The functional anatomy of time: what and when in the brain

Karl Friston and Gyorgy Buzsáki

Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom k.friston@ucl.ac.uk

NYU Neuroscience Institute, School of Medicine, New York University, New York, NY 10016; Institute of Experimental Medicine, Hungarian Academy of Sciences Gyorgy.Buzsaki@nyumc.org

Correspondence: k.friston@ucl.ac.uk (Karl Friston)

Keywords: hippocampus; ordinal; spatiotemporal; Bayesian; inference; sequences

Abstract

This article considers the implications – for functional anatomy – of how we represent temporal structure in our exchanges with the world. It offers a theoretical treatment that tries to make sense of the architectural principles seen in mammalian brains. Specifically, it considers a factorisation between representations of temporal succession and representations of content or, heuristically, a segregation into *when* and *what*. This segregation may explain the central role of the hippocampus in neuronal hierarchies, while providing a tentative explanation for recent observations of how ordinal sequences are encoded. The implications for neuroanatomy and physiology may have something important to say about how self-organized cell assembly sequences enable the brain to exhibit purposeful behaviour that transcends the here and now.

The principles of functional anatomy

There are certain architectural principles of neuroanatomy that seem amenable to explanation from a purely theoretical perspective. These range from the existence of axonal processes that form neuronal connections, through to macroscopic organisational principles such as functional segregation [1]. A key example is the segregation of dorsal and ventral streams into *what* and *where* streams [2]. How might these architectural features be explained from a theoretical perspective? In what follows, we appeal to active inference and the Bayesian brain hypothesis [3, 4] to show that functional segregation is mandated for brains that navigate a world with deep hierarchical structure. We then consider the implications of this argument for a fundamental aspect of this navigation; namely, the trajectories or ordered sequences of states we encounter [5]. Our conclusion is that there should be a functional segregation between *what* and *when* – a conclusion that seems to explain a number of anatomical and physiological observations; particularly in the hippocampal system.

Good enough brains and good enough models

A key theoretical development in neurobiology is the appreciation of the brain as a predictive organ – generating predictions of its actions and sensations [4, 6-9]. These predictions rest on an internal or *generative model* of how sensory input unfolds. Indeed, one can understand much of neuronal dynamics and synaptic plasticity as an optimisation of (Bayesian) model evidence – as scored by proxies like free energy and prediction errors [9-11]. If one subscribes to this normative theory, the brain must be a good (enough) model of its environment, where recurring sequences of events are the rule. This is an old idea, dating back to notions of good regulators in self-organisation and cybernetics [12, 13]. In brief, the good regulator theorem states that any system that can control its environment must be a good model of that environment. So what constitutes a good enough model?

Mathematically, a good enough model is simply a model that has high evidence, in light of the (sensory) data it has to explain. Evidence is the probability of sensory samples, under a model of how those samples were generated (See Box 1). In this sense, any brain can be understood as (self) organising itself to maximise model evidence. Here, we are implicitly appealing to the Bayesian brain hypothesis [14], while gently sidestepping big questions about its utility and falsifiability: e.g., [15, 16]. In what follows, we assume that the

imperative to maximise model evidence is a truism and consider the implications for functional anatomy. Our focus is not on the Bayesian brain *per se* but on the notion of a *mean field approximation* that is an integral part of *approximate* Bayesian inference.

A key conclusion – that follows from the Bayesian brain – is that the structure of a good brain will recapitulate the (statistical) structure of how sensations are caused. A simple but remarkable example is the very existence of neuronal connections. Why does the brain have connections, while the liver seems to operate perfectly happy without them? The answer to this is almost obvious when we consider that the brain has to model sparse dependences induced by causal structure in the world. In other words, a good enough (or sufficient) explanation for our sensory inputs is that they are generated by a small number of underlying causes that act upon each other (usually at a distance), in a lawful and structured way. This lawful structure requires a relatively sparse dependency among the causes; such as gravity causing things to fall or visual objects causing sensory impressions. In short, the causal structure of our world should, in principle, provide a sufficient explanation for the structure and fabric of any brain that is trying to model that world. One could apply this argument to other aspects of neuronal architecture. For example, our sensations are generated in a way that conforms to logarithmic rules (e.g., Weber's law). These statistical rules may then be transcribed into the lognormal statistics of synaptic physiology [17] or the connectome that supports this physiology [18]. Simply noting that causal regularities in the world are transcribed into neuronal architectures is interesting (and perhaps self-evident). However, this conjecture does not get to the heart of principles such as functional segregation. To understand how maximising model evidence leads to functional segregation, we have to consider the constraints under which evidence is optimised. This brings us to the notion of *approximate Bayesian inference* (Box 1).

