


# SCIENTIFIC REPORTS



OPEN

## Neuronal message passing using Mean-field, Bethe, and Marginal approximations

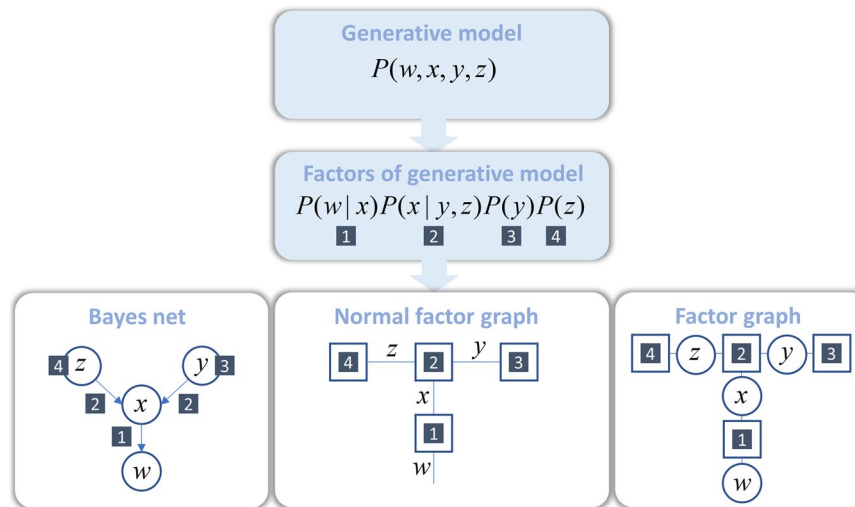
Thomas Parr<sup>1</sup>, Dimitrije Markovic<sup>2</sup>, Stefan J. Kiebel<sup>2</sup> & Karl J. Friston<sup>1</sup> 

Neuronal computations rely upon local interactions across synapses. For a neuronal network to perform inference, it must integrate information from locally computed messages that are propagated among elements of that network. We review the form of two popular (Bayesian) message passing schemes and consider their plausibility as descriptions of inference in biological networks. These are variational message passing and belief propagation – each of which is derived from a free energy functional that relies upon different approximations (mean-field and Bethe respectively). We begin with an overview of these schemes and illustrate the form of the messages required to perform inference using Hidden Markov Models as generative models. Throughout, we use factor graphs to show the form of the generative models and of the messages they entail. We consider how these messages might manifest neuronally and simulate the inferences they perform. While variational message passing offers a simple and neuronally plausible architecture, it falls short of the inferential performance of belief propagation. In contrast, belief propagation allows exact computation of marginal posteriors at the expense of the architectural simplicity of variational message passing. As a compromise between these two extremes, we offer a third approach – marginal message passing – that features a simple architecture, while approximating the performance of belief propagation. Finally, we link formal considerations to accounts of neurological and psychiatric syndromes in terms of aberrant message passing.

Recent work in theoretical neurobiology calls on the notion that the brain performs Bayesian inference<sup>1–5</sup>. This view treats perceptions as hypotheses about the causes of sensations<sup>6,7</sup>. Under this perspective, perceptual inference is the accumulation of evidence to confirm or refute various explanations for sensory data. As neuronal processing relies upon local signalling, the form of the inferences performed by the brain must involve the passing of local messages<sup>8</sup>. Here, we compare two forms of Bayesian message passing that have been used to explain cognitive phenomena. We consider their plausibility as accounts of neural processing, with a special focus on the anatomy of neural architectures that could implement these schemes. This calls for a set of criteria by which the plausibility of each scheme can be evaluated. Ultimately, this requires an evaluation of the evidence for each alternative process afforded by neurobiological data, considering prior constraints upon neural systems. Our focus here is upon the latter, and within this upon two important criteria. First, the computational architectures required for neuronal networks to perform inference should be as simple as possible. This is motivated by the spatial and metabolic constraints upon biological systems, and by Occam's razor (i.e. in trying to explain brain function, we should adopt the simplest explanation that is consistent with observed data). The second feature, which must be balanced against the first, is that these networks should be able to make accurate inferences about the causes of incoming sensory data.

This paper builds upon recent work that compares Bayesian message passing schemes in a simulated planning and decision making task<sup>9</sup>. The approach here complements this, but has a different focus. In this paper, our focus is upon the form and dynamics of the neuronal networks that are needed to perform inference under alternative message passing schemes. The novel aspects of this work include the specification of belief propagation in terms of a continuous gradient descent (for comparison with the dynamics previously used for variational message passing schemes<sup>10</sup>) and a neuronal network architecture that performs this gradient descent. This affords the opportunity to compare the dynamics of belief-updating under existing schemes. We then unpack a novel scheme –

<sup>1</sup>Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, WC1N 3BG, UK. <sup>2</sup>Chair of Neuroimaging, Psychology Department, Technische Universität Dresden, Dresden, Germany. Correspondence and requests for materials should be addressed to T.P. (email: [thomas.parr.12@ucl.ac.uk](mailto:thomas.parr.12@ucl.ac.uk))



**Figure 1.** A graphical representation of a probabilistic model. For a generative model, expressed as a joint probability distribution, it is possible to write down the associated factor graph by following a few simple steps. First, the model may be expressed in terms of the factors (prior and conditional distributions) that make up the joint probability. Square nodes are then associated with each of these factors. These nodes are connected whenever they are functions of the same random variable. The result is known as a normal factor graph<sup>13,20</sup>. For comparison, we present the same generative model expressed according to two alternative graphical representations. The Bayes net shown on the left places random variables in circular nodes and connects these with arrows corresponding to the conditional distributions. An alternative factor graph representation is shown on the right. This combines the normal factor graph formalism with that of the Bayes net; incorporating both factor and variable nodes. For the rest of this paper, we adopt the normal factor graph formalism as this provides a natural way to think about the form of local message passing.

marginal message passing, and argue that this offers a plausible compromise between the two criteria (simplicity and performance) considered above.

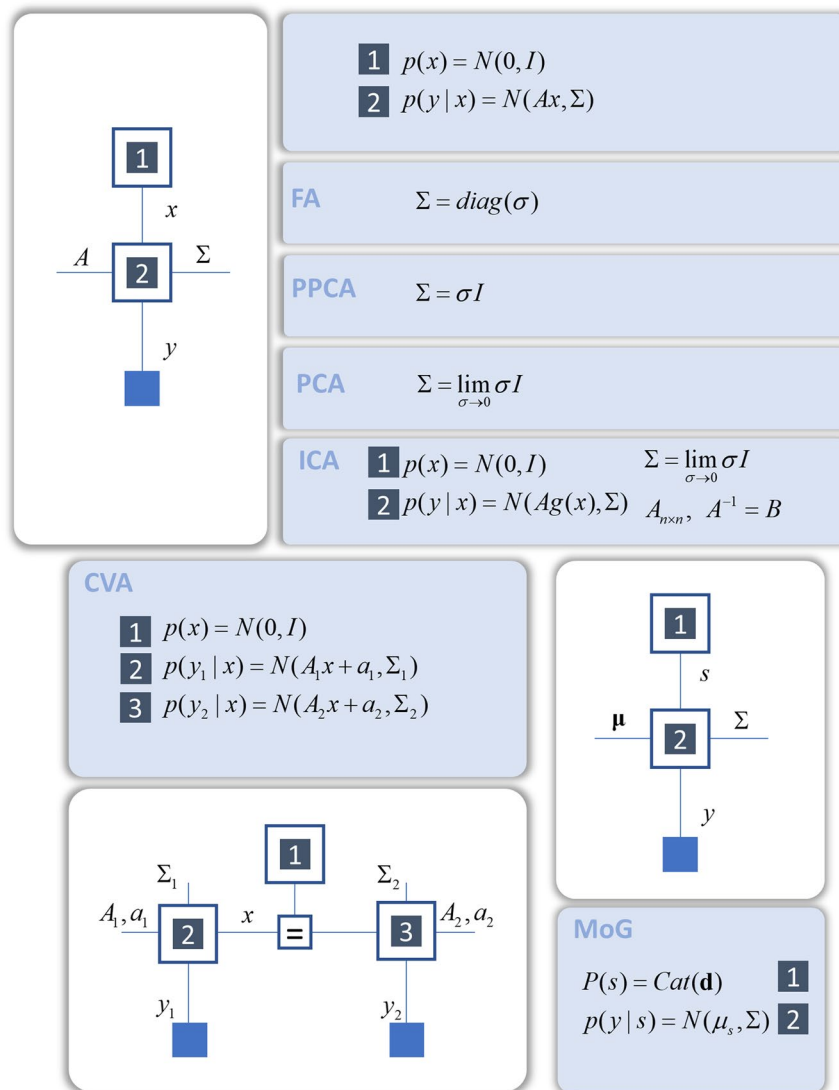
Local message passing schemes make use of simple update rules that can be applied to a probabilistic generative model<sup>11,12</sup>. The simplicity of these update rules can be illustrated using probabilistic graphical models – specifically, using normal factor graphs<sup>13–15</sup>. We will leverage the flexibility of these graphs in representing probabilistic models: this serves to illustrate that many common statistical inference procedures may be performed by passing local messages on factor graphs. However, different message passing algorithms differ in the approximate solution they provide – and in the computational complexity of the scheme. An interesting question then arises; which scheme might explain the ability of biological neural networks to perform statistical procedures such as blind source separation<sup>16</sup>. While we focus upon a hidden Markov model for illustrative purposes (as these have been used extensively in modelling perceptual inference e.g.<sup>17–19</sup>), the discussion here generalises to other generative models.

## Factor Graphs

We begin with an overview of inferential message passing, before specifying the message passing rules in detail. We consider plausible neuronal network architectures that could implement these schemes, and simulate the associated inferences. Finally, we consider the implications of each of these schemes for pathologies of neural computation.

To map a probabilistic generative model to a factor graph, we follow the steps illustrated in Fig. 1<sup>20</sup>. We start with a generative model that expresses a joint probability distribution over all the random variables in that model. We then factorise the joint probability distribution to show the conditional dependencies implied by the model. Each factor is represented graphically by a square node. If two factors are functions of the same random variable, we connect these factors with an edge representing that variable. Any probabilistic generative model can be specified in this way<sup>13</sup>. To illustrate the flexibility of factor graphs Figs 2 and 3 provides factor graph formulations of some commonly used generative models in data analysis and machine learning<sup>21</sup>. Inference about any of these models can be performed through local message passing algorithms.

