low-order (Taylor) approximation to the unknown neuronal system and explains measured BOLD signals as arising from synaptic coupling in large undifferentiated neuronal populations [60]. Despite this coarse representation, models of this type can be potentially useful, as demonstrated by recent studies. For example, in chronic schizophrenia (SZ), prefrontal–parietal coupling during working memory is reduced, regardless of performance or prefrontal activation [61]; across subjects at different disease states (from health via 'at risk mental state' to untreated first episode SZ), this coupling progressively declines but returns to levels indistinguishable from controls in treated first-episode patients [62]. Other DCM studies on SZ have scrutinized the effective connectivity during other tasks than working memory, including at risk subjects [63,64] and chronic patients [65].

DCMs for electrophysiological data provide a much finer conceptual resolution than for fMRI, distinguishing different types of neurons and synaptic connections. Following earlier validation studies in rodents (e.g., [66]), a recent proof-of-concept study in humans employed a dopaminergic drug challenge to demonstrate the feasibility of inferring DA-induced changes in NMDA and AMPA conductances in a prefrontal microcircuit [67••]. Another study used DCM to examine the contributions of NMDAR dependent short-term synaptic plasticity and neuronal adaptation to the reduced amplitude of the mismatch negativity (a model of impaired perceptual inference in SZ) under the NMDAR antagonist ketamine, finding a selective reduction of estimated short-term plasticity at auditory connections [68]. Provided these results can be confirmed in replication studies, models of this sort might serve as blueprints of clinically relevant assays for quantifying the status of transmitters systems in specific circuits.

The importance of generative models for differential diagnosis and subgroup detection

There are two reasons why generative models are important for computational psychiatry. First, as the name implies, generative models describe how observed data (brain activity or clinical symptoms) were generated by hidden mechanisms and causes (cf. Box 1). They thus force us to think mechanistically and be explicit about our

Figure 1



Graphical summary of key methodological building blocks for future extensions of psychiatric diagnostics through computational modeling.

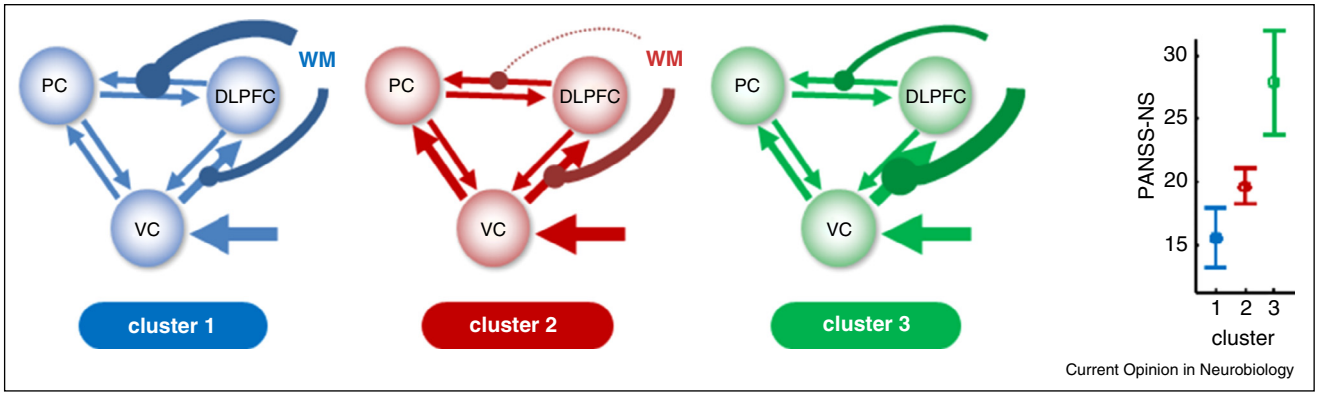
pathophysiological theories. Second, for any given measurement, different explanations are conceivable, that is, different models of the underlying (cognitive or neuronal) processes. These models can be compared using the (log) model evidence, approximated either during model inversion through variational Bayes (Box 2), or using classical approximations as the Bayesian Information Criterion (BIC). The evidence is a principled index of the trade-off between model fit and model complexity [69] which can be used to adjudicate between competing models. This Bayesian model selection (BMS) approach has seen increasing application in computational and neurophysiological modeling in recent years; for example [26,39,40,41,50,62,68,70]. Furthermore, subjects may differ in the processes generating their behavior, that is, the model itself may be a random variable in the population. This issue is particularly relevant for the heterogeneous spectrum disorders psychiatry deals with and has been addressed by the development of random effects BMS methods [71].