Box 1 about here

Good enough brains and approximate Bayesian inference

Any system or procedure that optimises (maximises) Bayesian model evidence can be regarded as implementing Bayesian inference. However, exact Bayesian inference is

generally impossible in the real world; especially when modelling data generated by hierarchically deep, dynamic and nonlinear processes. Almost invariably, this problem is solved with something called *approximate Bayesian inference*. Approximate Bayesian inference refers to optimisation in which an approximate representation (technically, a posterior probability distribution or 'belief') is made as similar as possible to the exact (Bayes-optimal) belief. There are many examples of approximate Bayesian inference. One popular example is Bayesian filtering (a.k.a. predictive coding [11]) that calls on a number of approximations. These include the assumption that probabilistic beliefs have a particular distributional form (usually a Gaussian or bell shaped distribution). Another important assumption – that is ubiquitous in statistical physics and data analysis – is referred to as a *mean field approximation* [19, 20]. Combining these two approximations leads to something called variational Bayes. The mathematical details here are unimportant – the key concept is that the brain is faced with an important choice in the way that it optimises the very structure of its generative model (the fabric of its connections) and associated beliefs (the physiology supported by this fabric).

Put simply, the mean field approximation approximates dependencies among multiple factors with a product of marginal distributions that is much easier to deal with – in terms of encoding and updating. A key challenge for approximate Bayesian inference is to find the right factorisation or marginalisation of beliefs about the causes of sensory input. Each possible factorisation or marginal representation corresponds to a different mean field approximation and a different way of 'carving nature at its joints' [21, 22]. As scientists, we use this judicious 'carving' whenever we design a factorial experiment and test for interactions. In this case, the two factors represent a parsimonious hypothesis about how our data are caused, where the interaction reflects how one factor influences the expression of the other. Can this basic tenet of good statistical modelling be applied to neurobiology?

There are two levels that immediately come to mind. The first is the perspective afforded by the Bayesian brain – and, in particular, the notion of perception as hypothesis testing [7]. In this instance, efficient perceptual synthesis reduces to an efficient and good factorisation of the putative causes of sensations. In other words, the brain has to learn about statistical independencies (technically, conditional independencies) to properly approximate the underlying causes of the sensorium. There is ample evidence to suggest that experience-dependent plasticity and associated learning plays a huge role in this process [23]. However,

one can also regard evolution as playing exactly the same game: it is becoming increasingly evident that evolution can be formulated as learning statistical structure in the environment and distilling that structure into the phenotype [24-26]. Indeed, formal treatments of replicator dynamics and Fisher's fundamental theorem demonstrate that these evolutionary processes are nothing more or less than Bayesian belief updating. Indeed, natural selection itself has been likened to Bayesian model selection, where adaptive fitness corresponds to (variational) free energy [9, 27] .

Functional segregation and carving nature at its joints

The second level at which a good (enough) factorisation might be expressed is in terms of functional anatomy and segregation. In short, millennia of evolutionary (Bayesian belief) updates have shaped the brain into an efficient (minimum free energy) mean field approximation that we know and study as *functional segregation* [1, 28, 29].

A compelling example – of the implicit division of labour – is the factorisation of syntax and lexico-semantic statistics of language [30]. Ample evidence demonstrates that Brodmann areas 45 and 47 respond not only to natural sentences in fMRI experiments but also to grammatically correct sentences without semantic content or meaning. This suggests a specialised role of these brain areas in syntactical organization of semantic information [31]. In contrast, several other neocortical areas respond selectively to meaningful sentences but not to grammatically correct sentences without semantic information [32].

Perhaps the most celebrated example of transcribing statistical independencies into neuroanatomy is the segregation of dorsal and ventral visual processing streams [2, 33, 34]. The argument here is straightforward: if the causes of our visual sensations are visual objects that can be in different positions, the optimal way to factorise these causes is into *where* an object is and *what* an object is. The implicit conditional independence is simply a reflection of the fact that knowing where an object is does not (generally) tell you what it is. Technically, installing this conditional independence into functional anatomy enables the brain to maximise (Bayesian) model evidence. The alternative would be to have neuronal representations of every object in every location. Clearly, this would lead to a complex generative model with redundant degrees of freedom (connections) – provided our world does indeed comprise objects in various locations (Box 2).

Box 2 about here

Technically, finding the right way to carve nature into the best factors ensures that the variational free energy is a better approximation to model evidence. In this view, the functional segregation of *what* and *where* streams embodies the fact that it is more efficient to encode *where* an object is and *what* an object is – as opposed to encoding every combination of *what* and *where*. The predictions of current sensations then involve multiplying the probability distribution over where an object is by the probability distribution over what an object is (we will return to the importance of this multiplication or interaction later). One could take this sort of argument much further, in terms of hierarchal representations and special cases of variational inference cast in terms of information theory; leading to the principle of minimum redundancy, the principle of maximum efficiency, imperatives for sparse coding and so on [35-38].

From a neurobiological perspective, this statistical carving (factorisation) corresponds to functional segregation [1, 28]. If correct, this means that conditionally independent causes of our sensations correspond to the attributes that define functional specialisation; for example, motion, colour, form and so on [39]. In other words, natural selection, epigenetics and experience-dependent plasticity equip the brain with the right sort of mean field approximation to infer the factors causing sensations. For example, knowing an object's colour does not (generically) determine its motion and so on. One could pursue this approach right down to the level of classical receptive fields [38, 40] and their contextual modulation (extra classical receptive field effects) implied by the multiplication of marginal distributions to form precise posterior (probabilistic) beliefs. However, here, we want to consider another potentially more fundamental carving of statistical independencies that speaks, not to *what* and *where* streams but to *what* and *when* systems – a dissection that may provide organising principles for the brains of higher animals.