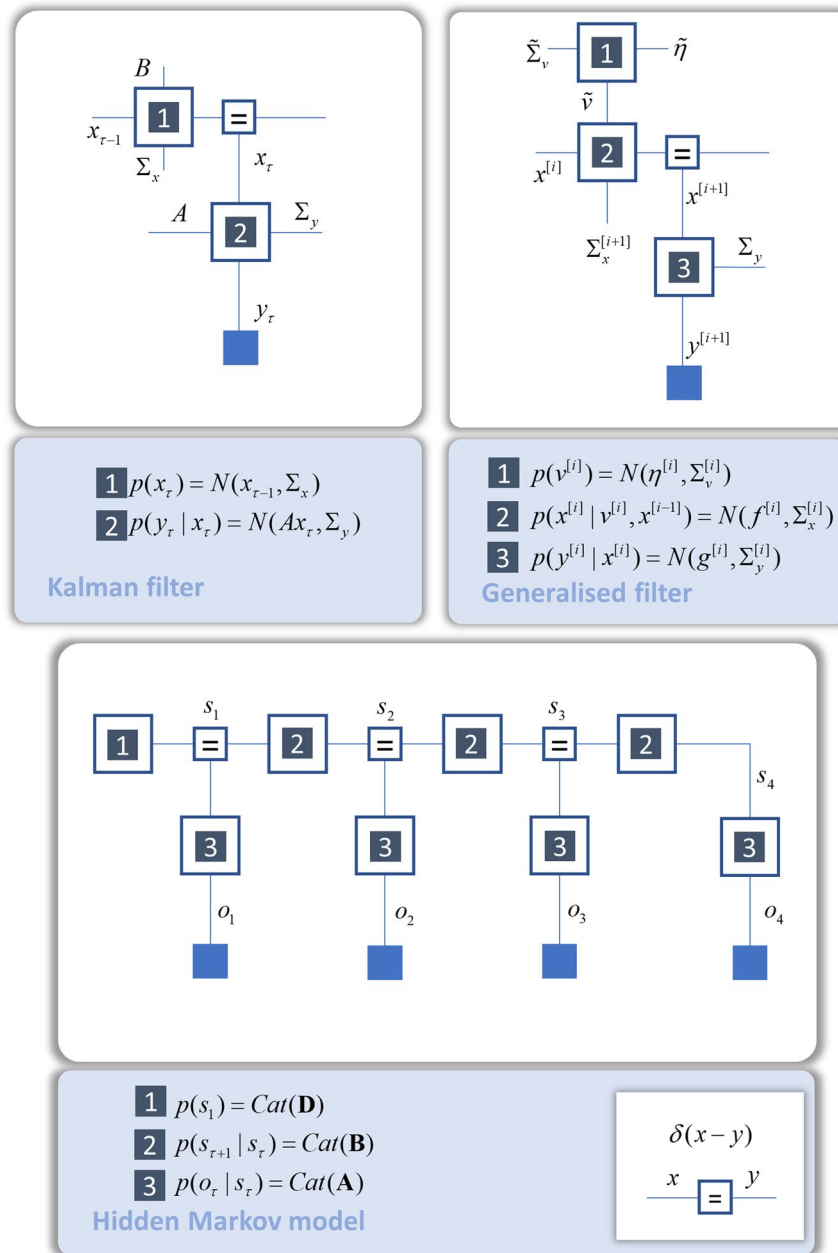
By message passing, we mean that each factor (square node) can synthesise the information coming in from one (or more) edge, and pass on this information in the form of a message along another edge. It is interesting to consider the relationship of this formal description of inference as message passing, and the informal notion that one part of the brain may communicate with another by sending a message. For the purposes of this paper, we equate the two, and associate the passing of messages between populations of neurons with inferential messages. These populations may be within a brain region, or may sit in different cortical areas. If the latter, the factor across which messages are passed is associated with the white matter tracts containing the axons that enable inter-areal communication.



**Figure 2.** Commonly used (static) generative models as factor graphs. This figure illustrates the generative models that underwrite many common inferential procedures. In each factor graph, small blue squares indicate observable data, while squares with an equality sign relate their adjoining edges via a delta function factor (shown explicitly in lower right inset in Fig. 3 - Hidden Markov model). Numbered squares relate factors to those in the probabilistic models in the blue panels. This Figure illustrates static models of the sort that underlie factor analysis (FA), probabilistic principle component analysis (PPCA), and principal component analysis (PCA)<sup>21</sup>. Each of these dimensionality reduction techniques relies upon the same generative model, but with different assumptions about the covariance structure of latent causes or sources. Adding in non-linear functions allows this generative model to be extended to incorporate independent component analysis (ICA)<sup>94</sup>, while using two different linear transforms leads to probabilistic canonical variates analysis (CVA)<sup>95</sup>. Incorporating discrete random variables gives mixture models including mixtures of Gaussians (MoG), which form the basis of many clustering algorithms<sup>96</sup>. The notation  $N$  indicates a normal (Gaussian) distribution, while  $Cat$  means a categorical distribution.

The connection between message passing on a factor graph and effective connectivity in the brain may seem a little difficult to intuit, in relation to some theoretical accounts of cortical communication. Communication between brain areas is sometimes framed in terms of oscillatory processes, where coherence of oscillations determines the influence of one set of neurons over another<sup>22</sup>. While the link between this and a message passed across a given factor may not be immediately apparent, the two may be reconciled if we treat the coherence as parametrising the precision (inverse variance) associated with this factor. If we imagine the connection between two brain regions represents factor 1 in Fig. 1, greater coherence between those regions representing  $x$  and those representing  $w$  would enhance the precision of  $P(w|x)$ , increasing the amount of information transmitted from factor 1 to the  $x$ -edge following an observation  $w$ .

Figure 2 shows the generative models that underwrite inferences about static latent variables. This includes dimensionality reduction techniques such as factor analysis and principal component analysis. From a generative



**Figure 3.** Commonly used (dynamic) generative models. This shows the dynamic generative models that support Kalman filtering<sup>97</sup> and generalised filtering<sup>98</sup> (the basis of predictive coding schemes<sup>99</sup> and inversion of dynamic causal models<sup>100</sup>). A discrete state space model that exhibits temporal dynamics is a hidden Markov model. This is the generative model we will take as our example for the remainder of this paper and, for this reason, we have expressed this in full, with a sequence of transitions over time. The notation  $N$  indicates a normal (Gaussian) distribution, while  $Cat$  means a categorical distribution.

modelling approach, these may be thought of as inferences about a low dimensional latent variable that generates relatively high dimensional data (with different assumptions about the covariance structure of this process). Similarly, independent components analysis, used to separate out data into different sets of causes, may be thought of as inferring the parameters of the mapping from a non-linear transform of a latent variable to data of the same dimension. Canonical variates analysis, a technique used to find linear transforms of two sets of multivariate data that render them maximally correlated, may be thought of as inferring the set of hidden states that best explain (i.e. could have generated) both datasets. Finally, clustering procedures, which are used to separate data into distinct clusters, are often based upon a ‘mixture-of-Gaussians’ generative model, that assumes data are generated from several different Gaussian distributions. Operations of this sort are important in inferring (and learning) the structure of our environment from sensory data. By framing these operations in terms of probabilistic inference, we can express them as local message passing procedures across appropriate factor graphs.

The procedures above may underwrite inferences the brain can draw about unchanging aspects of its environment. However, much of our environment is not static. As such, the brain must also make use of models like those of Fig. 3 that account for temporal dynamics. In the following, we focus upon Hidden Markov Models, as these provide a simple example of a dynamical generative model. Although we have chosen to illustrate the ideas in this paper using this example, we do not mean to imply that this is the best or only form of generative model used by the brain. Figures 2 and 3 make the point that, if the brain uses a particular local update rule defined on a factor graph, a whole range of inferential operations may be performed simply by applying these local rules to alternative factor graphs.

## Two Bayesian Message Passing Schemes

Bayesian message passing schemes work by passing messages from factor nodes (computed from the information at edges adjoining to that node) to each edge. The two messages arriving at each edge are multiplied together to obtain the posterior probability associated with the random variable at that edge. While these methods have found applications in engineering and machine learning<sup>23,24</sup>, we focus upon their biological implications. In the following, we consider two sorts of message. The first are those associated with belief propagation<sup>25,26</sup>. Often referred to as the ‘sum-product’ approach, belief propagation is a method used to perform exact Bayesian inference for marginal distributions on acyclic graphical models and approximate inference on cyclic graphs.

Belief propagation is at the heart of the circular inference account of neuronal computation<sup>4,27</sup>. Circular inference offers a biological implementation of belief propagation, and derives its name from the circular patterns of inhibitory connections it requires. This scheme additionally underwrites some theoretical accounts of the anatomy of cortical micro-circuitry<sup>28</sup>, and has been implemented in populations of simulated (spiking) neurons<sup>15,29</sup>.

A generic account of brain function that subsumes the Bayesian brain – and various forms of predictive processing such as predictive coding – is active inference<sup>30</sup>. This account derives from the imperative for living creatures to maximise Bayesian model evidence or, more simply, engage in self-evidencing<sup>31</sup>. This is equivalent to minimising their variational free energy<sup>3</sup>. One process theory associated with active inference<sup>32</sup> proposes that communication between populations of neurons occurs through an architecture based upon variational message passing<sup>12,33</sup>.

Both belief propagation and variational message passing have had some success in reproducing aspects of cognitive function, e.g.<sup>4,27,34–39</sup> but lead to rather different interpretations of false inference in neurological and psychiatric disorders.

## Inferential Message Passing

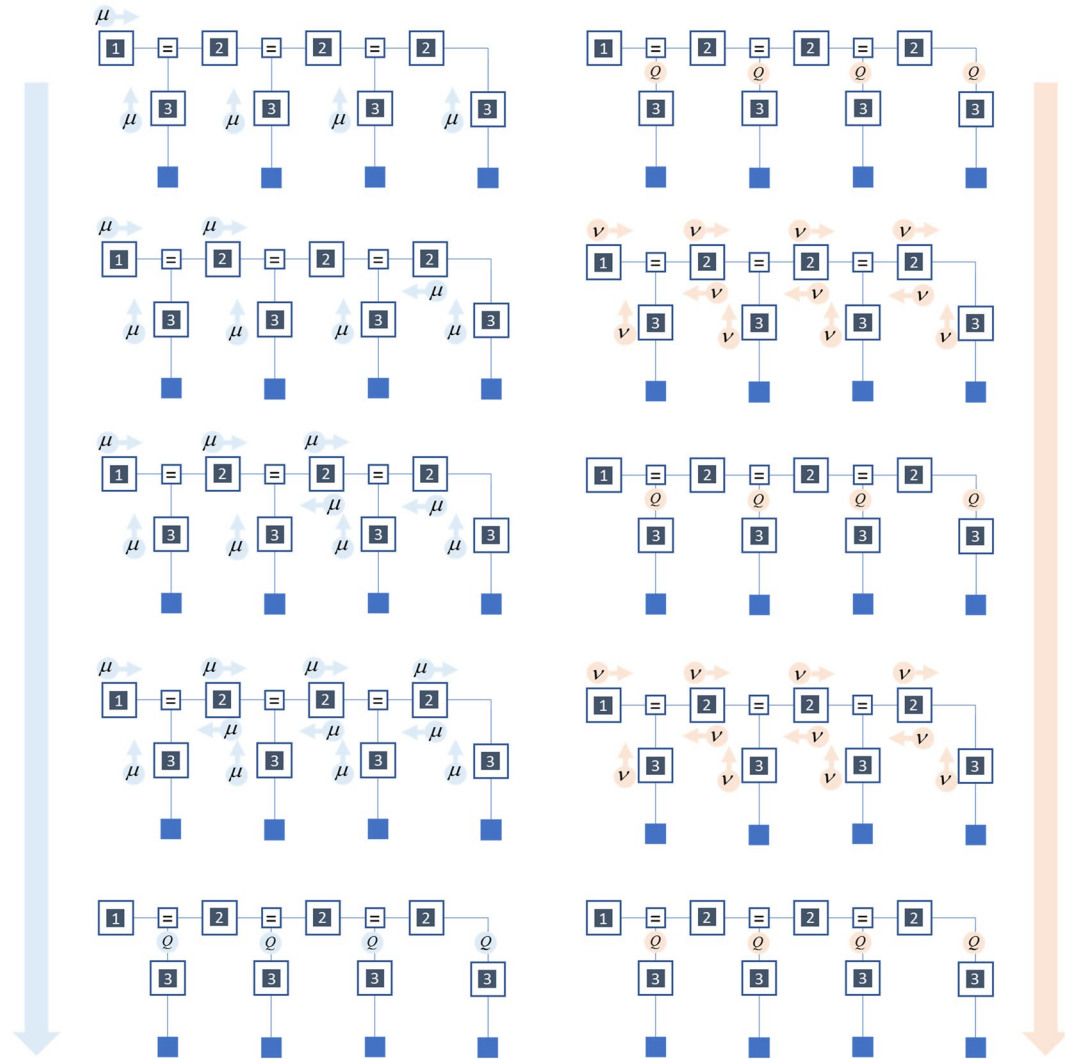
To illustrate the concepts we describe here, we use a Hidden Markov Model (HMM). This is a ubiquitous discrete state space model that also forms an important part of a Markov Decision Process. We chose the HMM as a showcase example for this study as it represents an important class of generative models used both in reinforcement learning<sup>40,41</sup> and active inference<sup>19,30,36</sup>. Both techniques have been used when modelling behaviour and, in general, are suited to describe processes that evolve through time – something that is crucial for biological (as well as robotic<sup>42</sup>) systems. The inferential message passing in an HMM takes a simple form that is related to schemes used in engineering, such as the Baum-Welch (forward-backward) algorithm<sup>23</sup>. The key aspect of these generative models is that hidden states are represented at each point in time over a sequence of outcomes. In other words, the hidden state at the beginning of a sequence is distinct from the same state at the end. Although we have selected an HMM for illustrative purposes, the two message passing methods we review here are applicable to any probabilistic generative model.

