

MAX PLANCK UCL CENTRE  
for Computational Psychiatry and Ageing Research



# Introduction to hierarchical probabilistic inference

Christoph Mathys

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

Zurich Computational Psychiatry Course



# A spectacular piece of information

The screenshot shows the BBC News Magazine homepage. At the top, there's a navigation bar with the BBC logo, a sign-in link, and categories like News, Sport, Weather, iPlayer, TV, and Radio. Below this is a large red banner with the words 'NEWS MAGAZINE' in white. Underneath the banner is a secondary navigation bar with links for Home, World, UK, England, N. Ireland, Scotland, Wales, Business, Politics, Health, Education, Sci/Ec, Video & Audio, Magazine (which is highlighted in orange), Editors' Blog, In Pictures, Also in the News, Have Your Say, and Special Report. The main headline 'Does chocolate make you clever?' is displayed in large, bold, black text. Below it, the author is listed as 'By Charlotte Pritchard' from 'BBC News'. A timestamp '19 November 2012 Last updated at 18:19' is shown, along with sharing options for 44K, Share, and social media platforms (Facebook, Twitter, Email, Print).

## Does chocolate make you clever?

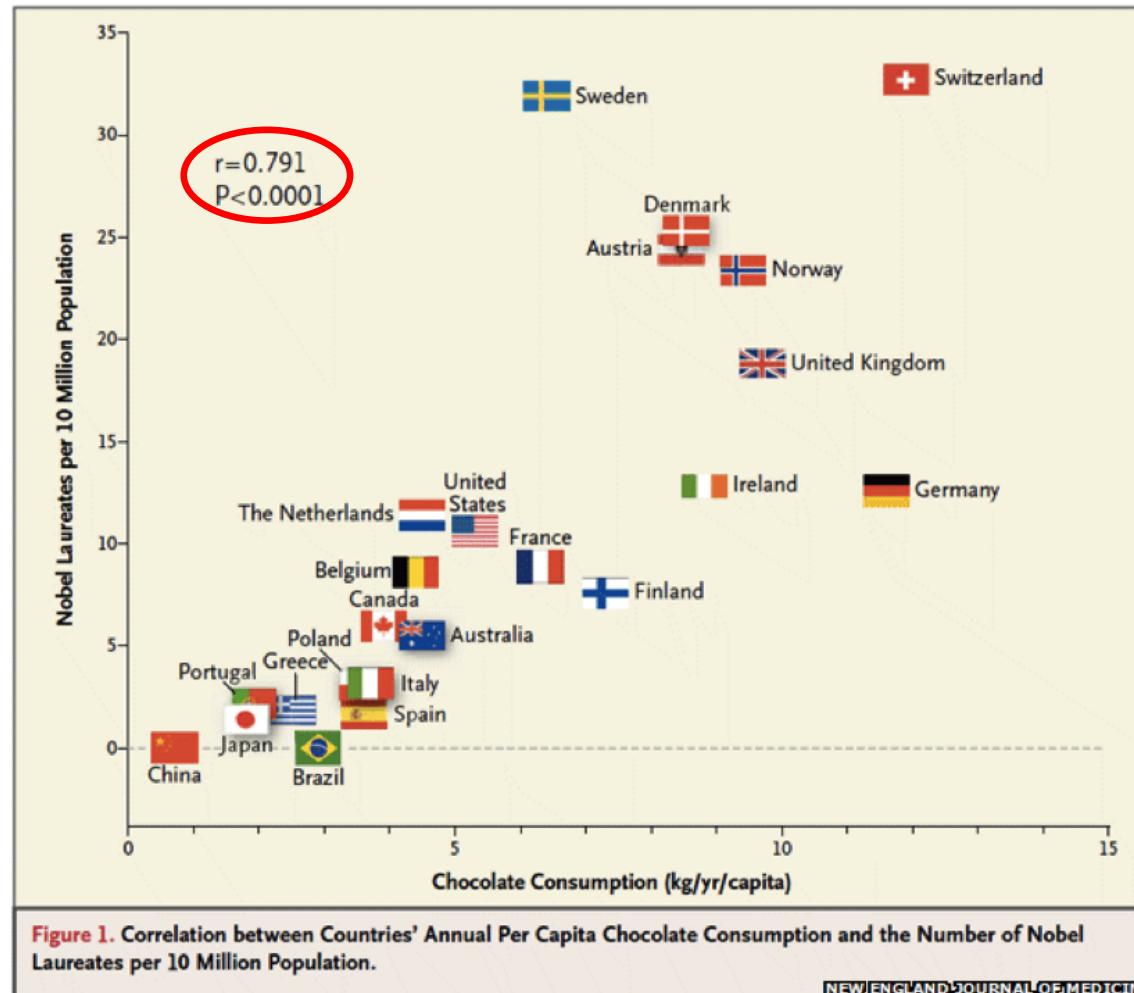
By Charlotte Pritchard

BBC News

Eating more chocolate improves a nation's chances of producing Nobel Prize winners - or at least that's what a recent study appears to suggest. But how much chocolate do Nobel laureates eat, and how could any such link be explained?

# A spectacular piece of information

- Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, 367(16), 1562–1564.





# So will I win the Nobel prize if I eat lots of chocolate?

- This is a question referring to uncertain quantities. Like almost all scientific questions, it cannot be answered by deductive logic. Nonetheless, quantitative answers can be given – but they can only be given in terms of probabilities.

- Our question here can be rephrased in terms of a conditional probability:

$$p(\text{Nobel} \mid \text{lots of chocolate}) = ?$$

- To answer it, we have to learn to calculate such quantities. The tool for this is Bayesian inference.



# «Bayesian» = logical and logical = probabilistic

«The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.»

— James Clerk Maxwell, 1850

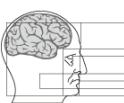


# «Bayesian» = logical and **logical = probabilistic**

But in what sense is probabilistic reasoning (i.e., reasoning about uncertain quantities according to the rules of probability theory) «logical»?

R. T. Cox showed in 1946 that the rules of probability theory can be derived from three basic desiderata:

1. Representation of degrees of plausibility by real numbers
2. Qualitative correspondence with common sense (in a well-defined sense)
3. Consistency



# The rules of probability

By mathematical proof (i.e., by deductive reasoning) the three desiderata as set out by Cox imply the rules of probability (i.e., the rules of inductive reasoning).

This means that anyone who accepts the desiderata must accept the following rules:

1.  $\sum_a p(a) = 1$  (Normalization)
2.  $p(b) = \sum_a p(a, b)$  (Marginalization – also called the **sum rule**)
3.  $p(a, b) = p(a|b)p(b) = p(b|a)p(a)$  (Conditioning – also called the **product rule**)

«Probability theory is nothing but common sense reduced to calculation.»

— Pierre-Simon Laplace, 1819



# Conditional probabilities

The probability of  **$a$  given  $b$**  is denoted by

$$p(a|b).$$

In general, this is different from the probability of  $a$  alone (the *marginal* probability of  $a$ ), as we can see by applying the sum and product rules:

$$p(a) = \sum_b p(a, b) = \sum_b p(a|b)p(b)$$

Because of the product rule, we also have the following rule (**Bayes' theorem**) for going from  $p(a|b)$  to  $p(b|a)$ :

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} = \frac{p(a|b)p(b)}{\sum_{b'} p(a|b')p(b')}$$

## The chocolate example

In our example, it is immediately clear that  $P(\text{Nobel}|\text{chocolate})$  is very different from  $P(\text{chocolate}|\text{Nobel})$ . While the first is hopeless to determine directly, the second is much easier to find out: ask Nobel laureates how much chocolate they eat. Once we know that, we can use Bayes' theorem:

can use Bayes' theorem:

$$p(\text{Nobel}|\text{chocolate}) = \frac{p(\text{chocolate}|\text{Nobel}) p(\text{Nobel})}{p(\text{chocolate})}$$

posterior

likelihood

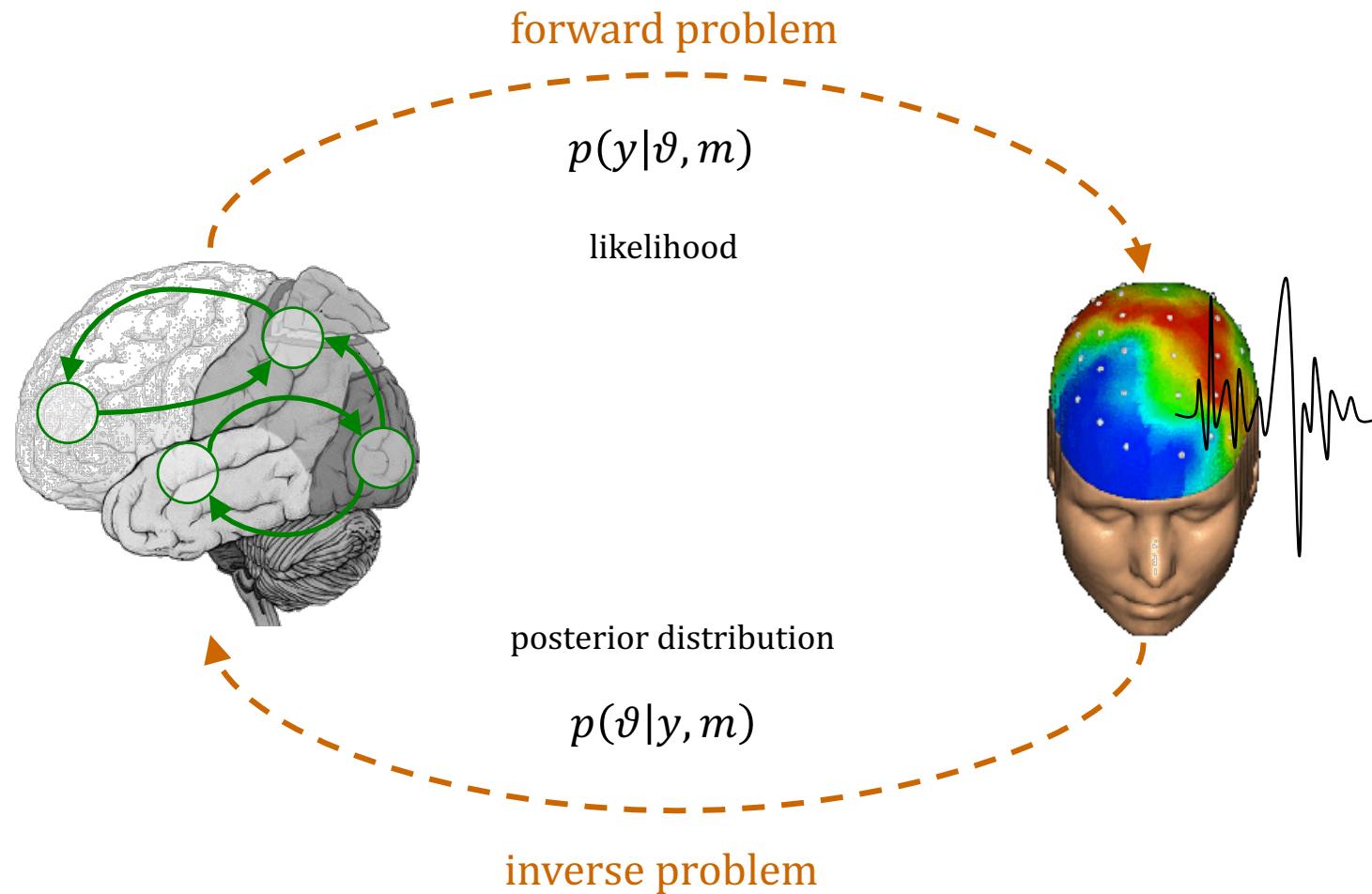
model

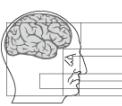
prior

evidence

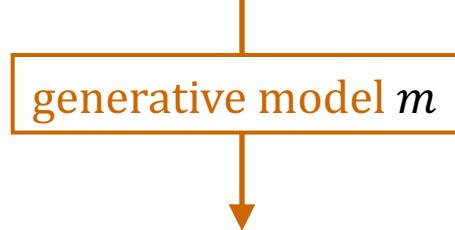
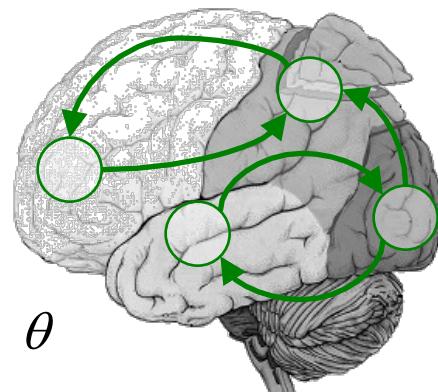
Inference on the quantities of interest in neuroscientific studies has exactly the same general structure.

# Inference in computational psychiatry





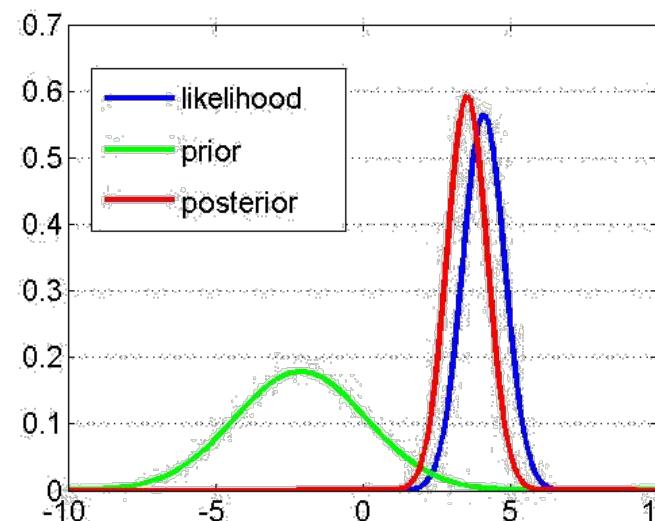
## Inference on neural processes



Likelihood:  $p(y|\vartheta, m)$

Prior:  $p(\vartheta|m)$

Bayes' theorem:  $p(\vartheta|y, m) = \frac{p(y|\vartheta, m)p(\vartheta|m)}{p(y|m)}$





# A simple example of Bayesian inference

(adapted from Jaynes (1976))

Two manufacturers, A and B, deliver the same kind of components that turn out to have the following lifetimes (in hours):

A:

59.5814  
37.3953  
47.5956  
40.5607  
48.6468  
36.2789  
31.5110  
31.3606  
45.6517

B:

48.8506  
48.7296  
59.1971  
51.8895

Assuming prices are comparable, from which manufacturer would you buy?



