7

How Could We Get Nosology from Computation?

Christoph Mathys

Abstract

Psychiatry has found it difficult to develop a nosology that allows for the targeted treatment of disorders of the mind. The historic inability of the field to agree on a nosology based on clinical experience has led it to retreat to diagnoses based on symptom checklists as laid down in the Diagnostic and Statistical Manual of Mental Disorders (DSM). While this has increased the reliability of diagnoses, hopes that biological findings would lead to the emergence of mechanistically founded diagnostic entities have not been realized despite considerable advances in neurobiology. This article sets out a possible way forward: harnessing systems theory to provide the conceptual constraints needed to link clinical phenomena with neurobiology. This approach builds on the insight that the mind is a system which, to regulate its environment, needs to have a model of that environment and needs to update predictions about it using the rules of inductive logic (i.e., Bayesian inference). The application of the rules of inductive logic is called Bayesian inference because Bayes's theorem is the most important consequence of these rules, prescribing how beliefs need to be updated in response to new information. Importantly, while Bayesian inference is by definition consistent with the rules of inductive logic, it can still be false (to the point of being pathological), in the sense of leading to false predictions, because the model underlying the inference is inadequate. Further, it can be shown that Bayesian inference can be reduced to updating beliefs based on precision-weighted prediction errors, where a prediction error is the difference between actual and predicted input, and precision is the confidence associated with the input prediction. Precision weighting of prediction errors entails that a given discrepancy between outcome and prediction means more, and leads to greater belief updates, the more confidently the prediction was made. This provides a conceptual framework linking clinical experience with the pathophysiology underlying disorders of the mind. Limitations of this approach are discussed and ways to work around them illustrated with examples. Finally, initial steps and possible future directions toward a nosology based on failures of precision weighting are discussed.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Introduction

The State of Psychiatric Nosology

Before DSM-III, the state of psychiatric nosology was widely seen as unsatisfactory. The main point of criticism was the lack of diagnostic reliability (e.g., Robins and Guze 1970). Given a dearth of biological or conceptual constraints on nosological speculation, clinical experience had to serve as the main guide in developing a nosology of the mind. Clinical experience came—and comes—in two forms: (a) each clinical practitioner has his or her own immediate experience with patients, but (b) he or she is also acculturated into the thinking of the field, whose collective clinical experience has been condensed into nosological categories that are passed on as a traditional body of knowledge. While neither of these sources of nosological insight is to be scoffed at, it is not surprising that, owing to the diversity of individual experience and nosological traditions, the inter-rater reliability of psychiatric diagnoses was low. DSM-III was a conscious effort to address this problem by shifting the focus of diagnosis to lists of easily observable or reportable symptoms. However, despite decades of efforts, and despite an increase in the reliability of diagnoses (Clarke et al. 2013; Freedman et al. 2013; Narrow et al. 2013; Regier et al. 2013), the state of psychiatric nosology is still widely held to be dire (Craddock and Owen 2010; Kapur et al. 2012). Critics focus mostly on the missing biological foundation of the existing diagnostic categories (Hyman 2012; Insel 2012; Owen 2014), and they express hope that directing research efforts toward the biological mechanisms underlying psychiatric disorders will enable a new, mechanistically grounded nosology and uncover new targets for pharmacological treatment. In mitigation, one might say that the current system produces reliable categories that have been useful for clinical care up to a point, but that to make further nosological advances which can help guide neurobiological research, improve predictions regarding what treatments will be most efficacious, and ultimately identify new treatment targets, computational approaches will be essential.

Constraints on Nosology from Systems Theory

There is relatively little appreciation for the fact that in addition to constraints on psychiatric nosology rooted in clinical experience and human biology, there are also constraints originating in systems theory. While systems theory is complicated, these constraints are simple, which makes them an important guide to the interpretation of the clinical and biological patterns observed in disorders of the mind. My aim here is to explain these constraints, their implications, and their simplicity.

The main systems theoretic constraint, from which all others follow, is the good regulator theorem (Conant and Ashby 1970): every good regulator of a