Figure 3 (lower half) shows the form of an HMM as a normal factor graph<sup>10,13,14,33</sup>. This is a representation of a joint probability distribution in terms of its factors. It involves two types of random variable – observable outcomes ( $o_t$ ) and hidden states ( $s_t$ ). Hidden states evolve through time in a Markov chain. This means that each state depends only upon the state at the previous time. At each time, hidden states give rise to an outcome. The sparsity of conditional dependencies in this (and other) generative models allows for efficient local message passing schemes to be derived. This is because the messages used to compute beliefs about a variable come only from the constituents of the variable’s Markov blanket<sup>11</sup>. The Markov blanket of a given hidden state in an HMM contains the state in the immediate past, the state in the immediate future, and observable data at the present.

This is illustrated in Fig. 4, where messages are indicated as arrows across factor nodes (large squares) to the edges (lines connecting factors) that represent the random variables. Each message can be computed from locally available information. The normalised product of incoming messages to an edge is the approximate posterior probability distribution over the random variable represented by that edge. The following sections overview two established message passing schemes – *belief propagation* and *variational message passing*. In addition, we introduce a third scheme – *marginal message passing* – that combines some of the key advantages of the previous two. We consider biologically plausible neuronal networks that could realise these schemes and simulate their behaviour when presented with sequential observations.

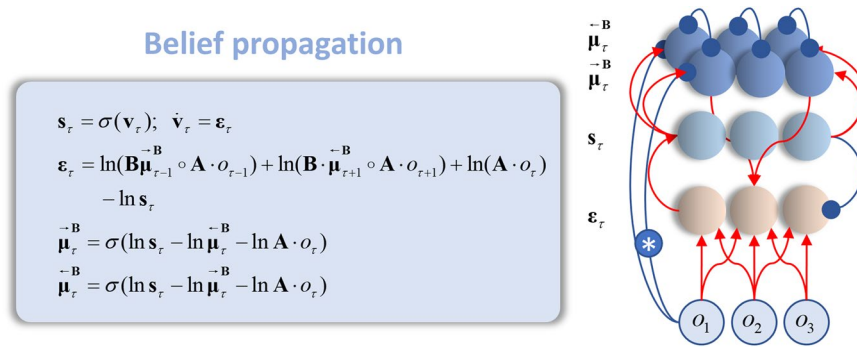
## Belief Propagation

Sum-product belief propagation arises naturally in directed acyclic graphs<sup>25</sup> but can also be applied to cyclic graphs as the belief update rules correspond to the fixed points of the Bethe free energy<sup>26</sup>, see the appendix for details. The message passing can be expressed in the following way (with messages  $\mu$  used to compute approximate posterior beliefs  $Q$  about states  $s$  at time  $\tau$ )



**Figure 4.** Message passing in a hidden Markov model. This schematic illustrates the scheduling of belief propagation (left) and variational message passing (right) in a hidden Markov model. Each row shows a single step in a round of message passing, ordered from top to bottom. Under belief propagation, for a message to be sent across a factor to an edge, the factor requires all of the other adjoining edges to provide a message. Initially, in the first step, this is only true for likelihood factors (that compute their message from the data), and priors (that are associated with just one edge). This enforces a strict scheduling that starts with the computation of messages at the extremities of the graph. More proximal factors then use these messages from the extremes to compute their own. Eventually, when all factors have passed their message, the incoming messages to each edge can be combined to compute the marginal posterior belief ( $Q$ ) about the associated random variable (last row). For directed acyclic graphs, one round of message passing is sufficient. For cyclic graphs, multiple rounds may be required. In contrast, variational message passing computes messages from the current beliefs associated with each edge, and not from other incoming messages. This means that variational message passing simply alternates between message passing (of all messages in parallel) and updating posterior expectations.

$$\begin{aligned}
 Q(s_\tau) &\approx P(s_\tau|\tilde{o}) \propto P(o_\tau|s_\tau) \sum_{s_1, s_2, \dots, s_{\tau-1}} P(s_{t<\tau}, o_{t<\tau}) \sum_{s_{\tau+1}, s_{\tau+2}, \dots, s_T} P(s_\tau, s_{t>\tau}, o_{t>\tau}) \\
 Q(s_\tau) &\propto \mu_A(s_\tau) \cdot \vec{\mu}_B(s_\tau) \cdot \overleftarrow{\mu}_B(s_\tau) \\
 \mu_A(s_\tau) &= P(o_\tau|s_\tau) \\
 \vec{\mu}_B(s_\tau) &= \sum_{s_{\tau-1}} P(s_\tau|s_{\tau-1}) \vec{\mu}_B(s_{\tau-1}) \mu_A(s_{\tau-1}) \\
 \overleftarrow{\mu}_B(s_\tau) &= \sum_{s_{\tau+1}} P(s_{\tau+1}|s_\tau) \overleftarrow{\mu}_B(s_{\tau+1}) \mu_A(s_{\tau+1})
 \end{aligned} \tag{1}$$



**Figure 5.** Belief propagation as neuronal message passing. The equations on the left show the form of belief propagation (Equations 2 and 3) when expressed in a neuronally plausible form. These equations are written in terms of the sufficient statistics of the probability distributions and auxiliary variables representing prediction errors ( $\epsilon_\tau$ ) and membrane potentials ( $v_\tau$ ). The softmax ( $\sigma$ ) functions act as neuronal transfer functions, converting presynaptic potentials to firing rates ( $s_\tau$ ), which represent the sufficient statistics of the posterior beliefs. Forwards and backwards messages across the transition factors ( $\mathbf{B}$ ) are written as  $\overset{\leftarrow}{\mu}_\tau, \overset{\rightarrow}{\mu}_\tau$  respectively. Red indicates an excitatory connection, while blue is inhibitory. The starred connection represents the subtraction of the ascending message from that passed on to other neuronal populations. This plays the role of an ascending ‘loop’, as in circular inference accounts of neuronal computation<sup>4,37</sup>. The analogous descending loops are the inhibitory connections between the neurons representing messages in opposite directions. This formulation assumes that the neurons representing the messages have much shorter time constants than those representing marginal beliefs, allowing the former to be ‘enslaved’ by the latter<sup>15,101</sup>. Although omitted here (and in later figures) for simplicity, the normalisation induced by the softmax functions could be mediated via recurrent inhibitory connections within a layer of neurons.

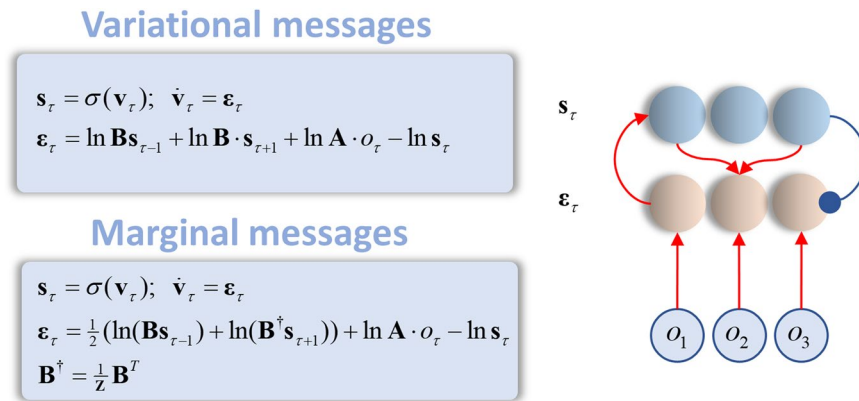
The first line of this equation uses an approximate equality, as  $Q$  is not always equal to the marginal posterior (although it is in the absence of cycles in a graph). The second line expresses a proportional relationship, as the product of messages needs normalisation. For consistency with the HMM shown in Fig. 3, we use a subscript A to indicate the messages being passed across a likelihood factor, and subscript B to indicate those passed across those factors representing transition probabilities (with a right-pointing arrow indicating a message derived from beliefs about the past, and a left-pointing arrow indicating a message from the future). One way to think about this scheme is that a message from a given region of the factor graph is the partition function of that region<sup>43</sup>. Each partition function is computed using the partition function of a sub-region within that region, and so on. The recursive computation of these messages would be problematic for a network of neurons representing marginal beliefs, as messages are derived from other messages to an edge (neuronal population), not from the marginal belief (neuronal activity) itself. This imposes strict constraints on the scheduling of message passing as illustrated in Fig. 4, where a factor does not pass a message on to a given edge until it has received messages from all the other edges connected to it. Technically, this constraint can be removed by using a ‘loopy’ belief propagation scheme<sup>44</sup>, or equivalently to rearrange the equations above so that messages depend upon the marginal<sup>4</sup>.

$$\overset{\rightarrow}{\mu}_B(s_\tau) \propto \exp(\ln Q(s_\tau) - \ln \mu_A(s_\tau) - \ln \overset{\leftarrow}{\mu}_B(s_\tau)) \tag{2}$$

To retain the conditional dependencies of the generative model, we update the marginals through the following equation (which is obtained from the second line of Equation 1 by substituting in the message definitions on the final three lines)

$$Q(s_\tau) \propto \exp\left(\ln P(o_\tau | s_\tau) + \ln E_{\overset{\rightarrow}{\mu}_B(s_{\tau-1}) \mu_A(s_{\tau-1})} [P(s_\tau | s_{\tau-1})] + \ln E_{\overset{\leftarrow}{\mu}_B(s_{\tau+1}) \mu_A(s_{\tau+1})} [P(s_{\tau+1} | s_\tau)]\right) \tag{3}$$

The ‘expectation’ notation is used heuristically here, as the messages are not probability distributions. We use this notation to mean a summation, where every element of the sum is weighted by the subscripted term (i.e. a linear combination of the terms within the expectation). Note that we could have written the above more simply as a product of the terms within the logarithms (similar to the second line of Equation 1). However, we have opted to express this as an exponential of the sum of three logarithms for consistency with the form of the equations for the other message passing schemes presented later. Once we have expressed the equations in this form, the marginals begin to play an important part in the message passing. Expressing the updates in the form of a gradient descent, we arrive at neuronally plausible updates as shown in Fig. 5. To obtain these equations we compute an error term ( $\epsilon$ ) that is the difference between (the log of) the current estimate of the posterior probability ( $s$ ) and the right hand side of Equation 3. We then construct a differential equation that changes  $s$  and has an  $\epsilon$  of zero at its fixed (attracting) point. The softmax (normalised exponential) functions have a high degree of biophysical plausibility, as the density dynamics of spiking neuron populations have a similar form<sup>45</sup>, where synaptic input is converted into firing rates that can be propagated along axons to other neural populations. For an interpretation of belief propagation in terms of spiking neurons, see<sup>46,47</sup>. The probability matrices now become connectivity matrices, lending a clear biological interpretation to the inferential equations above. Note that the gradient



**Figure 6.** Variational (and marginal) message passing. The equations on the left show variational message passing (upper panel) and marginal message passing (lower panel) expressed as gradient descents on the variational (or marginal) free energy. These equations are implemented by the neuronal network shown on the right. Notably, this is much simpler than the network of Fig. 5, requiring fewer neurons and a simpler connectivity structure. The primary reason for the simplicity of this structure is that these schemes take into account the current marginal beliefs of adjacent variables. The messages do not need to be recursively computed from other messages, Fig. 5. This limits the number of auxiliary variables required. We have introduced the notation  $\mathbf{B}^\dagger$  for the transpose of the transition matrix with normalised columns.

descent described by the differential equation occurs over a faster time scale than the frequency at which observations change. Biologically, this is consistent with things like the fast neuronal processing (gradient descent) that intervenes between saccadic eye movements (mediating changes in sensory input).