The inversion of competing generative models, each of which provides a different explanation for a measured behavioral or brain response, and their subsequent comparison by BMS could provide a formal framework for differential diagnosis in psychiatry. This requires, however, that the alternative disease pathways expressed across the spectrum of patients are known. A complementary approach is generative embedding, where the posterior estimates of relevant model parameters serve to construct a feature space for subsequent classification or clustering [72]. Initial proof-of-concept studies in aphasic and SZ patients, employing DCM for fMRI, have demonstrated excellent performance of generative embedding [72,73•], compared to standard approaches. Most importantly, generative embedding could aid in detecting pathophysiological subgroups in spectrum disorders. Such model-based definition of subgroups was recently demonstrated for SZ [73•]: here, clustering of connectivity estimates from a simple DCM of interactions between cortical areas during working memory revealed three patients subgroups that were distinguished by different visual–parietal–prefrontal connectivity (see Figure 2). Critically, these purely physiologically defined subgroups exhibited significantly different levels of clinical symptoms. The hope for the future is that the delineation of patient subgroups characterized by different disease processes, as indexed by mechanistically interpretable models, will allow for principled predictions about individual treatment and, eventually, pave the way towards a new nosology.

Summary and future challenges

This article has summarized some of the recent progress in establishing the methodology needed for establishing model-based assays as novel diagnostic tools for psychiatry. Over the next years, further progress might enable the practical implementation of a translational strategy for neuromodeling (Figure 1): first, establishing generative models that can be applied to data from optimally patient-friendly tasks; second, differential diagnosis for a given clinical symptom or measurement, based on a hypothesis set of competing neuronal and/or cognitive mechanisms, each of which is represented by a particular generative model; third, dissecting heterogeneous spectrum diseases into subgroups defined along mechanistic dimensions.

Figure 2



This figure, which is reproduced from [73•], illustrates how spectrum disorders, such as schizophrenia, can be partitioned into distinct subgroups using generative models. Here, fMRI measurements from 41 schizophrenic patients were analyzed using a simple dynamic causal model of interactions between visual (VC), parietal (PC) and dorsolateral prefrontal cortex (DLPFC). The connectivity estimates then served as input to an unsupervised clustering procedure based on a variational Gaussian mixture model. This showed that the most plausible partitioning of the connectivity data corresponded to three distinct subgroups. The connectivity architectures of these subgroups are shown iconically in different colors on the left (solid lines: positive connectivity estimates, broken lines: negative connectivity estimates; line width is proportional to the magnitude of the respective estimates). Critically, these three purely physiologically defined subgroups were distinguished by significantly different expression of clinical symptoms (right panel), that is, 'negative symptoms' as measured by the positive and negative syndrome scale (PANSS).

Clearly, many challenges lie ahead. Critically, whatever models are proposed, their assumptions and robustness must be carefully evaluated in basic validation studies, including initial pharmacological and stimulation (e.g., optogenetics) studies in animals and humans; for examples, see [66,67••,74,75•]. Subsequently, the most important challenge is to conduct longitudinal studies in patients with well-defined clinical problems (such as outcome or treatment response) that serve as real-world benchmarks against which the clinical utility of our models can (and must) be tested. While there are no such studies yet which prove that computational modeling can have a real practical impact on clinical decision-making in psychiatry, the many ongoing efforts in this regard instill hope that by the next time this topic features in an issue of this journal, first studies will have expressed an initial verdict on the practical utility of computational approaches to psychiatry.

