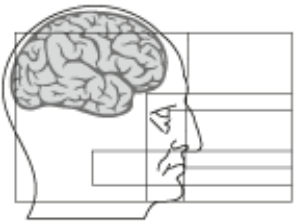


MAX PLANCK UCL CENTRE
for Computational Psychiatry and Ageing Research

UCL

Uncertainty, precision, prediction errors – and their relevance to computational psychiatry

Christoph Mathys

Wellcome Trust Centre for Neuroimaging at UCL, London, UK
Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK

Black Dog Institute, University of New South Wales
March 6, 2015

Uncertainty: a shamelessly artificial example

Imagine the following situation:

You're on a boat, you're lost in a storm and trying to get back to shore. A lighthouse has just appeared on the horizon, but you can only see it when you're at the peak of a wave. Your GPS etc., has all been washed overboard, but what you can still do to get an idea of your position is to measure the angle between north and the lighthouse. These are your measurements (in degrees):

76, 73, 75, 72, 77

What number are you going to base your calculation on?

Right. The mean: 74.6. How do you calculate that?

Uncertainty: updates to the mean

The usual way to calculate the mean $\bar{x}$ of $x_1, x_2, \ldots, x_n$ is to take

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This requires you to remember all x_i, which can become inefficient. Since the measurements arrive sequentially, we would like to update $\bar{x}$ sequentially as the x_i come in – without having to remember them.

It turns out that this is possible. After some algebra (see next slide), we get

$$\bar{x}_n = \bar{x}_{n-1} + \frac{1}{n}(x_n - \bar{x}_{n-1})$$

Uncertainty: updates to the mean

Proof of sequential update formula:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_n}{n} + \frac{1}{n}\sum_{i=1}^{n-1} x_i = \frac{x_n}{n} + \frac{n-1}{n}\underbrace{\frac{1}{n-1}\sum_{i=1}^{n-1} x_i}_{=\bar{x}_{n-1}} =$$

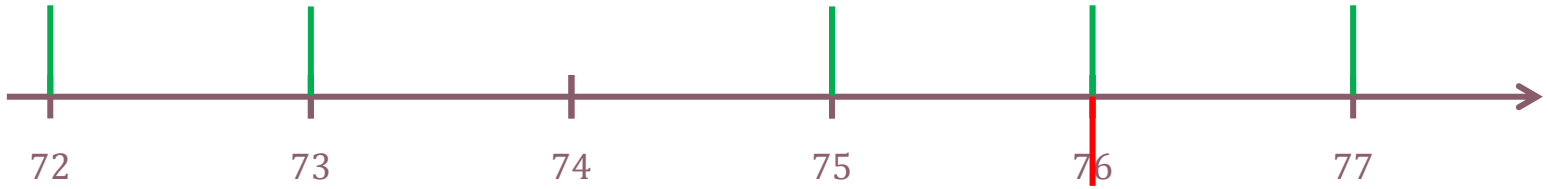
$$= \frac{x_n}{n} + \frac{n-1}{n}\bar{x}_{n-1} = \bar{x}_{n-1} + \frac{x_n}{n} + \frac{n-1}{n}\bar{x}_{n-1} - \frac{n}{n}\bar{x}_{n-1} =$$

$$= \bar{x}_{n-1} + \frac{1}{n}(x_n + (n-1-n)\bar{x}_{n-1}) = \bar{x}_{n-1} + \frac{1}{n}(x_n - \bar{x}_{n-1})$$

q.e.d.

Uncertainty: updates to the mean

The seqential updates in our example now look like this:



$$\bar{x}_1 = 76$$

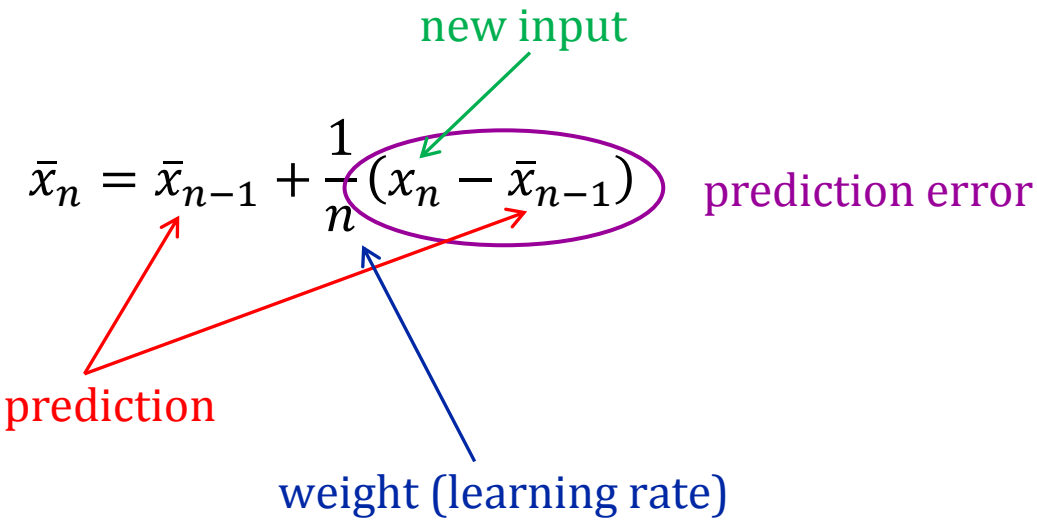
$$\bar{x}_2 = 76 + \frac{1}{2}(73 - 76) = 74.5$$

$$\bar{x}_3 = 74.5 + \frac{1}{3}(75 - 74.5) = 74.\overline{6}$$

$$\bar{x}_4 = 74.\overline{6} + \frac{1}{4}(72 - 74.\overline{6}) = 74$$

$$\bar{x}_5 = 74 + \frac{1}{5}(77 - 74) = 74.6$$

What are the building blocks of the updates we've just seen, and where does uncertainty enter?



So where is uncertainty?

The **learning rate** reflects uncertainty:
the less we know, the higher the learning rate.

Is this a general pattern?

- More specifically, does it generalize to Bayesian inference?
- «Bayesian inference» simply means inference on uncertain quantities according to the rules of probability theory (i.e., according to logic).
- Agents who use Bayesian inference will make better predictions (provided they have a good model of their environment), which will give them an evolutionary advantage.
- We may therefore assume that evolved biological agents use Bayesian inference, or a close approximation to it.
- So is Bayesian inference based on predictions that are updated using uncertainty-weighted prediction errors?

Updates in a simple Gaussian model

- Think boat, lighthouse, etc., again, but now we're doing Bayesian inference.