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

system needs to be a model of that system. This forms the basis of the reasoning in this article. Setting out from this theorem, I argue that any mind striving successfully to preserve its existence fits the description of a good regulator and therefore needs to be, in a well-defined sense, a model of its environment. To go about regulating its environment, the model (i.e., the mind) performs inference, learns, and selects actions in line with the laws of probability, or in other words, according to Bayesian inference (Jaynes 2003). At the heart of Bayesian inference is Bayes's theorem (Bayes and Price 1763; Laplace 1774). It prescribes the only way that, given a model, beliefs (i.e., probability distributions) can be updated in response to new information without violating the laws of probability. It is important to note that the word "belief" is used here as shorthand for "probability distribution of a state or parameter of the mind's model of its environment." Beliefs need not be conscious or even consciously accessible. In most cases, including almost all interesting ones, the equation governing probabilistic belief updates given by Bayes's theorem has no closed-form solution, meaning that it is mathematically impossible to write down an equation giving the solution. For example, the very simple equation $2^x = x + 1$ cannot be solved for x in closed form (i.e., it cannot algebraically be transformed into an equivalent equation of the form $x = \ldots$). However, solutions exist ($x = 0$ and $x = 1$), but short of guessing them, the lack of a closed-form solution forces us to find them by using approximations, which always involve assumptions. Importantly, approximations do not diminish the complexity and richness of the models being used to perform inference. To the contrary, approximate methods allow us to perform inference using much richer models than would be the case if we were restricted to cases where closed-form inference is possible. I will show that if these assumptions underlying approximate inference are chosen in the right way, Bayesian inference can be reduced to the application of simple update rules that all have one canonical form: precision-weighted prediction errors. A prediction error is the difference between actual and predicted input, and precision is the confidence associated with the input prediction. Precision weighting of prediction errors entails that a given discrepancy between outcome and prediction means more, and leads to greater belief updates, the more confidently the prediction was made. If the mind is a model of its environment and, in order to be a good regulator, has to take recourse to Bayesian inference, which is implemented as belief updating by precision weighting of prediction error, then disorders of the mind can be described as false inference based on broken precision weighting or prediction error processing. This gives us an additional set of constraints within which to develop a scientific psychiatric nosology. To be complete, any description of a disorder in such a nosology would have to address three questions: Is this concept of a disorder able to explain the patterns seen in clinical practice? How do these patterns emerge from false inference in terms of precision-weighted prediction errors? In what way is the biological machinery underlying belief updating broken? These questions illustrate

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

that, instead of complicating the picture, the additional constraints from systems theory simplify the task of relating clinical manifestations of disorders to underlying biological mechanisms because they serve as a conceptual bridge between them.

In what follows, I will sketch a tentative construction plan for this conceptual bridge. Its body will be formed by an elaboration of the preliminary systems theoretic reasoning given above, while the bridge ends will be formed by recent attempts to tie false inference to psychiatric symptomatology (Edwards et al. 2012; Adams et al. 2013; Lawson et al. 2014; Quattrocki and Friston 2014) and to pin down the neurobiology of belief updating (Bastos et al. 2012; Shipp et al. 2013).

Theory

The Good Regulator Theorem, Generative Models, and Inductive Reasoning

In systems theoretic terms, the mind of an organism is a system which, using its brain and other body parts, strives to survive by regulating a second system, its environment. This makes it subject to the good regulator theorem, as stated in the title of Conant and Ashby's article: "Every good regulator of a system must be a model of that system" (Conant and Ashby 1970:89). The authors explain what this means in their abstract:

> The design of a complex regulator often includes the making of a model of the system to be regulated. The making of such a model has hitherto been regarded as optional, as merely one of many possible ways.

They go on to construct a theorem which shows, under very broad conditions, that any regulator which is maximally both successful and simple must be isomorphic (i.e., of the same structure and equipped with the same properties) with the system being regulated. (The exact assumptions are given.) Making a model is thus necessary. The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival, must proceed, in learning, by forming a model (or models) of its environment.

This model of the brain's (or rather mind's) environment is a *generative model* of its sensory input, u; that is, a model of how the environment generates u. For example, u could be a slight crackling noise somebody hears while speaking on the telephone. Generative models consist of two parts. The first, called the likelihood, is the probability $p(u|x,\vartheta)$ of input u given state x of the environment and parameter ϑ. The difference between state and parameter is simply that the state changes with time while the parameter is constant. In this example, there are many possible causes that could have generated this crackling noise, some of them more plausible than others. One possible cause is that the telephone has been bugged and the crackling is caused by a listening device. The state x would then be that of the telephone, bugged or not bugged,

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

while the parameter ϑ would govern how exactly the telephone being bugged (or not) translates into the crackling noise u. Note that both state and parameter are sets that can have many elements. The second part of the model is called the *prior distribution*, or simply *prior*. This is the probability $p(x,\vartheta)$ of state and parameter in the absence (usually before, hence *prior*) of input u. In a clinical setting, different patients can attach very different prior probabilities to telephones being bugged, to bugging leading to crackling, etc. Such a (possibly largely unconscious) view of how the environment generates sensory input is formally described by the *joint distribution* $p(u,x,\vartheta) = p(u|x,\vartheta)\,p(x,\vartheta)$ of input, state, and parameter; that is, by the product of likelihood and prior, which constitutes a full generative model.