What and when – functional segregation of the neocortex and hippocampus

It is almost self-evident that the most pervasive and simplest conditional independence – that we deal with at all the time in perceptual synthesis and spatial navigation – is the temporal or ordinal succession of events [41]. Here, succession *per se* can be separated from the constituent events. In other words, the very concepts of "first", "last", "quick" and "fast" do

not specify what is happening and are not content-bound. This suggests a fundamental conditional independence between the temporal structure of succession (i.e. when) and the events that succeed each other (i.e., what). In exactly the same way that the brain may factorise hidden causes of sensations into *what* and *where*, it may apply the same marginalisation to *what* and *when*. This distinction may be more pervasive than *what* and *where* – it might apply at multiple levels of abstraction – and the very unfolding of experience itself. In other words, the attribute of *where* is limited to certain causes of our sensations; however, social and physiological narratives (which may not be located to a particular point in extrapersonal space) always have a sequential aspect; e.g., music and language.

Box 3 about here

To paint this picture heuristically, consider two ways of encoding sequences. First, we could have a repertoire of sequential states for every sequence encountered; in other words, a separate representation for every state at each point in a sequence. This would be like having a library of sentences that we could call on to make sense of written text. The alternative to activating sequences of representations would be to have representations of sequences whose content could be read sequentially: see Box 3 and [42]. This distinction is exactly the same as the distinction between the joint and marginal representations of what and where considered above. This distinction may sound subtle; however, the marginal (mean field) approximation is substantially less complex. This is because instead of having to represent hidden states or causes for every sentence (i.e., number of words in the sentence *times* the number of sentences) we just have to represent a preconfigured sequence and each sentence (i.e., the number of words per sentence *plus* the number of sentences). If a *what* and *when* distinction holds, there are some important predictions about the encoding of sequences in the hippocampus that we now briefly unpack.

If we call on a mean field approximation (functional segregation) of *what* and *when*, one would anticipate a generic architecture embodying the associated functional segregation. The natural candidate for this architecture is the distinction between the brain structures of temporal succession [43], such as the hippocampus and cerebellum, from (neocortical) structures encoding content. It also suggests that structures such as the hippocampus (*when*) should have the greatest divergent and convergent connectivity with representations of content (*what*). This may explain why the hippocampus is a hub with far-reaching

connectivity [44]; as opposed to more modular neocortical areas. This connectivity places the hippocampus and paralimbic cortex at the centre of (centrifugal) hierarchical cortical connectivity [45, 46].

Physiological support for model predictions

If temporal succession or ordinal structure is a (conditionally) independent statistical construct, one would expect to see sequential dynamics encoded by hippocampal neurons that are not bound to their content [47]. Self-generated sequences of neuronal firing patterns have been reported in the hippocampus [48], prefrontal cortex [49] and parietal cortex [50]. Thus, it appears that the brain is genetically equipped with neuronal architectures that encode canonical or *preconfigured* sequences, prior to those sequences being associated or imbued with (bound to) any particular content [51]. This provides a somewhat counterintuitive prediction that one should see sequential dynamics prior to any particular experience, in systems like the hippocampus [51]. This fits comfortably with recent observations that the neurons showing the greatest (sequential) firing rate modulations are impervious to the particular sequence of events experienced in the recent past [52]. Furthermore, experience with multiple sequences with different content, may be expected to engage the same canonical sequences, in the same way that the encoding of a spatial target in terms of its location is independent of its attributes [53].

These predictions also fit heuristically with the notion of fast firing units (‘choristers’) providing a canonical tempo for temporal succession, while slow firing neurons (‘soloists’) provide a context-specific content that may mediate the (plastic and context-sensitive) mapping to extra-hippocampal representations [54]. This perspective also explains the emergence of multiple place cells in the sequential encoding the pure attributes of temporal succession; i.e., the temporal order or sequence [55]. See [56] and [57] for compelling treatments of context in the Bayesian setting. In short, the picture that emerges here is of a neuronal representation of temporal succession that adumbrates any particular sequence, such that content-free sequences are associated with a particular content – through the use of auxiliary units that show a greater plasticity and context-sensitivity [58].

The statistics of neuronal encoding

An interesting aspect of a mean field approximation is that marginal probabilities have to be multiplied to generate joint distributions or distributions over outcomes. Indeed, in statistics, a ubiquitous scheme for evaluating marginal distributions – known as *belief propagation* – is also called *sum-product* message passing. This is important because the realization of the product of independent positive random variables is a lognormal process (this follows from the central limit theorem in the log domain). The implication for the statistics of neuronal encoding is that we might expect to see lognormal distributions of synaptic strengths, firing rates and burst probabilities – under the assumption that spiking encodes the probability or expectation of occupying hidden states [18].