- Before we make the next observation, our belief about the true angle ϑ can be described by a Gaussian prior:

$$p(\vartheta) \sim \mathcal{N}(\mu_\vartheta, \pi_\vartheta^{-1})$$

- The likelihood of our observation is also Gaussian, with precision π_ε:

$$p(x|\vartheta) \sim \mathcal{N}(\vartheta, \pi_\varepsilon^{-1})$$

- Bayes' rule now tells us that the posterior is Gaussian again:

$$p(\vartheta|x) = \frac{p(x|\vartheta)p(\vartheta)}{\int p(x|\vartheta')p(\vartheta')\mathrm{d}\vartheta'} \sim \mathcal{N}\left(\mu_{\vartheta|y}, \pi_{\vartheta|y}^{-1}\right)$$

Updates in a simple Gaussian model

- Here's how the updates to the sufficent statistics μ and π describing our belief look like:

$$\pi_{\vartheta|x} = \pi_\vartheta + \pi_\varepsilon$$

$$\mu_{\vartheta|x} = \mu_\vartheta + \frac{\pi_\varepsilon}{\pi_{\vartheta|x}}(x - \mu_\vartheta)$$

prediction error

prediction

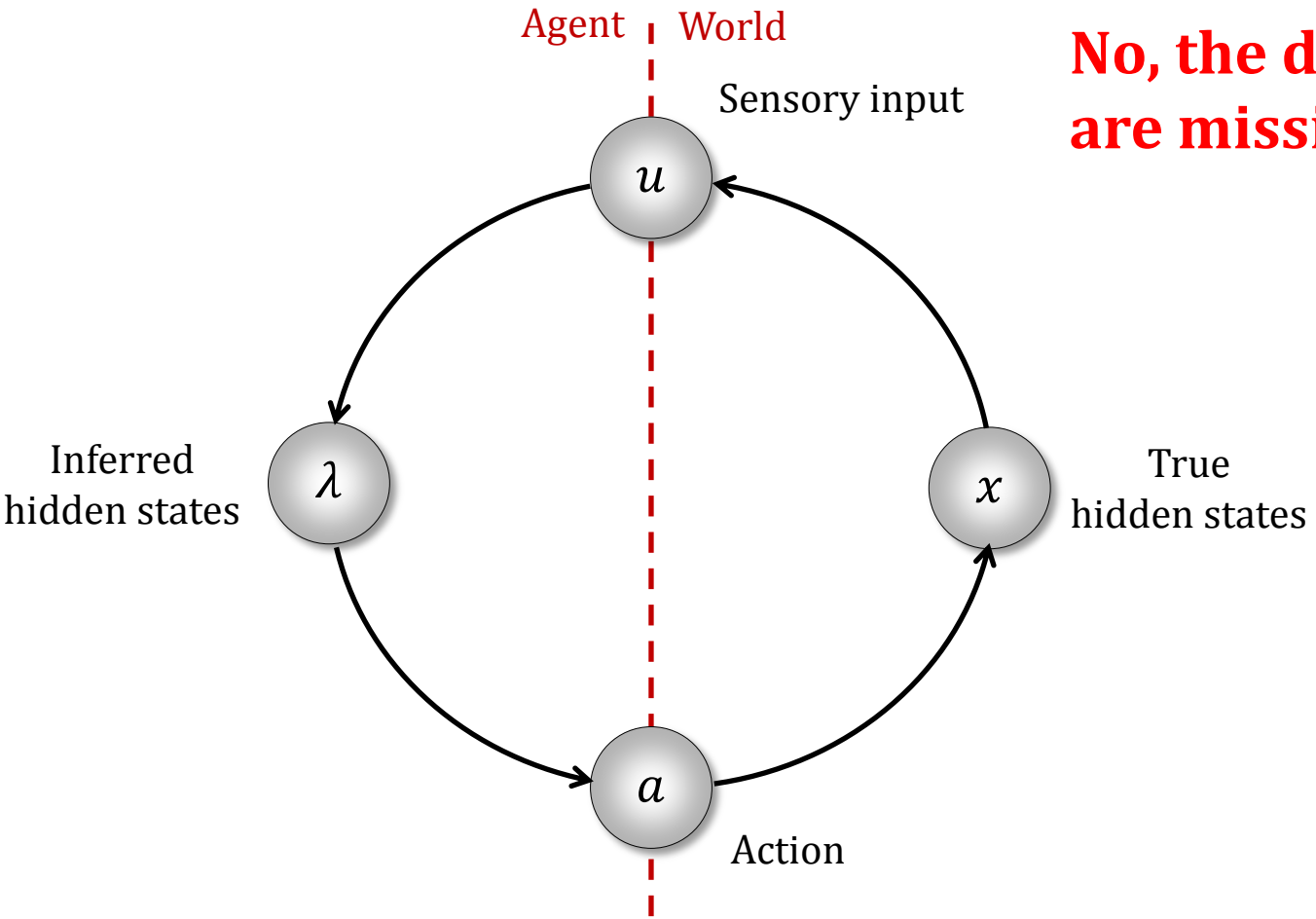
weight (learning rate)$=\frac{\text{how much we're learning here}}{\text{how much we already know}}$

- So it's the same story all over again: the mean is updated by an uncertainty-weighted (more specifically: prediction-weighted) prediction error.

- The size of the update is proportional to the likelihood precision and inversely proportional to the posterior precision.

- This pattern is not specific to the univariate Gaussian case, but generalizes to Bayesian updates for all exponential families of likelihood distributions with conjugate priors (i.e., to all formal descriptions of inference you are ever likely to need).

Does inference as we've described it adequately describe the situation of actual biological agents?



No, the dynamics are missing!

What about dynamics?

- Up to now, we've only looked at inference on static quantities, but biological agents live in a continually changing world.
- In our example, the boat's position changes and with it the angle to the lighthouse.
- How can we take into account that old information becomes obsolete? If we don't, our learning rate becomes smaller and smaller because our eqations were derived under the assumption that we're accumulating information about a stable quantity.

What's the simplest way to keep the learning rate from going too low?

- Keep it constant!
- So, taking the update equation for the mean of our observations as our point of departure...

$$\bar{x}_n = \bar{x}_{n-1} + \frac{1}{n}(x_n - \bar{x}_{n-1}),$$

- ... we simply replace $\frac{1}{n}$ with a constant α:

$$\mu_n = \mu_{n-1} + \alpha(x_n - \mu_{n-1}).$$

- This is called *Rescorla-Wagner learning* [although it wasn't this line of reasoning that led Rescorla & Wagner (1972) to their formulation].

Does a constant learning rate solve our problems?

- Partly: it implies a certain rate of forgetting because it amounts to taking only the $n = \frac{1}{\alpha}$ last data points into account. But...
- ... if the learning rate is supposed to reflect uncertainty in Bayesian inference, then how do we
- (a) know that α reflects the right level of uncertainty at any one time, and
- (b) account for changes in uncertainty if α is constant?
- What we really need is an adaptive learning that accurately reflects uncertainty.

An adaptive learning rate that accurately reflects uncertainty

- This requires us to think a bit more about what kinds of uncertainty we are dealing with.

- A possible taxonomy of uncertainty is (cf. Yu & Dayan, 2003; Payzan-LeNestour & Bossaerts, 2011):

- (a) **outcome uncertainty** that remains unaccounted for by the model, called *risk* by economists (π_ε in our Bayesian example); this uncertainty remains even when we know all parameters exactly,

- (b) **informational** or *expected* uncertainty about the value of model parameters ($\pi_{\vartheta|x}$ in the Bayesian example),

- (c) **environmental** or *unexpected* uncertainty owing to changes in model parameters (not accounted for in our Bayesian example, hence unexpected).