Given a generative model and input u, the mind can then, in principle, calculate the posterior distribution of state and parameter by the application of Bayes's theorem:

$$p(x,\vartheta \mid u) = \frac{p(u \mid x,\vartheta)\,p(x,\vartheta)}{\int p(u \mid x',\vartheta')\,p(x',\vartheta')\,\mathrm{d}x'\,\mathrm{d}\vartheta'} \quad . \tag{7.1}$$

This is the probability distribution of state and parameter given the input u, and the transition from $p(x,\vartheta)$ to $p(x,\vartheta|u)$ in response to u is a *belief update* in the sense that probability distributions constitute beliefs. Crucially, the update as given by Equation 7.1 is the only way to update the belief on x and ϑ that does not violate elementary requirements of inductive reasoning (Cox 1946; Jaynes 2003). Inductive reasoning is reasoning about uncertain quantities, as opposed to deductive reasoning, which deals with certain quantities. For example, if we know that every cat is an animal, we can deduce with certainty that A (an animal from the information) is a cat. However, if we at first know nothing about A and are then told it is an animal, this merely increases the probability that A is a cat without making it certain. In other words, A being a cat becomes more plausible. This increase and decrease in the plausibility of statements as a result of new information is what inductive reasoning addresses. Cox (1946) showed that the only rational way to update beliefs about the plausibility of statements is by applying the known rules of probability theory (e.g., Bayes's theorem; cf. Flagel et al., this volume). He proved this by showing that the rules of probability can be derived from three basic desiderata concerning inductive reasoning:

1. The plausibility of a statement can be represented by a real number (and the plausibilities of different statements compared this way).
2. Information that makes a statement more plausible increases the number associated with it.
3. Different ways to calculate the same plausibility should always give the same result.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

This means that reasoning incompatible with the rules of probability implies a violation of Cox's three desiderata (i.e., of common sense). This may seem like a restrictive constraint, but in reality it is anything but. According to the complete class theorem (Robert 2007:411), under mild conditions, there is always a prior accounting for any given combination of posterior, likelihood, and input. This means that no conclusion that could be drawn from a given observation is impossible in the sense of violating the rules of probability. Bayesian belief updating in no way constrains the inferences the mind can make about its environment. However, it constrains the way the mind is described. A full account of a belief update has only been given once the generative model it is based upon has been fully described; that is, once a likelihood and a prior have been specified. While the good regulator theorem states that the mind will have to be a model of its environment, Bayes's theorem provides the framework within which to describe belief updating in accordance with the rules of inductive reasoning but without constraining its substance.

In what follows, I will take a fairly didactic but technical walk through some of the formal aspects of Bayesian inference in the brain. In other words, we will look at the mathematical structure of how beliefs are encoded and updated and what this tells us about neuronal processes. Although this treatment is a bit mathematical, the end point of the analysis will be something that is central to a theoretical and neurobiologically grounded understanding of false inference in psychiatry. This is the central role of gain control or neuromodulation in the brain in weighting neuronal messages that are passed from one part of the brain to the other. Neuromodulatory mechanisms are invariably implicated in both the pathophysiology and pharmacology of psychiatric conditions (e.g., implicating classical neuromodulators like dopamine and serotonin). Furthermore, this form of gain control implicates fast spiking inhibitory interneurons and synchronized neuronal activity of the sort that can be measured noninvasively using EEG and, potentially, correlated with symptoms of false inference, and response to treatment.

Sequential Updating of a Time Series Mean

Before returning to Bayesian belief updating, let us first look at ways to update the mean of a series of sequentially observed numbers. This might at first seem a distraction, but will turn out to be fundamental.

Given N observations $\{u_1, u_2, \ldots, u_N\}$, it is simple to calculate their mean $\bar{u}_n$:

$$\bar{u}_N = \frac{1}{N}\sum_{n=1}^{N} u_n. \tag{7.2}$$

However, if $\{u_1, u_2, \ldots, u_N\}$ is a time series, keeping track of the mean as new observations arise requires all observations, if the calculation is to be made

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

according to Equation 7.2. Fortunately, there is a less memory-intensive way to achieve the same end. The following update equation can be applied sequentially:

$$\bar{u}_{n+1} = \bar{u}_n + \frac{1}{n+1}\left(u_{n+1} - \bar{u}_n\right). \tag{7.3}$$

Starting with $\bar{u}_1 = u_1$ and applying Equation 7.3 to all observations until $\bar{u}_n$ is reached, we get:

$$\bar{u}_N = \bar{u}_{N-1} + \frac{1}{N}\left(u_N - \bar{u}_{N-1}\right), \tag{7.4}$$

which gives the same result as Equation 7.2.