Acknowledgements

We are grateful to Quentin Huys for helpful comments on the manuscript. We would like to acknowledge support by the René and Susanne Braginsky Foundation (KES), the NCCR 'Neural Plasticity and Repair' (CM, KES), the joint initiative by University College London and the Max Planck Society on 'Computational Psychiatry and Ageing Research' (CM), and the Clinical Research Priority Programs (CRPP) 'Molecular Imaging' and 'Multiple Sclerosis' (KES).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. •• Kapur S, Phillips AG, Insel TR: **Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?** *Mol Psychiatry* 2012, **17**:1174-1179.
An excellent summary of the reasons why psychiatric research has so failed to deliver clinically useful tests of disease mechanisms.

2. •• Cuthbert BN, Insel TR: **Toward the future of psychiatric diagnosis: the seven pillars of RDoC**. *BMC Med* 2013, **11**:126.
A comprehensive outline of the problems associated with standard phenomenological classification schemes and of a research framework for establishing a revised nosology.

3. Tansey KE, Guipponi M, Perroud N, Bondolfi G, Domenici E, Evans D, Hall SK, Hauser J, Henigsberg N, Hu X *et al.*: **Genetic predictors of response to serotonergic and noradrenergic antidepressants in major depressive disorder: a genome-wide analysis of individual-level data and a meta-analysis**. *PLoS Med* 2012, **9**:e1001326.

4. • Stephan KE, Baldeweg T, Friston KJ: **Synaptic plasticity and dysconnection in schizophrenia**. *Biol Psychiatry* 2006, **59**:929-939.
The first proposal to use generative models of behavior and brain activity as computational assays for detecting pathophysiological mechanisms and predicting treatment response.

5. •• Maia TV, Frank MJ: **From reinforcement learning models to psychiatric and neurological disorders**. *Nat Neurosci* 2011, **14**:154-162.
A thoughtful analysis of how to proceed from theoretical models to clinical applications in psychiatry and neurology.

6. Huys QJ, Moutoussis M, Williams J: **Are computational models of any use to psychiatry?** *Neural Netw* 2011, **24**:544-551.

7. Friston KJ, Dolan RJ: **Computational and dynamic models in neuroimaging**. *Neuroimage* 2010, **52**:752-765.

8. •• Montague PR, Dolan RJ, Friston KJ, Dayan P: **Computational psychiatry**. *Trends Cogn Sci* 2012, **16**:72-80.
An important proposal of key questions, goals and methods for computational psychiatry.

9. http://www.mpib-berlin.mpg.de/en/research/research-initiative-mps-ucl.

10. http://computationalpsychiatry.org.

11. http://www.translationalneuromodeling.org.

12. Yoshida W, Dziobek I, Kliemann D, Heekeren HR, Friston KJ, Dolan RJ: **Cooperation and heterogeneity of the autistic mind**. *J Neurosci* 2010, **30**:8815-8818.

13. Moutoussis M, Bentall RP, El-Deredy W, Dayan P: **Bayesian modelling of Jumping-to-Conclusions bias in delusional patients**. *Cogn Neuropsychiatry* 2011, **16**:422-447.

14. Frank MJ, Seeberger LC, O'Reilly RC: **By carrot or by stick: cognitive reinforcement learning in parkinsonism**. *Science* 2004, **306**:1940-1943.

15. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR: **The rupture and repair of cooperation in borderline personality disorder**. *Science* 2008, **321**:806-810.

16. Kishida KT, King-Casas B, Montague PR: **Neuroeconomic approaches to mental disorders**. *Neuron* 2010, **67**:543-554.

17. Bullmore E, Vertes P: **From lichtheim to rich club: brain networks and psychiatry**. *JAMA Psychiatry* 2013, **70**:780-782.