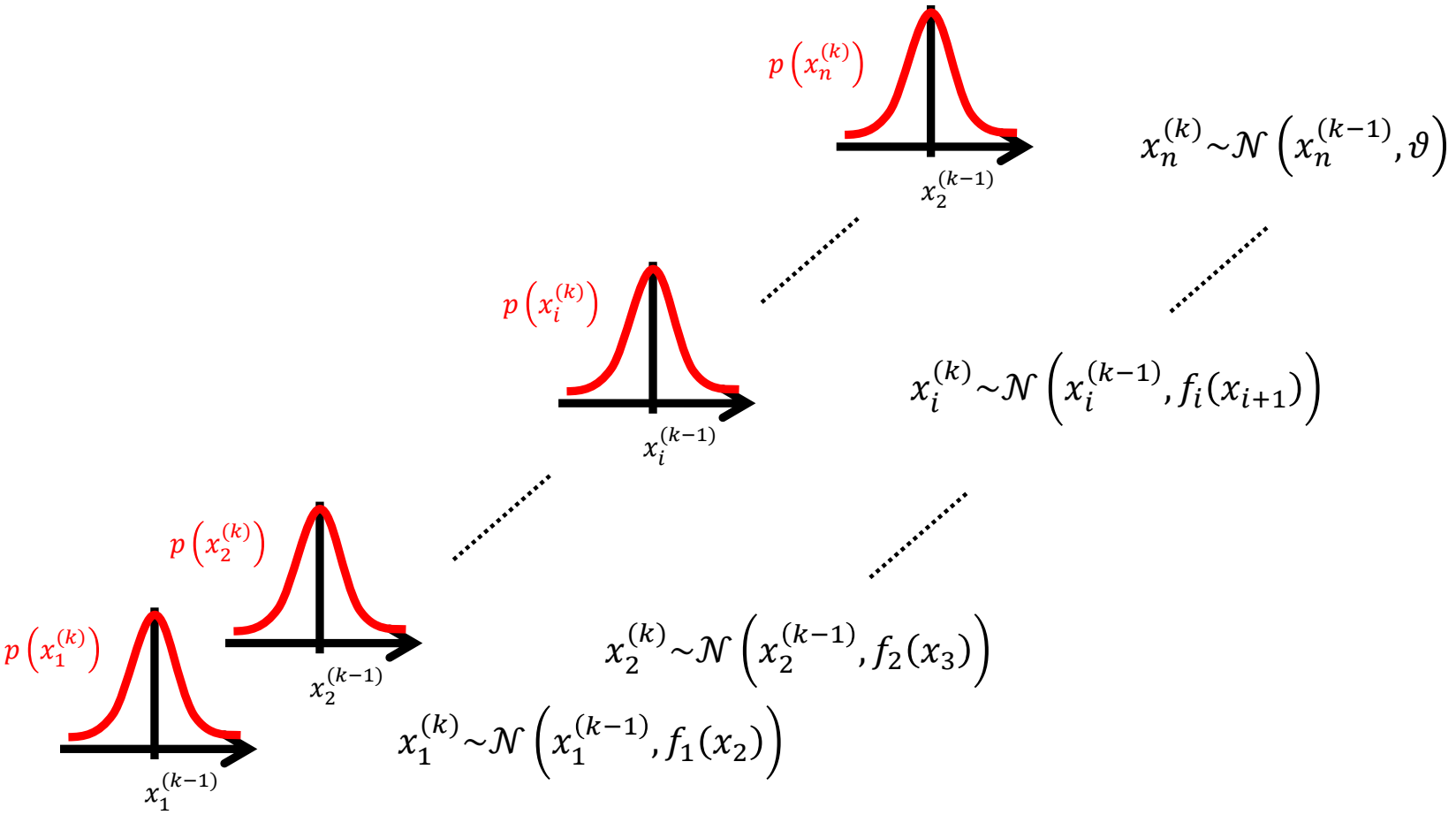
An adaptive learning rate that accurately reflects uncertainty

- Various efforts have been made to come up with an adaptive learning rate:
  - Kalman (1960)
  - Sutton (1992)
  - Nassar et al. (2010)
  - Payzan-LeNestour & Bossaerts (2011)
  - Mathys et al. (2011)
  - Wilson et al. (2013)

- The Kalman filter is optimal for linear dynamical systems, but realistic data usually require non-linear models.

- Mathys et al. use a generic non-linear hierarchical Bayesian model that allows us to derive update equations that are optimal in the sense that they minimize surprise.

The hierarchical Gaussian filter (HGF)



The hierarchical Gaussian filter (HGF)

- At the outcome level (i.e., at the very bottom of the hierarchy), we have

$$u^{(k)} \sim \mathcal{N}\left(x_1^{(k)}, \hat{\pi}_u^{-1}\right)$$

- This gives us the following update for our belief on x_1 (our quantity of interest):

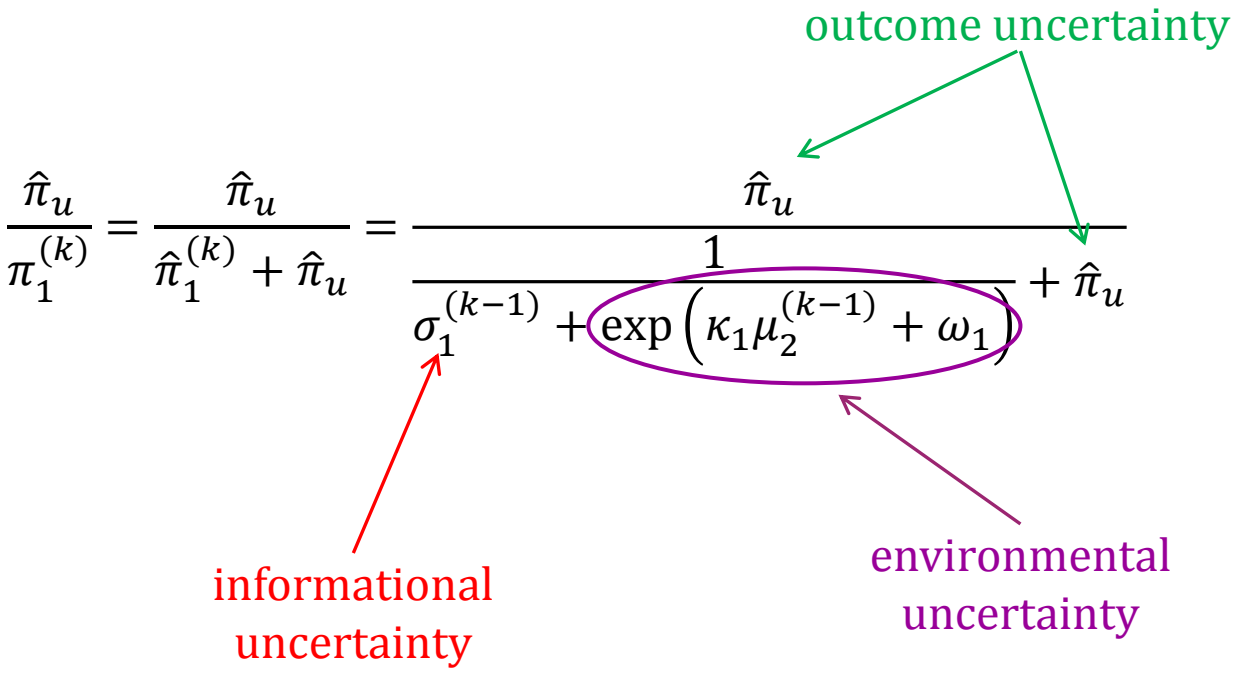
$$\pi_1^{(k)} = \hat{\pi}_1^{(k)} + \hat{\pi}_u$$

$$\mu_1^{(k)} = \mu_1^{(k-1)} + \frac{\hat{\pi}_u}{\pi_1^{(k)}}\left(u^{(k)} - \mu_1^{(k-1)}\right)$$