## A simple example of Bayesian inference

How do we compare such samples?

- By comparing their arithmetic means

Why do we take means?

- If we take the mean as our estimate, the error in our estimate is the mean of the errors in the individual measurements
- Taking the mean as maximum-likelihood estimate implies a **Gaussian error distribution**
- A Gaussian error distribution appropriately reflects our **prior** knowledge about the errors whenever we know nothing about them except perhaps their variance



# A simple example of Bayesian inference

What next?

- Let's do a t-test (but first, let's compare variances with an F-test):

```
>> [fh,fp,fci,fstats] = vartest2(xa,xb)
```

fh = 0                  fp = 0.3297                  fci = 0.2415                  fstats = 19.0173

fstat: 3.5114  
df1: 8  
df2: 3

Variances not significantly different!

```
>> [h, p, ci, stats]= ttest2(xa,xb)
```

h = 0                  p = 0.0665                  ci = -21.0191                  stats = 0.8151

tstat: -2.0367  
df: 11  
sd: 8.2541

Means not significantly different!

Is this satisfactory? No, so what can we learn by turning to probability theory (i.e., Bayesian inference)?



# A simple example of Bayesian inference

The procedure in brief:

- Determine your question of interest («What is the probability that...?»)
- Specify your models (likelihood and prior – and models, that's plural!)
- Calculate or approximate the full posterior using Bayes' theorem
- Compare your models (check their predictions, look at model evidence, etc.)
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

(Ok, sometimes you'll encounter a nasty integral – but that's a technical difficulty, not a conceptual one).

-> See <https://github.com/chmathys/bayesian-inference-example>



# Model comparison: negative variational free energy $F$

**log – model evidence** :=  $\log p(y|m)$

$$\begin{aligned}
 & \xrightarrow{\text{sum rule}} = \log \int p(y, \vartheta | m) d\vartheta \\
 & \xrightarrow{\text{multiply by } 1 = \frac{q(\vartheta)}{q(\vartheta)}} = \log \int q(\vartheta) \frac{p(y, \vartheta | m)}{q(\vartheta)} d\vartheta \\
 & \xrightarrow{\text{Jensen's inequality}} \geq \int q(\vartheta) \log \frac{p(y, \vartheta | m)}{q(\vartheta)} d\vartheta \\
 & =: F = \text{negative variational free energy}
 \end{aligned}$$

a lower bound on the log-model evidence

$$\begin{aligned}
 F &:= \int q(\vartheta) \log \frac{p(y, \vartheta | m)}{q(\vartheta)} d\vartheta \\
 &\xrightarrow{\text{product rule}} = \int q(\vartheta) \log \frac{p(y|\vartheta, m)p(\vartheta|m)}{q(\vartheta)} d\vartheta \\
 &= \underbrace{\int q(\vartheta) \log p(y|\vartheta, m) d\vartheta}_{\text{Accuracy (expected log-likelihood)}} - \underbrace{KL[q(\vartheta), p(\vartheta|m)]}_{\text{Complexity}}
 \end{aligned}$$

Kullback-Leibler divergence



## Model comparison: $F$ in relation to Bayes factors, AIC, BIC

$$\text{Bayes factor} := \frac{p(y|m_1)}{p(y|m_0)} = \exp\left(\log \frac{p(y|m_1)}{p(y|m_0)}\right) = \exp(\log p(y|m_1) - \log p(y|m_0))$$

Bayes factor

$$\approx \exp(F_1 - F_0)$$

[Meaning of the Bayes factor:

$$\frac{p(m_1|y)}{p(m_0|y)} = \frac{p(y|m_1)}{p(y|m_0)} \frac{p(m_1)}{p(m_0)}$$

Posterior odds

Prior odds

$$F = \int q(\vartheta) \log p(y|\vartheta, m) d\vartheta - KL[q(\vartheta), p(\vartheta|m)]$$

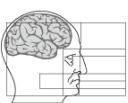
= Accuracy – Complexity

$$\text{AIC} := \text{Accuracy} - p$$

Number of parameters

$$\text{BIC} := \text{Accuracy} - \frac{p}{2} \log N$$

Number of data points



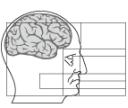
## A note on informative priors

- Any model consists of two parts: likelihood and prior.
- The choice of likelihood requires as much justification as the choice of prior because it is just as «subjective» as that of the prior.
- The data never speak for themselves. They only acquire meaning when seen through the lens of a model. However, this does not mean that all is subjective because models differ in their validity.
- In this light, the widespread concern that informative priors might bias results (while the form of the likelihood is taken as a matter of course requiring no justification) is misplaced.
- Informative priors are an important tool and their use can be justified by establishing the validity (face, construct, and predictive) of the resulting model as well as by model comparison.



## A note on uninformative priors

- Using a flat or «uninformative» prior doesn't make you more «data-driven» than anybody else. It's a choice that requires just as much justification as any other.
- For example, if you're studying a small effect in a noisy setting, using a flat prior means assigning the same prior probability mass to the interval covering effect sizes -1 to +1 as to that covering effect sizes +999 to +1001.
- Far from being unbiased, this amounts to a bias in favor of implausibly large effect sizes. Using flat priors exposes you to a great risk of chasing noise, especially if your dataset is small and your measurements are noisy.
- One way to address this is to collect enough data to swamp any inappropriate priors. A cheaper way is to use more appropriate priors.
- Disclaimer: if you look at my papers, you will find flat priors. I'll try to do better in future. So do as I say, not as I do.



# Hierarchical models

- A non-hierarchical model consists of a likelihood and a prior. The likelihood gives the probability of the observations given parameter values, and the prior gives the probability of parameter values.

$$p(y|\vartheta)p(\vartheta)$$

- The prior  $p(\vartheta)$  on the parameter  $\vartheta$  may itself have parameters  $\varphi$  (called hyperparameters) with a hyperprior  $p(\varphi)$ .

$$p(y|\vartheta)p(\vartheta|\varphi)p(\varphi)$$

- And so on.
- Models with multiple layers like this are called hierarchical models.
- They are useful in various contexts. We will here deal with the hierarchical Bayesian modelling of time series.



# Time series: 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?



# Updates to the mean

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

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



# Updates to the mean

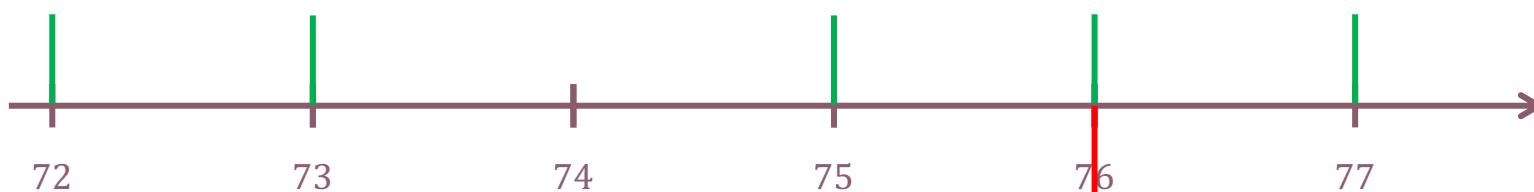
Proof of sequential update formula:

$$\begin{aligned}\bar{x}_{n+1} &= \frac{1}{n+1} \sum_{i=1}^{n+1} x_i = \frac{x_{n+1}}{n+1} + \frac{1}{n+1} \sum_{i=1}^n x_i = \frac{x_{n+1}}{n+1} + \frac{n}{n+1} \underbrace{\frac{1}{n} \sum_{i=1}^n x_i}_{=\bar{x}_n} = \\ &= \frac{x_{n+1}}{n+1} + \frac{n}{n+1} \bar{x}_n = \bar{x}_n + \frac{x_{n+1}}{n+1} + \frac{n}{n+1} \bar{x}_n - \frac{n+1}{n+1} \bar{x}_n = \\ &= \bar{x}_n + \frac{1}{n+1} (x_{n+1} + (n-n-1) \bar{x}_n) = \bar{x}_n + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)\end{aligned}$$

q.e.d.

# Uncertainty: updates to the mean

The sequential updates in our example now look like this:



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

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

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

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

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

# What are the building blocks of the updates we've just seen?

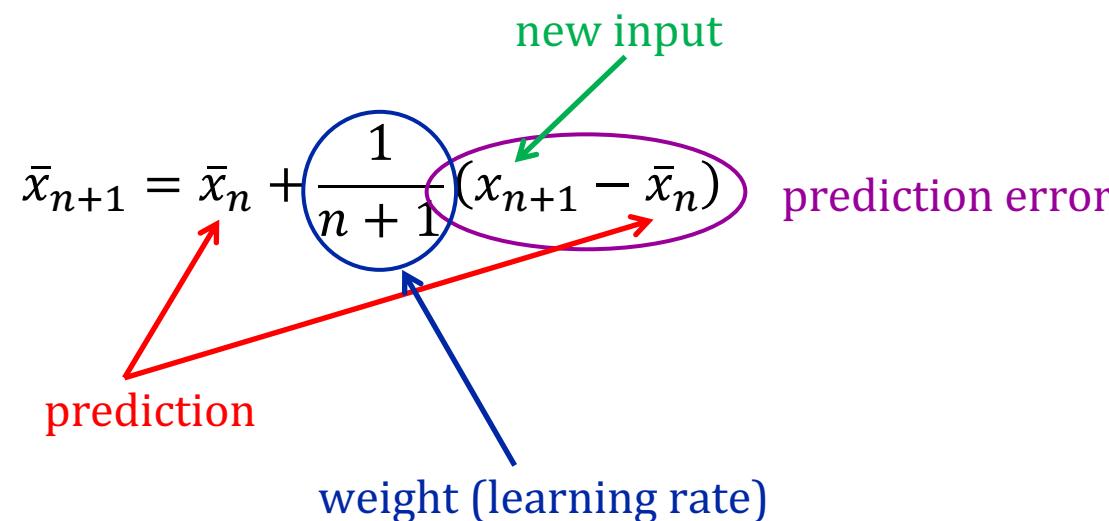
$$\bar{x}_{n+1} = \bar{x}_n + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)$$

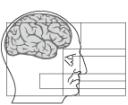
new input

prediction error

prediction

weight (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  $\vartheta$  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  $\pi_\varepsilon$  :

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

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

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

# Updates in a simple Gaussian model

- Here's how the updates to the sufficient statistics  $\mu$  and  $\pi$  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** ← (purple oval)  
**prediction** →  $\mu_\vartheta$   
**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: precision-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).



# The analogy with simple mean updating goes further

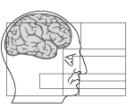
- Reminder (Gaussian update):

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

- Reducing by  $\pi_\varepsilon$  the fraction of precisions that make the learning rate, we get

$$\mu_{\vartheta|x} = \mu_\vartheta + \frac{1}{\frac{\pi_\vartheta}{\pi_\varepsilon} + 1}(x - \mu_\vartheta)$$

- This is again our equation for updating an arithmetic mean, but with  $n$  replaced by  $\frac{\pi_\vartheta}{\pi_\varepsilon}$ .
- This shows that Bayesian inference on the mean of a Gaussian distribution entails nothing more than updating the arithmetic mean of observations with  $\frac{\pi_\vartheta}{\pi_\varepsilon} =: \nu$  as a proxy for the number of prior observations, i.e. for the **weight of the prior relative to the observation**.



# “Every good regulator of a system must be a model of that system” (Conant & Ashby, 1970)

Abstract:

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

*In this paper a theorem is presented which shows, under very broad conditions, that any regulator that is maximally both successful and simple must be isomorphic with the system being regulated. (The exact assumptions are given.) Making a model is thus necessary.*

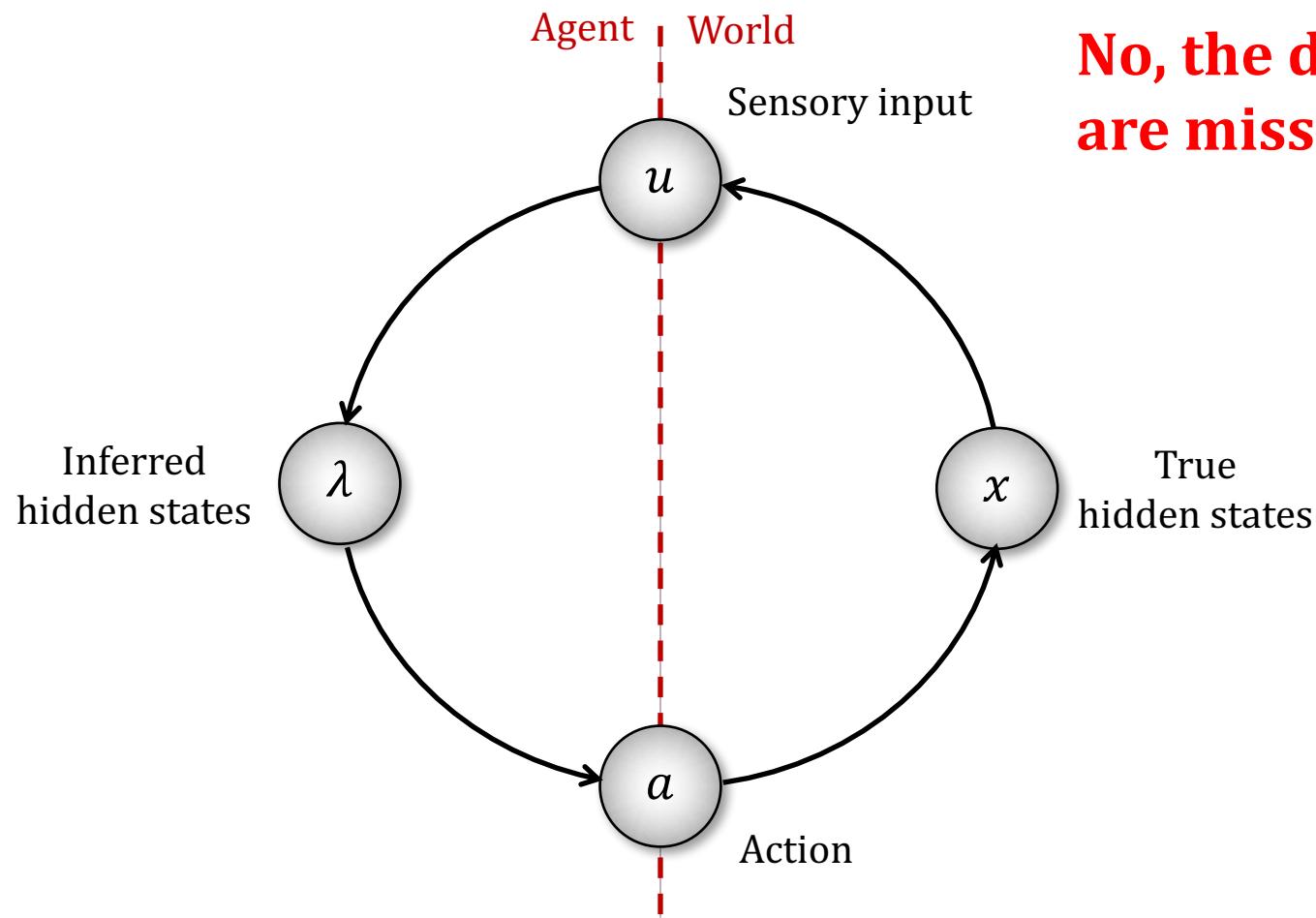
*The theorem has the interesting corollary that **the living brain**, so far as it is to be successful and efficient as a regulator for survival, **must proceed, in learning, by the formation of a model (or models) of its environment.**»*