The use of separate populations of neurons that represent messages and marginals resembles previous accounts of belief propagation in neuronal networks<sup>15,48</sup>. An alternative is that one abandons any explicit representation of marginal probabilities, and that neurons represent the messages<sup>15,49</sup> only. This interpretation has two drawbacks. First, it runs into the same scheduling issues described above. Second, explicit representations of the marginal beliefs are essential in performing inferential operations including model comparison, model selection, and model averaging. This is because these operations require evaluation of approximations to model evidence (and expected model evidence) that depend upon these posteriors. Importantly, to properly estimate model evidence as a minimum of the Bethe free energy (see below) besides the singleton marginals, a neuronal network implementing belief propagation would also have to represent the pairwise marginals, which would add additional degrees of complexity to the network illustrated in Fig. 5. Model comparison, model selection, and model averaging are thought to underwrite the evaluation of behavioural policies that support active engagement with the sensorium<sup>50,51</sup> and inference with hierarchical models<sup>10</sup>.

### Variational Message Passing

Variational message passing takes a superficially similar form to belief propagation. Marginals of the posterior distribution are computed by the product of messages from neighbouring factors<sup>12,33</sup>.

$$\begin{aligned} Q(s_\tau) &\propto \nu_A(s_\tau) \cdot \vec{\nu}_B(s_\tau) \cdot \overleftarrow{\nu}_B(s_\tau) \\ \ln \nu_A(s_\tau) &= \ln P(o_\tau | s_\tau) \\ \ln \vec{\nu}_B(s_\tau) &= E_{Q(s_{\tau-1})} [\ln P(s_\tau | s_{\tau-1})] \\ \ln \overleftarrow{\nu}_B(s_\tau) &= E_{Q(s_{\tau+1})} [\ln P(s_{\tau+1} | s_\tau)] \end{aligned} \quad (4)$$

In contrast to belief propagation, the messages ( $\nu$ ) here are derived from the posterior marginal beliefs at each edge. This means that we do not need to wait for all of the incoming messages. Instead, we can iterate between computing messages (at all factor nodes in parallel) and updating posterior beliefs. Using a gradient ascent ensures we can combine these steps into a single differential equation, as in Fig. 6, without any need to manipulate the form of the messages as with belief propagation. This leads naturally to the simple neuronal network illustrated here; for which connectivity matrices are log probabilities. The structure in Fig. 6 forms part of the cortical microcircuit proposed for active inference in Markov decision processes<sup>52</sup>, and can be extended for generative models associated with precision parameters<sup>53</sup>, and for continuous state space models<sup>10</sup>.

If we used a generative model that contained  $n$  types of hidden state, that could take on  $m$  possible values, for  $t$  time-steps, the architecture of Fig. 6 would require  $2 \times n \times m \times t$  neuronal populations. On comparing this to the  $4 \times n \times m \times t$  populations required for the architecture of Fig. 5, it is clear that there is a substantial saving to adopting the architecture of Fig. 6 for any sizable generative model. Appealing to minimum wire length principles<sup>54</sup> and taking note of the metabolic<sup>55</sup> and therefore informational<sup>56</sup> costs of individual neurons, this network has a substantial structural advantage over that given by belief propagation. However, dendritic computations<sup>57</sup> may have the potential to rescue belief propagation in this regard; i.e., it is possible that forwards and backwards



messages could be represented in different parts of the dendritic tree, eliminating the need for additional ‘message’ neurons.

### Marginal Message Passing

A third approach to inference combines the simplicity offered by variational message passing with the sophistication of belief propagation. This aims to approximate the messages from the former, but to do so using only the locally available marginal beliefs. This was first described as an alternative to variational message passing for active inference in Appendix C of<sup>32</sup>. Marginal message passing recapitulates the pattern from Equations 1 and 4, with marginal posteriors expressed as the product of messages ( $\eta$ ) from their Markov blanket:

$$\begin{aligned} Q(s_\tau) &\propto \eta_A(s_\tau) \cdot \overrightarrow{\eta}_B(s_\tau) \cdot \overleftarrow{\eta}_B(s_\tau) \\ \ln \eta_A(s_\tau) &= \ln P(o_\tau | s_\tau) \\ \ln \overrightarrow{\eta}_B(s_\tau) &= \frac{1}{2} \ln E_{Q(s_{\tau-1})}[P(s_\tau | s_{\tau-1})] \\ \ln \overleftarrow{\eta}_B(s_\tau) &= \frac{1}{2} \ln E_{Q(s_{\tau+1})}[P(s_\tau | s_{\tau+1})] \end{aligned} \quad (5)$$

Like belief propagation, the expectation sits within the logarithm. Like variational message passing, the messages are derived from adjacent marginals. A comparison with equation 4 shows that this approach implicitly assumes the following relation.

$$\frac{1}{2} \ln E_{Q(s_{\tau-1})}[P(s_\tau | s_{\tau-1})] \approx \ln \frac{E_{P(s_{\tau-1}|o_1, o_2, \dots, o_{\tau-1})}[P(s_\tau | s_{\tau-1})]}{\overrightarrow{\mu}_B(s_\tau)} \quad (6)$$

Intuitively, as  $Q(s_{\tau-1})$  approximates the posterior following all available observations, it will be more precise (i.e. have a lower Shannon entropy) than  $P(s_{\tau-1}|o_1, o_2, \dots, o_{\tau-1})$ . This is because the latter represents a partial posterior that considers only past observations. Halving the log message computed using the approximate posterior reduces the precision of the resulting message, better approximating the belief propagation message. The motivation for using  $\frac{1}{2}$  will become more apparent when we describe the underlying free energy functional in the next section. However, it would also be possible to treat this corrective factor as a parameter that itself could be optimised in relation to data. Variational message passing fails to attenuate the precision of this message and leads to overconfidence in estimating posteriors, as we will illustrate below using simulations. The marginal approach retains the simplicity of the variational architecture but eludes this overconfidence issue. An interesting feature of this scheme is that backwards messages use transitions from the future to the past – something not seen in the other two approaches. We will unpack this in more detail in the following section.

### Model Evidence and Free Energies

Both variational message passing, and belief propagation can be shown to represent fixed points for approximations to model evidence<sup>26</sup>. In each case, these approximations take the form of free energy functions. In this section, we briefly outline the relationship between free energy and model evidence. We then specify the free energies that act as the landscapes upon which these inferential optimisations take place. In brief, the variational free energy (under the mean-field approximation) approximates model evidence using a relatively simple form for the approximate posterior, in which one assumes no interactions between random variables. For belief propagation, the Bethe approximation uses a more sophisticated approximate posterior which takes into account pairwise interaction, but under specific conditions sometimes found in cyclic graphs may lead to erroneous estimates of the free energy.

Given that belief propagation may be motivated as in Equation 1, it might seem a little redundant to additionally motivate it in terms of the Bethe approximation. However, the Bethe approximation is crucial in understanding how these different message passing schemes relate to one another – as all are free energy minimising schemes that maximise a lower bound on model evidence. It is also important in understanding the approximations that belief propagation makes in a general setting, and in justifying the generalisation of belief propagation to settings beyond acyclic graphs.

In a typical inference problem, one is interested in determining posterior beliefs over hidden states  $\vec{s} = (s_1, \dots, s_T)$  given some set of observations  $\vec{o} = (o_1, \dots, o_T)$  using Bayes’ rule

$$P(\vec{s} | \vec{o}) = \frac{P(\vec{o}, \vec{s})}{P(\vec{o})} \quad (7)$$

For a general inference problem, the above relation is analytically intractable. First, the denominator on the right-hand side (also known as model evidence or marginal likelihood) can be only estimated using approximate numerical methods. Second, the posterior probability distribution  $P(\vec{s} | \vec{o})$  might not have a known analytic form. Variational inference resolves these difficulties using the following approximate scheme for probabilistic inference: (i) Map the true posterior to a tractable parametric family of probability distributions  $Q(\vec{s})$ . (ii) Find the approximate estimate of the true posterior at the minimum of a free energy approximation to the negative model evidence.

Model evidence is related to free energy through Jensen’s inequality<sup>58</sup>.

$$\underbrace{-F}_{\text{Negative Free Energy}} = \underbrace{E_Q \left[ \ln \frac{P(\tilde{o}, \tilde{s})}{Q(\tilde{s})} \right]}_{\text{Jensen's inequality}} \leq \ln E_Q \left[ \frac{P(\tilde{o}, \tilde{s})}{Q(\tilde{s})} \right] = \underbrace{\ln P(\tilde{o})}_{\text{logevidence}} \quad (8)$$

The first of these equations indicates that the (negative) free energy is a lower bound on the evidence for a generative model (known in machine learning as an evidence lower bound or ELBO).

We are concerned with the optimisation of this bound, that is, with finding  $Q(\tilde{s})$  which minimises free energy. The minimum of the free energy corresponds to the best approximation to the true posterior and closest estimate of the model evidence, within the given family of probability distributions. A rearrangement of the terms in the left-hand side of the inequality above gives

$$\begin{aligned} F &= \underbrace{-\ln P(\tilde{o})}_{\text{Evidence}} + \underbrace{D_{KL}[Q(\tilde{s})||P(\tilde{s}|\tilde{o})]}_{\text{Divergence}} \\ &= \underbrace{-E_Q[\ln P(\tilde{o}, \tilde{s})]}_{\text{Energy}} - \underbrace{H[Q(\tilde{s})]}_{\text{Entropy}} \end{aligned} \quad (9)$$

The first line shows that the bound between the free energy and model evidence is the KL-Divergence between  $Q$  and the posterior distribution (i.e. the best approximation to the posterior is at the free energy minimum). The second shows the free energy expressed as an energy minus entropy (Shannon entropy of the approximate posterior). The minimum of the free energy  $F$  is obtained for  $Q(\tilde{s}) = P(\tilde{s}|\tilde{o})$ , in which case the free energy is equal to the negative log-model evidence. However, a difficulty here is to find a good approximation to the true posterior which makes computation of both the energy and the entropy analytically tractable<sup>59,60</sup>.