Predictions of the 'what and when' distinction – the remembered present

There is something quite distinct about representations with and without factorisation over time and content. An inspection of the figure in Box 3 reveals that *the representations of context do not change with time*. In contrast, with an exhaustive representation of both *what* and *when*, there is no temporal invariance and expectations cascade with the progress of time. Put simply, the first word in it sentence is always the same word before or after reading it. This means that sequential (*when*) states do all the heavy lifting incurred by temporal succession, endowing contextual (*what*) representations with a form of translational invariance, not in space but over time. Effectively, this converts a sequence of representations into the representation of a sequence. Heuristically, this means the representation of a narrative, trajectory or sequence of states is no longer tied to the present; enabling – or indeed mandating – an explicit representation of the past (i.e. memory) and future [59-63]. This intuition might explain why brain structures associated with memory are also implicated in planning [64-66]. This fits comfortably with the fact that mental travel into the past and future engages the same anatomical substrates and algorithms deployed for spatial navigation in the present [67-70].

In short, the factorisation into *what* and *when* necessarily entails a working memory that can accommodate postdiction and prediction. In this setting, postdiction corresponds to the accumulation of evidence for any particular content (sequence); namely, updating beliefs about the past – and, simultaneously, predictions about the future. For example, this predicts

that there are neuronal populations in the brain that encode the current sentence, in a way that necessarily predicts its conclusion. This representation changes much more slowly than the (e.g., predictive coding) processing of graphemes and word forms that are engaged by saccadic eye movements [71-73]. In this sense, carving the world into canonical sequences – and the context under which those sequences unfold – provides a deep and hierarchical representation of time, as exemplified by the nested nature of the multitude of brain rhythms [74]. See also [75, 76]. See Figure 1 for simulated hippocampal responses during saccadic eye movements, under a mean field assumption.

Figure 1 about here

An important insight that can be drawn from *what and where* and *what and when* formulations (c.f., Box 2 and Box 3) is that the representational roles of *where* and *when* become conflated in navigation (i.e., spatial sequencing). This is remarkable since every principal neuron in the hippocampus can be regarded as either a 'place cell' [77] or 'time cell' [78] – as opposed to assigning time or space to distinct subsets of neurons. Whether cells in the hippocampus and entorhinal cortex 'code' for position versus absolute time – or distance versus duration – depends largely on the testing conditions and the theoretical perspective of the observer [79-82]. Indeed, one might anticipate that the *what* versus *where* distinction is (statistically and anatomically) conflated with the *what* versus *when* distinction; especially when dealing with trajectories in extrapersonal space. Whether its space or time, the ordinal sequences in the hippocampal system can 'index' the items (*what*) in the neocortex [83]. A marginal encoding of ordinal sequences (*where* and *when*) and the semantic meaning of the ordered items (*what*), make the division of labour analogous to the role of a librarian (hippocampus; pointing to the items) in a library (neocortex; where accumulated semantic knowledge is stored). The organized access (in spatiotemporal trajectories) to the neocortex-stored stored items (*what*) then becomes episodic information [84].

One could argue that we are simply putting a Bayesian gloss on the fact that the hippocampus encodes sequences of events. However, our proposition is somewhat simpler and subtler: the hippocampus – in contrast to other organs of succession such as the basal ganglia and

cerebellum – has a privileged role; it encodes the very essence of sequences, without reference to particular events. The content of the sequence depends on how events are 'bound' to content-free sequences, through context-sensitive and activity-dependent changes in synaptic efficacy. If this view is right, one would expect to see intrinsic (sequential) dynamics in hippocampal activity, even "in the absence of any external memory demand or spatiotemporal boundary" [82] – a prediction that is now attracting empirical attention [47, 82]. The broader empirical implication presents an avenue for falsifying the mean field hypothesis (see also [85]); namely, if a subset of hippocampal neurons encode the marginal probability of where they are in a sequence, one should be able to identify cells that are 'repurposed' for trajectories (e.g., in linear mazes) that insensitive to the particular environment or direction of travel. In other words, they should show a context-invariance that speaks to conditional independence.

Concluding remarks – Active inference and narratives

Many interesting predictions follow from this perspective. For example, place cell activity is typically identified by correlating neuronal responses with the current location of an animal. However, if the hippocampus encodes both time (*when*) and space (*what*), the activity of cells encoding the first and subsequent places visited should accumulate evidence over the duration of the sequence. This means one should be able to find neurons whose activity is predicted not by the current location but by where the animal started – and where it is going.

A second interesting corollary of this perspective on mnemonic representation of sequences is that beliefs about the future are tied to beliefs about the past. If we act upon these beliefs, then we create a (non-Markovian, i.e., history-dependent) world with rich temporal structure. This follows because, in the absence of any action of the brain on the world, the succession of worldly states can be predicted completely by the laws of nature (e.g., Hamilton's principle of least action, classical mechanics, and so on). Crucially, these laws are compatible with a Markovian world in which the next state depends only on the previous state. However, if we now put mnemonic agents into the mix – whose action depends upon the past – the world becomes much more interesting. Indeed, hippocampal firing sequences continue to evolve even in the absence of continuous sensory inputs [47]. In short, the way we represent

temporal succession and the implicit narratives that predict and explain our senses leads inevitably to behaviour that transcends the rules of classical physics.