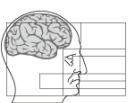
# Systems theory as the conceptual bridge between clinical phenomena and neuronal pathophysiology

- Belief updating by precision-weighted prediction errors provides a conceptual framework in which both clinical phenomena and neurobiological findings can be interpreted.
- For examples of this approach, see Adams et al. (2013) (psychosis), or Lawson et al. (2014), Quattrocki & Friston (2014) (autism).
- Summary: the mind needs to be a model of its environment  $\Rightarrow$  needs to perform Bayesian inference  $\Rightarrow$  needs to use precision-weighting of prediction errors  $\Rightarrow$  if that's all the mind does, it's also all that can go wrong  $\Rightarrow$  both clinical manifestations and the neurobiology of psychiatric disorders must be interpretable in these terms.
- Now that we have this conceptual framework, we can start filling it with content.

# 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 equations 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  $\alpha$ :

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

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



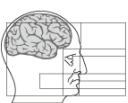
# Does a constant learning rate solve our problems?

- 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  $\alpha$  reflects the right level of uncertainty at any one time, and
  - (b) account for changes in uncertainty if  $\alpha$  is constant?
- What we really need is an adaptive learning that accurately reflects uncertainty.



# Needed: 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 ( $\pi_\varepsilon$  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).

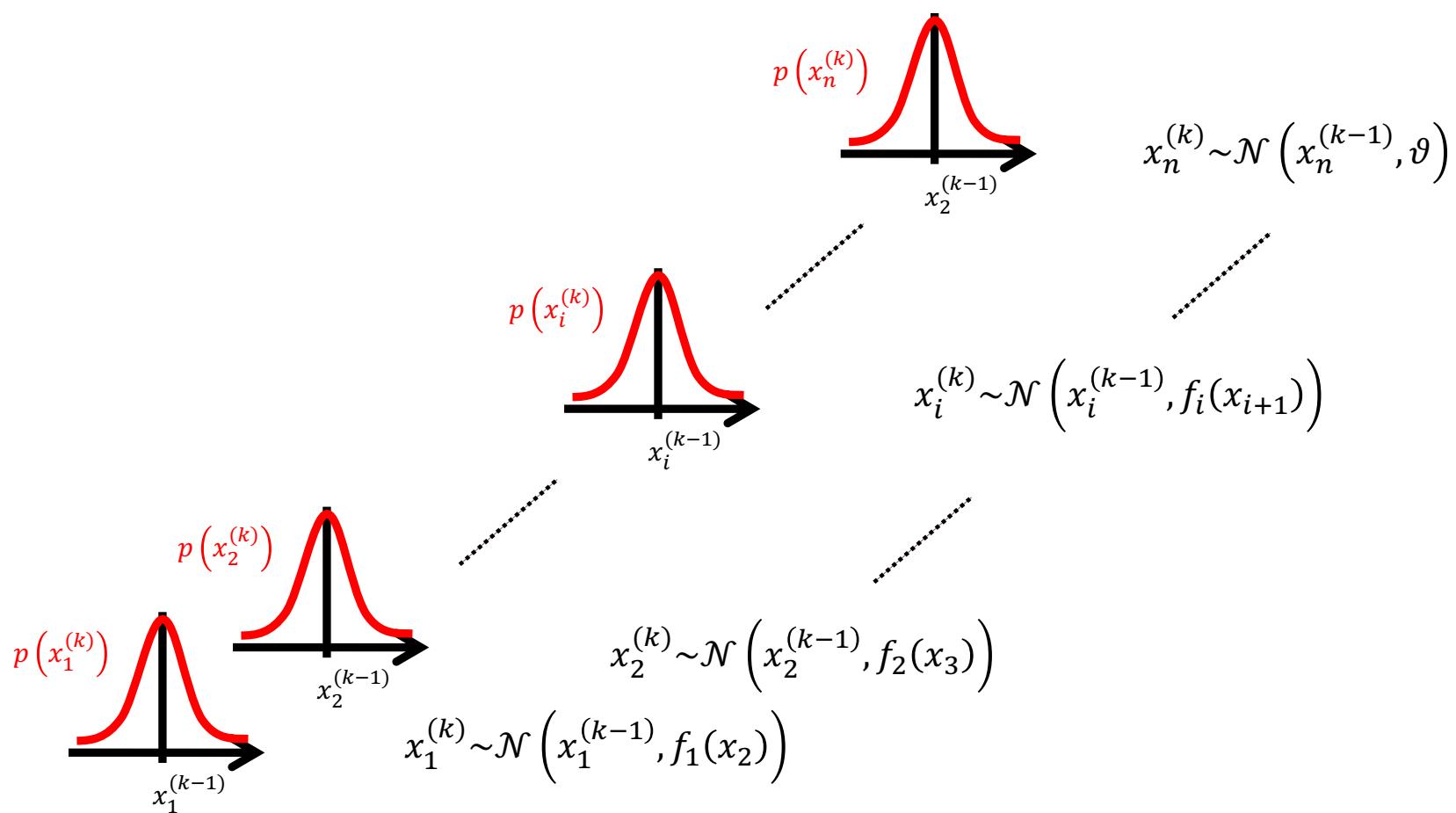


# An adaptive learning rate that accurately reflects uncertainty

- 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:

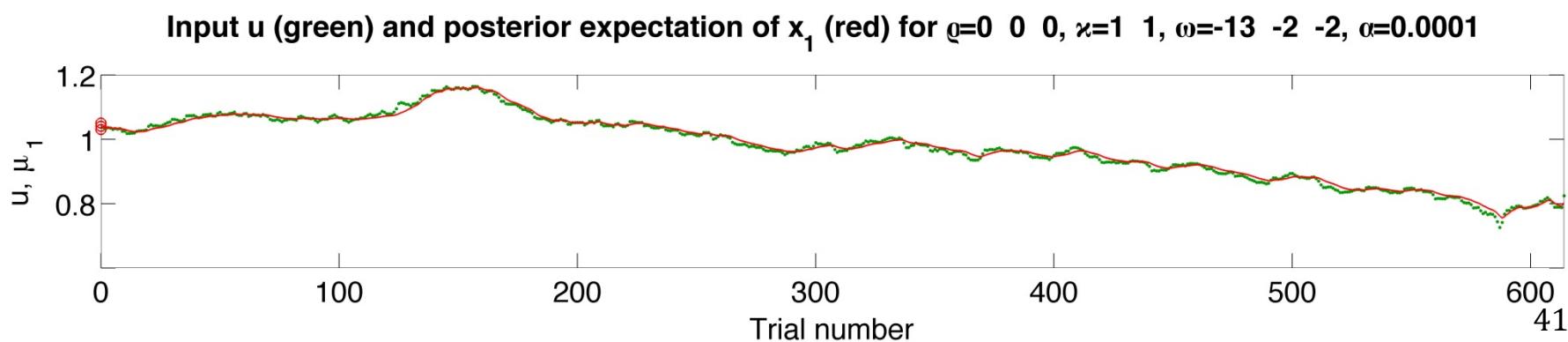
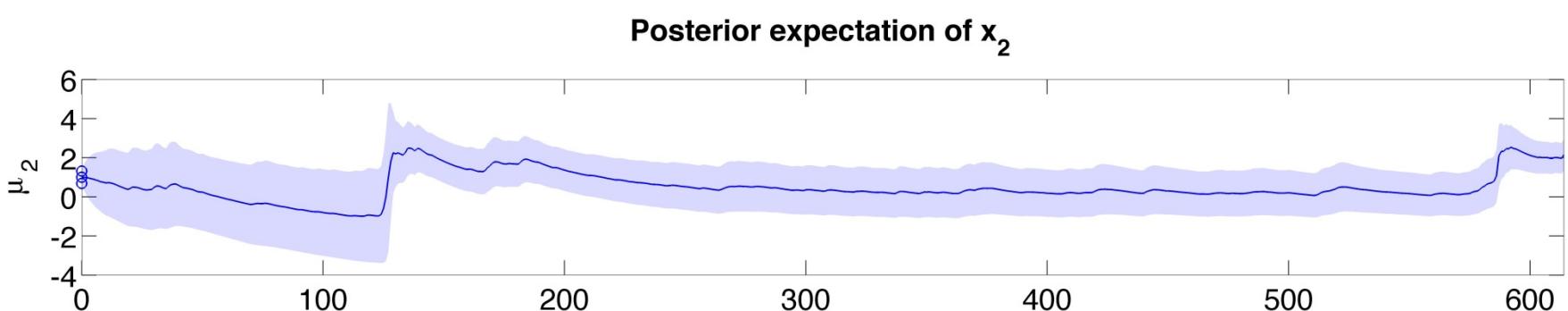
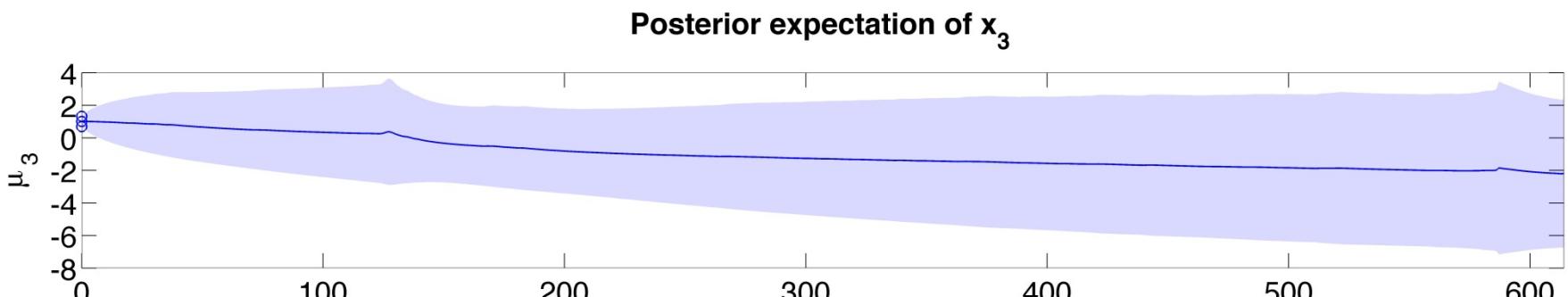
$$\frac{\hat{\pi}_u}{\pi_1^{(k)}} = \frac{\hat{\pi}_u}{\hat{\pi}_1^{(k)} + \hat{\pi}_u} = \frac{\hat{\pi}_u}{\frac{1}{\sigma_1^{(k-1)} + \exp(\kappa_1 \mu_2^{(k-1)} + \omega_1)} + \hat{\pi}_u}$$