18. Klöppel S, Abdulkadir A, Jack CR Jr, Koutsouleris N, Mourao-Miranda J, Vemuri P: **Diagnostic neuroimaging across diseases**. *Neuroimage* 2012, **61**:457-463.

19. • Mathys C, Daunizeau J, Friston KJ, Stephan KE: **A Bayesian foundation for individual learning under uncertainty**. *Front Hum Neurosci* 2011, **5**:39.
A hierarchical Bayesian model for inferring individual mechanisms of (approximate) Bayes-optimality from measured behavior.

20. Botvinick MM: **Hierarchical reinforcement learning and decision making**. *Curr Opin Neurobiol* 2012, **22**:956-962.

21. Lee D, Seo H, Jung MW: **Neural basis of reinforcement learning and decision making**. *Annu Rev Neurosci* 2012, **35**:287-308.

22. Dayan P: **How to set the switches on this thing**. *Curr Opin Neurobiol* 2012, **22**:1068-1074.

23. Ito M, Doya K: **Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit**. *Curr Opin Neurobiol* 2011, **21**:368-373.

24. Gershman SJ, Niv Y: **Learning latent structure: carving nature at its joints**. *Curr Opin Neurobiol* 2010, **20**:251-256.

25. Doll BB, Simon DA, Daw ND: **The ubiquity of model-based reinforcement learning**. *Curr Opin Neurobiol* 2012, **22**:1075-1081.

26. Huys QJ, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP: **Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees**. *PLoS Comput Biol* 2012, **8**:e1002410.

27. Robbins TW, Gillan CM, Smith DG, de Wit S, Ersche KD: **Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry**. *Trends Cogn Sci* 2012, **16**:81-91.

28. Gläscher J, Daw N, Dayan P, O'Doherty JP: **States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning**. *Neuron* 2010, **66**:585-595.

29. Beierholm UR, Anen C, Quartz S, Bossaerts P: **Separate encoding of model-based and model-free valuations in the human brain**. *Neuroimage* 2011, **58**:955-962.

30. •• Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ: **Model-based influences on humans' choices and striatal prediction errors**. *Neuron* 2011, **69**:1204-1215.
A thorough examination of potentially competing learning mechanisms in humans using fMRI.

31. Dayan P, Hinton GE, Neal RM, Zemel RS: **The Helmholtz machine**. *Neural Comput* 1995, **7**:889-904.

32. Knill DC, Pouget A: **The Bayesian brain: the role of uncertainty in neural coding and computation**. *Trends Neurosci* 2004, **27**:712-719.

33. •• Friston K: **The free-energy principle: a unified brain theory?** *Nat Rev Neurosci* 2010, **11**:127-138.
A comprehensive theoretical framework for understanding adaptive cognition which has provided a foundation for many computational and physiological studies of mental disease.

34. Friston K, Kilner J, Harrison L: **A free energy principle for the brain**. *J Physiol Paris* 2006, **100**:70-87.

35. Fletcher PC, Frith CD: **Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia**. *Nat Rev Neurosci* 2009, **10**:48-58.

36. •• Corlett PR, Taylor JR, Wang XJ, Fletcher PC, Krystal JH: **Toward a neurobiology of delusions**. *Prog Neurobiol* 2010, **92**:345-369.
A compelling framework for understanding and experimentally studying delusions.

37. Adams RA, Perrinet LU, Friston K: **Smooth pursuit and visual occlusion: active inference and oculomotor control in schizophrenia**. *PLoS ONE* 2012, **7**:e47502.

38. Daunizeau J, den Ouden HE, Pessiglione M, Kiebel SJ, Stephan KE, Friston KJ: **Observing the observer (I): meta-bayesian models of learning and decision-making**. *PLoS ONE* 2010, **5**:e15554.

39. Lieder F, Daunizeau J, Garrido MI, Friston KJ, Stephan KE: **Modelling trial-by-trial changes in the mismatch negativity**. *PLoS Comput Biol* 2013, **9**:e1002911.

40. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HE, Stephan KE: **Hierarchical prediction errors in midbrain and basal forebrain during sensory learning**. *Neuron* 2013, **80**:519-530.

41. Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE: **Spatial attention, precision, and bayesian inference: a study of saccadic response speed**. *Cereb Cortex* 2013 http://dx.doi.org/10.1093/cercor/bhs418.

42. • Joffily M, Coricelli G: **Emotional valence and the free-energy principle**. *PLoS Comput Biol* 2013, **9**:e1003094.
A compelling example how computational frameworks can be used to formally define clinically relevant mental states.

43. Payzan-LeNestour E, Bossaerts P: **Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings**. *PLoS Comput Biol* 2011, **7**:e1001048.

44. Nassar MR, Wilson RC, Heasly B, Gold JI: **An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment**. *J Neurosci* 2010, **30**:12366-12378.

45. Dayan P: **Twenty-five lessons from computational neuromodulation**. *Neuron* 2012, **76**:240-256.

46. Doya K: **Modulators of decision making**. *Nat Neurosci* 2008, **11**:410-416.

47. Yu AJ, Dayan P: **Uncertainty, neuromodulation, and attention**. *Neuron* 2005, **46**:681-692.

48. Preuschoff K, t Hart BM, Einhauser W: **Pupil dilation signals surprise: evidence for noradrenaline's role in decision making**. *Front Neurosci* 2011, **5**:115.

49. Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI: **Rational regulation of learning dynamics by pupil-linked arousal systems**. *Nat Neurosci* 2012, **15**:1040-1046.

50. Payzan-LeNestour E, Dunne S, Bossaerts P, O'Doherty JP: **The neural representation of unexpected uncertainty during value-based decision making**. *Neuron* 2013, **79**:191-201.

51. Bunzeck N, Dayan P, Dolan RJ, Duzel E: **A common mechanism for adaptive scaling of reward and novelty**. *Hum Brain Mapp* 2010, **31**:1380-1394.

52. Schultz W, Preuschoff K, Camerer C, Hsu M, Fiorillo CD, Tobler PN, Bossaerts P: **Explicit neural signals reflecting reward uncertainty**. *Philos Trans R Soc Lond B Biol Sci* 2008, **363**:3801-3811.

53. Fiorillo CD, Tobler PN, Schultz W: **Discrete coding of reward probability and uncertainty by dopamine neurons**. *Science* 2003, **299**:1898-1902.

54. Anticevic A, Gancsos M, Murray JD, Repovs G, Driesen NR, Ennis DJ, Niciu MJ, Morgan PT, Surti TS, Bloch MH *et al.*: **NMDA receptor function in large-scale anticorrelated neural systems with implications for cognition and schizophrenia**. *Proc Natl Acad Sci USA* 2012, **109**:16720-16725.

55. Rolls ET, Deco G: **A computational neuroscience approach to schizophrenia and its onset**. *Neurosci Biobehav Rev* 2011, **35**:1644-1653.

56. •• Murray JD, Anticevic A, Gancsos M, Ichinose M, Corlett PR, Krystal JH, Wang XJ: **Linking microcircuit dysfunction to cognitive impairment: effects of disinhibition associated with schizophrenia in a cortical working memory model**. *Cereb Cortex* 2012 http://dx.doi.org/10.1093/cercor/bhs370.
A nice demonstration how predictions from biophysical models can be tested in human pharmacological studies.

57. Breakspear M, Heitmann S, Daffertshofer A: **Generative models of cortical oscillations: neurobiological implications of the kuramoto model**. *Front Hum Neurosci* 2010, **4**:190.

58. Olier I, Trujillo-Barreto NJ, El-Deredy W: **A switching multi-scale dynamical network model of EEG/MEG**. *Neuroimage* 2013, **83C**:262-287.

