# Applications of the HGF to attention and expectation and HGF message passing within and between brain regions

**Christoph Mathys**


Icahn School of Medicine at Mount Sinai, New York

31 July 2019

# A simple example of sequential observations in a noisy setting

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. How exactly would you proceed?

Right. You would make multiple measurements. Why?

Right. To reduce uncertainty due to observation noise. How does that work?

Right again. Let's worry about this later and first look at an example.

# Observations in a noisy setting

You're still on the boat and 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. Why?

Right.

- 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

# Updating the mean of a series of observations

By the way: how do you calculate the mean $\bar{u}$ of $u_1, u_2, \ldots, u_n$?

The usual way is to take

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

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

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

$$\bar{u}_{n+1} = \bar{u}_n + \frac{1}{n+1}(u_{n+1} - \bar{u}_n)$$

# Updating the mean of a series of observations

Proof of sequential update equation:

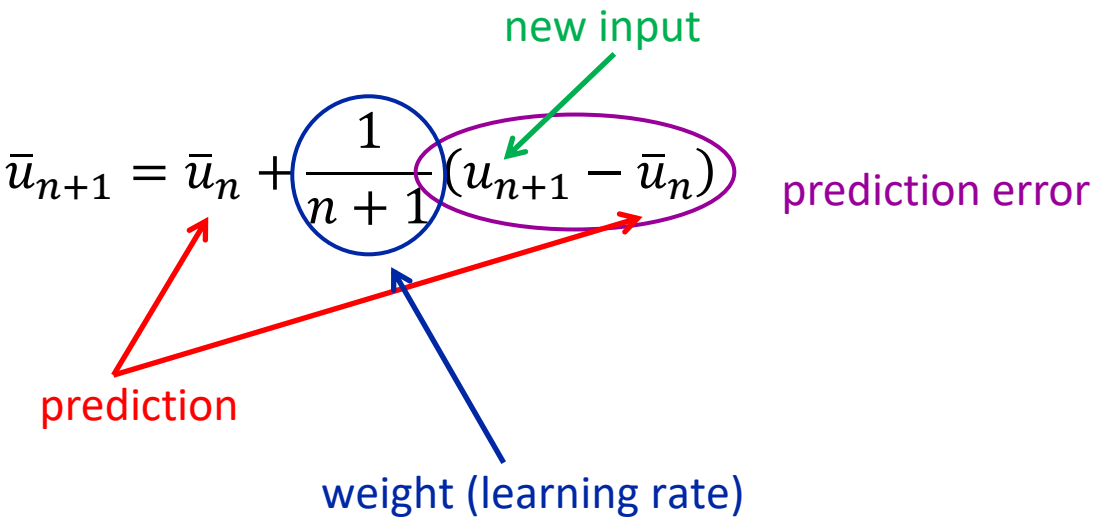
$$\bar{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}\left(x_{n+1} + n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right) =$$

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

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

q.e.d.

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




Is this a general pattern?

More specifically, does it generalize to Bayesian inference?

Indeed, it turns out that in many cases, Bayesian inference can be based on parameters that are updated using **precision-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 value of the parameter $\vartheta$ can be described by a Gaussian prior:

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

The likelihood of an observation $y$ is also Gaussian, with precision $\pi_\varepsilon$ :

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

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

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

# Updates in a simple Gaussian model

Here's how the updates to the sufficent statistics $\mu$ and $\pi$ describing our belief look like:

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

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

prediction error

prediction

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

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).

# Generalization to all exponential families of distributions

Many of the most widely used probability distributions are families of exponential distributions.

For example, the Gaussian distribution is an exponential family of distributions (and so are the beta, gamma, binomial, Bernoulli, multinomial, categorical, Dirichlet, Wishart, Gaussian-gamma, log-Gaussian, multivariate Gaussian, Poisson, and exponential distributions, among others). This means it can be written the following way:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta})) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

with

$$\boldsymbol{x} = x, \qquad \boldsymbol{\vartheta} = (\mu, \sigma)^{\mathrm{T}}, \qquad h(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}}, \qquad \boldsymbol{\eta}(\boldsymbol{\vartheta}) = \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma}\right)^{\mathrm{T}}, \qquad \boldsymbol{T}(\boldsymbol{x}) = (x, x^2)^{\mathrm{T}}, \qquad A(\boldsymbol{\vartheta}) = \frac{\mu^2}{\sigma} + \frac{\ln \sigma}{2}$$

This allows us to look at Bayesian belief updating in a very general way for all exponential families of distributions.

# Generalization to all exponential families of distributions (Mathys, 2016)

Our likelihood is an exponential family in its general form:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))$$

The vector $\boldsymbol{T}(\boldsymbol{x})$ (a function of the observation $\boldsymbol{x}$) is called the sufficient statistic.