outcome uncertainty

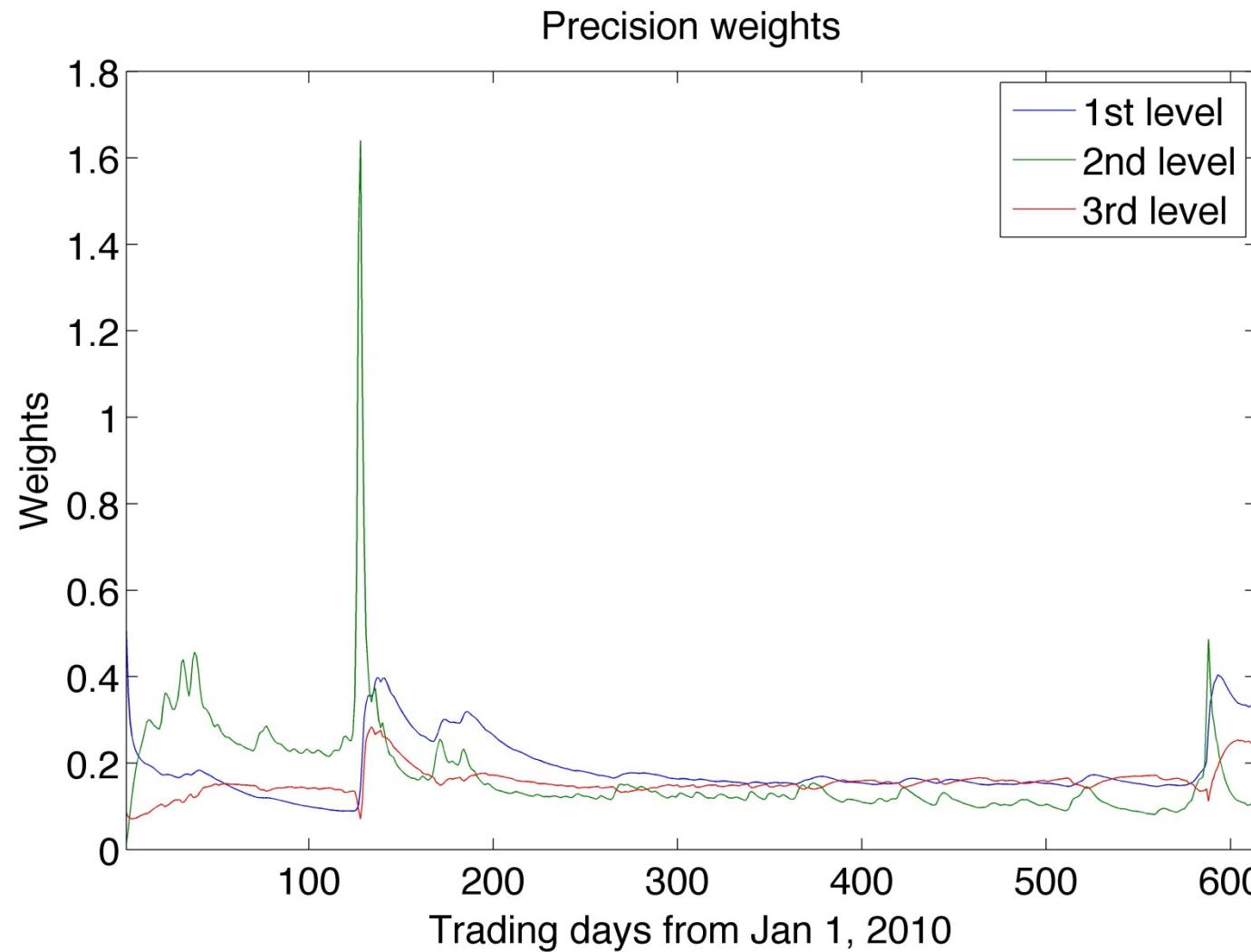
informational uncertainty

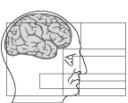
environmental uncertainty

# 3-level HGF for continuous observations



# 3-level HGF for continuous observations





# VAPEs and VOPEs

The updates of the belief on  $x_1$  are driven by value prediction errors (VAPEs)

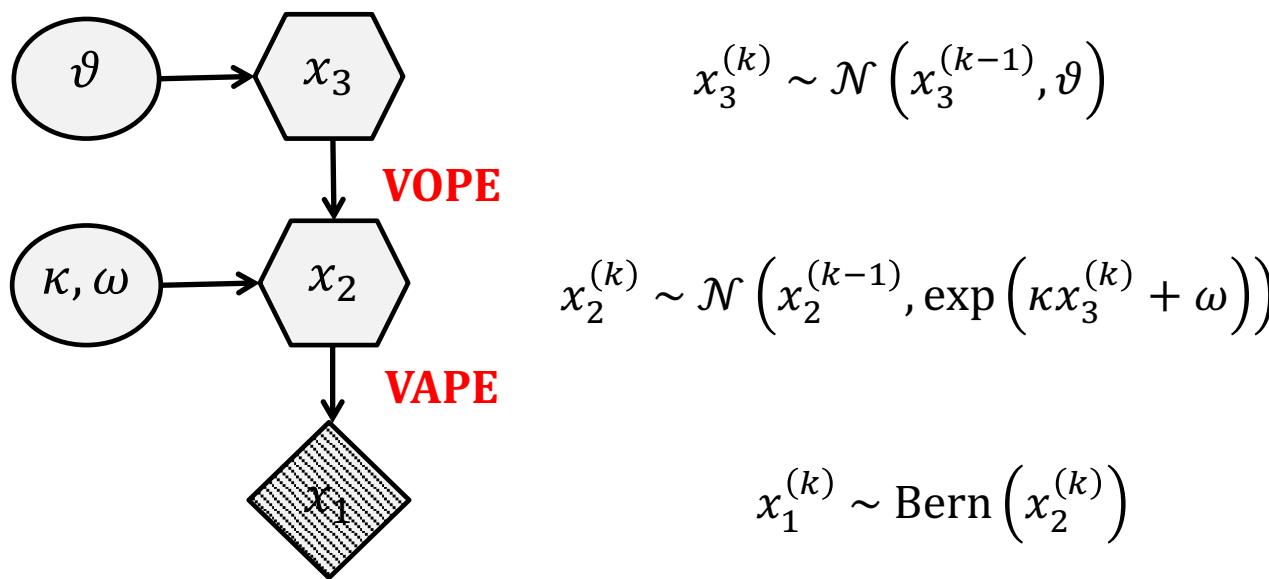
$$\mu_1^{(k)} = \mu_1^{(k-1)} + \frac{\hat{\pi}_u}{\pi_1^{(k)}} (u^{(k)} - \mu_1^{(k-1)}), \text{ VAPE}$$

while the  $x_2$ -updates are driven by volatility prediction errors (VOPEs)

$$\mu_2^{(k)} = \mu_2^{(k-1)} + \frac{1}{2} \kappa_1 v_1^{(k)} \frac{\hat{\pi}_1^{(k)}}{\pi_2^{(k)}} \delta_1^{(k)}, \text{ VOPE}$$

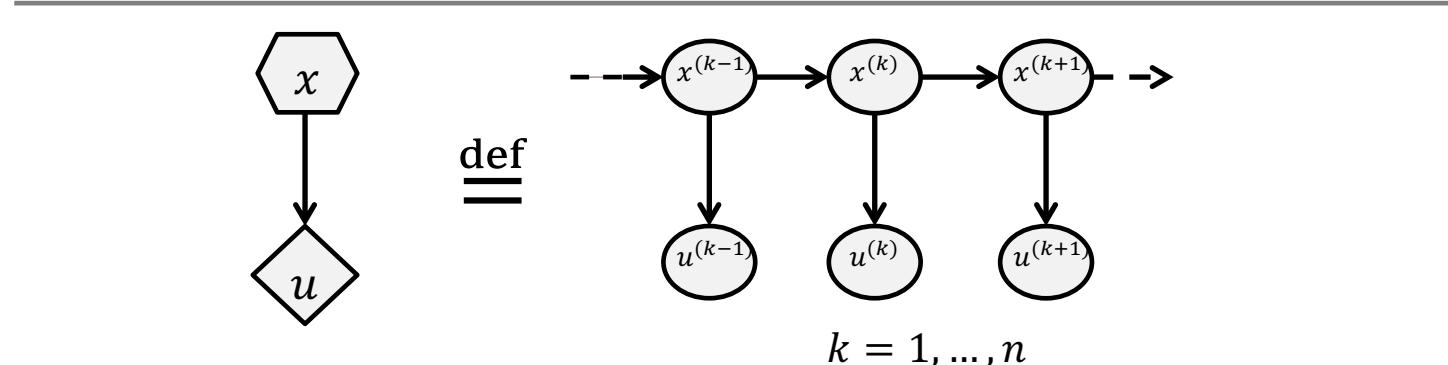
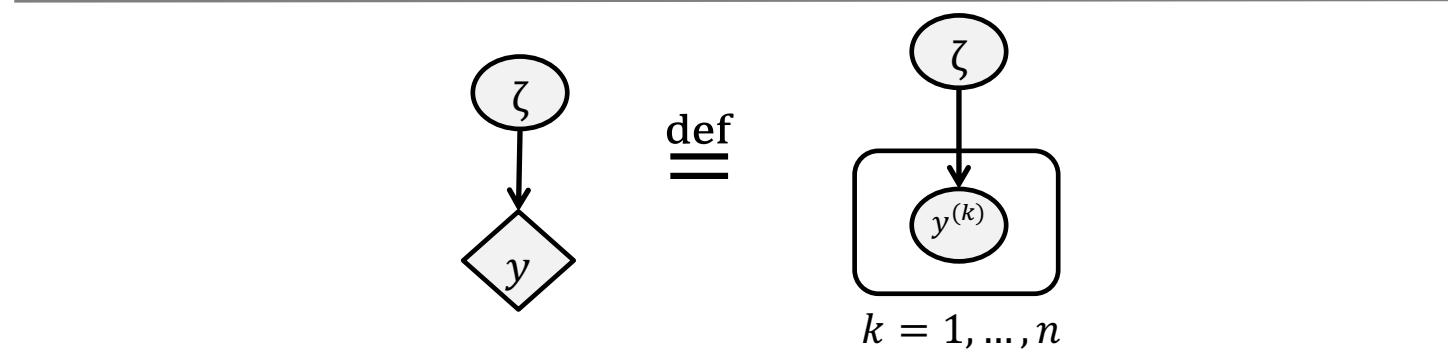
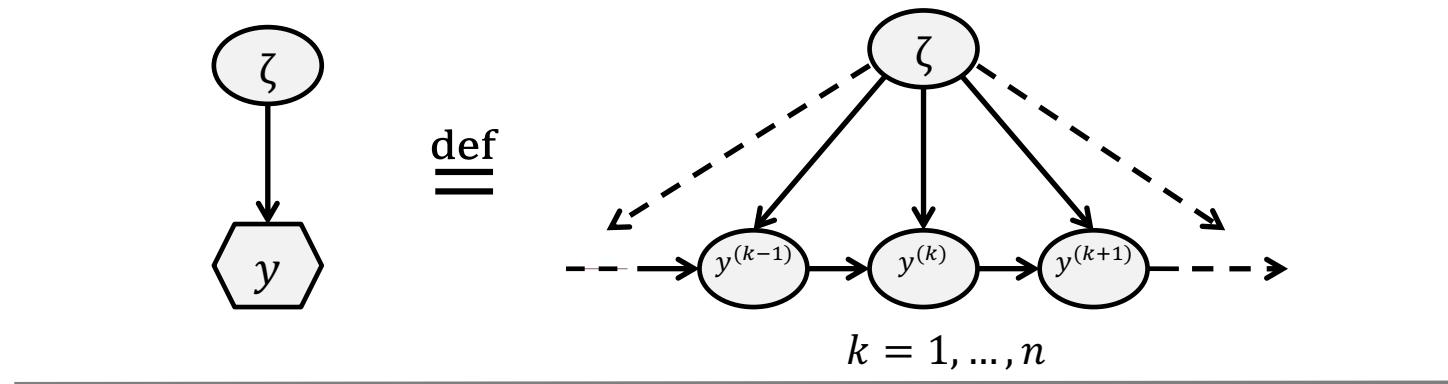
$$\delta_1^{(k)} \stackrel{\text{def}}{=} \frac{\sigma_1^{(k)} + (\mu_1^{(k)} - \mu_1^{(k-1)})^2}{\sigma_1^{(k-1)} + \exp(\kappa_1 \mu_2^{(k-1)} + \omega_1)} - 1$$

# 3-level HGF for binary observations

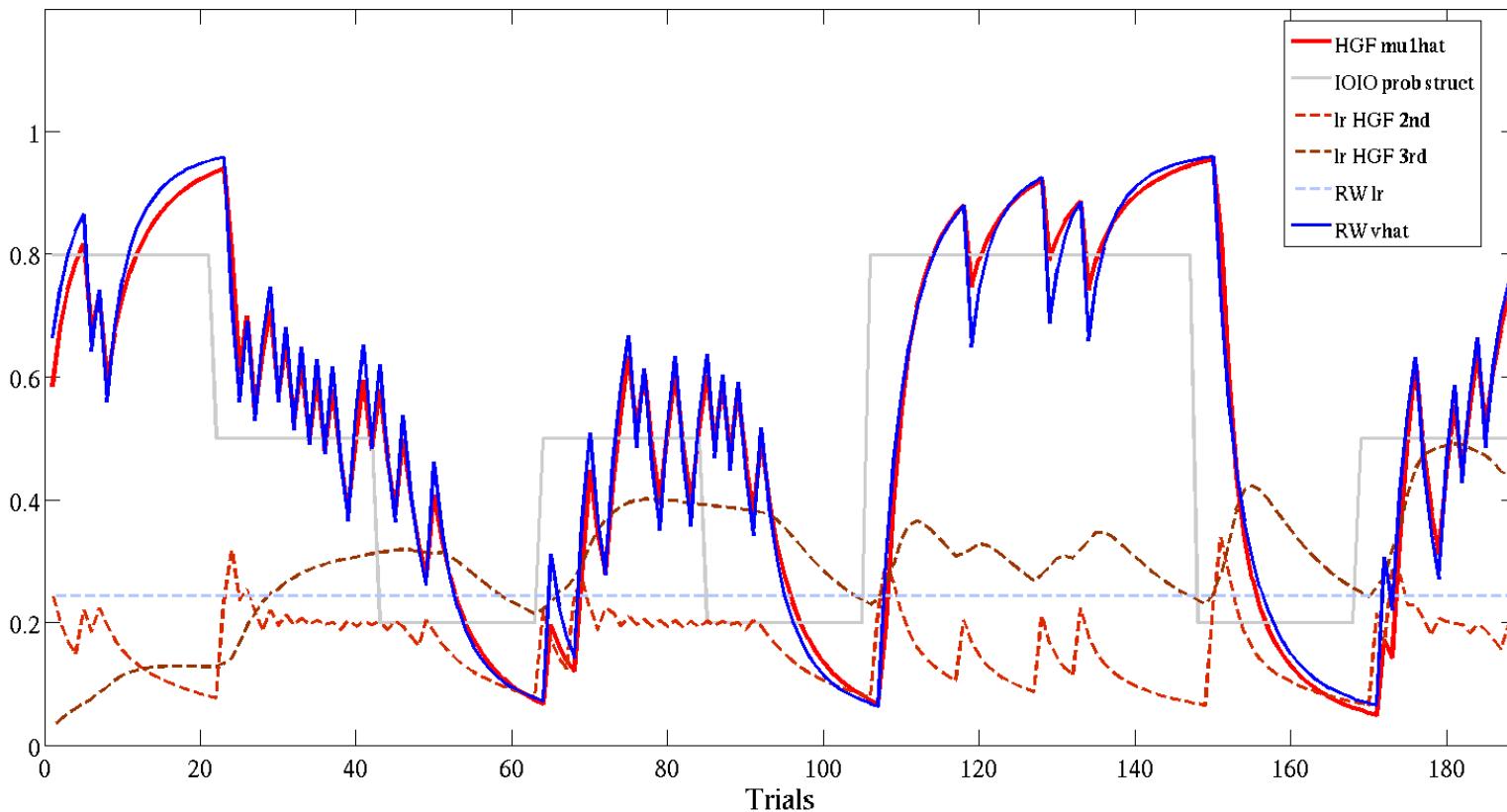


Mathys et al., 2011; Iglesias et al., 2013; Vossel et al., 2014a; Hauser et al., 2014; Diaconescu et al., 2014; Vossel et al., 2014b; ...