The sequential updating of Equation 7.3 has the advantage that it requires remembering only two numbers: the previous mean $\bar{u}_n$ and the number of previous observations n.

Since the update rule of Equation 7.3 is of fundamental importance, it is worth looking at its components. There is the previous mean $\bar{u}_n$, representing the state of belief before the new observation u_{n+1}. Since the current state of belief corresponds to the best possible prediction for any new observation, the difference $u_{n+1} - \bar{u}_n$ between the new observation and the current belief is a *prediction error*. This means that the update has the form:

$$\text{new mean} = \text{old mean} + \text{weight} \cdot \text{prediction error}, \tag{7.5}$$

where the weight of the prediction error depends on how many previous observations there have been. The more observations that have already been made, the less a new observation will be able to move the mean.

Bayesian Belief Updating

A simple example of Bayesian belief updating is the case where the likelihood $p(u \mid \vartheta) = \mathcal{N}\left(u; \vartheta, \pi_\varepsilon^{-1}\right)$ is Gaussian (i.e., follows a normal distribution) with known precision (i.e., inverse variance) π_ε and the prior $p(\vartheta) = \mathcal{N}\left(x; \mu_\vartheta, \pi_\vartheta^{-1}\right)$ is also Gaussian with mean μ_ϑ and precision π_ϑ. There is no time-varying state x here, the parameter ϑ is a simple scalar, and the prior hyperparameter $\{\mu_\vartheta, \pi_\vartheta\}$ (i.e., the parameter governing the prior distribution of the parameter) is taken to be known. The posterior now also turns out to be Gaussian: $p(\vartheta \mid u) = \mathcal{N}\left(x; \mu_{\vartheta|u}, \pi_{\vartheta|u}^{-1}\right)$, where the updated precision and mean are:

$$\begin{aligned} \pi_{\vartheta|u} &= \pi_\vartheta + \pi_\varepsilon, \\ \mu_{\vartheta|u} &= \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|u}}\left(u - \mu_\vartheta\right). \end{aligned} \tag{7.6}$$

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

Remarkably, the update of the mean has the same structure (i.e., that of Equation 7.5) as the update of the mean of Equation 7.3. This similarity becomes even more obvious if we rearrange Equation 7.6 to read:

$$\mu_{\vartheta|u} = \mu_{\vartheta} + \frac{1}{\pi_{\vartheta} / \pi_{\varepsilon} + 1}(u - \mu_{\vartheta}). \tag{7.7}$$

This shows the correspondence between n, the number of previous observations in Equation 7.3, and the relative precision $\pi_{\vartheta}/\pi_{\varepsilon}$ of the prior with respect to the likelihood. In both cases, this represents the weight of previous evidence relative to new information. In what follows, this relative weight will be called v to emphasize that it can be any positive number, whereas n was a natural number.

This update structure is not restricted to the simple Gaussian example used above. All generative models we are ever likely to need to describe the brain (or equivalently, all generative models the brain is ever likely to need to describe its environment) will only involve exponential families of likelihoods with conjugate priors. These are families of likelihood distributions that can all be written in one canonical form, which is a generic representation of all families. A conjugate prior is one that gives rise to a posterior of the same family when combined with a given likelihood. For example, the Gaussian distribution is an exponential family, and it is its own conjugate prior. As we saw in Equation 7.6, this means that a Gaussian likelihood with a Gaussian prior leads to a Gaussian posterior. In addition to the Gaussian distribution, this includes the beta, gamma, binomial, Bernoulli, multinomial, categorical, Dirichlet, Wishart, Gaussian-gamma, log-Gaussian, multivariate Gaussian, Poisson, and exponential distributions, and many others. For all of these distributions, the Bayesian belief update has the following form:

$$\xi' = \xi + \frac{1}{\nu + 1}\big(T(u) - \xi\big), \tag{7.8}$$

where $T(u)$ is a function of the input u called the *sufficient statistic*, ξ is the hyperparameter governing the prior, and ξ' is the updated hyperparameter, which governs the posterior.

In the case of exponential families with conjugate priors, this means that Bayesian inference reduces to tracking the mean of the sufficient statistics of observations. The weight of the prior in this update is determined by the positive number v, which can be interpreted as the number of observations preceding u, whose weight is 1. In light of Equation 7.7, v is the precision of the prior relative to that of the observation. Bayesian belief updating thus takes place by precision-weighting prediction errors on the sufficient statistics of observations.