For the prior, we may assume that we have made $\nu$ observations with sufficient statistic $\boldsymbol{\xi}$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\xi}, \nu) = z(\boldsymbol{\xi}, \nu) \exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big)$$ (where $z(\xi, \nu)$ is a normlization constant)

It then turns out that the posterior has the same form, but with an updated $\boldsymbol{\xi}$ and $\nu$ replaced with $\nu + 1$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{x}, \boldsymbol{\xi}, \nu) = z(\boldsymbol{\xi}', \nu + 1) \exp\big((\nu + 1)(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi}' - A(\boldsymbol{\vartheta}))\big)$$

$$\boldsymbol{\xi}' = \boldsymbol{\xi} + \frac{1}{\nu + 1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

# Proof of the update equation

$$\overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu)}^{\text{posterior}} \propto \overbrace{p(\boldsymbol{x}|\boldsymbol{\vartheta})}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu)}^{\text{prior}}$$

$$= h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))z(\boldsymbol{\xi},\nu) \exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big)$$

$$\propto \exp\big(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot (\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - (\nu+1)A(\boldsymbol{\vartheta})\big)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - A(\boldsymbol{\vartheta})\right)\right)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi} - (\nu+1)\boldsymbol{\xi})\right) - A(\boldsymbol{\vartheta})\right)\right)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \left(\underbrace{\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})}_{=:\boldsymbol{\xi}'}\right) - A(\boldsymbol{\vartheta})\right)\right)$$

$$\Longrightarrow \quad p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu) = z(\boldsymbol{\xi}',\nu') \exp\Big(\nu'\big(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi}' - A(\boldsymbol{\vartheta})\big)\Big)$$

$$\text{with } \nu' \coloneqq \nu + 1, \qquad \boldsymbol{\xi}' \coloneqq \boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

q.e.d.

# … so in sum:

… we have dealt (among others) with beliefs about states that are

- binary (Bernoulli)
- categorical
- bounded on both sides (beta)
- bounded on one side (gamma)
- and unbounded (Gaussian)

This includes most kinds of states we can have beliefs about. Notably,

- **All Bayesian (i.e., probabilistic, rational) updates of such beliefs take the form of precision-weighted prediction errors.**

- **These prediction errors and their precision weights are easy to compute.**

- **The prediction errors are simple functions of inputs.**

# Limitations

- Examples of distributions that are not exponential families: Student's *t*, Cauchy

- These distributions are popular because of their «fat tails». However, fat tails can also be achieved with appropriate hierarchies of Gaussians (cf. the hierarchical Gaussian filter, HGF)

- A further kind of distributions that are not exponential families are found in mixture models.

- Such models are popular because of they provide multimodal distributions. But again, appropriate hierarchies of distributions may save the day.

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

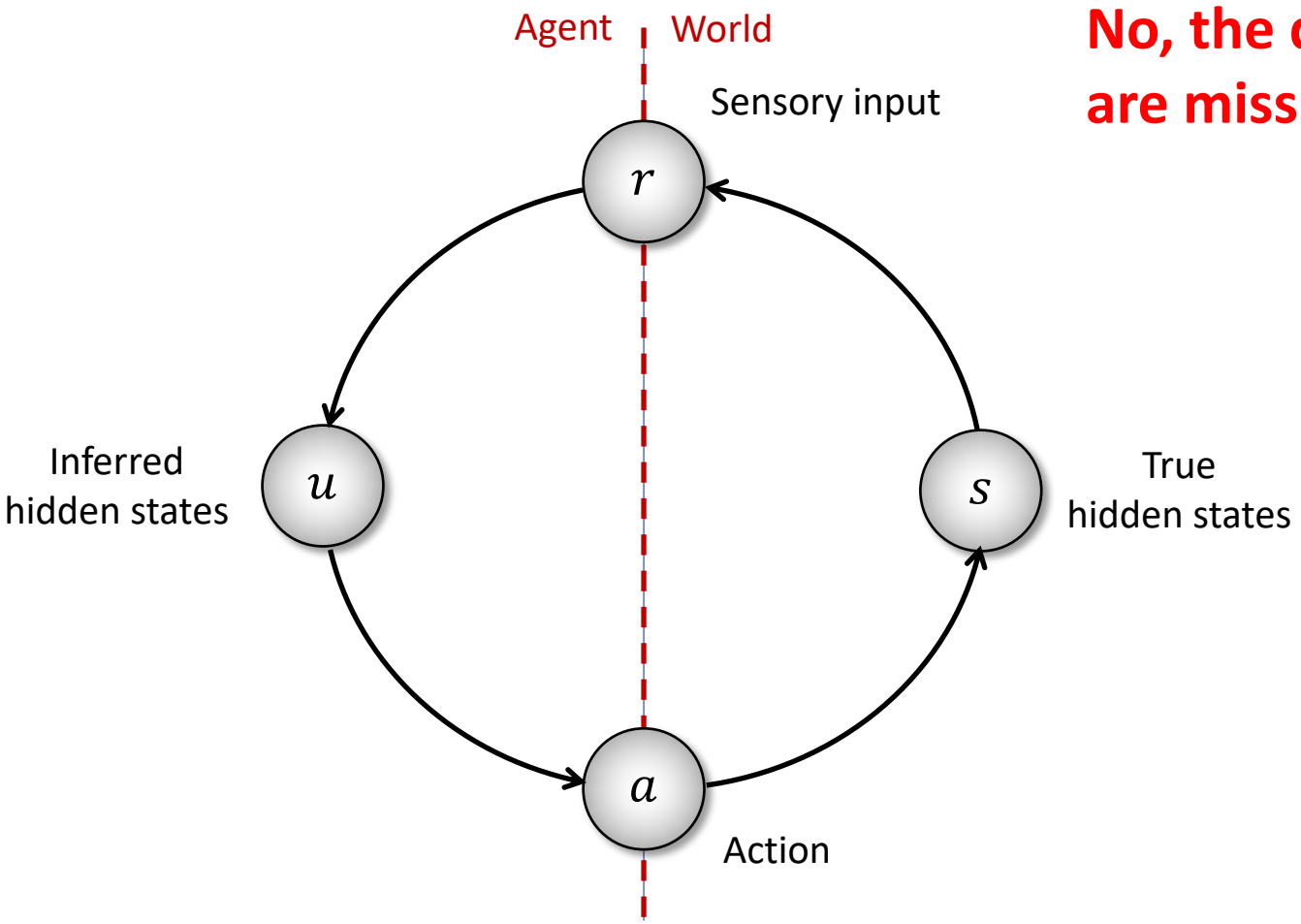
- Formulate the problem hierarchically (i.e., imitate evolution: when it built a brain that supports a mind which is a model of its environment, it came up with a (largely) hierarchical solution)