# Taking it all together: notation



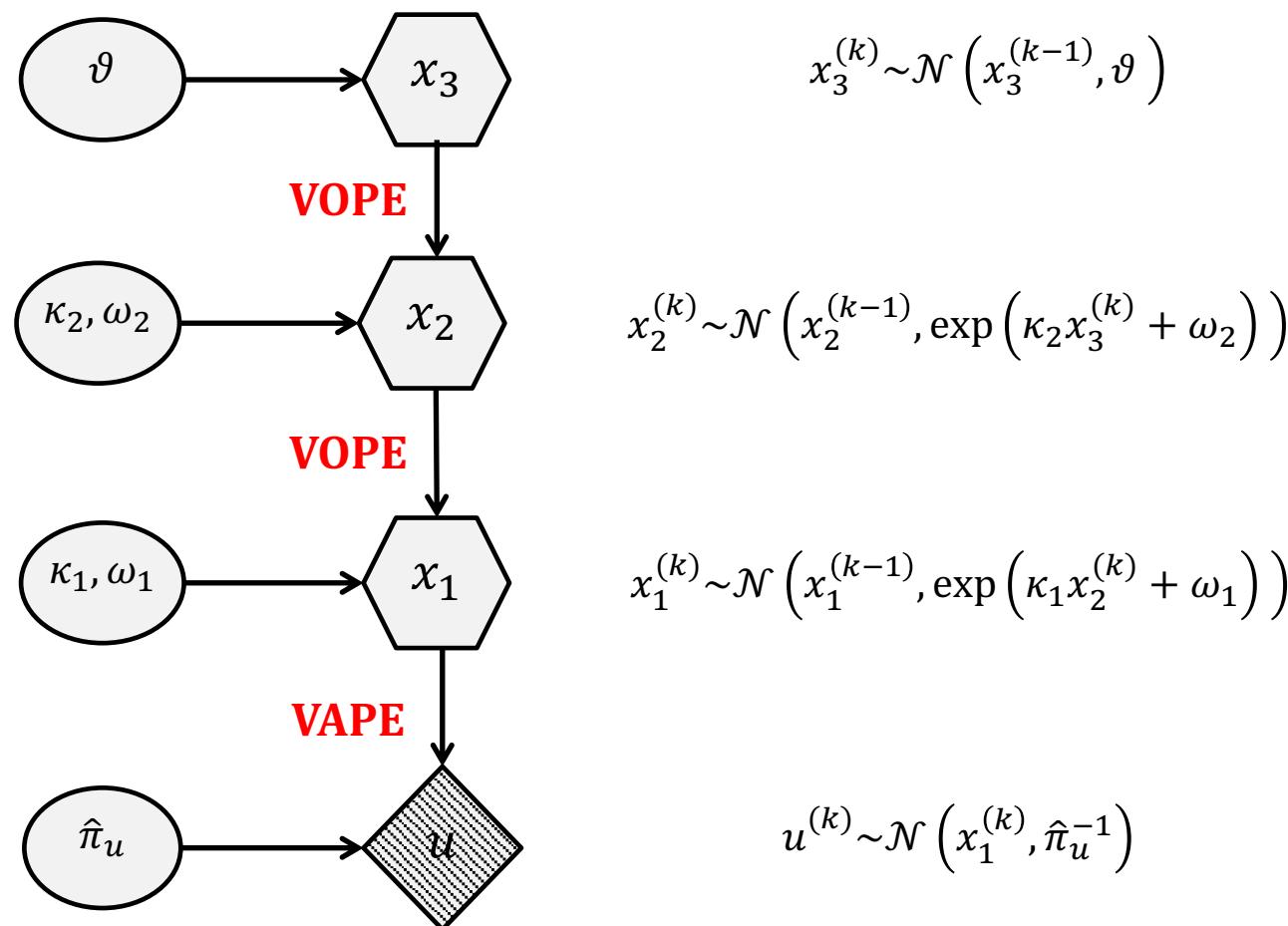
# The learning rate in the HGF

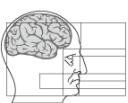


Andreea Diaconescu

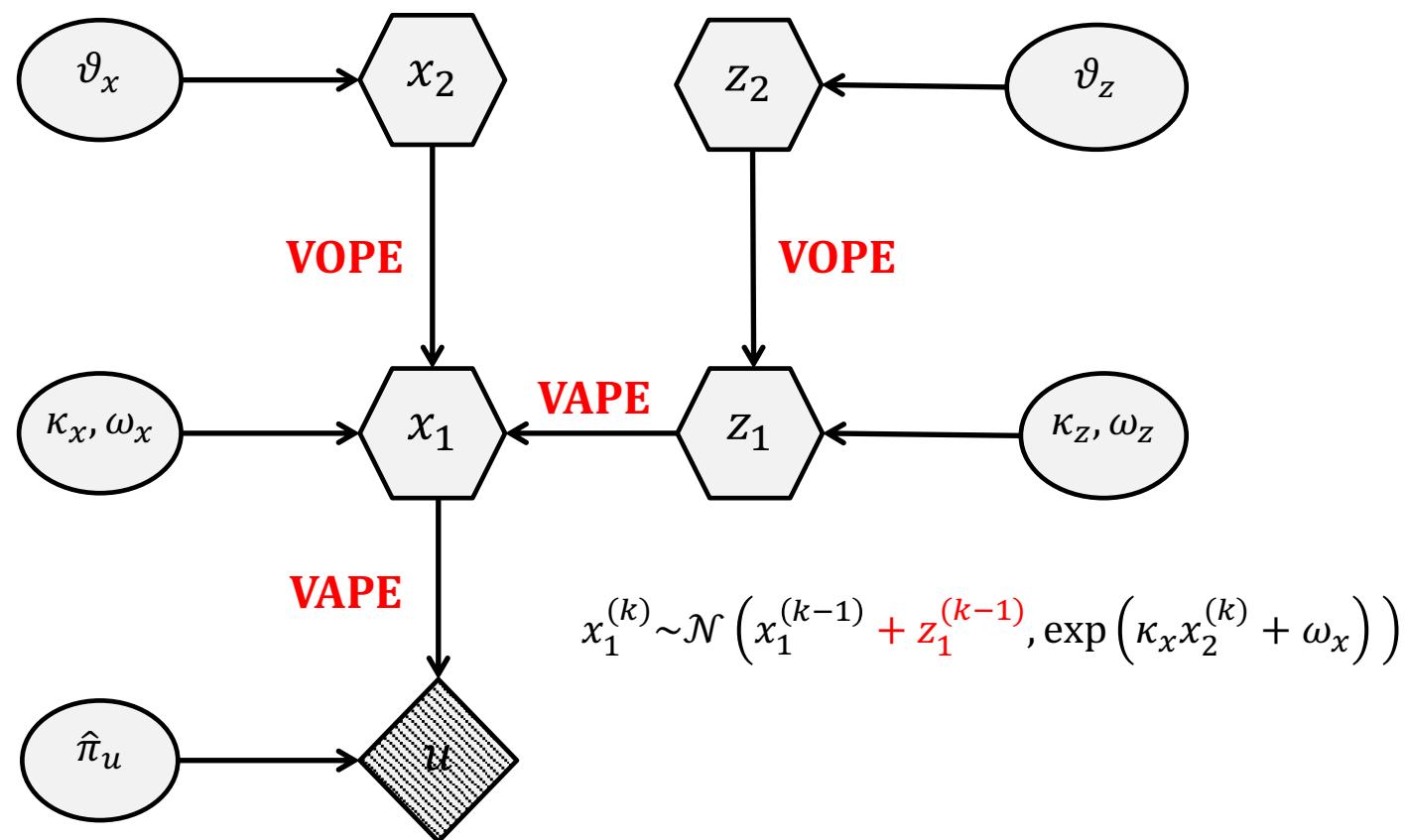


# 3-level HGF for continuous observations

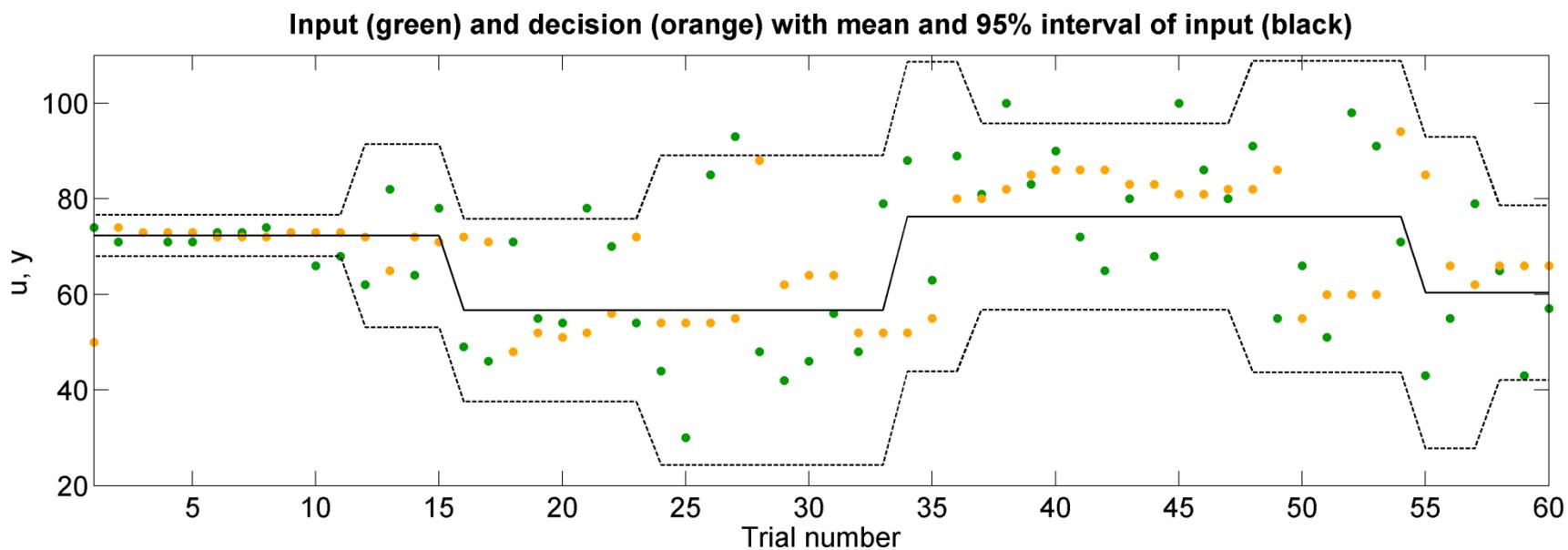




# Variable drift

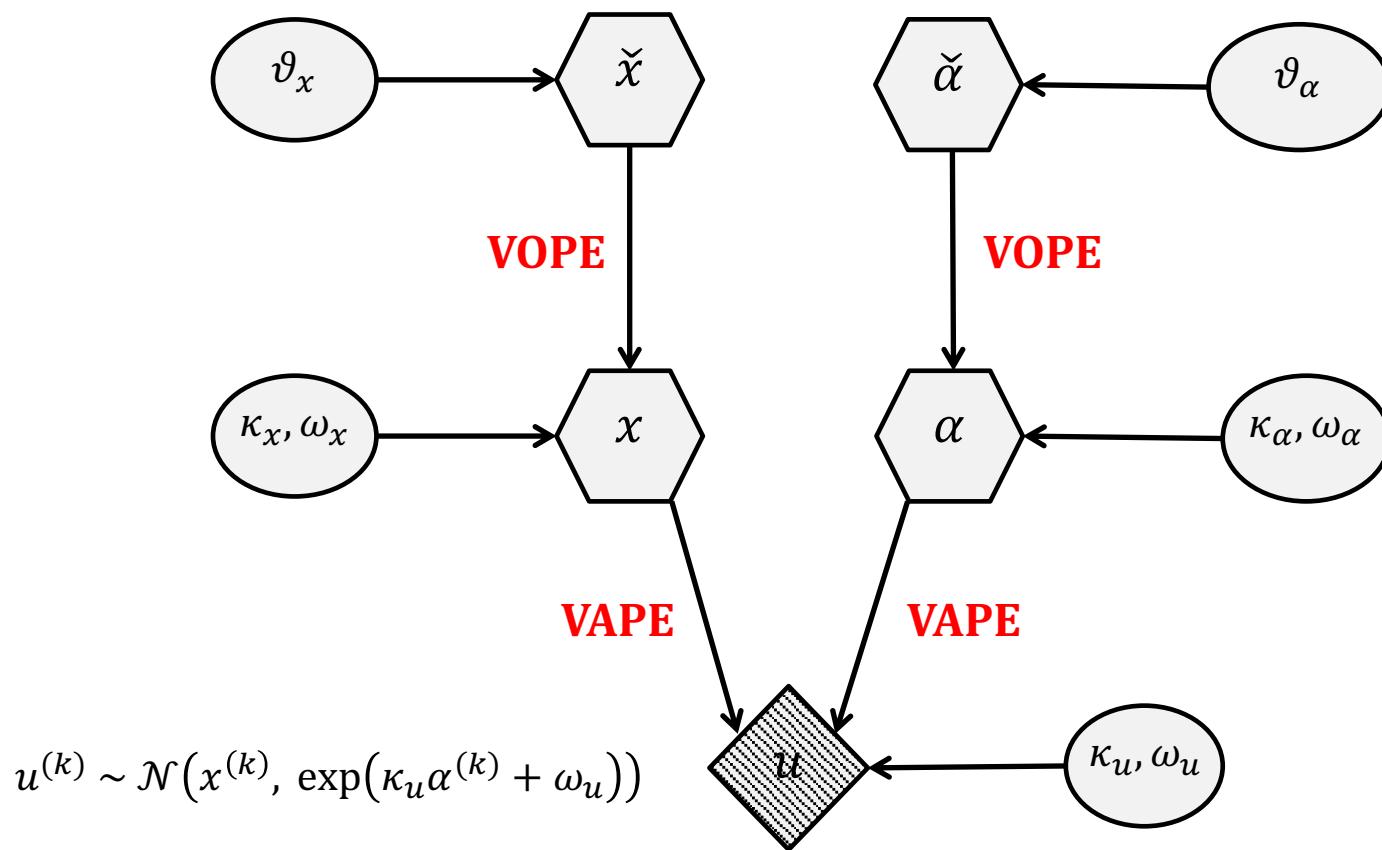


# Jumping Gaussian estimation task



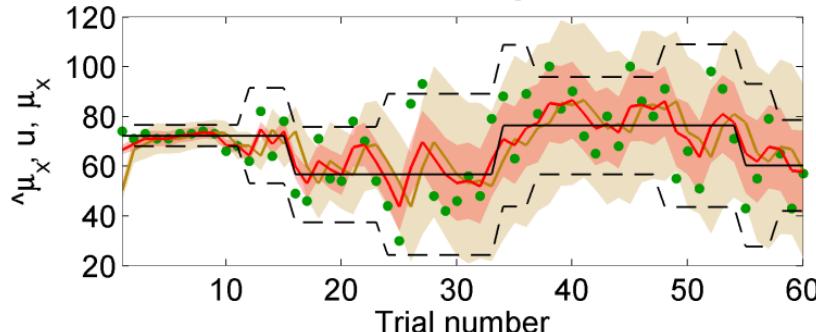
Chaohui Guo

# Independent mean and variance model

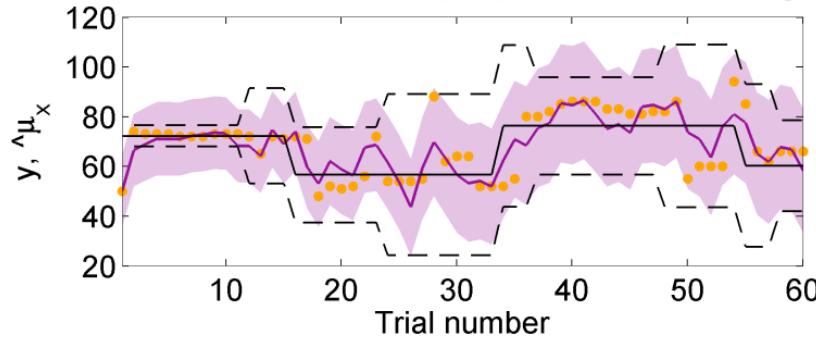


# Jumping Gaussian estimation task

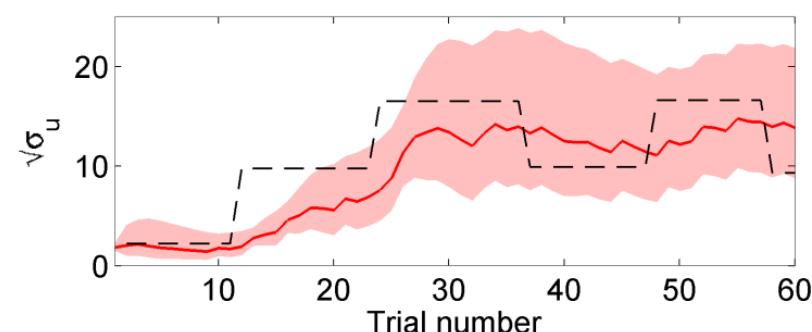
Prediction of input (brown), input (green), posterior belief (red)



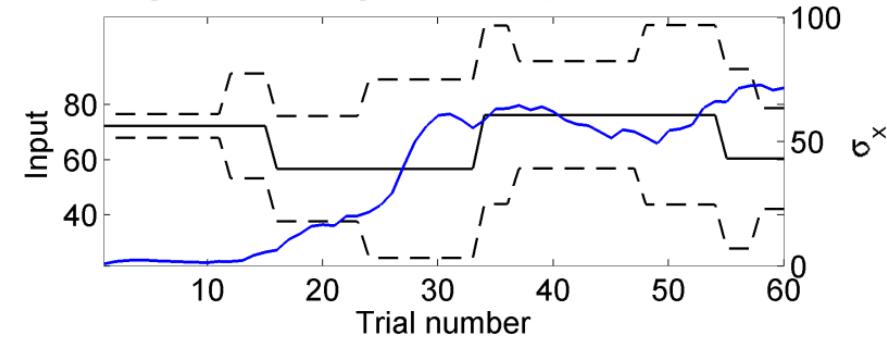
Prediction of decision (purple), decision (orange)



Belief on noise (red), true noise (dashed black)



Learning rate (blue; right scale), input (black; left scale)

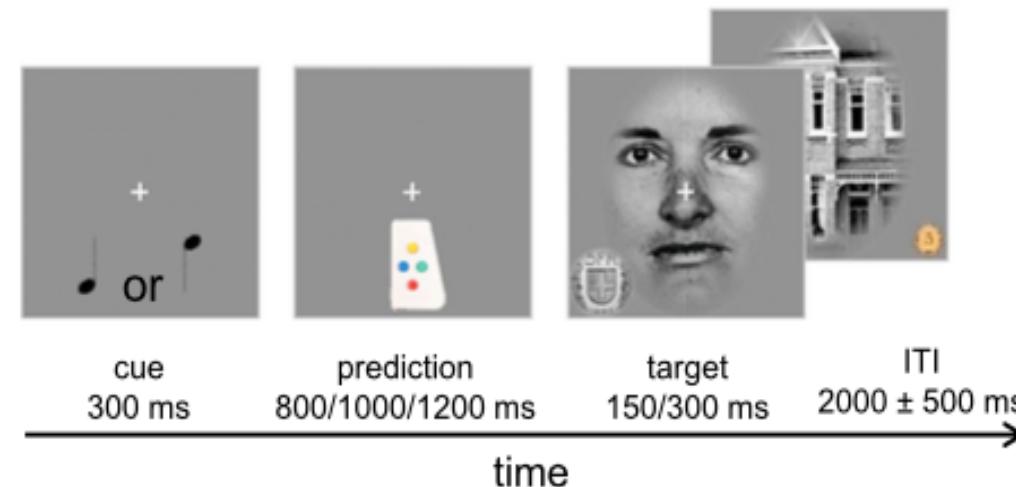




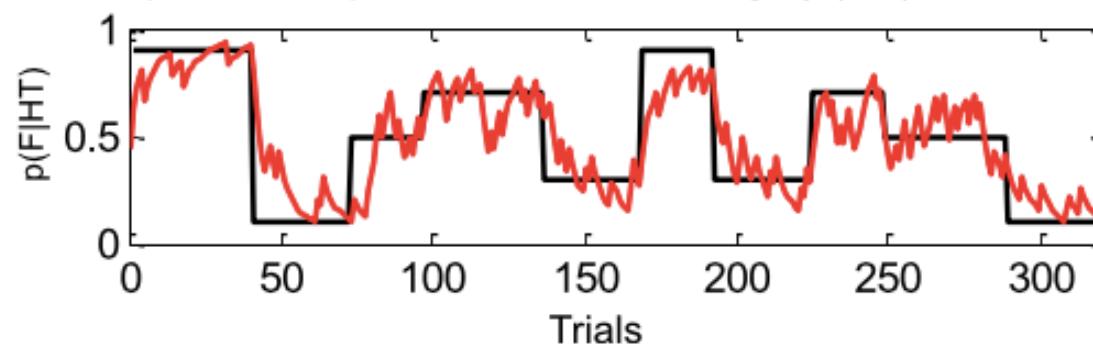
# How to reveal the precision-weighting of prediction errors when simple exponential-family likelihoods will not do

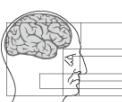
- Formulate the problem hierarchically (i.e., imitate evolution: when it built a brain that supports a mind which is a model of its environment, it came up with a (largely) hierarchical solution)