To conclude this discussion, I will give a few examples of a more technical nature. (Less technically inclined readers can skip this without missing anything essential.) In the case of a Gaussian model with unknown mean and

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

known precision (as in Equation 7.6), $T(u)=u$; if both mean and precision are unknown, $T(u)=(u, u^2)^{\mathrm{T}}$. This generalizes to $T(\mathbf{u})=(\mathbf{u}, \mathbf{u}\mathbf{u}^{\mathrm{T}})^{\mathrm{T}}$ for a multivariate Gaussian. In the case of a gamma model, $T(u)=(\ln u, u)^{\mathrm{T}}$. In the beta case, $T(u)=(\ln u, \ln(1-u))^{\mathrm{T}}$, and in the categorical case, $T(\mathbf{u})=\mathbf{u}$. Between these models, situations are addressed where observations are on an unbounded continuum, on a continuum bounded on one side or on both sides, all multivariate generalizations of these, and situations where observations are categorical. In all of these models, and in many more cases, Bayesian inference reduces to Equation 7.8 (i.e., to mean updating).

Discussion

The Bridge Ends: Clinical Phenomena and Neurobiology

If the mind is necessarily a model of the environment it regulates, and if using a model to regulate the environment entails updating beliefs according to the laws of probability (i.e., according to the rules of Bayesian inference), and if Bayesian inference entails precision weighting of prediction errors, then disorders of the mind will have to be interpretable in terms of precision weighting of prediction errors. Further, if the brain is the organ of the mind, then the brain's physiology will also have to be interpretable in terms of precision weighting of prediction errors. This is how systems theory can serve as a bridge between clinical manifestations of disorders of the mind and the disordered biological mechanisms underlying them, connecting them in a way that allows us to make sense of both.

Turning first to the side of clinical manifestations, there have been many recent attempts to understand disorders of the mind in terms of precision weighting of prediction errors. Relating to psychosis, Adams et al. (2013) give a broad overview and explain many of the manifestations of psychosis, such as hallucinations, delusions, catatonia, and sensory attenuation deficits, as the result of aberrant precision weighting of exteroceptive sensory input. For example, patients with schizophrenia show abnormalities in smooth pursuit eye movements (Thaker et al. 1999). When following a dot as it moves smoothly back and forth, right to left, they are less able to predict where it will reappear after it has been occluded by a vertical bar for a short while. When the dot disappears, patients slow down their eye movement more than healthy controls, forcing them to accelerate more to catch up with the dot once it reappears. Conversely, when the dot makes unexpected jerky movements, patients with schizophrenia are better able to follow it than healthy controls (Hong et al. 2005). Aberrant precision weighting of prediction errors can explain this apparent paradox. The observed effects are predicted by models where healthy controls rely more on top-down predictions from an internal model of dot motion to follow the dot, while patients rely more on the immediate bottom-up

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

input. This relative disregard by healthy controls of sensory input in favor of model-based predictions is called sensory attenuation, and the patients' behavior is explained by a failure of sensory attenuation. Specifically, when the precision weight on the prediction error regarding sensory input is high relative to that regarding the model-based prediction of dot position, then the eye will remain as if glued to the stimulus, enabling quick reactions to unexpected jerks, but losing its bearings whenever the dot disappears. Failure of sensory attenuation can also explain other abnormalities in patients with schizophrenia, for example resistance to the hollow-mask illusion and to the force-matching illusion. Furthermore, aberrant precision weighting leading to a failure of sensory attenuation can explain the emergence of delusions and hallucinations (cf. Adams et al. 2013). This is because action (e.g., moving a hand or an eyeball) is impossible without sensory attenuation. When confronted with prediction errors, a biological agent existing under the constraints of the good regulator theorem (i.e., trying to make good predictions) has two general ways to reduce them: it can either update its beliefs or act to change the environment so that sensory input matches predictions. For example, if we feel cold outside, we can go inside, thereby regulating our environment to conform to a temperature range that evolution has hardwired us to find pleasant (i.e., that minimizes prediction error with respect to an unconscious model we have of how our environment will be) because it makes our survival and reproduction most likely. In this example, simply updating our beliefs about which kind of environment we will encounter would lead us to stay out in the cold, which would make our reproductive success less likely. However, starting to walk inside is also associated with prediction errors. If I have a correct proprioceptive model of myself standing still, then the way to minimize prediction errors about that is to keep standing still. For me to act, I need to attenuate proprioceptive prediction errors so that the prediction error about myself being in a cold environment can become dominant and trigger the action of going inside. If I am unable to attenuate my proprioception, I will either become catatonic or I will have to try to override the power of my proprioceptive prediction errors by ascribing my own intentions and predictions about sensory input to external forces. In other words, I will develop delusions or hallucinations (for details of these mechanisms, see Adams et al. 2013; Brown et al. 2013).