The mean-field and the Bethe approximations choose different forms for the distribution  $Q$ . The mean-field approximation<sup>61</sup> assumes fully factorised posterior probability (although the same principles apply to structured mean-field factorisations<sup>33</sup>).

$$Q(\tilde{s}) = \prod_{\tau} Q(s_{\tau}) \quad (10)$$

The Bethe approximation is more nuanced, and accounts for the pairwise interactions between variables.

$$Q(\tilde{s}) = \prod_{\tau} Q(s_{\tau}) \prod_{\tau, \tau-1} \frac{Q(s_{\tau}, s_{\tau-1})}{Q(s_{\tau})Q(s_{\tau-1})} \quad (11)$$

Returning to the energy-entropy expression, we can write the variational free energy (for a HMM) as

$$\begin{aligned} F &= \underbrace{-\sum_{\tau} (E_{Q(s_{\tau})Q(s_{\tau-1})}[\ln P(s_{\tau}|s_{\tau-1})] + E_{Q(s_{\tau})}[\ln P(o_{\tau}|s_{\tau})])}_{\text{Energy}} \\ &\quad - \underbrace{\sum_{\tau} H[Q(s_{\tau})]}_{\text{Entropy}} \end{aligned} \quad (12)$$

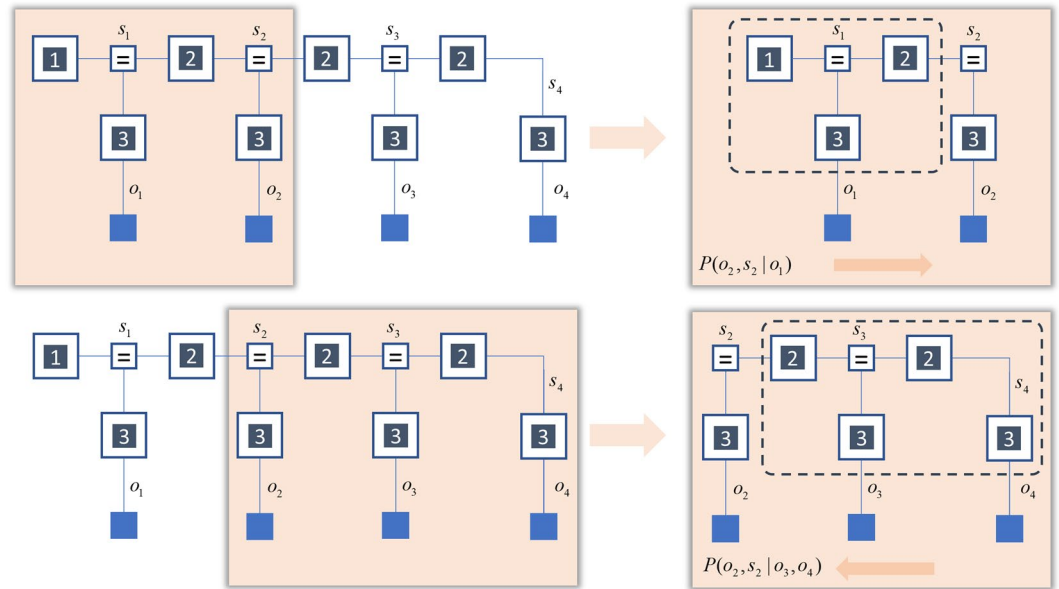
We then write the Bethe free energy (free energy under the Bethe approximation) as

$$\begin{aligned} F &= \underbrace{-\sum_{\tau} (E_{Q(s_{\tau}, s_{\tau-1})}[\ln P(s_{\tau}|s_{\tau-1})] + E_{Q(s_{\tau})}[\ln P(o_{\tau}|s_{\tau})])}_{\text{Energy}} \\ &\quad - \underbrace{\sum_{\tau} (H[Q(s_{\tau})] - D_{KL}[Q(s_{\tau}, s_{\tau-1})||Q(s_{\tau})Q(s_{\tau-1})])}_{\text{Bethe entropy}} \end{aligned} \quad (13)$$

Although the energy component of the Bethe free energy preserves pairwise interactions between temporally proximate states in the expectation, this makes the entropy term a little more complicated. We can express the Bethe entropy in terms of the entropy of the marginal factors, but then subtract the mutual information between these and the joint factors. This approximates the true entropy (entropy of the exact posterior) by taking into account only the pairwise interactions between the variables and ignoring any higher order dependencies. For directed acyclic graphs (of the sort considered in this paper), the Bethe energy is always exact and the Bethe entropy, although approximate, will always correspond to the entropy of the joint probability distribution  $Q(\tilde{s})$ .

However, for a cyclic graph the Bethe energy is approximate, and the Bethe entropy might return suboptimal estimates of the entropy of the joint posterior, under certain conditions. This can happen when the solutions that satisfy the relation between the singleton  $Q(s_t)$  and pairwise marginals  $Q(s_t, s_{t-1})$  are implausible for a given cyclic graph. Importantly, in such cases, the Bethe free energy might produce strange behaviour (e.g. convergence to a limit cycle or improbable configurations of the marginal posterior) and anomalous free energy estimates. To mitigate these issues, higher order approximations have been proposed that are based on cluster variational methods and the Kikuchi approximation<sup>59,62,63</sup>.

The reason for the differences in convergence behaviours in mean-field and Bethe approaches is related to the convexity (or non-convexity) of their respective free energy functionals, specifically the entropy terms. A negative entropy is a convex functional, with a positive curvature. However, the negative Bethe entropy has contributions



**Figure 7.** Marginal (forward-backward) models. This schematic illustrates the steps that motivate the marginal free energy. On the left, we show an HMM that is divided in two different ways. For the future part (lower row), we reverse the direction of the transition probabilities by normalising with respect to the earlier times. On the right, we take these partitioned generative models and sum over all variables within the dashed boxes. This leads to two, marginal, generative models – one that progresses from the past to the future, and one that reverses this. By approximating the free energies for each model, and mixing them in equal parts, we define a marginal free energy with empirical priors that are independently constrained by the future and the past.

from the negative pairwise (convex) and positive singleton (non-convex) entropies. While overlapping pairwise marginals are sufficient to characterise the posterior, the Bethe entropy will always be convex. If interactions between three or more variables contain information that cannot be captured in overlapping pairwise interactions, the singleton entropies could dominate the pairwise entropies in some parts of the free energy landscape, inducing non-convexities that impede convergence. The mean-field entropy is not subject to this problem, as it comprises only positive entropy terms. This is a slightly simplistic explanation, to aid intuition. Interested readers are referred to<sup>60,64</sup> for more formal treatments of this issue.

Despite these convergence issues, the Bethe free energy is often a better approximation to the log evidence than the variational free energy (when a mean-field approximation is employed). Under the mean-field approximation both the energy and entropy terms are approximations of the energy and entropy terms that would be obtained by setting the approximate posterior equal to the true posterior. For this reason, we ideally want to make inferences that are as close as possible to those obtained using belief propagation. The marginal free energy offers a way to do this, while retaining the architecture of variational message passing. Marginal message passing (Equation 5) is the scheme obtained at the fixed point of the marginal free energy.

Unlike the mean-field or Bethe approaches, marginal message passing makes no claim as to the form of the full posterior belief. Instead, it relies upon locally defined free energy functionals to optimise marginals of the posterior at each time, while remaining agnostic about how these combine to form a global free energy. Figure 7 illustrates the idea behind this functional. First, we divide the generative model into two overlapping parts – past and future – around the variable we wish to estimate. We then sum (or integrate) over all other hidden states. This leads to two marginal generative models, the first with an empirical prior derived from the past and the second with an empirical prior derived from the future.

$$\begin{aligned}
 P(o_\tau, s_\tau | o_1, \dots, o_{\tau-1}) &= P(o_\tau | s_\tau) \underbrace{E_{P(s_{\tau-1} | o_1, \dots, o_{\tau-1})}}_{P(s_\tau | o_1, \dots, o_{\tau-1})} [P(s_\tau | s_{\tau-1})] \\
 P(o_\tau, s_\tau | o_{\tau+1}, \dots, o_T) &= P(o_\tau | s_\tau) \underbrace{E_{P(s_{\tau+1} | o_{\tau+1}, \dots, o_T)}}_{P(s_{\tau+1} | o_{\tau+1}, \dots, o_T)} [P(s_\tau | s_{\tau+1})]
 \end{aligned}
 \tag{14}$$

To ensure that the empirical prior from the future sums to one, we have normalised the transition probabilities so that the future causes the past. This implies the HMM could be run in reverse, consistent with the conservation of probability mass. Although the empirical priors cannot be computed without resorting to the recursive approach of belief propagation, we can approximate these, to give the following (forwards and backwards) free energies.

$$\begin{aligned}
 F_F(\tau) &= \underbrace{-E_{Q(s_\tau)}[\ln P(o_\tau|s_\tau) + \ln E_{Q(s_{\tau-1})}[P(s_\tau|s_{\tau-1})]]}_{\text{Energy}} - \underbrace{H[Q(s_\tau)]}_{\text{Entropy}} \\
 F_B(\tau) &= \underbrace{-E_{Q(s_\tau)}[\ln P(o_\tau|s_\tau) + \ln E_{Q(s_{\tau+1})}[P(s_\tau|s_{\tau+1})]]}_{\text{Energy}} - \underbrace{H[Q(s_\tau)]}_{\text{Entropy}}
 \end{aligned}
 \tag{15}$$

We conjecture, but offer no proof for, the inequality:

$$F_B(\tau) + F_F(\tau) \geq -E_{Q(s_\tau)}[\ln P(o_\tau, s_\tau|\delta_{\setminus\tau})] - H[Q(s_\tau)] \tag{16}$$

This suggests we can define an (approximate) marginal free energy as the mixture of forwards and backwards free energies.

$$F(\tau) = -E_{Q(s_\tau)}\left[\frac{1}{2} \ln E_{Q(s_{\tau-1})}[P(s_\tau|s_{\tau-1})] + \frac{1}{2} \ln E_{Q(s_{\tau+1})}[P(s_\tau|s_{\tau+1})] + \ln P(o_\tau|s_\tau)\right] - H[Q(s_\tau)] \tag{17}$$

This can then be optimised at each time-step. Minimising the marginal free energy can be thought of as applying variational filters in forwards and backwards directions and combining the results. This is subtly different to the mean-field and Bethe approaches, that each apply a single Bayesian smoother. The mixture in the marginal approach ensures we do not overestimate the precision of forwards or backwards messages, as could happen when optimising a mean-field posterior.

We hope that we will be able to provide a formal proof of Equation 16 in future work, and to be able to specify the conditions under which (if any) the inequality may fail. However, even in the absence of this, it is possible to motivate Equation 17 on heuristic grounds. The overconfidence of variational message passing depends upon the way in which beliefs about factors of the approximate posterior constrain one another. In Equation 17, the contribution of terms that depend upon other factors has been attenuated relative to those that do not mediate this influence. Notably, this means that the entropy of posterior beliefs (final term) offers a greater contribution to the free energy gradients than it would under a mean-field approach. This favours solutions with more uncertainty (i.e. minima with lower curvature in the free energy landscape<sup>65</sup>) than can be achieved through variational message passing; thereby, finessing the overconfidence problem.

Belief propagation represents the message passing scheme obtained at the stationary point of the Bethe free energy, variational message passing is the scheme found at the stationary point of the variational free energy, and marginal message passing represents the minimum of the marginal free energies at each time (see Appendix). On acyclic graphs they all act as lower bounds for the evidence for a model and could be utilised by a self-evidencing biological system. As such, all three forms of neuronal message passing are consistent with the principles that underwrite active inference<sup>9</sup>. The efficiency of these algorithms makes them especially suitable in biological settings, as they can ensure the minimisation of the free energy, and its time integral in an efficient manner.

## Simulations

To compare the behaviour of each of these message passing schemes during online inference, we simulated their responses while presenting data sequentially. In other words, the scheme accumulates evidence for different hidden states by assimilating successive outcomes into posterior beliefs. We used an HMM employing the probability distributions specified in Fig. 8. This contains two hidden state factors (shapes of different colours) with data conditionally dependent upon only one. The purpose of this is to illustrate the behaviour of each scheme in the presence of informative and uninformative sensory input. Each of these hidden states starts with a defined shape (blue triangle, green square), but undergoes stochastic transitions. This means that the future should always be more uncertain than the past. In what follows, we use belief propagation as a gold standard for inferential performance, against which the other two schemes are compared. Our aim here is to illustrate the overconfidence of mean-field inference relative to the exact marginal inference of the Bethe approach, and to show how marginal approximations mitigate this, achieving a better approximation to belief propagation.