- Separate levels using a mean-field approximation
- Derive update equations

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



Changes in cue strength (black), and posterior expectation of visual category (red)





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

Model comparison:

BMS results	Behavioral study		fMRI study 1		fMRI study 2	
	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

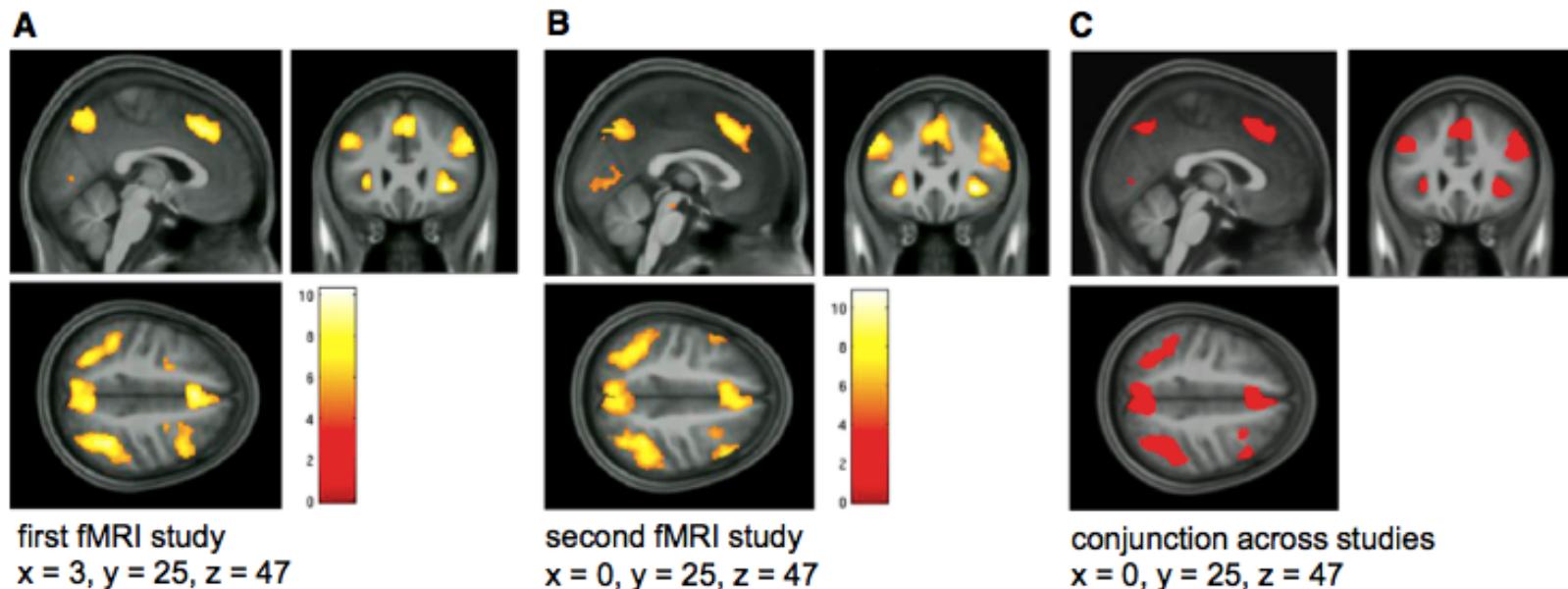


# HGF: empirical evidence (Iglesias et al., 2013)

Model comparison:

BMS results	Behavioral study		fMRI study 1		fMRI study 2	
	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

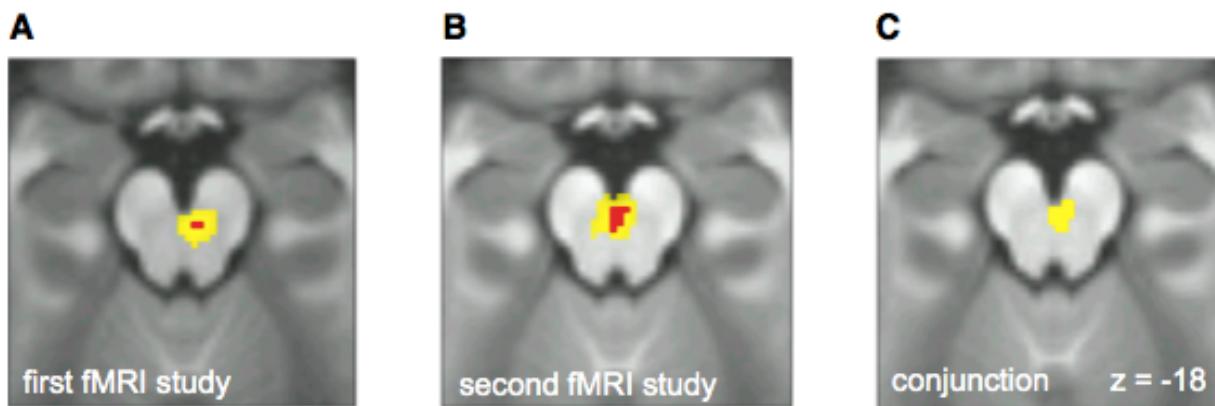
# HGF: empirical evidence (Iglesias et al., 2013)



**Figure 2. Whole-Brain Activations by  $e_2$**

Activations by precision-weighted prediction error about visual stimulus outcome,  $e_2$ , 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  $\varepsilon_2$**

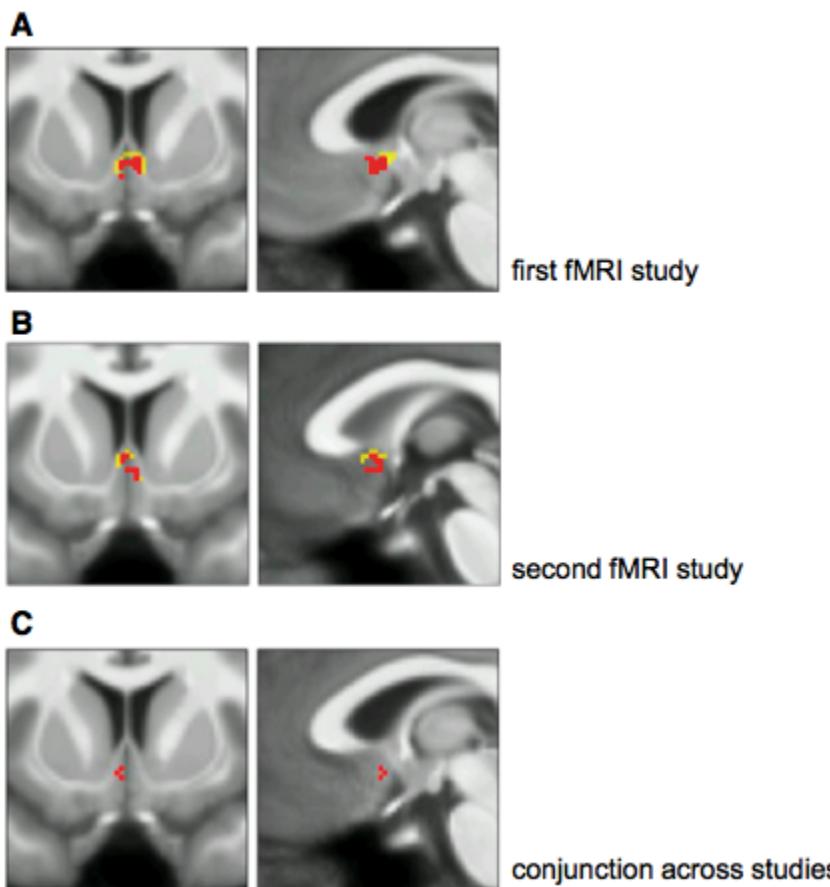
Activation of the dopaminergic VTA/SN associated with precision-weighted prediction error about stimulus category,  $\varepsilon_2$ . This activation is shown both at  $p < 0.05$  FWE whole-brain corrected (red) and  $p < 0.05$  FWE corrected for the volume of our anatomical mask comprising both dopaminergic and cholinergic nuclei (yellow).