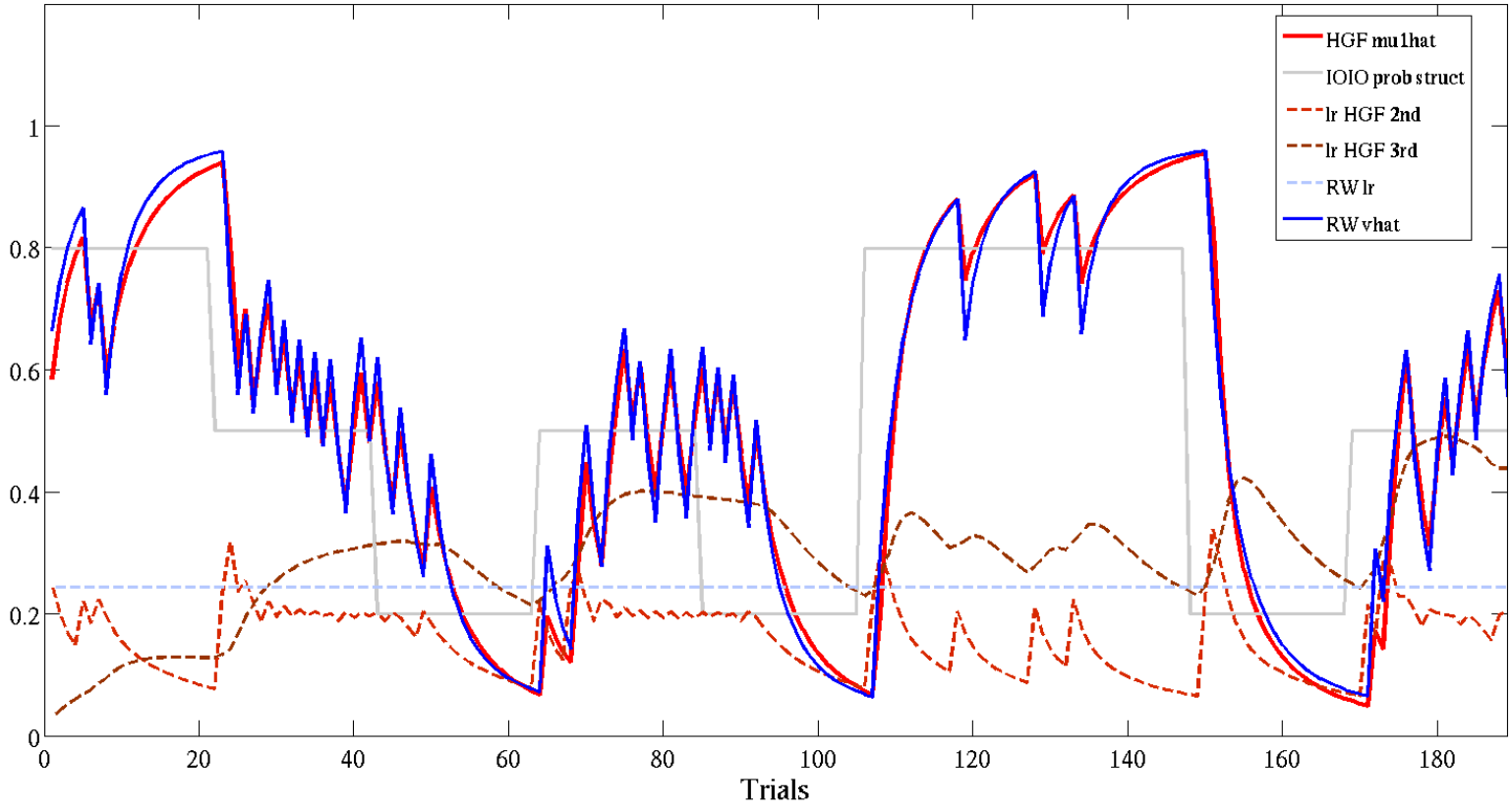
- The familiar structure again – but now with a learning rate that is responsive to all kinds of uncertainty, including environmental (unexpected) uncertainty.

The learning rate in the HGF

Unpacking the learning rate, we see:



The learning rate in the HGF



Andreea Diaconescu

HGF: empirical evidence (Iglesias et al., 2013)

Model comparison:

	Behavioral study		fMRI study 1		fMRI study 2	
BMS results	PP	XP	PP	XP	PP	XP
HGF1	0.8435	1	0.7422	1	0.7166	1
HGF2	0.0259	0	0.0200	0	-	-
HGF3	0.0361	0	0.1404	0	0.1304	0
Sutton	0.0685	0	0.0710	0	0.0761	0
RW	0.0260	0	0.0264	0	0.0769	0

HGF: empirical evidence (Iglesias et al., 2013)

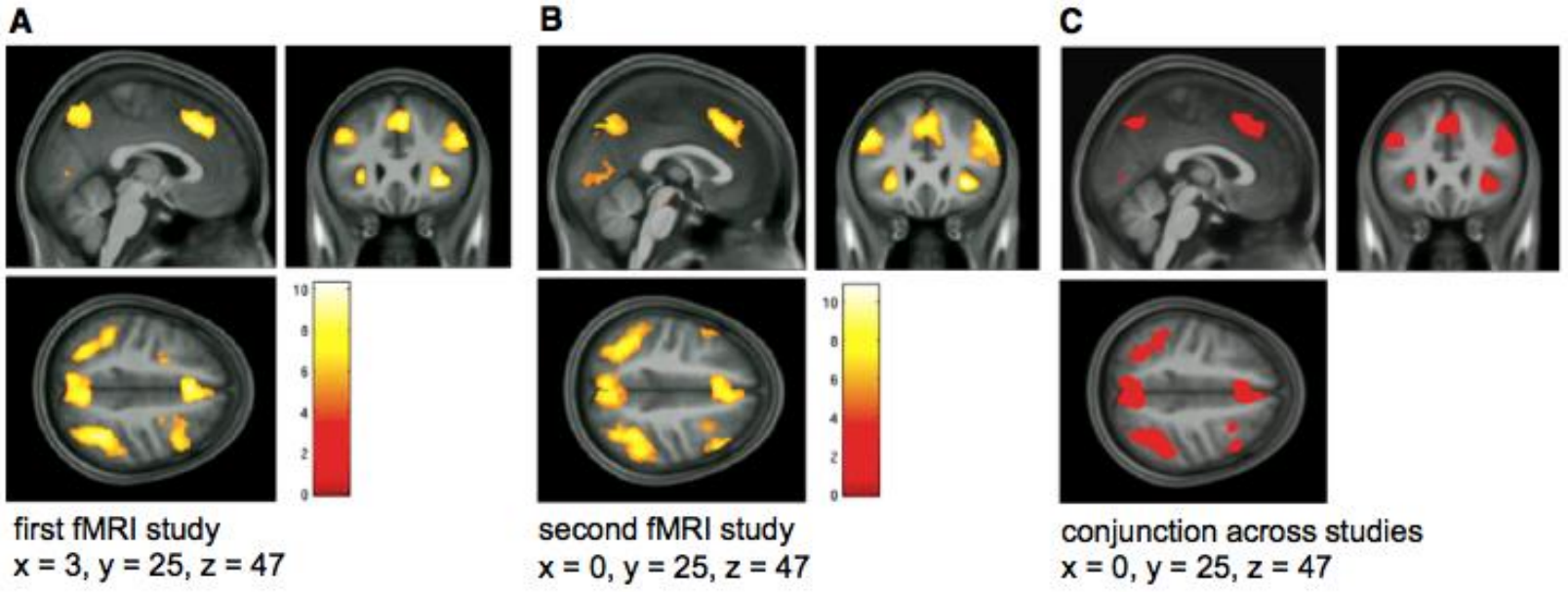


Figure 2. Whole-Brain Activations by ε_2

Activations by precision-weighted prediction error about visual stimulus outcome, ε_2, in the first fMRI study (A) and the second fMRI study (B). Both activation maps are shown at a threshold of $p < 0.05$, FWE corrected for multiple comparisons across the whole brain. To highlight replication across studies, (C) shows the results of a "logical AND" conjunction, illustrating voxels that were significantly activated in both studies.