Figure 9 shows the results of simulating inference via the three forms of neuronal message passing outlined above. This illustrates some cardinal features of the three schemes. The trajectories of beliefs following each outcome show that the majority of belief updating occurs very early in variational message passing, before the presentation of most of the data. While a few revisions to these beliefs occur at later stages, it does not take long to arrive at highly confident beliefs about future states – this over-confidence of posterior beliefs is a well-recognised feature of variational inference under the mean-field approximation<sup>66</sup>. In contrast, belief propagation and marginal message passing take a more restrained approach, with each new observation driving updating. This more tentative approach pays off, as they make fewer errors in estimating the true states that generated the data. This is consistent with the fact that belief propagation offers an exact estimate of marginal beliefs for these models, while the variational approach is only ever approximate.

The over-confidence of the variational approach manifests clearly in the posterior beliefs about the green shapes. Given the stochastic transitions, and the absence of any informative data about these states, posterior beliefs about the green shapes should become increasingly uncertain with distance from the (deterministic) initial state. The Bethe approach clearly shows this, but the variational scheme does not, with highly confident beliefs about even the penultimate state. Marginal message passing compensates for this overconfidence issue, providing a much better approximation to an exact inference scheme than under the mean-field approach. In fact, it slightly overcompensates in the absence of precise data, leading to posteriors that are less confident than the belief propagation marginals. The temporal dynamics of belief updating (the upper plots) further illustrate the overconfidence of variational message passing relative to the other two schemes. Within the first time-step, the sufficient statistics of beliefs about the states over time (each represented as a line) approach extreme (zero or one) values. This means



**Figure 8.** Probability distributions for simulated HMM. The probability distributions here make up the generative model we used to simulate neuronal message passing under both schemes. These probabilities have been chosen to make several points. Firstly, we separated the hidden state into two distinct state factors (light blue shapes and light green shapes). This allowed us to generate data (darker blue shapes) that depends upon only one of these (light blue). The likelihood mapping illustrates how data are (probabilistically) generated from these states ( $P(o_\tau = i | s_\tau = j) = \mathbf{A}_{ij}$ ). We used deterministic priors for the initial states ( $P(s_1 = i) = \mathbf{D}_i$ ) and stochastic transitions ( $P(s_{\tau+1} = i | s_\tau = j) = \mathbf{B}_{ij}$ ).

that, with only one observation, the mean-field variational approach exhibits an excessive confidence about present and future states, that is maintained as new observations are made. In contrast, the belief propagation and marginal message passing schemes afford more modest belief updates – following the first observation – that become more confident as new data are acquired.

Notably, the three schemes share some of the same errors (four errors in steps 2, 4, 9 and 10). By errors, we mean that the inferred state (darkest shade) at a given time-step does not match the state that actually generated the data (red dot). These errors happen when very unlikely events occur, such as a dark blue square generated by a light blue triangle. Although incorrect, an inference that the light blue square caused the dark blue one is still Bayes optimal under the generative model we employed. In contrast, the additional four errors of variational message passing in steps 5, 8, 12 and 13 occur even though the data are highly consistent with the hidden states (e.g., a dark blue circle generated by a light blue circle). These errors reflect the excessive weight given to the empirical priors in variational message passing – it assumes the most probable *a priori* transition, ignoring the conflicting observation. It is important to note that increasing the dimensionality of the space state would further emphasise the failings of the mean-field approximation relative to the other two approximations.

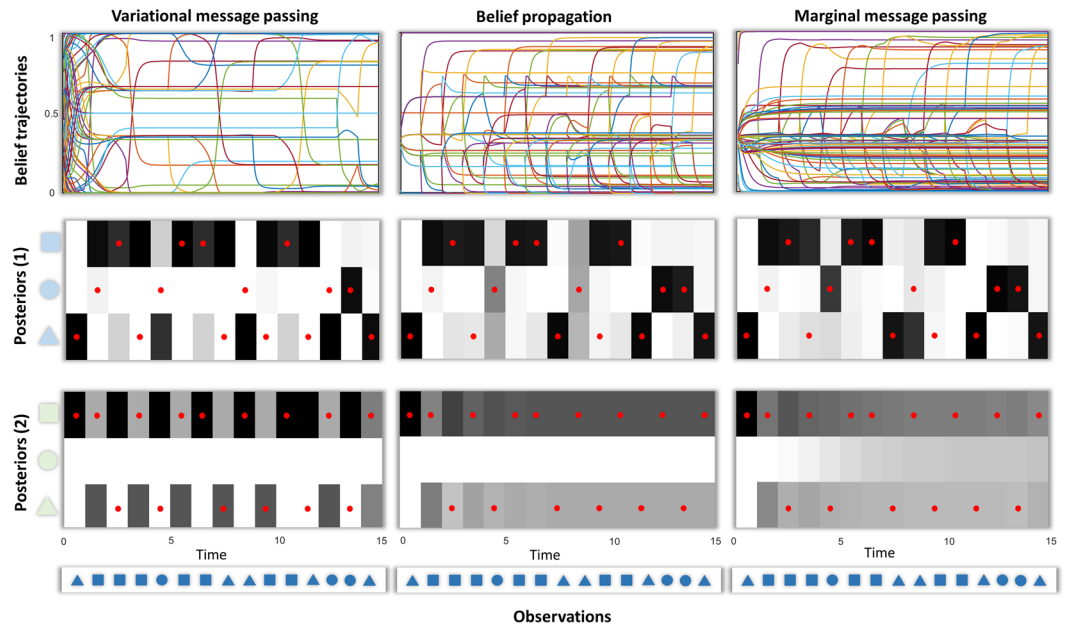
To quantify the performance of the mean-field and marginal approaches, we can exploit the fact that the Bethe approach is exact for the marginal posteriors for this inference problem. A simple way to do this is to compute the KL-Divergence between the marginal posteriors obtained through belief propagation and the solutions of the other two schemes. The smaller this divergence, the better the approximation to exact marginal beliefs. For the simulations of Fig. 9, the divergences summed over marginal posteriors give the following:

$$\sum_{\tau} D_{KL}[Q_{BP}(s_{\tau}) || Q_{VMP}(s_{\tau})] = 86.0563 \text{ nats}$$

$$\sum_{\tau} D_{KL}[Q_{BP}(s_{\tau}) || Q_{MMP}(s_{\tau})] = 3.7874 \text{ nats}$$

This demonstrates quantitatively that, even for the relatively simple inference problem used here, there is a much greater divergence between the exact marginal posterior beliefs and those obtained using variational message passing, relative to marginal message passing.

Although we have presented this as a single simulation, the way in which the generative model is defined, and the sequential presentation of the data, actually induce several distinct inference problems that we have implicitly appealed to above in characterising these schemes. First, the factorisation of the hidden state-space into two different types of latent variable (the light blue and light green shapes of Figs 8 and 9) allows us to compare the extreme case in which data are uninformative about the latent variable (light green shapes) with the case in which there is only moderate uncertainty about the relationship between (light blue) states and the (dark blue) data. Figure 9 shows that, while marginal and Bethe approaches attenuate their confidence – when data is uninformative compared to informative – the mean-field approach furnishes confident inferences in both cases. Note that these differences rely upon there being some uncertainty in the transitions from one state to the next. If we were



**Figure 9.** Simulated neuronal message passing for the three schemes: variational message passing, belief propagation and Marginal message passing. These plots show the consequences of generating data from the model described in Fig. 8 and solving the equations of Figs 5 and 6 for these data. The upper plots show the beliefs ( $\bar{s}$ ) throughout the trial in terms of the sufficient statistics of the categorical distributions (i.e. probability of each alternative state at each time). These depict belief updating in terms of expectations about the two hidden factors, each with three levels. Crucially, these beliefs are about each hidden state at (15) different points in time. Each line is then the posterior probability that a given hidden state at a given time takes on a specific value. The colour-coding of these lines is consistent between the plots along the upper row. These plots are important in that they give a sense of the time-course of belief updating. While variational message passing shows rapid changes at the very start of the simulation (nearly every line reaches an extreme value within the first time-step) and few thereafter, the updates of belief propagation and marginal message passing happen over a much longer time scale. The beliefs after each successive outcome are shown in the second and third rows (with black = 1 and white = 0). Each line in the plots in the first row therefore represents a cell in the second or third row – drawn at the time point encoded by each line. Red dots indicate the ‘true’ states used to generate the data. Note that there are no red dots associated with green circles, as these never occur when the initial green state is a square (see the transition matrix in Fig. 8). The final row shows the sequence of observations presented to each message passing scheme. The second and third rows show the posterior beliefs about the two factors comprising the hidden state space as shown in Fig. 8.

to use deterministic transition probabilities, the differences between these schemes would be largely abolished as all would make confident inferences.

The second comparison we have used relies upon sequential presentation of the outcomes. This means that each time-step represents a distinct inference problem, with more data available at later times than earlier. This is where the dynamics shown in the upper row of Fig. 9 are revealing. At each successive time point, the inference problem becomes more constrained, as an additional observation is made. This allows us to compare the confidence using a small amount of data (at the start of the trial) with the confidence after more data have been seen (near the end of the trial). After making the first observation, variational message passing shows a fairly consistent level of confidence until the end of the trial. This can be seen in the plot by noting that the distribution of lines in the vertical direction is relatively constant throughout the horizontal (temporal) axis. This contrasts with the other two schemes that show a greater proportion of lines reaching extreme values with each new observation.

## Discussion

This paper considers local computations that lend a biological plausibility to a range of inference schemes. We initially outlined two approaches to solving inference problems – variational message passing and belief propagation. Both can be expressed in a neurobiologically plausible form, but the former features a much simpler neuronal network structure. This simplicity comes at the cost of overconfidence, in comparison with the exact solutions obtained using belief propagation. The relative advantages of each motivated a third scheme; namely, ‘marginal message passing’ that uses a simple architecture but compensates for the problem of overconfidence. The temporal dynamics of inference reflect this, with variational message passing showing much earlier and exuberant changes in beliefs about those variables that are yet to give rise to observations, compared to the latter two schemes.

Crucially, we have expressed all three schemes in terms of differential equations describing evolution of beliefs over time (Figs 5 and 6). In addition to exposing the temporal dynamics of these approaches, and enabling a direct comparison, the resulting schemes also resemble the sorts of expressions that underwrite neural mass

models<sup>67</sup>. In other words, the neuronal dynamics implied by all three schemes are determined by a mixture of the activity of other populations of neurons. This mixture of input determines the rate of change of a log probability, which may be thought of as a membrane potential. The membrane potential itself determines the influence of a neuronal population on other populations through a softmax function, and this can be seen as analogous to the translation of a membrane potential into an average firing rate. In short, all three schemes have a potential biological validity in that belief dynamics can be expressed in a form closely related to that of neuronal mass dynamics (please see<sup>32</sup> for further details).

Note that the classical ‘forward-backward’ algorithm<sup>68</sup> is a special case of the belief propagation applied to an HMM. This algorithm computes a set of forward probabilities and backwards probabilities before combining the results to give marginal posterior probability distributions. Marginal message passing can therefore be regarded as an approximation to the inference steps used in these schemes.

The ‘forward-backward’ algorithm (hence belief propagation) is also used in the inference step of the Baum-Welch algorithm<sup>69</sup>. However, the Baum-Welch algorithm has an additional step that distinguishes it from belief propagation and other variational approaches. This is a maximum likelihood update of the parameters of the generative model. The alternation between an inference step and a maximum likelihood update in the Baum-Welch algorithm underwrites its formal equivalence with the Expectation-Maximisation<sup>70</sup> algorithm (as applied to an HMM), although the latter is rarely articulated in terms of message passing.