- Separate levels using a mean-field approximation

- Derive update equations

- Example: HGF

# But: 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...

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

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

$$v_n = v_{n-1} + \varepsilon(r_n - v_{n-1}).$$

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

# How are we treating observations in Rescorla-Wagner learning?

Rewriting the RW learning rule reveals that $q_n$ is an average weighted by $\alpha$ of the observation $x_n$ and the previous value $q_{n-1}$

$$
\begin{aligned}
q_n &= q_{n-1} + \alpha(u_n - q_{n-1}) \\
&= (1-\alpha)q_{n-1} + \alpha u_n \\
&= (1-\alpha)\big((1-\alpha)q_{n-2} + \alpha u_{n-1}\big) + \alpha u_n \\
&= (1-\alpha)^2 q_{n-2} + (1-\alpha)\alpha u_{n-1} + \alpha u_n \\
&= (1-\alpha)^3 q_{n-3} + (1-\alpha)^2 \alpha u_{n-2} + (1-\alpha)\alpha u_{n-1} + \alpha u_n \\
&= (1-\alpha)^n q_0 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} u_i
\end{aligned}
$$

Recursively unpacking the content of $q_n$ reveals that **observations $u_i$ are exponentially discounted into the past.**

# How are we treating observations in Rescorla-Wagner learning?

Taking $q_0 = 0$ and $\gamma := 1 - \alpha$, we get

$$q_n = q_{n-1} + \alpha(u_n - q_{n-1})$$

$$= (1 - \alpha)q_{n-1} + \alpha u_n$$

$$= \alpha \sum_{i=1}^{n} (1 - \alpha)^{n-i} u_i$$

$$= (1 - \gamma) \sum_{i=1}^{n} \gamma^{n-i} u_i$$

$$= (1 - \gamma) \sum_{i=0}^{n-1} \gamma^{i} u_{n-i}$$

# Does a constant learning rate solve our problems?

Partly: it implies a certain rate of forgetting because it amounts to taking only the $n = \frac{1}{\varepsilon}$ 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 $\varepsilon$ reflects the right level of uncertainty at any one time, and

(b) account for changes in uncertainty if $\varepsilon$ 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 ($\pi_r$ 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 ($\pi_{s|r}$ 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)
- Et cetera

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.

# 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)
- Et cetera

We will look at two of these:

- **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.

# Dealing with nonstationary environments: the Kalman filter

So far, so good: we have learnt how to model uncertainty resulting from observation noise explicitly

However, we face the same problem as when we updated the mean of a time series: in making updates from sequential observations in this way, **we are assuming that the underlying hidden state $x$ is stationary.**

Relaxing this assumption and replacing it with a Gaussian random walk gives us the **Kalman filter:**

$$p\left(x^{(k)}|x^{(k-1)},\vartheta\right)=\mathcal{N}\left(x^{(k)};x^{(k-1)},\vartheta\right)$$

$$p\left(u^{(k)}|x^{(k)},\sigma\right)=\mathcal{N}\left(u^{(k)};x^{(k)},\sigma\right)$$

Combining this with the **prior**

$$p\left(x^{(k-1)}\right)=\mathcal{N}\left(x^{(k-1)};\mu_x^{(k-1)},1/\pi_x^{(k-1)}\right),\ldots$$

# Dealing with nonstationary environments: the Kalman filter

... and doing some algebra, we get the **posterior**

$$p\left(x^{(k)}\right) = \mathcal{N}\left(x^{(k)}; \mu_x^{(k)}, 1/\pi_x^{(k)}\right)$$

with

$$\pi_x^{(k)} = \frac{1}{\sigma_x^{(k-1)} + \vartheta} + \frac{1}{\sigma} = \hat{\pi}_x^{(k-1)} + \hat{\pi}_u$$

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

$$= \mu_x^{(k-1)} + \frac{\hat{\pi}_u}{\frac{1}{\sigma_x^{(k-1)} + \vartheta} + \hat{\pi}_u}\left(u^{(k)} - \mu_x^{(k-1)}\right)$$