Acknowledgements

KJF is funded by the Wellcome trust (Ref: 088130/Z/09/Z). We would also like to thank the Hungarian Academy of Sciences for hosting a lunch, during which this article was conceived.

Box 1: approximate Bayesian inference

Bayesian inference refers to optimising beliefs about a model or its hidden states (s) in the light of outcomes (o) or evidence. Formally, this can be expressed as minimising a variational free energy bound on Bayesian model evidence [86], with respect to beliefs about hidden states encoded by a probability density $Q(s)$ (with expectation: $E[Q(s)] = \mathbf{s}$)

$$\begin{aligned} F(o,\mathbf{s}) &= \underbrace{D[Q(s)\|P(s|o)]}_{\text{relative entropy}} - \underbrace{\ln P(o)}_{\text{log evidence}} \geq \underbrace{\ln P(o)}_{\text{log evidence}} \\ &= \underbrace{D[Q(s)\|P(s)]}_{\text{complexity}} - \underbrace{\mathrm{E}_{Q}[\ln P(o|s)]}_{\text{accuracy}} \end{aligned}$$

Here, the model is specified by a joint distribution over outcomes and their causes or hidden states: $P(o,s) = P(o|s)P(s)$. The first expression for free energy shows that when free energy is minimised, the relative entropy or Kullback-Leibler (KL) divergence attains its minimum (zero) and free energy becomes the negative logarithm of model evidence. In other words, when free energy is minimised, the approximate posterior beliefs become the true posterior beliefs (i.e., the distribution of hidden states given outcomes) and free energy becomes negative log evidence.

Another way of conceptualizing free energy is in terms of accuracy and complexity – as shown in the second equality. This equality shows that minimising free energy minimises complexity. Here, complexity is the KL divergence between posterior beliefs and prior beliefs (prior to any outcomes). In other words, complexity reflects the degrees of freedom – above and beyond prior beliefs – needed to provide an accurate account of observed data. It is easy to show that when one is absolutely certain about the hidden states causing data, the complexity increases with the number of hidden states entertained by the model.

The imperative to minimise complexity is known as Occam's principle and is the basis of approximations to model evidence provided by the Akaike and Bayesian information criteria [87]. The role of complexity will become important later, when we consider models with a large number of states encoding joint distributions over two factors, relative to parsimonious models (with greater model evidence) that just encode the factors or marginal densities (see Box 2). In terms of the equations above, this distinction can be expressed as the mean field approximation $Q(s) = Q(s^{where})Q(s^{what})$

Box 2: mean field approximations in the brain

Box 1 figure about here

This schematic illustrates two ways of encoding a moving object in the visual field. In both cases, the visual input corresponds to an 'H' moving downwards. The left panels show a generative model encoding a joint representation over *what* an object is and *where* it is. To generate a sequence of observations, the generative model uses state transitions, from one state to the next – where each hidden state determines the observed outcome. The **B** matrices encode state transitions, while **A** encodes a probabilistic mapping from states to outcomes. In the right-hand model, there is a separate representation for each object in every position and object motion simply entails transitions from the current object in one location to another (usually the same) object in the next location.

The left panel shows the equivalent model but under a mean field approximation, in which the joint distribution is approximated by the product of marginal distributions over the factors *what* and *where*. Here, motion is generated by transitions from one location to the next, while the object's identity remains unchanged. Crucially, the outcome rests on a product or multiplication of the two marginal representations. This is denoted by the Kronecker tensor product $\otimes$.

So which is the better model? If observations are generated by a world in which objects are invariant, then the mean field approximation provides an accurate explanation for observed outcomes with the least complexity. This is because there are fewer hidden states (or degrees of freedom) than in the joint representation. Because the same accuracy is obtained with a lower complexity, this model will have more evidence and will be selected during natural (Bayesian model) selection (see Box 1). Conversely, in a magical or ambiguous world – in which the identity of a moving object can change instantaneously – the joint model will be necessary to generate accurate predictions. For example, an instantaneous switch from 'H' to 'T' after the first observation cannot be modelled under the mean field approximation (indicated by the red arrow). In this magical world, the joint model will justify its extra complexity by providing more accurate explanations for observations. However, in a real world, it is overly complex – with a redundant or inefficient parameterisation [35].

Box 3: what and when architectures

Box 2 figure about here

The schematic uses the same form as Box 2. However, we have replaced *where* with *when* (and letters with words). The argument for a mean field (factorised or marginal) representation is exactly the same but in this context we are generating sequences over time at the same location. The joint representation (left panel) has an explicit representation of every possible sequence (labelled A, B,…). A complicated probability transition matrix then mediates jumps among hidden states to generate a sequence of outcomes.

A more parsimonious generative model – that predicts the same sequences – is shown on the right. Here, there is no explicit representation of content but simply a representation of the ordinal structure or sequence *per se* (e.g., a sentence or context). All the heavy lifting – in terms of predicting the next outcome – is done by the connections from each representation of the sentence and their interactions with connections from representations encoding sequential transitions. As in the *what* and *where* example, the what or context factor (e.g., sentence) does not change in time. Crucially, this means the representation of a sequence is not a sequence of representations. It is this architecture (mean field approximation) that enables sequential representations to transcend the passage of time.