In contrast, the three variational schemes we have discussed here are (approximately) Bayesian, and so do not permit maximum likelihood updating. Instead, to draw inferences about parameters in these schemes, we need to specify prior distributions over the parameter values and use these to compute posterior beliefs<sup>58,71</sup>. From the perspective of Bayesian message passing, this just means extending the factor graph to include parameters, and passing additional messages. From the perspective of neurobiology, updates in beliefs about transition or likelihood probabilities would manifest as plastic changes in the efficacy of synapses connecting neuronal populations. This would allow for rewiring<sup>72</sup> of the network to deal with a different (HMM) generative model if the way in which data were generated changed. Note that, once the generative model has been learned (as is assumed in the simulations here), the inference task depends only upon the neuronal activities, and does not require changes in connectivity in response to new data.

The particular form of message passing has implications for empirical studies, as disambiguating between alternative mechanisms that underwrite biological inference may be important in understanding psychopathology; e.g., hallucinations and delusions. All three message passing schemes make clear predictions about the time-course of electrophysiological responses at different time points following an outcome; i.e., within a trial. To adjudicate along different belief updating schemes, one could present participants with a sequence like that above, after exposing them to previous sequences so that they have learned the probabilistic structure of the task. If the brain employs variational message passing, we would anticipate greater evoked responses for the first few stimuli, given the greater rate of belief updating at these times. On immediate recall of the sequence, one might expect participants to make errors consistent with those found here – overestimating the precision of transitions.

To better accommodate behavioural responses, we could equip the schemes above with an active component<sup>3</sup>, and fit the resulting model to human behaviour<sup>36</sup>. It is more difficult to distinguish between belief propagation and a scheme (like marginal message passing) that seeks to emulate it in a simpler architecture. However, there are several forms of data that could be brought to bear on this question. First, one could appeal to a similar task as that above and fit the evoked responses with neural mass models<sup>67</sup> that mimic the architectures of Figs 5 and 6. If additional neurons with fast time constants, representing the messages, improve the accuracy of the fit in excess of any increase in complexity, this would offer evidence in favour of belief propagation. Decreased accuracy, or preserved accuracy in the presence of an increased complexity, would instead favour a simpler architecture like marginal message passing. Further evidence could be garnered from tract tracing studies; e.g.<sup>73</sup>, or from single unit recordings – to ask whether they are better explained as representing messages rather than the sufficient statistics of marginal beliefs. Alternatively, one could compare the ability of simulated belief trajectories to explain electrophysiological responses. Current circuit-level research shows a high degree of consistency with the form of the neuronal networks presented here<sup>74–76</sup>, but these data are not sufficient to confidently disambiguate between the two architectures. Sensory input (via the thalamus) predominantly targets granular layers of cortex<sup>77,78</sup>, which excite more superficial cells and are disinaptically inhibited by them in turn<sup>79</sup>. This is consistent with Figs 5 and 6. The belief propagation architecture of Fig. 5 calls for another set of neurons (those representing the messages) that are inhibited by input from sensory streams and that excite the granular cells. Reversing the signs (excitation-inhibition), it is plausible that inhibitory interneurons in layer IV in receipt of sensory input<sup>74</sup> could play this role; inducing an inhibition in the granular cells in response to sensory input (as opposed to exciting them in its absence). Apart from the empirical question, which of the neuronal schemes is supported by empirical data, there is also an important theoretical issue: we have focussed upon the neuronal manifestations of inference, but there are many other examples of biological inference that depend upon similar local interactions. It will be interesting to see whether the same principles of local message passing can be scaled up to collective behaviour<sup>80,81</sup>, with individuals exchanging information with their neighbours; or whether it can be scaled down to the computations performed by biochemical networks<sup>82</sup>.

Many neurological and psychiatric syndromes can be thought of in terms of false inference<sup>83</sup>. The form of healthy computation may be important for understanding the types of pathology that might affect it. Broadly speaking, computational pathologies can be described in terms of optimal inference using a suboptimal generative model<sup>84–87</sup>, or as broken inferential machinery<sup>4,37</sup>. An account of a pathology in terms of a suboptimal generative model transcends specific Bayesian message passing schemes and could be reproduced using variational message passing or belief propagation (or any other scheme). Theories of schizophrenia<sup>86</sup>, autism<sup>88,89</sup>, and visual neglect<sup>97</sup> (among others) that appeal to pathological prior beliefs may be interpreted as deficits in any of the message passing schemes described here. In contrast, an account based on broken message passing commits to a specific neuronal implementation and is only meaningful if formulation of belief updating is correct.

An example of the latter is a recent theory that aims to account for inferential deficits that underwrite perceptual changes in schizophrenia. This posits ascending and descending inferential ‘loops’, and that disruption of these could lead to an ‘overcounting’ of a message<sup>4</sup>. To build some intuition for this idea, imagine we were to cut the starred connection in Fig. 5. This connection subtracts the ascending message from the forward message. A failure to subtract this message means the forward message will also contain the ascending message. The neurons representing the marginal then receive two copies of the ascending message; i.e., it is overcounted<sup>37</sup>. This could lead to an oversensitivity to sensory information<sup>90</sup>, and a failure to contextualise it using prior beliefs. That this is due to a failure of subtraction by an inhibitory interneuron is consistent with data suggesting disruptions in the balance of excitatory and inhibitory synaptic activity<sup>91–93</sup> in patients with psychosis. Crucially, this account relies upon belief propagation implemented in a specific way – and does not generalise to marginal or variational message passing schemes.

## Conclusion

Variational message passing and belief propagation both represent means of performing Bayesian inference using local computations, consistent with the computations of biological neuronal networks. Both minimise free energy functionals, so are consistent with active inference – or the minimisation of free energy through action and perception. There are notable differences between the two schemes. While belief propagation represents exact Bayesian inference (for many generative models), variational approaches yield approximate inference. The latter tends towards excessive confidence in the face of uncertainty, deviating from exact Bayesian optimality. However, the kinds of neuronal network that support belief propagation appear to require a greater number of neurons and axons to achieve biological plausibility. This suggests a trade-off between inferential accuracy (belief propagation) and a complexity cost (variational message passing). Drawing from the relative benefits and drawbacks, we have introduced a third possibility. The brain may approximate belief propagation, using circuitry consistent with variational message passing. We hope to disambiguate between these possibilities in future work. An interesting conceptual issue that arises from these considerations is how best to understand the false inferences that underwrite neurological and psychiatric disease. Appealing to broken generative models enables one to be agnostic about the exact form of message passing, while hypothetical lesions to the inferential machinery depend upon the validity of that machinery.

## Data Availability

The script used for the simulations is available at <https://github.com/tejparr/nmpassing>.

## References

- Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences* **27**, 712–719 (2004).
- Doya, K. *Bayesian brain: Probabilistic approaches to neural coding*. (MIT press, 2007).
- Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. Action and behavior: a free-energy formulation. *Biological Cybernetics* **102**, 227–260, <https://doi.org/10.1007/s00422-010-0364-z> (2010).
- Jardri, R. & Denève, S. Circular inferences in schizophrenia. *Brain* **136**, 3227–3241, <https://doi.org/10.1093/brain/awt257> (2013).
- Marković, D. & Kiebel, S. J. Comparative Analysis of Behavioral Models for Adaptive Learning in Changing Environments. *Frontiers in Computational Neuroscience*, **10**, <https://doi.org/10.3389/fncom.2016.00033> (2016).
- Gregory, R. L. Perceptions as Hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **290**, 181 (1980).
- Von Helmholtz, H. *Handbuch der physiologischen Optik*. Vol. 9 (Voss, 1867).
- Rao, R. P. Neural Models of Bayesian Belief Propagation. *Bayesian brain: Probabilistic approaches to neural coding*, **239** (2007).
- Schwöbel, S., Kiebel, S. & Marković, D. Active Inference, Belief Propagation, and the Bethe Approximation. *Neural computation*, 1–38 (2018).
- Friston, K. J., Parr, T. & Vries, B. D. The graphical brain: belief propagation and active inference. *Network Neuroscience* **0**, 1–78, [https://doi.org/10.1162/NETN\\_a\\_00018](https://doi.org/10.1162/NETN_a_00018) (2017).
- Pearl, J. Graphical models for probabilistic and causal reasoning (1997).
- Winn, J. & Bishop, C. M. Variational message passing. *Journal of Machine Learning Research* **6**, 661–694 (2005).
- Forney, G. D. Codes on graphs: Normal realizations. *IEEE Transactions on Information Theory* **47**, 520–548 (2001).
- Loeliger, H. A. et al. The Factor Graph Approach to Model-Based Signal Processing. *Proceedings of the IEEE* **95**, 1295–1322, <https://doi.org/10.1109/JPROC.2007.896497> (2007).
- Steimer, A., Maass, W. & Douglas, R. Belief Propagation in Networks of Spiking Neurons. *Neural Computation* **21**, 2502–2523, <https://doi.org/10.1162/neco.2009.08-08-837> (2009).
- Isomura, T., Kotani, K. & Jimbo, Y. Cultured Cortical Neurons Can Perform Blind Source Separation According to the Free-Energy Principle. *PLoS Computational Biology* **11**, e1004643, <https://doi.org/10.1371/journal.pcbi.1004643> (2015).
- Angela, J. Y. & Dayan, P. Acetylcholine in cortical inference. *Neural Networks* **15**, 719–730 (2002).
- Beck, J. M. & Pouget, A. Exact inferences in a neural implementation of a hidden Markov model. *Neural computation* **19**, 1344–1361 (2007).
- Friston, K. & Samothrakis, S. & Montague, R. Active inference and agency: optimal control without cost functions. *Biological Cybernetics* **106**, 523–541, <https://doi.org/10.1007/s00422-012-0512-8> (2012).
- Loeliger, H. A. An introduction to factor graphs. *IEEE Signal Processing Magazine* **21**, 28–41, <https://doi.org/10.1109/MSP.2004.1267047> (2004).
- Roweis, S. & Ghahramani, Z. A Unifying Review of Linear Gaussian Models. *Neural Computation* **11**, 305–345, <https://doi.org/10.1162/089976699300016674> (1999).
- Fries, P. Rhythms For Cognition: Communication Through Coherence. *Neuron* **88**, 220–235, <https://doi.org/10.1016/j.neuron.2015.09.034> (2015).
- Welch, L. R. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* **53**, 10–13 (2003).
- Winn, J. M. *Variational message passing and its applications*, Citeseer (2004).
- Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. (Elsevier, 2014).
- Yedidia, J. S., Freeman, W. T. & Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* **51**, 2282–2312 (2005).