A similar line of reasoning is applied by Lawson et al. (2014) and Quattrocki and Friston (2014) to the interoceptive domain, which allows them to describe many of the symptoms of autism as a consequence of aberrant precision weighting. Edwards et al. (2012) are concerned with hysteria (i.e., functional motor and sensory symptoms, sometimes described as "psychogenic" or "medically unexplained"). They use the precision-weighting framework in an effort to introduce more rigor (more precision, one might say) into the discussion of a disorder whose mention has been all but banned, yet has stubbornly refused to go away.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

On the neurobiological side of the conceptual bridge, formal models of a hierarchical, precision-weighted message passing in the brain have been developed (Friston 2008; Bastos et al. 2012; Shipp et al. 2013). These efforts build on work dating back to Helmholtz (1860/1962), who was the first to propose that the brain is a predictive machine that becomes active in response to input only insofar as the input is unexpected. This concept of the brain is theoretically reinforced by the good regulator theorem, which prescribes the presence of a model, and it underlies the Bayesian brain hypothesis (Dayan et al. 1995), which postulates that the brain uses Bayesian inference to make the predictions required in the Helmholtzian view. Neurobiologically, the Bayesian brain is taken to be implemented by predictive coding (Rao and Ballard 1999; Friston 2005, 2008), which postulates that bottom-up prediction errors and top-down predictions are processed in the cortical neuronal hierarchy by a message passing between different cortical layers at different hierarchical levels. Specifically, according to Bastos et al. (2012), there is a canonical cortical microcircuit (cf. Douglas and Martin 1991; Haeusler and Maass 2007) which receives forward connections into cortical layer 4, conveying precision-weighted prediction errors from lower levels of a message-passing hierarchy embodied in the hierarchical neuronal anatomy of the brain (Figure 7.1). These prediction errors are used to adjust predictions at the level in question and sent backward (i.e., down the hierarchy) from the deep cortical layers. In the superficial layers, backward connections from higher areas are received and compared to predictions. The resulting prediction errors are precision weighted and passed forward (i.e., up the hierarchy), where the same information processing occurs in a higher region.

Limitations

The perspective laid out here has its limitations. We pay a price for reducing Bayesian inference to the tracking of sufficient statistics by foregoing the use of (a) any likelihoods that are not exponential families and (b) any but conjugate priors. These restrictions, however, are much milder than they might appear at first sight.

I address the first limitation by looking at an example by Daunizeau et al. (2010), where a likelihood from an exponential family clearly will not do. This likelihood needs to describe the probability of sensory, in this case retinal, input where the object presented in a black-and-white image is a house or a face. Applying principal component analysis to their sixteen images (eight of each kind), Daunizeau et al. (2010) recover two clusters representing houses and faces which can adequately be described by a pair of two-dimensional Gaussian distributions in the first two principal components (Figure 7.2). While this mixture is a simple and adequate likelihood for the situation at hand, it is not from an exponential family, which means that at first sight, belief updates