(A) Results from the first fMRI study.

(B) Second fMRI study.

(C) Conjunction (logical AND) across both studies.

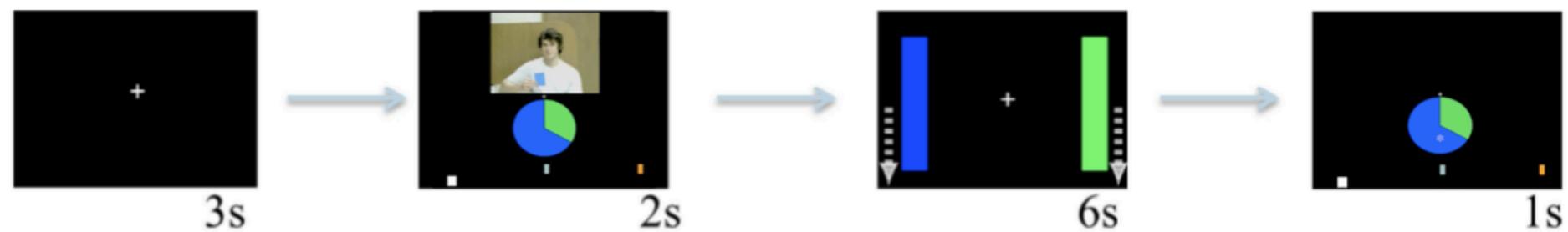
# HGF: empirical evidence (Iglesias et al., 2013)



**Figure 6. Basal Forebrain Activations by  $\varepsilon_3$**

Activation of the cholinergic basal forebrain associated with precision-weighted prediction error about stimulus probabilities  $\varepsilon_3$  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 ( $x = 2, y = 11, z = -8$ ).

# HGF: empirical evidence (Diaconescu et al., in preparation)

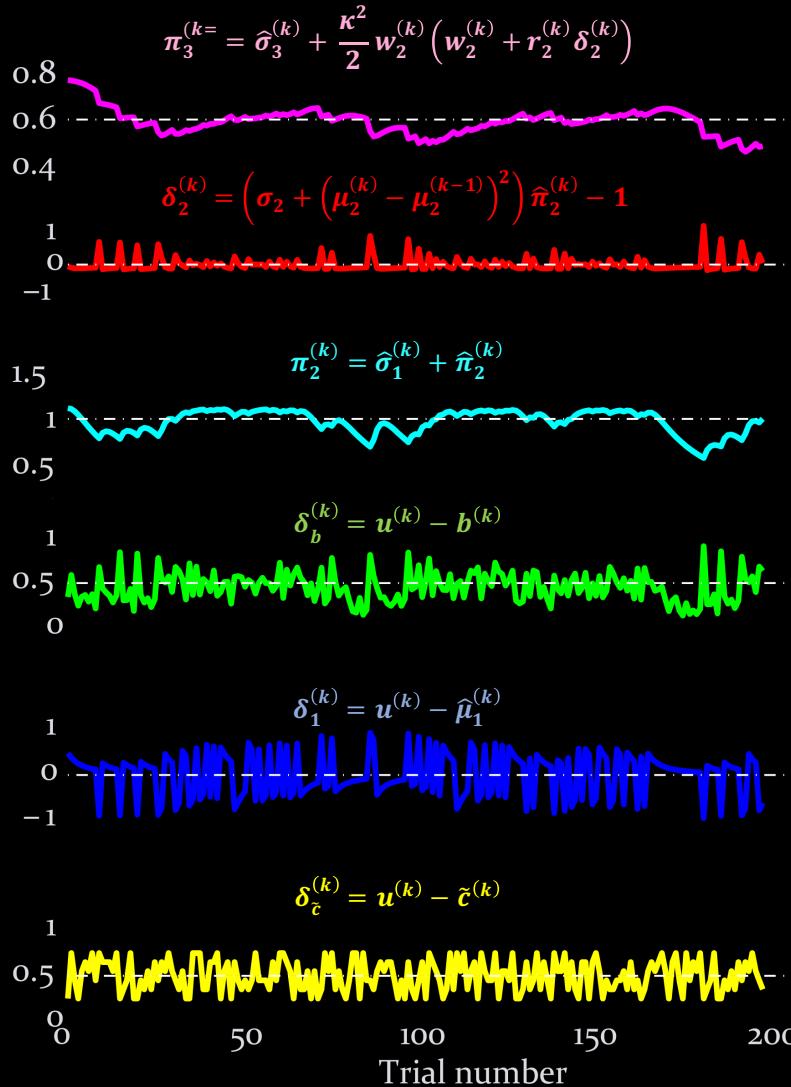


**Fig. 1.** Experimental Paradigm: 100 male volunteers played a binary lottery task and received advice about which option to choose from a more informed agent who was also incentivized to influence the participants' choices. To decide whether to take his advice into account, participants also inferred on the other's intentions and how they changed in time.