HGF: empirical evidence (Iglesias et al., 2013)

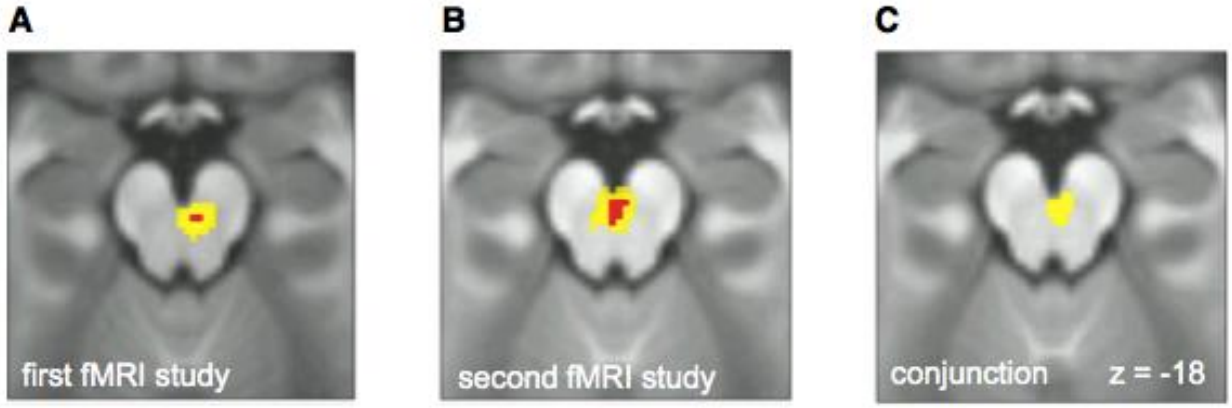


Figure 3. Midbrain Activation by ε_2
Activation of the dopaminergic VTA/SN associated with precision-weighted prediction error about stimulus category, ε_2. This activation is shown both at $p < 0.05$ FWE whole-brain corrected (red) and $p < 0.05$ FWE corrected for the volume of our anatomical mask comprising both dopaminergic and cholinergic nuclei (yellow).
(A) Results from the first fMRI study.
(B) Second fMRI study.
(C) Conjunction (logical AND) across both studies.

HGF: empirical evidence (Iglesias et al., 2013)

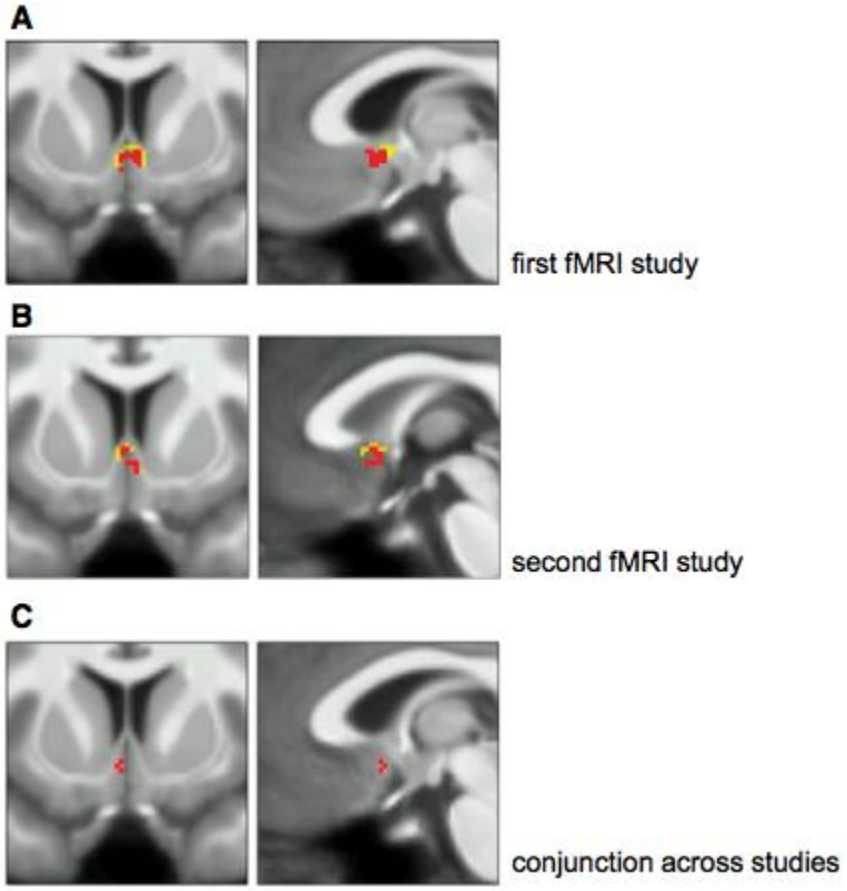


Figure 6. Basal Forebrain Activations by ε_3
Activation of the cholinergic basal forebrain associated with precision-weighted prediction error about stimulus probabilities ε_3 within the anatomically defined mask. For visualization of the activation area we overlay the results thresholded at $p < 0.05$ FWE corrected for the entire anatomical mask (red) on the results thresholded at $p < 0.001$ uncorrected (yellow) in the first (A: $x = 3, y = 9, z = -8$) and the second fMRI study (B: $x = 0, y = 10, z = -8$). (C) The conjunction analysis ("logical AND") across both studies ($x = 2, y = 11, z = -8$).

How to estimate and compare models: the HGF Toolbox

- Available at

  http://www.translationalneuromodeling.org/tapas

- Start with README, manual, and interactive demo
- Modular, extensible
- Matlab-based

Why is this important for computational psychiatry?

A failure of sensory or interoceptive attenuation may be at the root of many clinical phenomena.