27. Jardri, R., Duverne, S., Litvinova, A. S. & Denève, S. Experimental evidence for circular inference in schizophrenia. *Nature Communications* **8**, 14218, <https://doi.org/10.1038/ncomms14218>, <https://www.nature.com/articles/ncomms14218#supplementary-information> (2017).
28. George, D. & Hawkins, J. Belief propagation and wiring length optimization as organizing principles for cortical microcircuits. (Technical report, Numenta, <http://www.numenta.com>, 2006).
29. Deneve, S. In *Advances in neural information processing systems*. 353–360.
30. Friston, K. *et al.* Active inference and epistemic value. *Cognitive Neuroscience* **6**, 187–214, <https://doi.org/10.1080/17588928.2015.1020053> (2015).
31. Hohwy, J. The Self-Evidencing Brain. *Nous* **50**, 259–285, <https://doi.org/10.1111/nous.12062> (2016).
32. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active Inference: A Process Theory. *Neural Comput* **29**, 1–49, [https://doi.org/10.1162/NECO\\_a\\_00912](https://doi.org/10.1162/NECO_a_00912) (2017).
33. Dauwels, J. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. 2546–2550 (IEEE).
34. Parr, T. & Friston, K. J. Working memory, attention, and salience in active inference. *Scientific reports* **7**, 14678, <https://doi.org/10.1038/s41598-017-15249-0> (2017).
35. Perrinet, L. U., Adams, R. A. & Friston, K. J. Active inference, eye movements and oculomotor delays. *Biological Cybernetics* **108**, 777–801, <https://doi.org/10.1007/s00422-014-0620-8> (2014).
36. Mirza, M. B., Adams, R. A., Mathys, C. & Friston, K. J. Human visual exploration reduces uncertainty about the sensed world. *PLOS ONE* **13**, e0190429, <https://doi.org/10.1371/journal.pone.0190429> (2018).
37. Leptourgos, P., Denève, S. & Jardri, R. Can circular inference relate the neuropathological and behavioral aspects of schizophrenia? *Current Opinion in Neurobiology* **46**, 154–161, <https://doi.org/10.1016/j.conb.2017.08.012> (2017).
38. Kaplan, R. & Friston, K. J. Planning and navigation as active inference. *Biological Cybernetics*, <https://doi.org/10.1007/s00422-018-0753-2> (2018).
39. Friston, K. J. *et al.* Active inference, curiosity and insight. *Neural Computation* (2017).
40. Jaakkola, T., Singh, S. P. & Jordan, M. I. In *Advances in neural information processing systems*. 345–352.
41. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. Vol. 1 (MIT press Cambridge, 1998).
42. Tani, J. Self-Organization and Compositionality in Cognitive Brains: A Neurorobotics Study. *Proceedings of the IEEE* **102**, 586–605, <https://doi.org/10.1109/JPROC.2014.2308604> (2014).
43. Forney, G. D. Jr. & Vontobel, P. O. Partition functions of normal factor graphs. *arXiv preprint arXiv:1102.0316* (2011).
44. Heskes, T. In *Advances in neural information processing systems*. 359–366.
45. Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M. & Friston, K. The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields. *PLOS Computational Biology* **4**, e1000092, <https://doi.org/10.1371/journal.pcbi.1000092> (2008).
46. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS computational biology* **7**, e1002211 (2011).
47. Pecevski, D., Buesing, L. & Maass, W. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS computational biology* **7**, e1002294 (2011).
48. George, D. & Hawkins, J. Towards a Mathematical Theory of Cortical Micro-circuits. *PLOS Computational Biology* **5**, e1000532, <https://doi.org/10.1371/journal.pcbi.1000532> (2009).
49. Steimer, A. & Douglas, R. Spike-based probabilistic inference in analog graphical models using interspike-interval coding. *Neural computation* **25**, 2303–2354 (2013).
50. Mirza, M. B., Adams, R. A., Mathys, C. D. & Friston, K. J. Scene Construction, Visual Foraging, and Active Inference. *Frontiers in Computational Neuroscience* **10**, <https://doi.org/10.3389/fncom.2016.00056> (2016).
51. FitzGerald, T., Dolan, R. & Friston, K. Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.*, <https://doi.org/10.3389/fnhum.2014.00457> (2014).
52. Friston, K. J., Rosch, R., Parr, T., Price, C. & Bowman, H. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews* **77**, 388–402, <https://doi.org/10.1016/j.neubiorev.2017.04.009> (2017).
53. Parr, T. & Friston, K. J. Uncertainty, epistemics and active inference. *Journal of The Royal Society Interface* **14** (2017).
54. Laughlin, S. B. & Sejnowski, T. J. Communication in Neuronal Networks. *Science (New York, N.Y.)* **301**, 1870–1874, <https://doi.org/10.1126/science.1089662> (2003).
55. Lennie, P. The Cost of Cortical Computation. *Current Biology* **13**, 493–497, [https://doi.org/10.1016/S0960-9822\(03\)00135-0](https://doi.org/10.1016/S0960-9822(03)00135-0) (2003).
56. Landauer, R. Irreversibility and heat generation in the computing process. *IBM journal of research and development* **5**, 183–191 (1961).
57. London, M. & Häusser, M. DENDRITIC COMPUTATION. *Annual Review of Neuroscience* **28**, 503–532, <https://doi.org/10.1146/annurev.neuro.28.061604.135703> (2005).
58. Beal, M. J. (University of London United Kingdom, 2003).
59. Wainwright, M. J. & Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1**, 1–305 (2008).
60. Heskes, T. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research* **26**, 153–190 (2006).
61. Feynman, R. P. *Statistical Mechanics: A Set Of Lectures*. (Avalon Publishing, 1998).
62. Mohri, T. Cluster Variation Method. *Jom* **65**, 1510–1522 (2013).
63. Maren, A. J. The Cluster Variation Method: A Primer for Neuroscientists. *Brain Sciences* **6**, 44, <https://doi.org/10.3390/brainsci6040044> (2016).
64. Weller, A., Tang, K., Sontag, D. & Jebara, T. In *30th Conference on Uncertainty in Artificial Intelligence, UAI.* (AUAI Press, 2014).
65. Friston, K., Breakspear, M. & Deco, G. Perception and self-organized instability. *Frontiers in Computational Neuroscience* **6**, <https://doi.org/10.3389/fncom.2012.00044> (2012).
66. Consonni, G. & Marin, J.-M. Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics & Data Analysis* **52**, 790–798, <https://doi.org/10.1016/j.csda.2006.10.028> (2007).
67. Moran, R., Pinotsis, D. A. & Friston, K. Neural masses and fields in dynamic causal modeling. *Frontiers in Computational Neuroscience* **7**, 57, <https://doi.org/10.3389/fncom.2013.00057> (2013).
68. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**, 257–286, <https://doi.org/10.1109/5.18626> (1989).
69. Baum, L. E. & Eagon, J. A. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73**, 360–363 (1967).
70. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38 (1977).
71. Friston, K. *et al.* Active inference and learning. *Neuroscience & Biobehavioral Reviews* **68**, 862–879, <https://doi.org/10.1016/j.neubiorev.2016.06.022> (2016).
72. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99**, 609–623, <https://doi.org/10.1016/j.neuron.2018.07.003> (2018).
73. Vélez-Fort, M. *et al.* The Stimulus Selectivity and Connectivity of Layer Six Principal Cells Reveals Cortical Microcircuits Underlying Visual Processing. *Neuron* **83**, 1431–1443, <https://doi.org/10.1016/j.neuron.2014.08.001> (2014).

74. Haeusler, S. & Maass, W. A Statistical Analysis of Information-Processing Properties of Lamina-Specific Cortical Microcircuit Models. *Cerebral Cortex* **17**, 149–162, <https://doi.org/10.1093/cercor/bhj132> (2007).
75. Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711, <https://doi.org/10.1016/j.neuron.2012.10.038> (2012).
76. Shipp, S. Neural Elements for Predictive Coding. *Frontiers in Psychology* **7**, 1792, <https://doi.org/10.3389/fpsyg.2016.01792> (2016).
77. Miller, K. D. Understanding Layer 4 of the Cortical Circuit: A Model Based on Cat V1. *Cerebral Cortex* **13**, 73–82, <https://doi.org/10.1093/cercor/13.1.73> (2003).
78. Shipp, S. Structure and function of the cerebral cortex. *Current Biology* **17**, R443–R449, <https://doi.org/10.1016/j.cub.2007.03.044> (2007).
79. Thomson, A. M., West, D. C., Wang, Y. & Bannister, A. P. Synaptic Connections and Small Circuits Involving Excitatory and Inhibitory Neurons in Layers 2–5 of Adult Rat and Cat Neocortex: Triple Intracellular Recordings and Biocytin Labelling *In Vitro*. *Cerebral Cortex* **12**, 936–953, <https://doi.org/10.1093/cercor/12.9.936> (2002).
80. Herbert-Read, J. E. *et al.* Inferring the rules of interaction of shoaling fish. *Proceedings of the National Academy of Sciences* **108**, 18726–18731, <https://doi.org/10.1073/pnas.1109355108> (2011).
81. Mann, R. P. & Garnett, R. The entropic basis of collective behaviour. *Journal of The Royal Society Interface* **12** (2015).
82. Genot, A. J., Fujii, T. & Rondelez, Y. Computing with Competition in Biochemical Networks. *Physical Review Letters* **109**, 208102 (2012).
83. Parr, T., Rees, G. & Friston, K. J. Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience* **12**, <https://doi.org/10.3389/fnhum.2018.00061> (2018).
84. Daunizeau, J. *et al.* Observing the Observer (I): Meta-Bayesian Models of Learning and Decision-Making. *PLOS ONE* **5**, e15554, <https://doi.org/10.1371/journal.pone.0015554> (2010).
85. Schwartenbeck, P. *et al.* Optimal inference with suboptimal models: addiction and active Bayesian inference. *Medical hypotheses* **84**, 109–117, <https://doi.org/10.1016/j.mehy.2014.12.007> (2015).
86. Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D. & Friston, K. J. The Computational Anatomy of Psychosis. *Frontiers in Psychiatry* **4**, 47, <https://doi.org/10.3389/fpsyg.2013.00047> (2013).
87. Parr, T. & Friston, K. J. The Computational Anatomy of Visual Neglect. *Cerebral Cortex*, 1–14, <https://doi.org/10.1093/cercor/bhx316> (2017).
88. Lawson, R. P., Rees, G. & Friston, K. J. An aberrant precision account of autism. *Frontiers in Human Neuroscience* **8**, 302, <https://doi.org/10.3389/fnhum.2014.00302> (2014).
89. Lawson, R. P., Mathys, C. & Rees, G. Adults with autism overestimate the volatility of the sensory environment. *Nat Neurosci* **20**, 1293–1299, <https://doi.org/10.1038/nn.4615>, <http://www.nature.com/neuro/journal/v20/n9/abs/nn.4615.html#supplementary-information> (2017).
90. Shergill, S. S., Samson, G., Bays, P. M., Frith, C. D. & Wolpert, D. M. Evidence for Sensory Prediction Deficits in Schizophrenia. *American Journal of Psychiatry* **162**, 2384–2386, <https://doi.org/10.1176/appi.ajp.162.12.2384> (2005).
91. Lisman, J. E. *et al.* Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends in neurosciences* **31**, 234–242, <https://doi.org/10.1016/j.tins.2008.02.005> (2008).
92. Perry, T., Buchanan, J., Kish, S. & Hansen, S.  $\gamma$ -Aminobutyric-acid deficiency in brain of schizophrenic patients. *The Lancet* **313**, 237–239 (1979).
93. Blum, B. P. & Mann, J. J. The GABAergic system in schizophrenia. *International Journal of Neuropsychopharmacology* **5**, 159–179 (2002).
94. Bell, A. J. & Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation* **7**, 1129–1159 (1995).
95. Bach, F. R. & Jordan, M. I. A probabilistic interpretation of canonical correlation analysis (2005).
96. Nowlan, S. J. In *Advances in neural information processing systems*. 574–582.
97. Kalman, R. E. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* **82**, 35–45 (1960).
98. Friston, K., Stephan, K., Li, B. & Daunizeau, J. Generalised filtering. *Mathematical Problems in Engineering* **2010** (2010).
99. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 1211 (2009).
100. Li, B. *et al.* Generalised filtering and stochastic DCM for fMRI. *NeuroImage* **58**, 442–457, <https://doi.org/10.1016/j.neuroimage.2011.01.085> (2011).
101. Haken, H. Slaving principle revisited. *Physica D: Nonlinear Phenomena* **97**, 95–103, [https://doi.org/10.1016/0167-2789\(96\)00080-2](https://doi.org/10.1016/0167-2789(96)00080-2) (1996).

## Acknowledgements

TP is supported by the Rosetrees Trust (Award Number 173346). KJF is a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z). This work was supported by the Deutsche Forschungsgemeinschaft (SFB 940/2, Project A9) and by the TU Dresden Graduate Academy.

## Author Contributions

T.P., D.M., S.K., K.J.F. contribution to writing. T.P. and D.M. performed the simulations.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-38246-3>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019