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

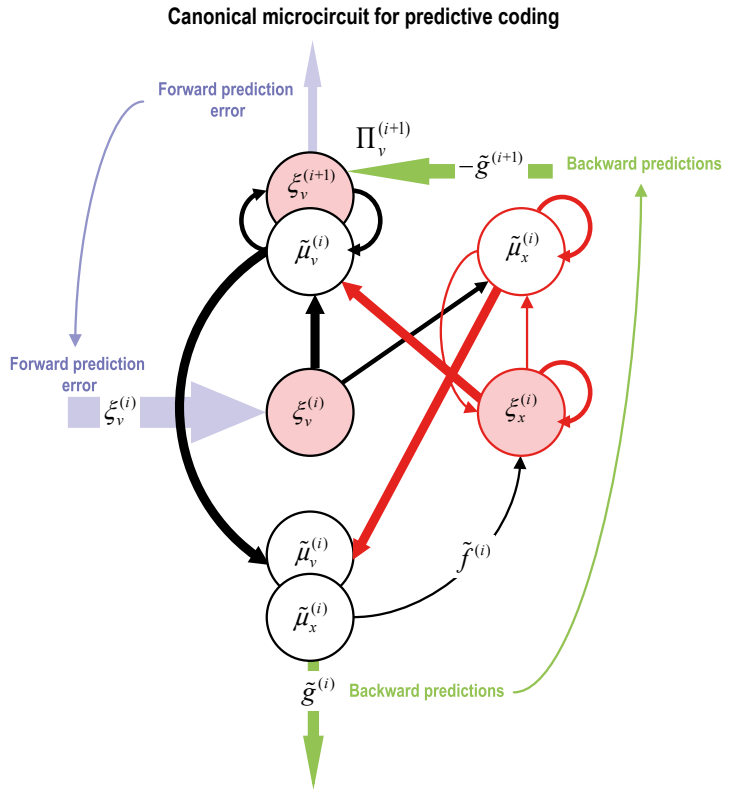

Figure 7.1 A proposed canonical microcircuit for predictive coding (reproduced with permission from Bastos et al. 2012). This is a schematic representation of a cortical column with supragranular layers at the top, infragranular layers at the bottom, and granular layers in the middle. Pink: prediction error populations. Red: inhibitory connections. Black: excitatory connections. Predictions ($\tilde{\mu}$) are encoded in supragranular excitatory and inhibitory interneurons and are passed to infragranular pyramidal cells. Prediction errors (ξ) enter granular layers from regions situated lower in the hierarchy. Prediction errors that are passed on to the next higher hierarchical level are computed in supragranular pyramidal cells. Crucially, they are weighted by the precision (Π) of the prediction ($\tilde{g}$) received from the higher level.

based on this model cannot be formulated as precision-weighted prediction errors. There is a way around this, however. The key is to formulate the problem hierarchically, with a prior on the probability of houses and faces, respectively. This allowed Daunizeau et al. (2010) to use variational Bayesian methods[1] to calculate belief updates by separating the levels of the hierarchy using a mean field approximation. Using a mean field approximation in this context means

[1] Briefly, variational Bayes is a method of model estimation that uses variational calculus to find the posterior distribution of parameters by maximizing the model evidence instead of calculating the posterior directly (these terms are explained in Flagel et al., this volume).

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

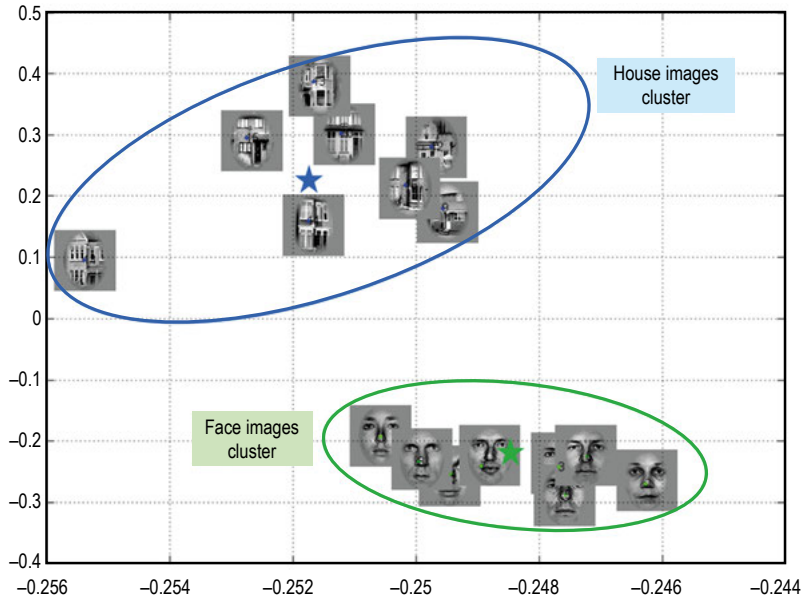

Figure 7.2 Projection of eight face and eight house stimuli onto their first two principal eigenvectors. Faces and houses form distinct clusters, and each cluster can be described by a two-dimensional Gaussian. Stars represent the means of the Gaussians, ellipses their covariances. Reproduced with permission from Daunizeau et al. (2010).

optimizing the parameters by iterating through separate subsets of them while the distributions of those not currently being optimized are assumed known (for details, see Friston et al. 2007). Building on this, we were later able to show that the belief updates for this model—and much more complicated ones—can be reduced to precision-weighted prediction errors (Mathys et al. 2011). The principles that we used are entirely general and have since been applied to many more models (e.g., Iglesias et al. 2013; Diaconescu et al. 2014; Hauser et al. 2014; Vossel et al. 2014). The procedure is straightforward: formulate the model hierarchically using mixtures of exponential family distributions and use a mean field approximation to separate the levels of the hierarchy, which then allows you to derive precision-weighted prediction error updates at each of the levels separately.