Cogn Process (2013) 14:411–427
DOI 10.1007/s10339-013-0571-3

RESEARCH REPORT

Active inference, sensory attenuation and illusions

Harriet Brown · Rick A. Adams · Isabel Parees · Mark Edwards · Karl Friston

frontiers in PSYCHIATRY

The computational anatomy of psychosis

Rick A. Adams[1]*, Klaas Enno Stephan[1,2,3], Harriet R. Brown[1], Christopher D. Frith[1] and Karl J. Friston[1]

doi:10.1093/brain/aws129

BRAIN
A JOURNAL OF NEUROLOGY

OCCASIONAL PAPER

A Bayesian account of 'hy

Mark J. Edwards,[1,*] Rick A. Adams,[2,*] Harriet Bro

frontiers in HUMAN NEUROSCIENCE

HYPOTHESIS

An aberrant precision account of autism

Rebecca P. Lawson[1]*, Geraint Rees[1,2] and Karl J. Friston[1]

[1] Wellcome Trust Centre for Neuroimaging, University College London, London, UK

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

frontiers in PSYCHOLOGY

Free-energy and

Harriet Brown* and Karl J. Fri

The Wellcome Trust Centre for Neuroimaging,

Edited by:
Lars Muckli, University of Glasgow,

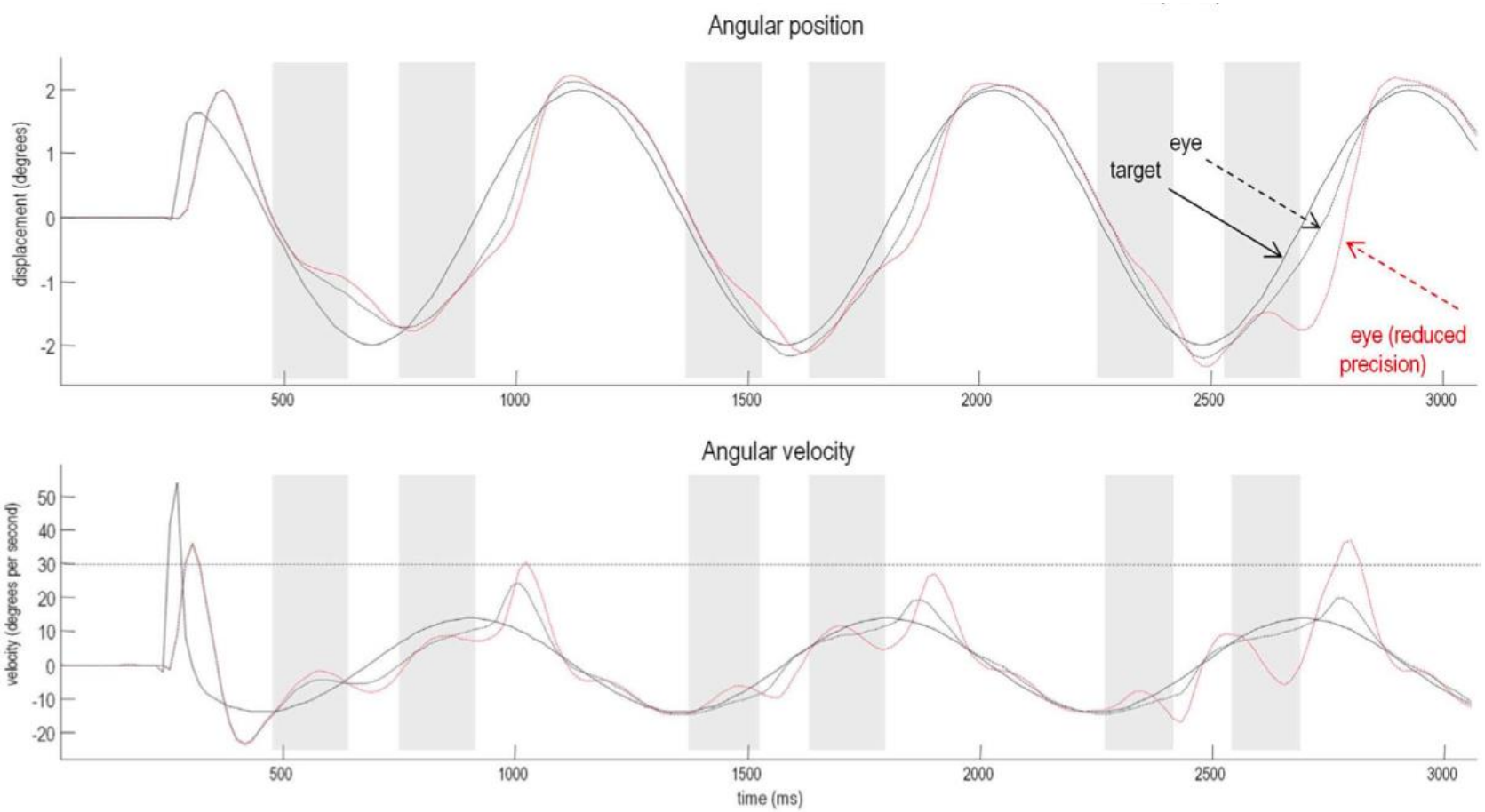
Review

Autism, oxytocin and interoception

E. Quattrocki*, Karl Friston[1]

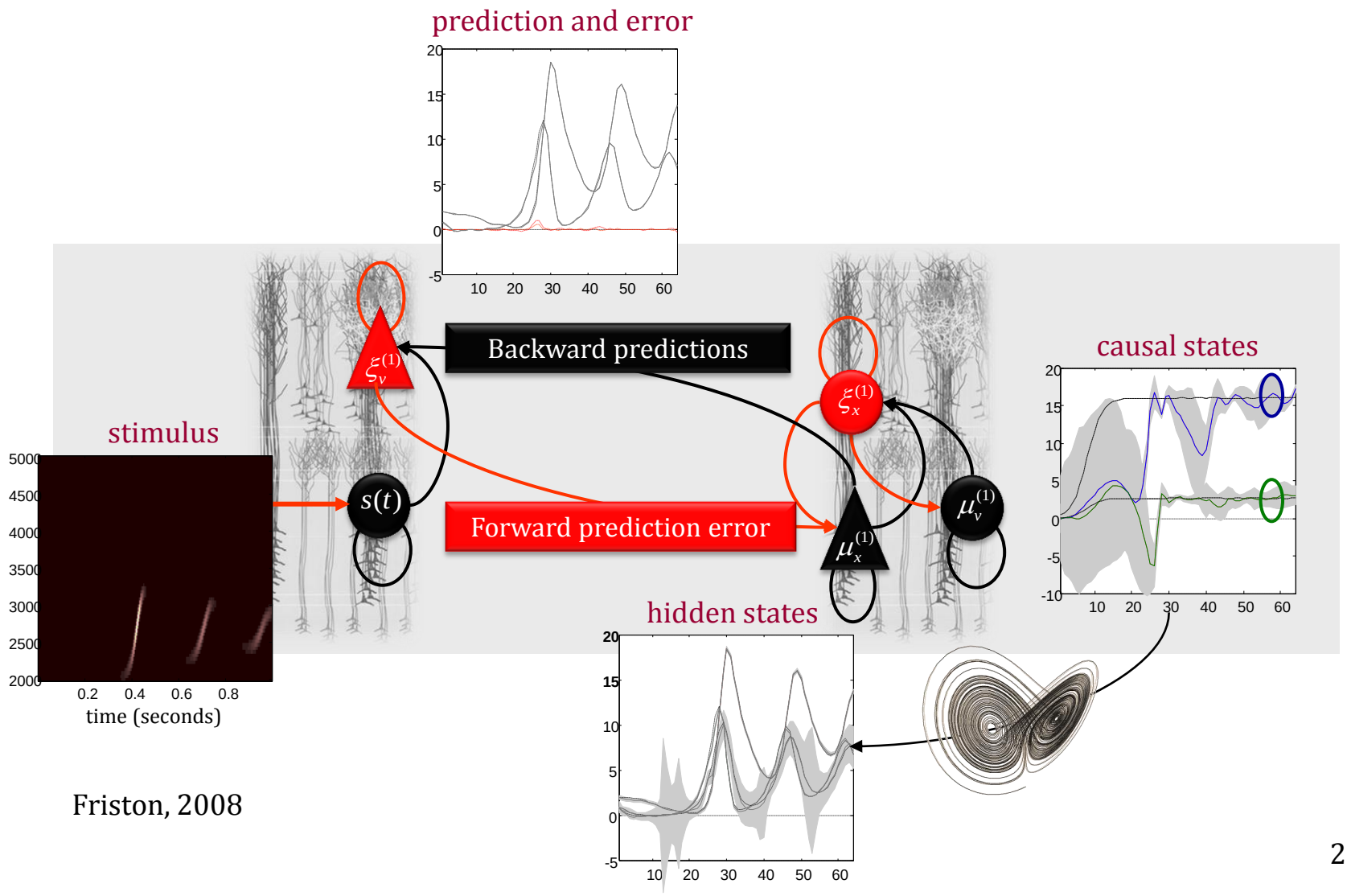
The Wellcome Trust Centre for Neuroimaging, UCL, 12 Queen Square, London WC1N 3BG, UK