**The Kalman filter is optimal for linear dynamic systems**.

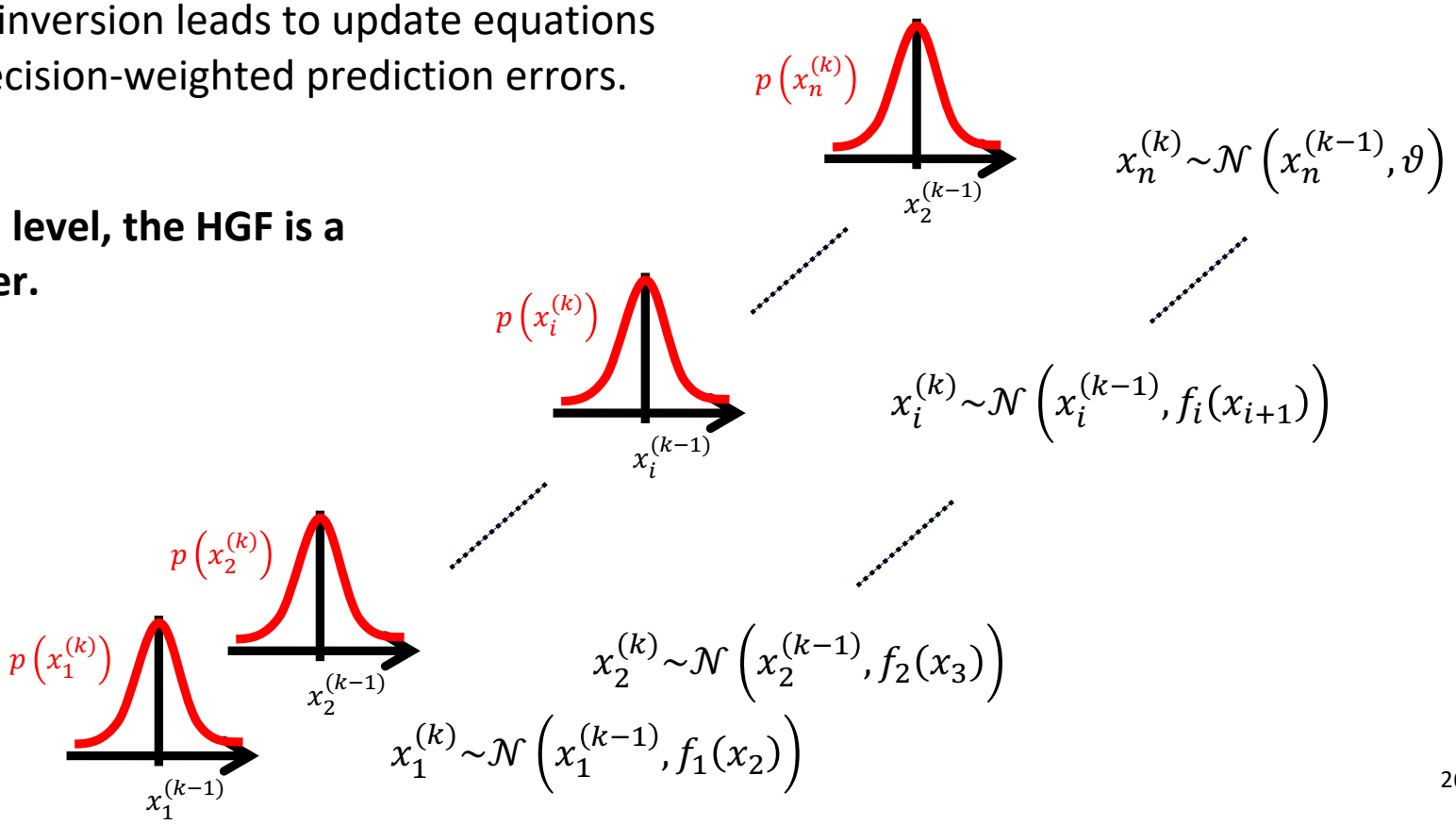
Unfortunately, except for simple physical systems, **the world is not linear**. Living organisms need to be able to filter inputs whose rate of change changes, in other words: **processes whose volatility is volatile.**

# The hierarchical Gaussian filter (HGF, Mathys et al., 2011; 2014)

The HGF provides a generic solution to the problem of adapting one's learning rate in a volatile environment.

Variational inversion leads to update equations that are precision-weighted prediction errors.

**With only 1 level, the HGF is a Kalman filter.**



# Variational inversion and update equations

- Inversion proceeds by introducing a mean field approximation and fitting quadratic approximations to the resulting variational energies (Mathys et al., 2011).
- This leads to **simple one-step update equations** for the sufficient statistics (mean and precision) of the approximate Gaussian posteriors of the states $x_i$.
- The updates of the means have the same structure as value updates in Rescorla-Wagner learning:

$$\Delta\mu_i \propto \frac{\hat{\pi}_{i-1}}{\pi_i}\delta_{i-1}$$

Prediction error

Precisions determine learning rate

- The updates are **precision-weighted prediction errors**.

# Precision-weighting of volatility updates

Comparison to the simple non-hierarchical Bayesian update:

**HGF:**

$$\mu_i^{(k)} = \mu_i^{(k-1)} + \frac{1}{2}\kappa_{i-1}\, v_{i-1}^{(k)} \cdot \frac{\hat{\pi}_{i-1}^{(k)}}{\pi_i^{(k)}} \cdot \delta_{i-1}^{(k)}$$

Precision-weighted prediction error

**Simple Gaussian:**

$$\mu_{s|r} = \mu_s + \frac{\pi_r}{\pi_{s|r}}(r - \mu_s)$$

# Updates at the outcome level

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):