Figure legends

Box 1 figure – no legend

Box 2 figure – no legend

Figure 1: **Simulated electrophysiological responses**. This figure illustrates the electrophysiological responses predicted by approximate Bayesian inference under a mean field assumption. These data report simulations of saccadic eye movements during reading, using the scheme described in [88]. In brief, we simulated saccadic eye movements sampling four successive 'words', under the hypotheses that the words were generated by one of six sentences. The generative model used to accumulate evidence was based on a mean field approximation that included marginal distributions over the order of words (*when*) and the six alternative sentences (*what*). **A**: (hippocampal responses) shows responses based upon a gradient descent on free energy for *when* expectations, while **B**: (cortical responses) shows the equivalent responses for *what* expectations. In this example, the first sentence was correctly inferred after the third word. The upper left panels show the activity (firing rate) of units encoding hidden states in image (raster) format, over the five epochs preceding saccades (**A**: four ordinal states. **B**: six sentences). The first column reports all hidden states, over all future time points, at the beginning of the sequence, while the rows encode each hidden states over time. This means the lower diagonal entries effectively encode the future, while the upper diagonal expectations encode beliefs about the past (i.e., memory). The upper right panels plot the same information to illustrate evidence accumulation and the resolution uncertainty about the context (i.e. sentence). The simulated local field potentials (i.e. the rate of change of neuronal firing) are shown in the lower right panel. The lower left panels show average local field potentials over all units before (dotted line) and after (solid line) bandpass filtering at 4 Hz, superimposed upon its time frequency decomposition. The important thing to take from these simulated neuronal responses is that they possess many features of empirical activity; for example, there is a natural theta-gamma coupling [89-91] due to fast (gamma) activity elicited by each cue that is sampled at a slower (theta) frequency (lower left panels). One can also see a characteristic phase precession [92] as predictions about the future are confirmed by sensory evidence. These simulations can be reproduced with the DEM toolbox, available from http://www.fil.ion.ucl.ac.uk/spm.

Glossary of (Bayesian) terms

Bayesian belief updating: the combination of prior beliefs about the causes of an observation and the likelihood of that observation to produce a posterior belief about its hidden causes. This updating conforms to Bayes rule.

Likelihood: the probability of an observation under a generative model, given its causes.

Prior belief: a probability distribution over the hidden causes of observations, before they are observed.

Posterior beliefs: a probability distribution over the hidden causes of observed consequences, after they are observed.

Hidden causes or states: the unobserved (possibly fictive) causes of observed data

Generative model: a probabilistic specification of the dependencies among causes and consequences; usually specified in terms of a prior belief and the likelihood of observations, given their causes.

Expectation: the mean or average (the first order moment of a probability distribution).

Approximate Bayesian inference: Bayesian belief updating in which approximate posterior distributions are optimized by minimizing variational free energy. The approximate posterior convergences to the true posterior when free energy is minimized.

Variational free energy: a functional of a probability distribution (and observations) that upper bounds (is always greater than) the negative log evidence for a generative model. This negative log evidence is also known as surprise or self information in information theory.

Bayesian model evidence: this is the probability that some observations were generated by a model. It is also known as the marginal or integrated likelihood because it does not depend upon the hidden causes.

Complexity: the difference or divergence between prior and posterior beliefs. The complexity of a model reflects the change in prior beliefs produced by Bayesian belief updating (also known as Bayesian surprise)

References:

1 Zeki, S. and Shipp, S. (1988) The functional logic of cortical connections. *Nature* 335, 311-317
2 Ungerleider, L.G. and Mishkin, M. (1982) Two cortical visual systems. In *Analysis of Visual Behavior* (.Ingle, D., *et al.*, eds), pp. 549-586, MIT Press
3 Kersten, D., *et al.* (2004) Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271-304
4 Dayan, P., *et al.* (1995) The Helmholtz machine. *Neural Computation* 7, 889-904
5 Hasselmo, M.E. and Stern, C.E. (2015) Current questions on space and time encoding. *Hippocampus* 25, 744-752
6 Helmholtz, H. (1866/1962) Concerning the perceptions in general. In *Treatise on physiological optics*, Dover
7 Gregory, R.L. (1980) Perceptions as hypotheses. *Phil Trans R Soc Lond B.* 290, 181-197
8 Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci.* 36, 181-204
9 Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci.* 11, 127-138
10 Ballard, D.H., *et al.* (1983) Parallel visual computation. *Nature* 306, 21-26
11 Rao, R.P. and Ballard, D.H. (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci.* 2, 79-87
12 Ashby, W.R. (1947) Principles of the self-organizing dynamic system. *J Gen Psychology.* 37, 125-128
13 Conant, R.C. and Ashby, W.R. (1970) Every Good Regulator of a system must be a model of that system. *Int. J. Systems Sci.* 1, 89-97
14 Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712-719
15 Davis, R.H. (2006) Strong Inference: rationale or inspiration? *Perspectives in biology and medicine* 49, 238-250
16 Bowers, J.S. and Davis, C.J. (2012) Bayesian just-so stories in psychology and neuroscience. *Psychol Bull* 138, 389-414
17 Reynolds, J.H. and Heeger, D.J. (2009) The normalization model of attention. *Neuron* 61, 168-185
18 Buzsaki, G. and Mizuseki, K. (2014) The log-dynamic brain: how skewed distributions affect network operations. *Nat Rev Neurosci* 15, 264-278
19 Jaakkola, T. and Jordan, M. (1998) Improving the Mean Field Approximation Via the Use of Mixture Distributions. In *Learning in Graphical Models* (Jordan, M., ed), pp. 163-173, Springer Netherlands
20 Buice, M.A. and Cowan, J.D. (2009) Statistical mechanics of the neocortex. *Progress in biophysics and molecular biology* 99, 53-86
21 Couchman, J.J., *et al.* (2010) Carving nature at its joints using a knife called concepts. *The Behavioral and brain sciences* 33, 207-208
22 Gershman, S.J. and Niv, Y. (2010) Learning latent structure: carving nature at its joints. *Curr Opin Neurobiol* 20, 251-256
23 Buzsaki, G. (1998) Memory consolidation during sleep: a neurophysiological perspective. *Journal of sleep research* 7 Suppl 1, 17-23
24 Paulin, M.G. (2005) Evolution of the cerebellum as a neuronal machine for Bayesian state estimation. *J Neural Eng.* 2, S219-234
25 Fernando, C., *et al.* (2012) Selectionist and evolutionary approaches to brain function: a critical appraisal. *Frontiers in computational neuroscience* 6, 24
26 Harper, M. (2011) Escort evolutionary game theory. *Physica D-Nonlinear Phenomena* 240, 1411-1415
27 Sella, G. and Hirsh, A.E. (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci.* 102, 9541-9546
28 Tononi, G., *et al.* (1994) A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc Natl Acad Sci U S A* 91, 5033-5037
29 Park, H.J. and Friston, K. (2013) Structural and functional brain networks: from connections to cognition. *Science* 342, 1238411
30 Lees, R.B. (1957) *Language* 33, 375-408
31 Tyler, L.K., *et al.* (2010) Preserving syntactic processing across the adult life span: the modulation of the frontotemporal language system in the context of age-related atrophy. *Cereb Cortex* 20, 352-364
32 Pallier, C., *et al.* (2011) Cortical representation of the constituent structure of sentences. *Proc Natl Acad Sci U S A* 108, 2522-2527
33 Ungerleider, L.G. and Haxby, J.V. (1994) 'What' and 'where' in the human brain. *Current Opinion in Neurobiology* 4, 157-165
34 Goodale, M.A., *et al.* (2004) Two distinct modes of control for object-directed action. *Progress in brain research* 144, 131-144

35 Barlow, H. (1961) Possible principles underlying the transformations of sensory messages. In *Sensory Communication* (Rosenblith, W., ed), pp. 217-234, MIT Press
36 Linsker, R. (1990) Perceptual neural organization: some approaches based on network models and information theory. *Annu Rev Neurosci.* 13, 257-281
37 Optican, L. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II Information theoretic analysis. *J Neurophysiol.* 57, 132-146
38 Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607-609
39 Zeki, S. (2005) The Ferrier Lecture 1995 behind the seen: The functional specialization of the brain in space and time. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 1145–1183.
40 Angelucci, A. and Bressloff, P.C. (2006) Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog Brain Res.* 154, 93-120
41 Zucker, H.R. and Ranganath, C. (2015) Navigating the human hippocampus without a GPS. *Hippocampus* 25, 697-703
42 Dehaene, S., *et al.* (2015) The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron* 88, 2-19
43 Verschure, P.F.M.J. and Edelman, G.M. (1992) The Remembered Present: A Biological Theory of Consciousness. *The American Journal of Psychology* 105, 477
44 Wittner, L., *et al.* (2007) Three-dimensional reconstruction of the axon arbor of a CA3 pyramidal cell recorded and filled in vivo. *Brain structure & function* 212, 75-83
45 Felleman, D. and Van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1, 1-47
46 Mesulam, M.M. (1998) From sensation to cognition. *Brain* 121, 1013-1052
47 Itskov, V., *et al.* (2011) Cell assembly sequences arising from spike threshold adaptation keep track of time in the hippocampus. *J Neurosci* 31, 2828-2834
48 Pastalkova, E., *et al.* (2008) Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322-1327
49 Fujisawa, S., *et al.* (2008) Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat Neurosci* 11, 823-833
50 Harvey, C.D., *et al.* (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* 484, 62-68
51 Mizuseki, K. and Buzsaki, G. (2013) Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex. *Cell reports* 4, 1010-1021
52 Grosmark, A.D. and Buzsaki, G. (2016) Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science* 351, 1440-1443
53 Stark, E., *et al.* (2015) Local generation of multineuronal spike sequences in the hippocampal CA1 region. *Proc Natl Acad Sci U S A* 112, 10521-10526
54 Okun, M., *et al.* (2015) Diverse coupling of neurons to populations in sensory cortex. *Nature* 521, 511-515
55 Buzsaki, G. (2015) Neuroscience. Our skewed sense of space. *Science* 347, 612-613
56 Fuhs, M.C. and Touretzky, D.S. (2007) Context learning in the rodent hippocampus. *Neural Comput* 19, 3173-3215
57 Gershman, S.J., *et al.* (2010) Context, learning, and extinction. *Psychol Rev* 117, 197-209
58 Dragoi, G., *et al.* (2003) Place representation within hippocampal networks is modified by long-term potentiation. *Neuron* 39, 843-853
59 Eichenbaum, H. and Fortin, N.J. (2009) The neurobiology of memory based predictions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364, 1183-1191
60 Manns, J.R., *et al.* (2007) Gradual changes in hippocampal activity support remembering the order of events. *Neuron* 56, 530-540
61 Schacter, D.L., *et al.* (2008) Episodic simulation of future events - Concepts, data, and applications. In *Year in Cognitive Neuroscience 2008* (Kingstone, A. and Miller, M.B., eds), pp. 39-60
62 Scoville, W.B. and Milner, B. (1957) LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *Journal of Neurology Neurosurgery and Psychiatry* 20, 11-21
63 Squire, L.R. (1992) MEMORY AND THE HIPPOCAMPUS - A SYNTHESIS FROM FINDINGS WITH RATS, MONKEYS, AND HUMANS. *Psychological Review* 99, 195-231
64 Buckner, R.L. (2010) The role of the hippocampus in prediction and imagination. *Annual review of psychology* 61, 27-48, c21-28
65 Epstein, R., *et al.* (1999) The parahippocampal place area: Recognition, navigation, or encoding? *Neuron* 23, 115-125