Example: Abnormalities of smooth pursuit eye movements in schizophrenia



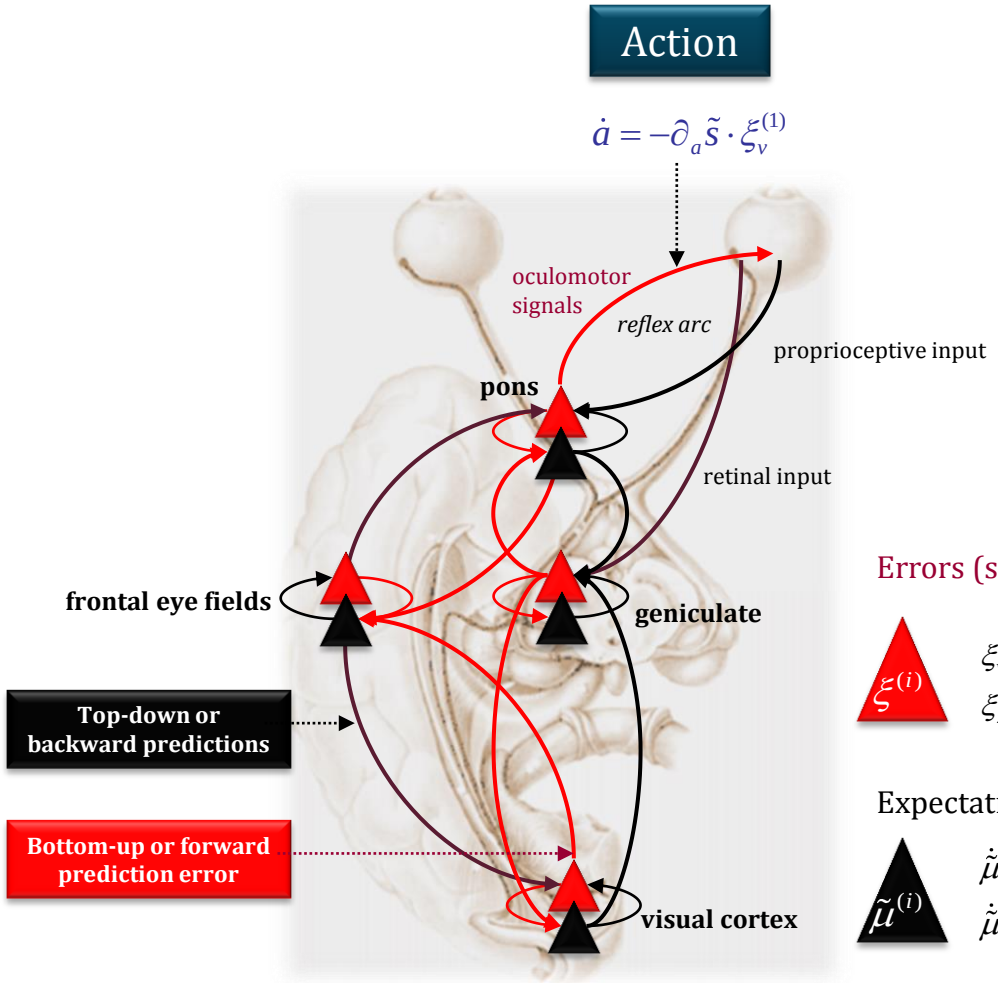
Thaker et al., 1999; Hong et al., 2005; Adams et al., 2013

The neurobiology of the Bayesian brain: predictive coding



Friston, 2008

Example: Hierarchical message passing in the visual-oculomotor system



Perception

Precision-weighted prediction errors

Errors (superficial pyramidal cells)

$$\xi^{(i)}$$

$$\xi_v^{(i)} = \tilde{\Pi}_v^{(i)} \tilde{\varepsilon}_v^{(i)} = \tilde{\Pi}_v^{(i)} (\tilde{\mu}_v^{(i-1)} - g^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}))$$
$$\xi_x^{(i)} = \tilde{\Pi}_x^{(i)} \tilde{\varepsilon}_x^{(i)} = \tilde{\Pi}_x^{(i)} (\mathcal{D}\tilde{\mu}_x^{(i)} - f^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}))$$