Apart from multimodality, there is another property which, at first sight, exponential families seem unable to deliver. It is sometimes desirable to have a distribution with "fat tails" (or more formally, positive excess kurtosis). Fat tails imply an increased probability of extreme values compared to a Gaussian of the same mean and variance. Distributions with fat tails are popular because the predictions they imply are conservative in the sense that they guard against underestimating the probability of extreme events. Examples of such distributions are Student's t-distribution and the Cauchy distribution, neither of which

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

is an exponential family. The second limitation is the requirement of conjugate priors, and again the solution is a hierarchical approach. As the example from Daunizeau et al. (2010) shows, multimodal distributions, including priors, can be approximated using mixtures of distributions from exponential families. This allows for precision-weighted prediction error updating just as in the case of multimodal likelihoods.

Taken together, these limitations are not severe. All of them can be overcome by using a hierarchical approach. It might, therefore, be that the brain has evolved its hierarchical organization to take advantage of the effective and efficient predictive power of hierarchical models that are updated by precision weighting of prediction errors.

Nosologies Based on Aberrant Precision Weighting

Reducing the mind to precision-weighted belief updating implies nosologies based on false inference, owing to maladaptive weighting of prediction errors. In these terms, each nosological entity has two sides: a clinical manifestation of a particular precision-weighting disorder and a neurobiological mechanism underlying it. Steps in this direction have already been taken: Adams et al. (2013) traced out a computational anatomy of psychosis, and Lawson et al. (2014) and Quattrocki and Friston (2014) did the same with respect to autism. To identify new targets for treatments, these efforts will have to be expanded and refined with the goal of going beyond traditional diagnoses of, say, schizophrenia and autism, which lump together many disparate clinical phenomena and, we may suppose, pathophysiological mechanisms. The precision-weighting framework will help accomplish this because it tells us which questions to ask for each clinical phenomenon: Where in the inferential hierarchy is precision weighting going awry to produce this? What neurophysiological mechanism underpins the disordered precision weighting? Possible nosologies could be based on widespread aberrations in precision weighting originating in the neuromodulator systems of the brainstem and midbrain, equally widespread aberrations originating in the thalamus, or more localized aberrations originating in particular regions of the cortex or the basal ganglia, etc.

The best example of such an approach to date has been Adams et al. (2013), where many of the symptoms of psychosis are explained by a failure of sensory attenuation. Other pathologies will not be at the level of sensory input, but at other levels of the inferential hierarchy. For example, some symptoms of posttraumatic stress disorder (PTSD) could be a result of aberrant precision weighting when inferring the different possible causes of events in the environment. A loud bang is a prediction error for all of us, but while most will assign little precision to any one of the many possible explanations, enabling us to wait for more information before reacting, a PTSD patient might have very high precision on a prediction of being under fire. If this reaches such an extent that the patient is—unconsciously—constantly slightly surprised not to be

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.

under fire, a loud bang will be an opportunity for him to reduce this prediction error. Of course, this comes at the cost of a rather large prediction error about being in a safe environment. However, it depends on their relative precision, which of these contradictory predictions ("I am under fire" vs. "I am in a safe environment") dominates inference. Now that we have formulated the clinical side of the symptom in terms of precision weighting, albeit still in a very cursory and informal way, this enables us to look at the neurobiological side and know which questions to ask and how to interpret what we see. Specifically, when we investigate which neural systems are activated in PTSD patients in response to nonspecific stimuli that are over- or misinterpreted as threats, we can interpret what we see in terms of precision weighting. This could then give us a handle on manipulating precision weighting pharmacologically or psychotherapeutically, while monitoring progress neurobiologically as well as clinically. Crucially, this could enable us to transfer our interventions to other domains and—because we know the general *conceptual* mechanism in terms of which to interpret the underlying biology—allow us to make predictions about the clinical changes we expect in other domains.

Summary

In summary, I argue that the mind can only exist as a successful regulator of its environment if it continually updates model-based predictions about its interactions with that environment based on precision-weighted prediction errors. This is because the optimal way to make predictions is Bayesian inference, which can be reduced to tracking of sufficient statistics of observations under certain conditions. These conditions are that the likelihood be from an exponential family and that the prior be conjugate. These conditions are not restrictive because multimodal and fat-tailed (or otherwise nonstandard) distributions can be built hierarchically from exponential family distributions and inverted level by level by means of a mean field approximation. This amounts to a radical reduction of the mind to belief updating by means of precision-weighted prediction errors. The advantage of this reduction is that it provides terms in which both clinical phenomena and their underlying neurobiology can be understood. This enables it to serve as a bridge between the two fields and allows for the interpretation of one field's findings in terms of those of the other.

From "Computational Psychiatry: New Perspectives on Mental Illness,"
A. David Redish and Joshua A. Gordon, eds. 2016. Strüngmann Forum Reports, vol. 20,
series ed. J. Lupp. Cambridge, MA: MIT Press. ISBN 978-0-262-03542-2.