59. David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ: **Dynamic causal modeling of evoked responses in EEG and MEG**. *Neuroimage* 2006, **30**:1255-1272.

60. Friston KJ, Harrison L, Penny W: **Dynamic causal modelling**. *Neuroimage* 2003, **19**:1273-1302.

61. Deserno L, Sterzer P, Wustenberg T, Heinz A, Schlagenhauf F: **Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia**. *J Neurosci* 2012, **32**:12-20.

62. Schmidt A, Smieskova R, Aston J, Simon A, Allen P, Fusar-Poli P, McGuire PK, Riecher-Rossler A, Stephan KE, Borgwardt S: **Brain connectivity abnormalities predating the onset of psychosis: correlation with the effect of medication**. *JAMA Psychiatry* 2013, **70**:903-912.

63. Dauvermann MR, Whalley HC, Romaniuk L, Valton V, Owens DG, Johnstone EC, Lawrie SM, Moorhead TW: **The application of nonlinear dynamic causal modelling for fMRI in subjects at high genetic risk of schizophrenia**. *Neuroimage* 2013, **73**:16-29.

64. Diwadkar VA, Wadehra S, Pruitt P, Keshavan MS, Rajan U, Zajac-Benitez C, Eickhoff SB: **Disordered corticolimbic interactions during affective processing in children and adolescents at risk for schizophrenia revealed by functional magnetic resonance imaging and dynamic causal modeling**. *Arch Gen Psychiatry* 2012, **69**:231-242.

65. Banyai M, Diwadkar VA, Erdi P: **Model-based dynamical analysis of functional disconnection in schizophrenia**. *Neuroimage* 2011, **58**:870-877.

66. Moran RJ, Jung F, Kumagai T, Endepols H, Graf R, Dolan RJ, Friston KJ, Stephan KE, Tittgemeyer M: **Dynamic causal models and physiological inference: a validation study using isoflurane anaesthesia in rodents**. *PLoS ONE* 2011, **6**:e22790.

67. •• Moran RJ, Symmonds M, Stephan KE, Friston KJ, Dolan RJ: **An in vivo assay of synaptic function mediating human cognition**. *Curr Biol* 2011, **21**:1320-1325.
A proof-of-concept study of model-based assays, using a dopaminergic drug challenge to demonstrate the feasibility of inferring drug-induced changes in NMDA and AMPA conductances in a prefrontal microcircuit.

68. Schmidt A, Diaconescu AO, Kometer M, Friston KJ, Stephan KE, Vollenweider FX: **Modeling ketamine effects on synaptic plasticity during the mismatch negativity**. *Cereb Cortex* 2013, **23**:2394-2406.

69. Penny WD: **Comparing dynamic causal models using AIC, BIC and free energy**. *Neuroimage* 2012, **59**:319-330.

70. Frank MJ, Badre D: **Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis**. *Cereb Cortex* 2012, **22**:509-526.

71. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ: **Bayesian model selection for group studies**. *Neuroimage* 2009, **46**:1004-1017.

72. Brodersen KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, Stephan KE: **Generative embedding for model-based classification of fMRI data**. *PLoS Comput Biol* 2011, **7**:e1002079.

73. • Brodersen KH, Deserno L, Schlagenhauf F, Lin Z, Penny WD, Buhmann JM, Stephan KE: **Dissecting psychiatric spectrum disorders by generative embedding**. *Neuroimage: Clin* 2013, **4**:98-111.
A first demonstration how generative models could be used for identifying mechanistically distinct subgroups in spectrum disorders.

74. David O, Guillemain I, Saillet S, Reyt S, Deransart C, Segebarth C, Depaulis A: **Identifying neural drivers with functional MRI: an electrophysiological validation**. *PLoS Biol* 2008, **6**:2683-2697.

75. • Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH: **A causal link between prediction errors, dopamine neurons and learning**. *Nat Neurosci* 2013, **16**:966-973.
An impressive example how optogenetics can be used to validate central assumptions of computational models.