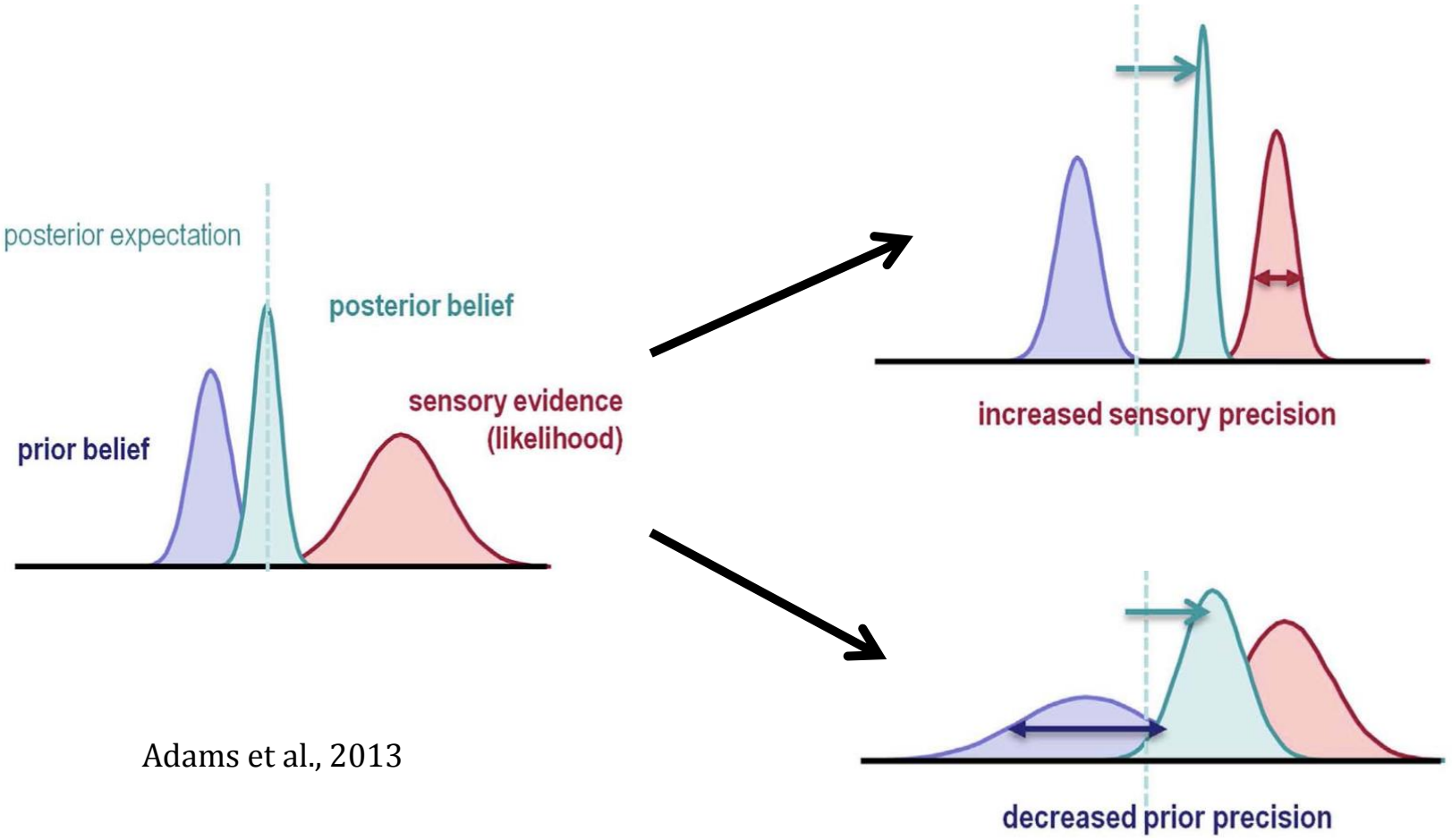
Expectations (deep pyramidal cells)

$$\tilde{\mu}^{(i)}$$

$$\dot{\tilde{\mu}}_v^{(i)} = D\tilde{\mu}_v^{(i)} - \partial_{\tilde{v}} \tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} - \xi_v^{(i+1)}$$
$$\dot{\tilde{\mu}}_x^{(i)} = D\tilde{\mu}_x^{(i)} - \partial_{\tilde{x}} \tilde{\varepsilon}^{(i)} \cdot \xi^{(i)}$$

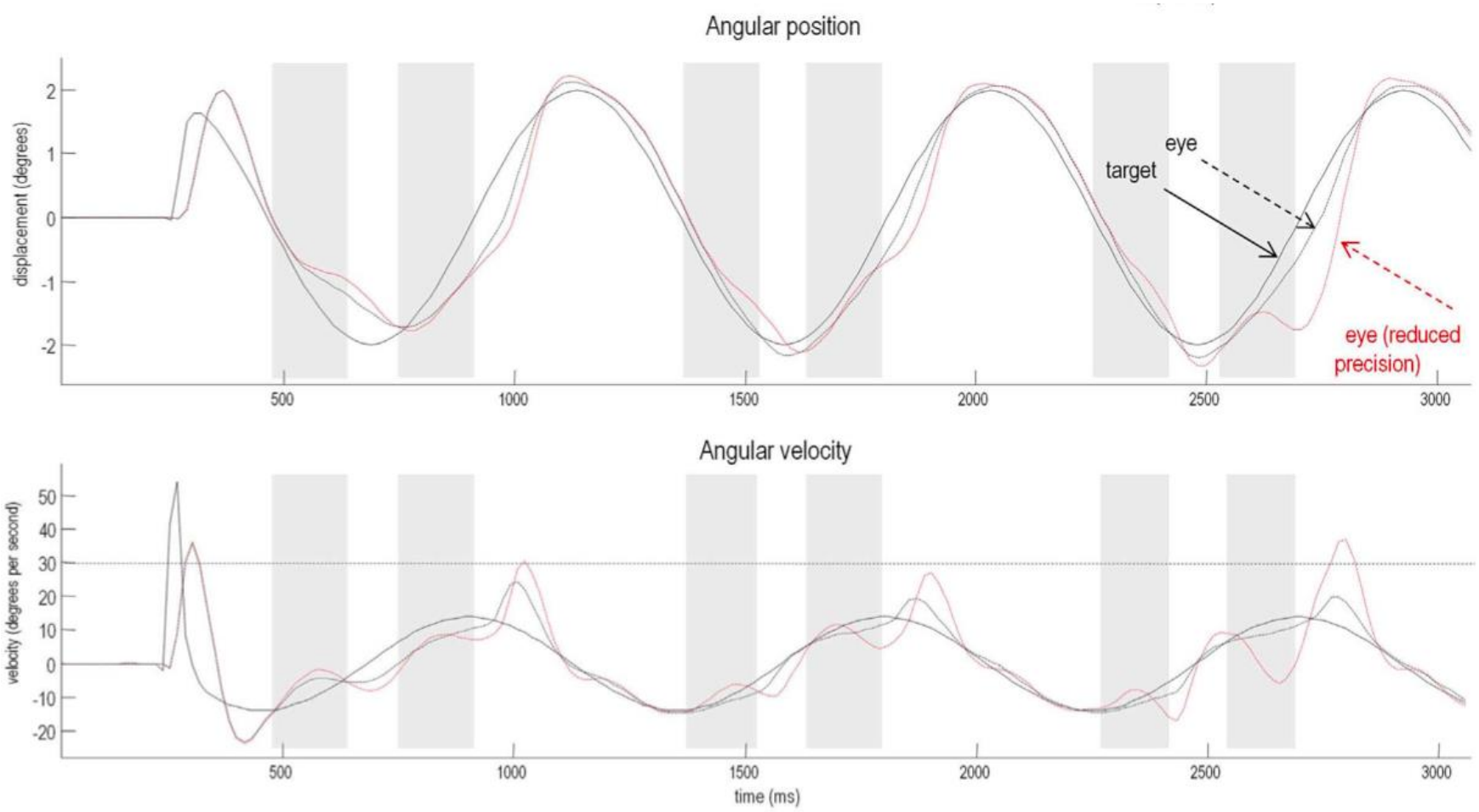
Adams et al., 2013

Overweighting of sensory evidence



Adams et al., 2013

Smooth pursuit of a partially occluded target with and without high-level precision



Adams et al., 2013

Failure of input attenuation: further examples

- Force matching illusion (Shergill et al., 2003; 2005; Teufel et al., 2010; Brown et al., 2013)

- Cornsweet effect (Brown & Friston, 2012)

- Hollow mask illusion (Dima et al., 2009; 2010)

  https://www.youtube.com/watch?v=ORoTCBrCKIQ

  https://www.youtube.com/watch?v=6YIPtJlCbIA

- Attenuation of interoceptive signals in autism (Lawson et al., 2014; Quattrocki & Friston, 2014)

Summary

- We have to make good predictions to avoid surprise and survive, that is we have to use probabilistic (i.e., Bayesian) inference based on a good model of our environment.

- Bayesian inference means updating beliefs by uncertainty- (i.e., precision-) weighted prediction errors.

- Precision-weighting has to take account of all forms of uncertainty.

- A breakdown in this may be the root of many psychopathological phenomena.

Thanks