66 Pastalkova, E., *et al.* (2008) Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322-1327
67 Buzsaki, G. and Moser, E.I. (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nat Neurosci* 16, 130-138
68 Hassabis, D. and Maguire, E.A. (2007) Deconstructing episodic memory with construction. *Trends Cogn Sci* 11, 299-306
69 Zeidman, P., *et al.* (2015) Investigating the functions of subregions within anterior hippocampus. *Cortex; a journal devoted to the study of the nervous system and behavior* 73, 240-256
70 Izquierdo, I. and Medina, J.H. (1997) Memory formation: The sequence of biochemical events in the hippocampus and its connection to activity in other brain structures. *Neurobiology of Learning and Memory* 68, 285-316
71 Rayner, K. (2009) Eye Movements in Reading: Models and Data. *J Eye Mov Res.* 2, 1–10
72 Friston, K., *et al.* (2012) Perceptions as hypotheses: saccades as experiments. *Front Psychol.* 3, 151
73 Pierrotdeseilligny, C., *et al.* (1995) CORTICAL CONTROL OF SACCADES. *Annals of Neurology* 37, 557-567
74 Buzsaki, G., *et al.* (2013) Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. *Neuron* 80, 751-764
75 George, D. and Hawkins, J. (2009) Towards a mathematical theory of cortical micro-circuits. *PLoS Comput Biol* 5, e1000532
76 Kiebel, S.J., *et al.* (2008) A hierarchy of time-scales and the brain. *PLoS Comput Biol.* 4, e1000209
77 O'Keefe, J. and Dostrovsky, J. (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171-175
78 Eichenbaum, H. (2014) Time cells in the hippocampus: a new dimension for mapping memories. *Nat Rev Neurosci* 15, 732-744
79 Buzsaki, G. (2013) Cognitive neuroscience: Time, space and memory. *Nature* 497, 568-569
80 Schiller, D., *et al.* (2015) Memory and Space: Towards an Understanding of the Cognitive Map. *J Neurosci* 35, 13904-13911
81 Kraus, B.J., *et al.* (2015) During Running in Place, Grid Cells Integrate Elapsed Time and Distance Run. *Neuron* 88, 578-589
82 Villette, V., *et al.* (2015) Internally Recurring Hippocampal Sequences as a Population Template of Spatiotemporal Information. *Neuron* 88, 357-366
83 Teyler, T.J. and DiScenna, P. (1986) The hippocampal memory indexing theory. *Behav Neurosci* 100, 147-154
84 Tulving, E. (1987) Multiple memory systems and consciousness. *Human neurobiology* 6, 67-80
85 Hasselmo, M.E. (2015) If I had a million neurons: Potential tests of cortico-hippocampal theories. *Prog Brain Res* 219, 1-19
86 Fox, C. and Roberts, S. (2011) A tutorial on variational Bayes. In *Artificial Intelligence Review*, pp. DOI 10.1007/s10462-10011-19236-10468, Springer
87 Penny, W.D. (2012) Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* 59, 319-330
88 Friston, K., *et al.* (2015) Active inference and epistemic value. *Cogn Neurosci*, 1-28
89 Canolty, R.T., *et al.* (2006) High gamma power is phase-locked to theta oscillations in human neocortex. *Science* 313, 1626-1628
90 Lisman, J. and Redish, A.D. (2009) Prediction, sequences and the hippocampus. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364, 1193-1201
91 Lisman, J. and Buzsaki, G. (2008) A neural coding scheme formed by the combined function of gamma and theta oscillations. *Schizophr Bull* 34, 974-980
92 Skaggs, W.E., *et al.* (1996) Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* 6, 149-172