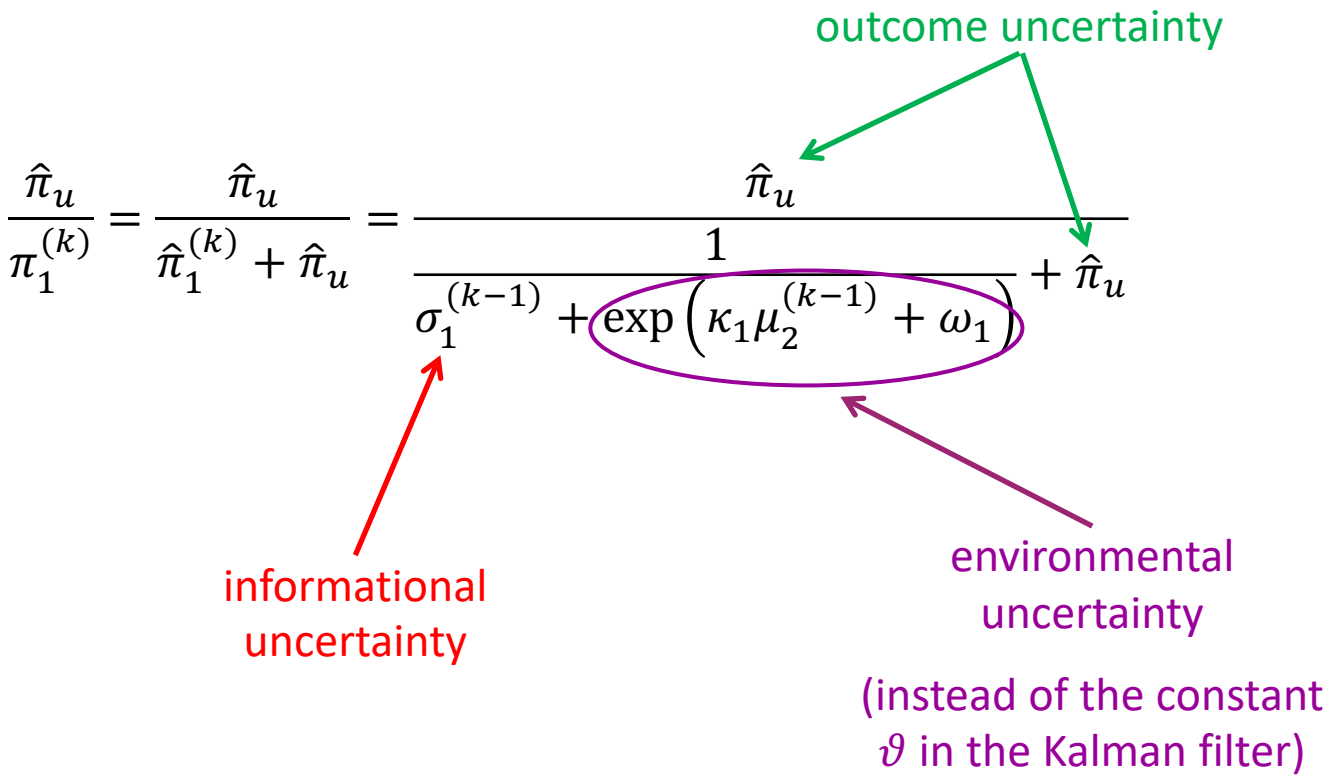
$$\pi_1^{(k)} = \hat{\pi}_1^{(k)} + \hat{\pi}_u$$

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

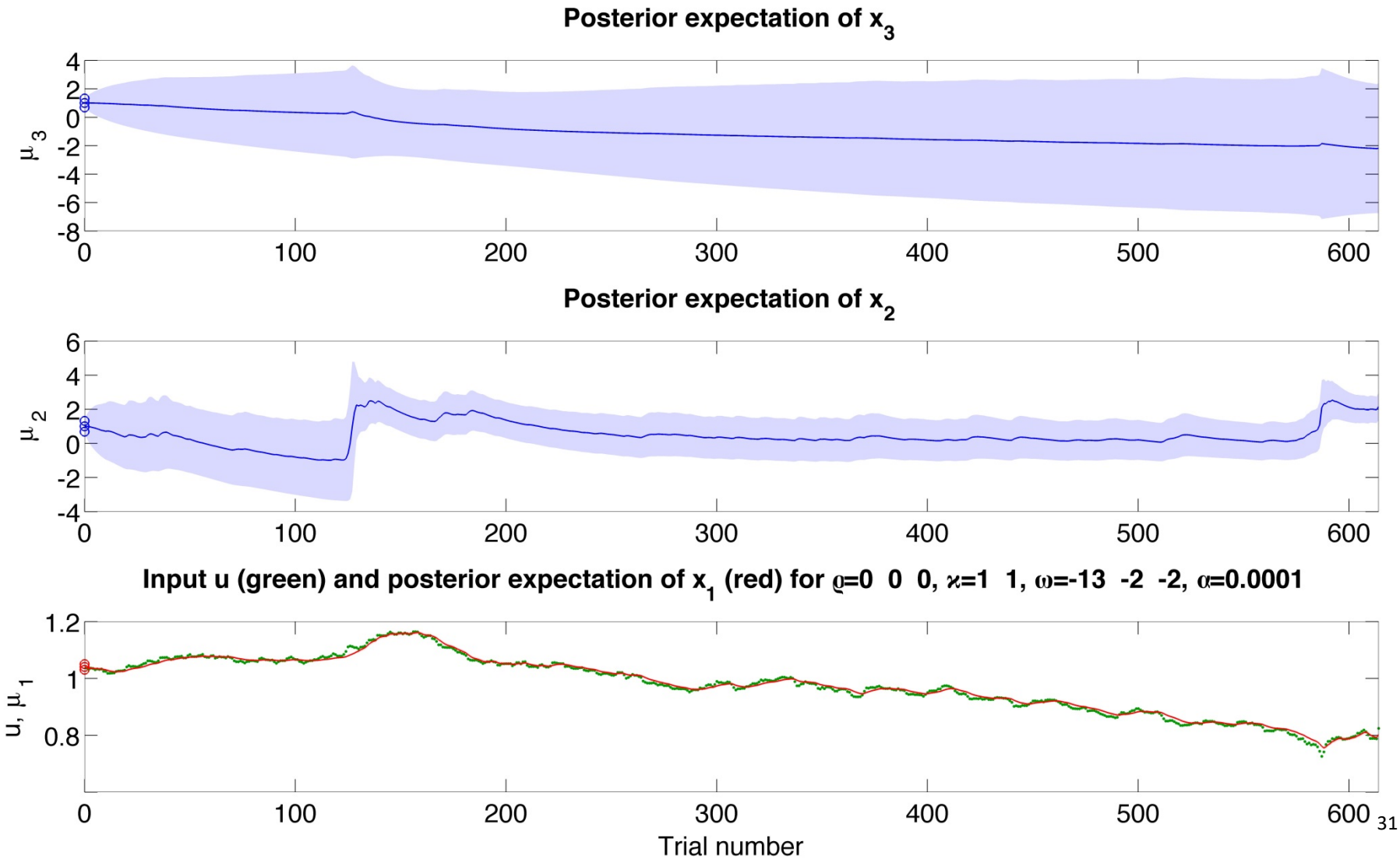
The familiar structure again – but now with a learning rate that is responsive to all kinds of uncertainty, including environmental (unexpected) uncertainty.