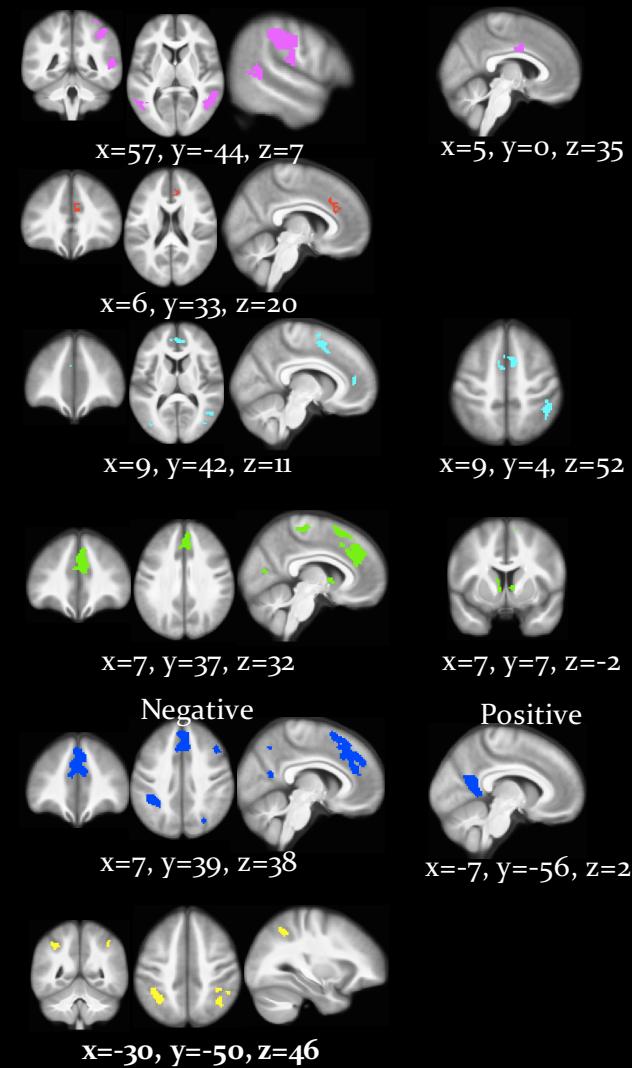


# HGF: empirical evidence (Diaconescu et al., in preparation)

## 6. Volatility Precision



## 4. Advice Precision



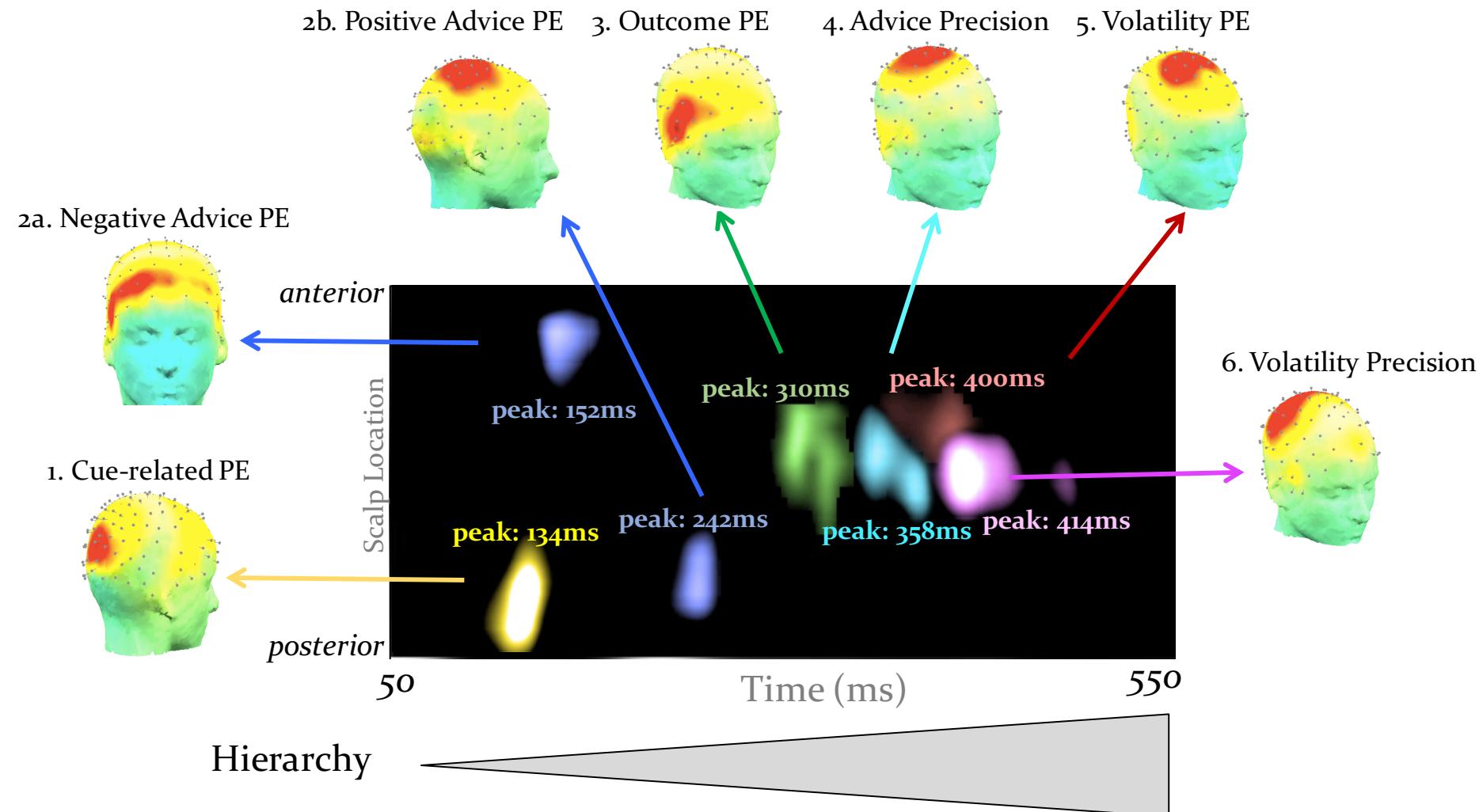
## 3. Outcome PE

## 2. Advice PE

## 1. Cue-Related PE

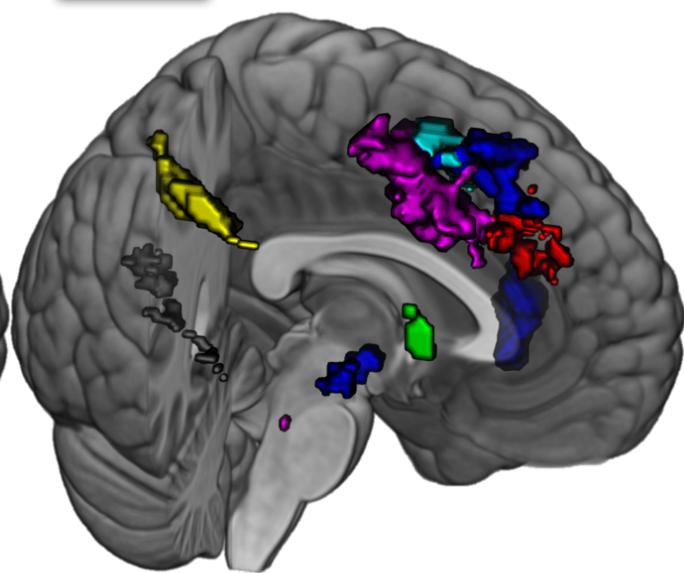
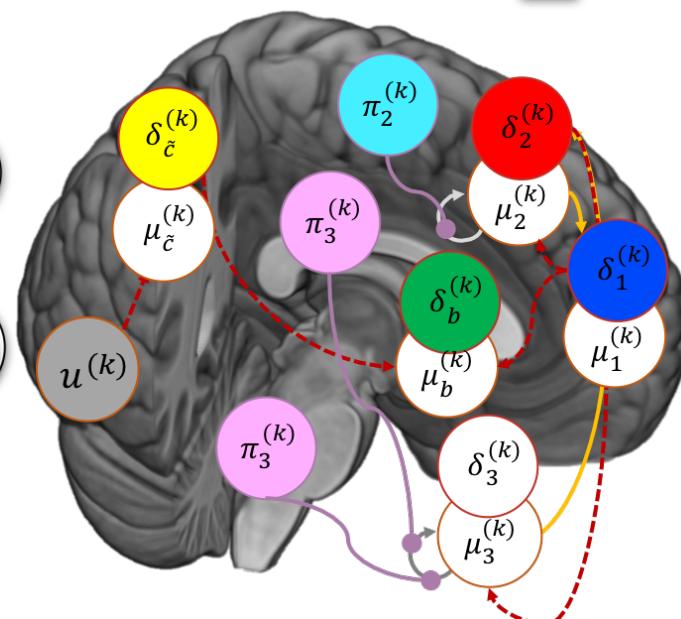
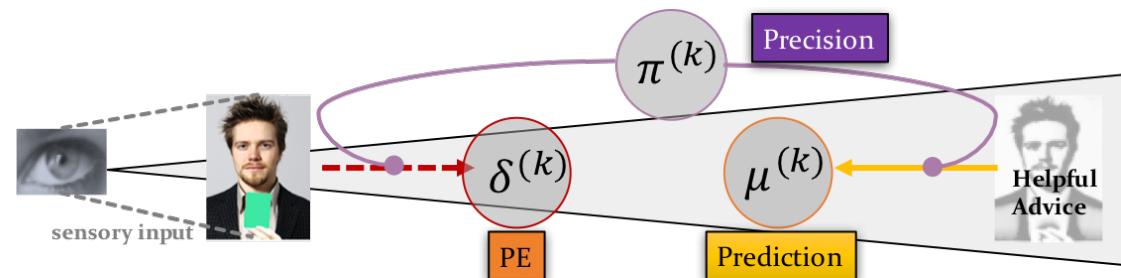
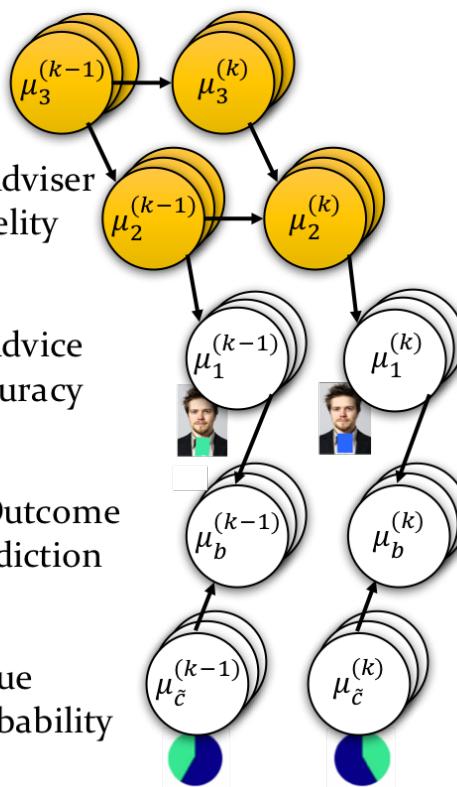
## Hierarchy

# HGF: empirical evidence (Diaconescu et al., in preparation)



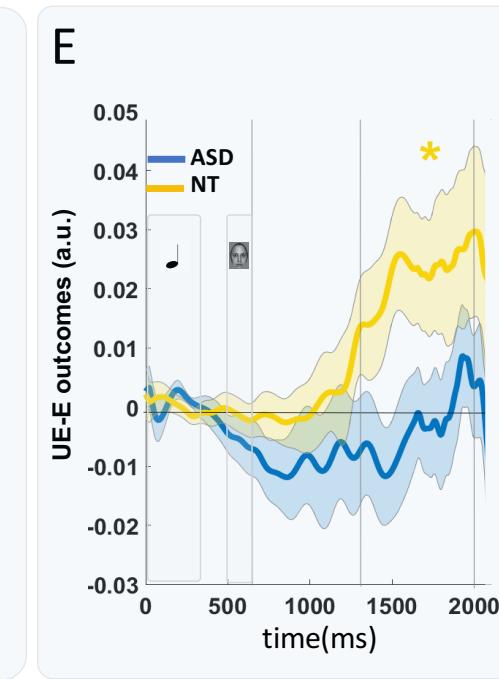
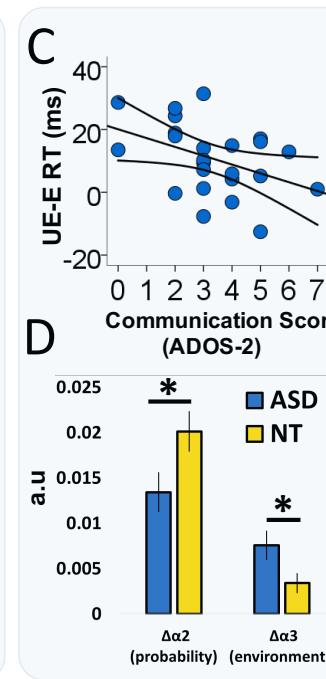
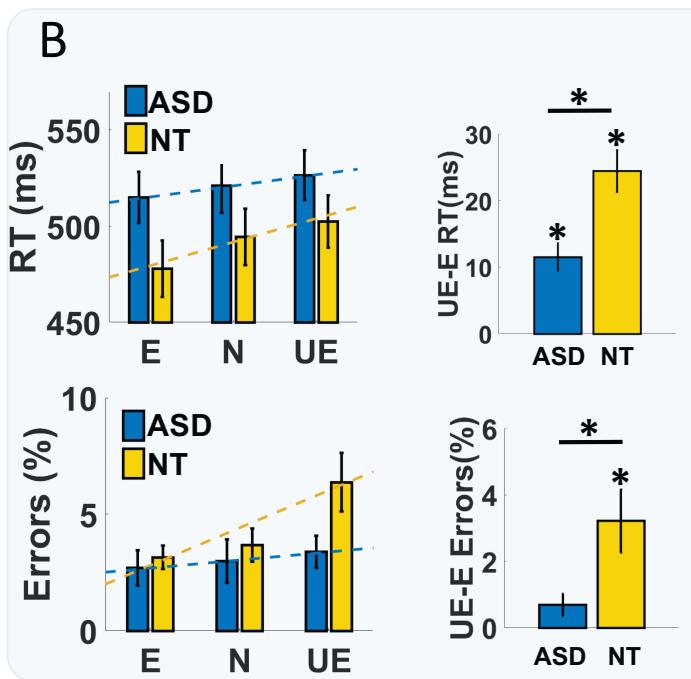
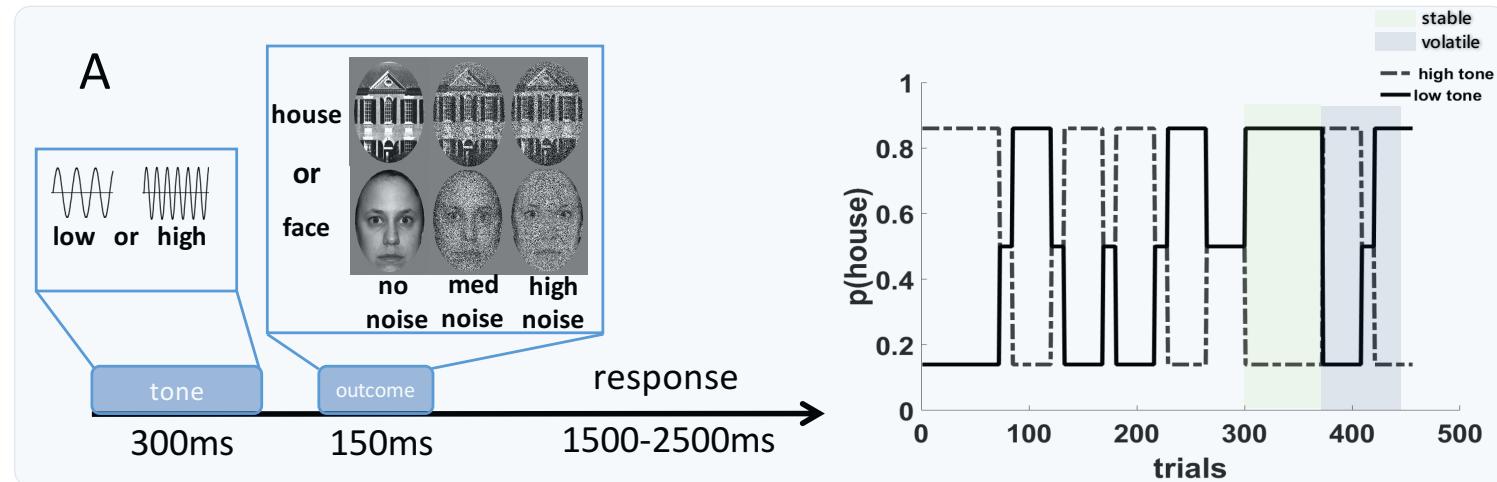
# HGF: empirical evidence (Diaconescu et al., in preparation)

5. Intention Volatility



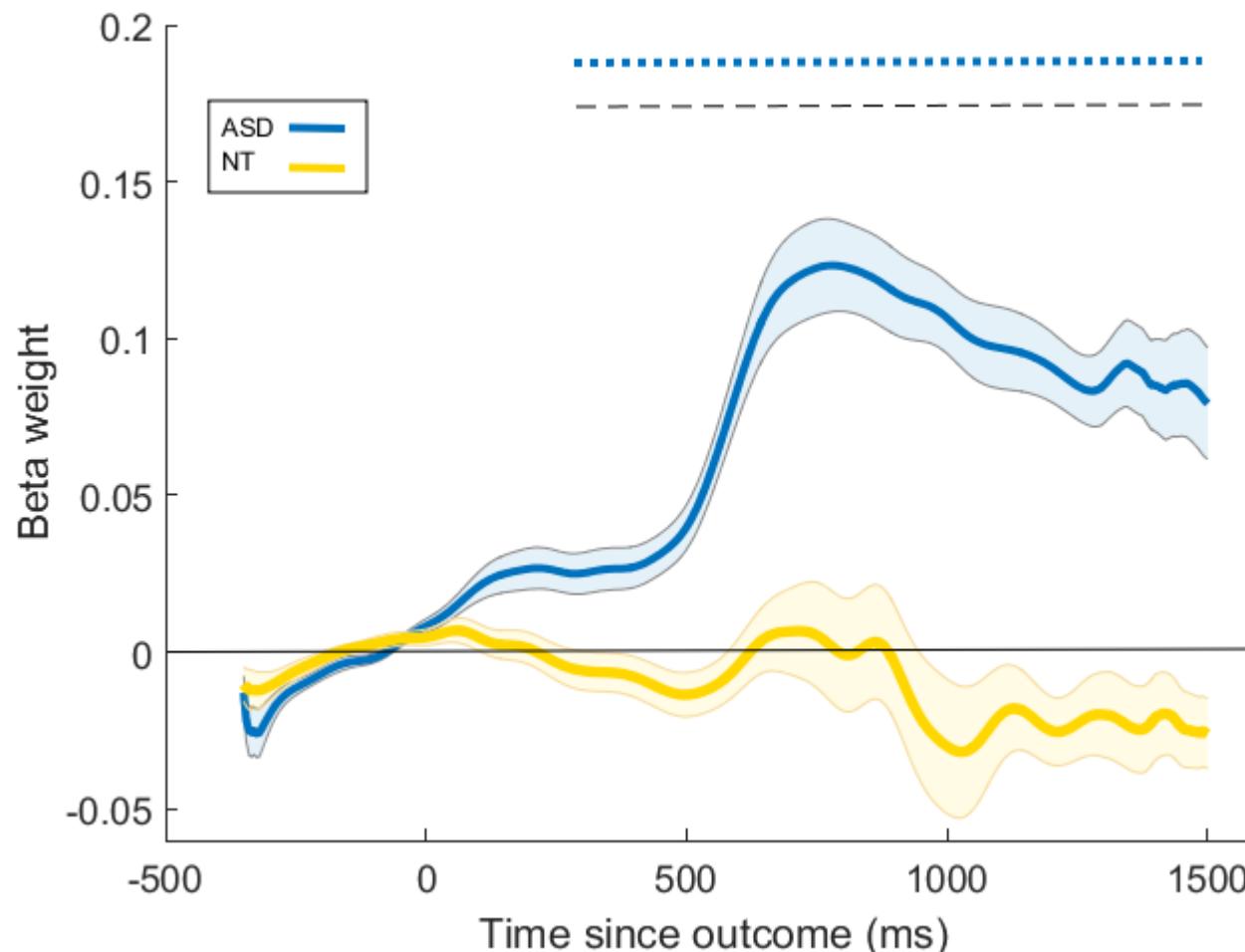


# HGF: empirical evidence (Lawson et al., in preparation)



# HGF: empirical evidence (Lawson et al., in preparation)

Effect of precision-weighted volatility prediction error  $\varepsilon_3$  on pupil diameter:





# 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



# 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

- Rick Adams
- Kay Brodersen
- Jean Daunizeau
- Andreea Diaconescu
- Chaohui Guo
- Karl Friston
- Sandra Iglesias
- Lars Kasper
- Ekaterina Lomakina
- Saeed Paliwal
- Klaas Enno Stephan
- Simone Vossel
- Lilian Weber