# The learning rate ('Kalman gain') in the HGF

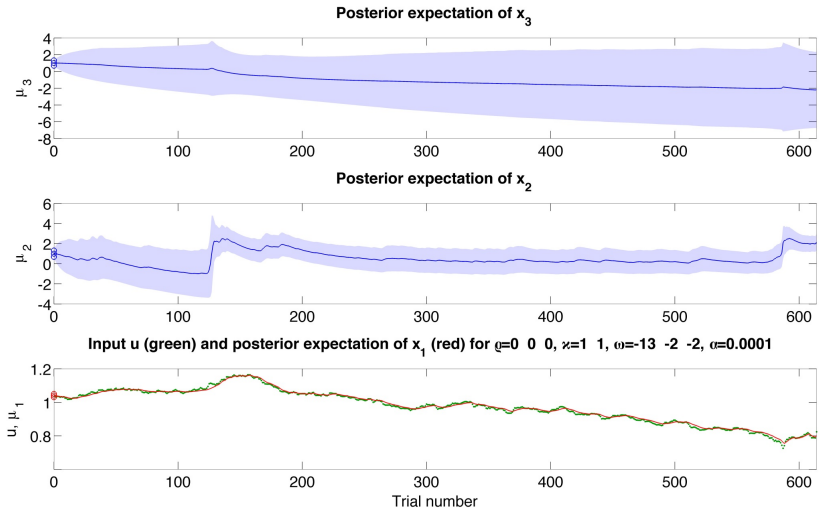
Unpacking the learning rate, we see:

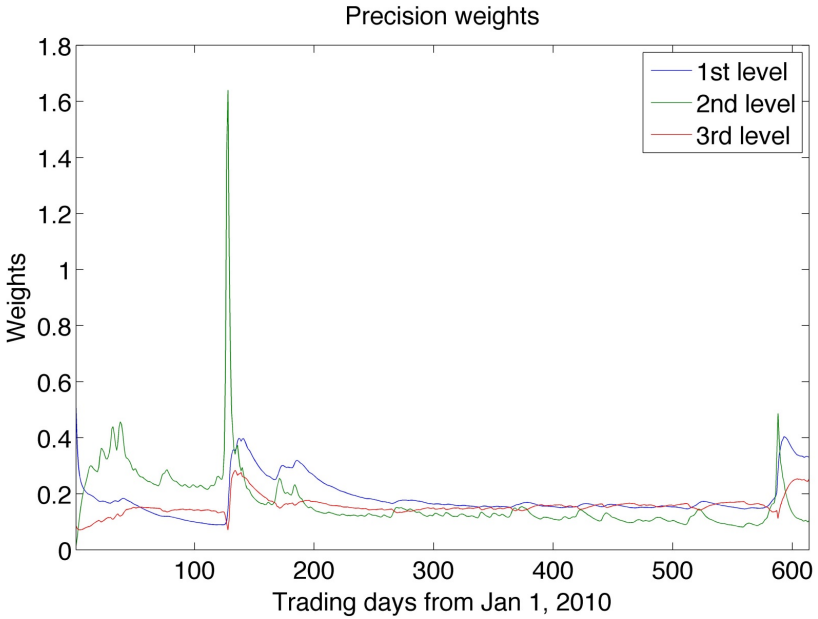


# 3-level HGF for continuous observations

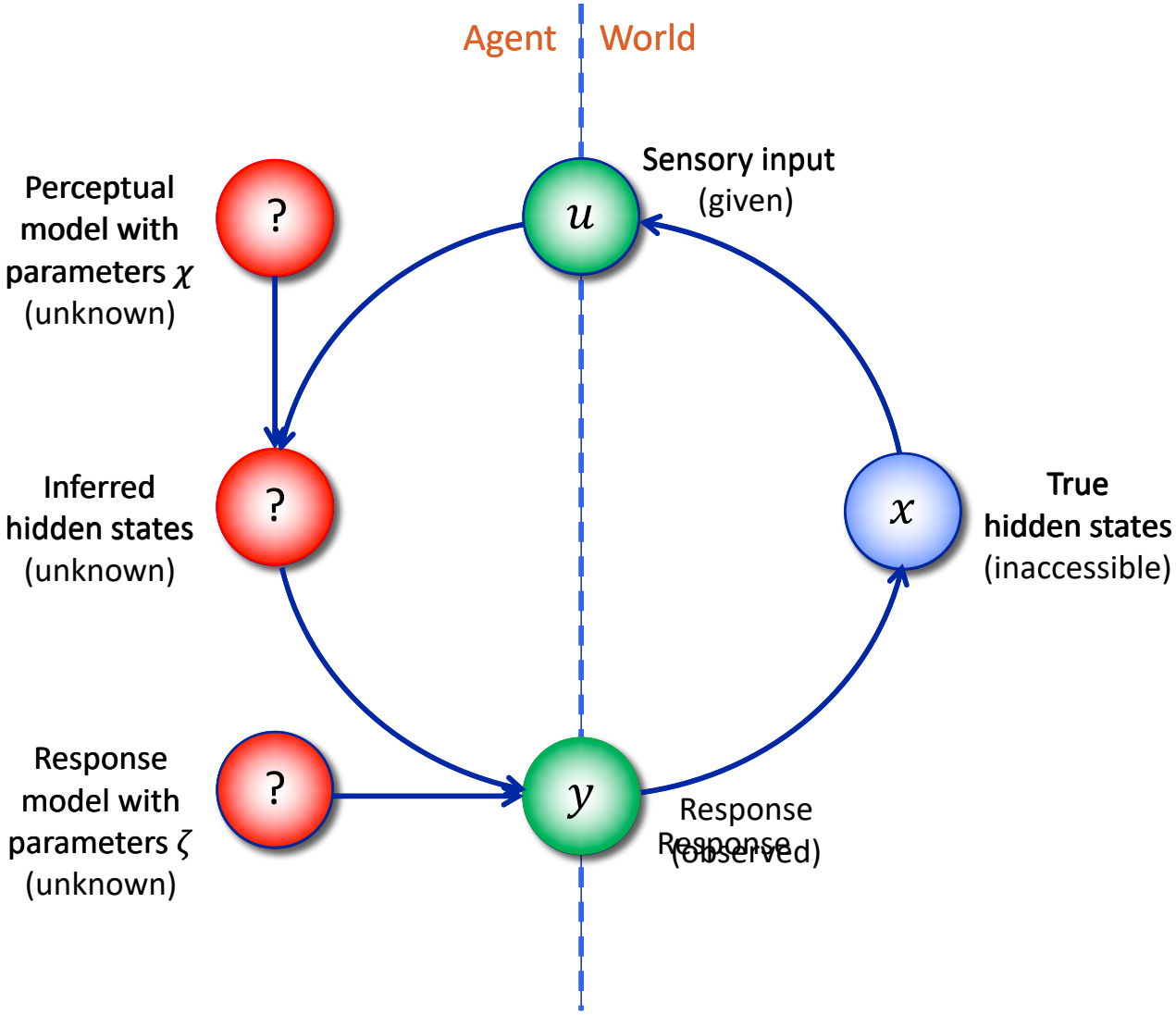


# Example of precision weight trajectory

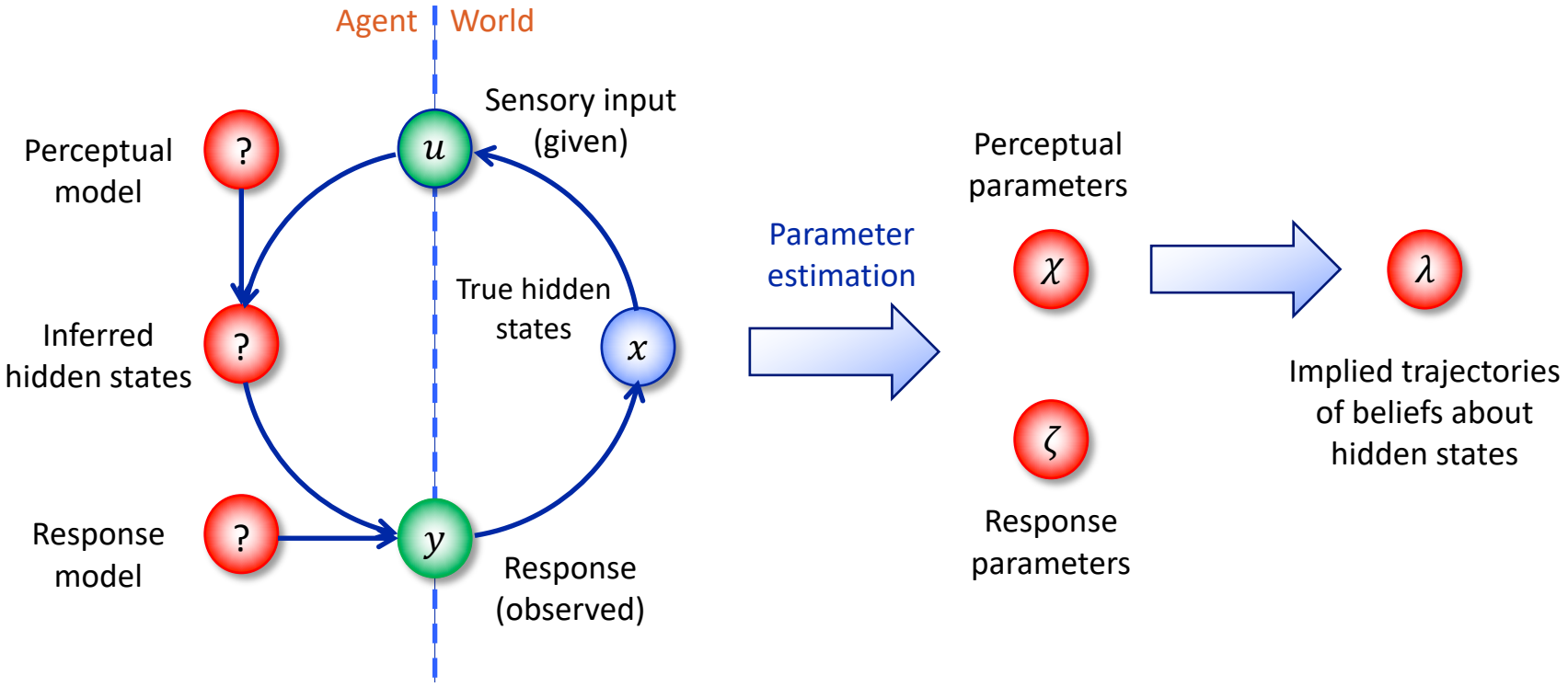


# Application to experimental data



# Application to experimental data: parameter estimation



# Parameter estimation with the HGF Toolbox

- Available at

  **https://translationalneuromodeling.github.io/tapas**

- Start with README, manual, and interactive demo

- Modular, extensible, Matlab-based

```
est2 = tapas_fitModel(sim2.y,...
                      usdchf,...
                      'tapas_hgf_config',...
                      'tapas_gaussian_obs_config',...
                      'tapas_quasinewton_optim_config');
```

```
Parameter estimates for the perceptual model:
    mu_0: [1.0352 1]
    sa_0: [3.7101e-05 0.0996]
     rho: [0 0]
      ka: 1
      om: [-12.8680 -1.8689]
    pi_u: 9.8449e+03
```

```
Parameter estimates for the observation model:
    ze: 2.3970e-05
```

# Studies of attention and expectation

Cerebral Cortex June 2014;24:1436–1450
doi:10.1093/cercor/bhs418
Advance Access publication January 14, 2013

## Spatial Attention, Precision, and Bayesian Inference: A Study of Saccadic Response Speed

Simone Vossel[1], Christoph Mathys[2,3], Jean Daunizeau[2,3], Markus Bauer[1], Jon Driver[1], Karl J. Friston[1] and Klaas E. Stephan[1,2,3]

**Posner paradigm with varying % cue validity:**

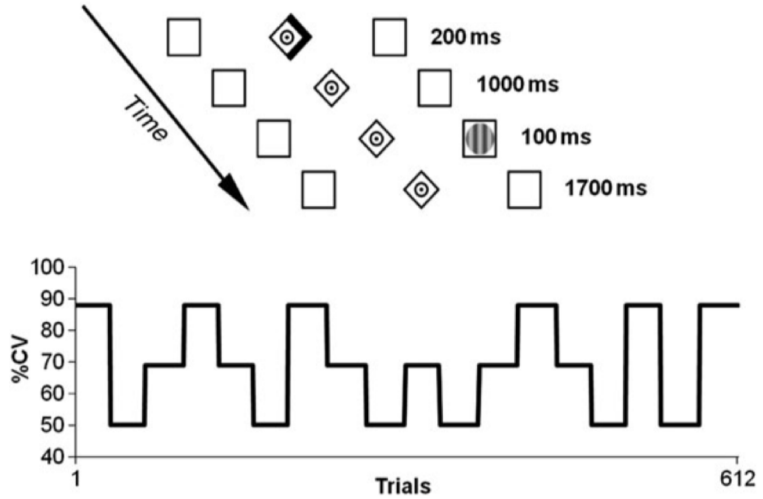



**Figure 1.** Illustration of the experimental task and the manipulation of %CV over the 612 trials.

# Studies of attention and expectation

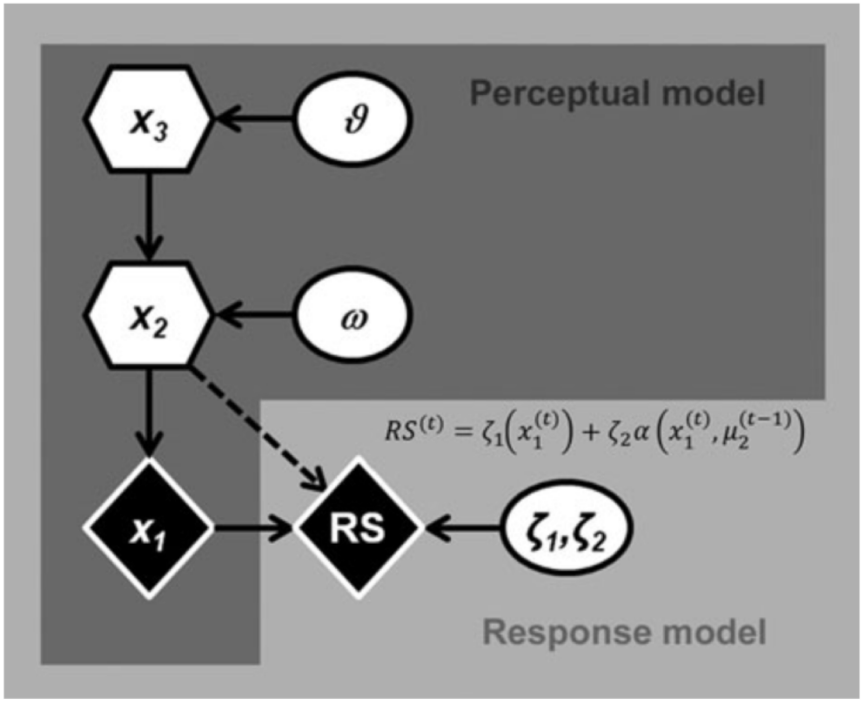



**Figure 2.** Graphical illustration of the perceptual (generative) model with states $x_1$, $x_2$, and $x_3$. The model parameters $\omega$ and $\vartheta$ impact on the time course of subjects' inferred belief about the states $x$ and are estimated from the individual subject RS data. Circles represent constants, while diamonds represent quantities that change in time (i.e., that carry a time (or trial) index). Hexagons, like diamonds, represent quantities that change in time but that additionally depend on their previous state in time in a Markovian fashion.

# Studies of attention and expectation

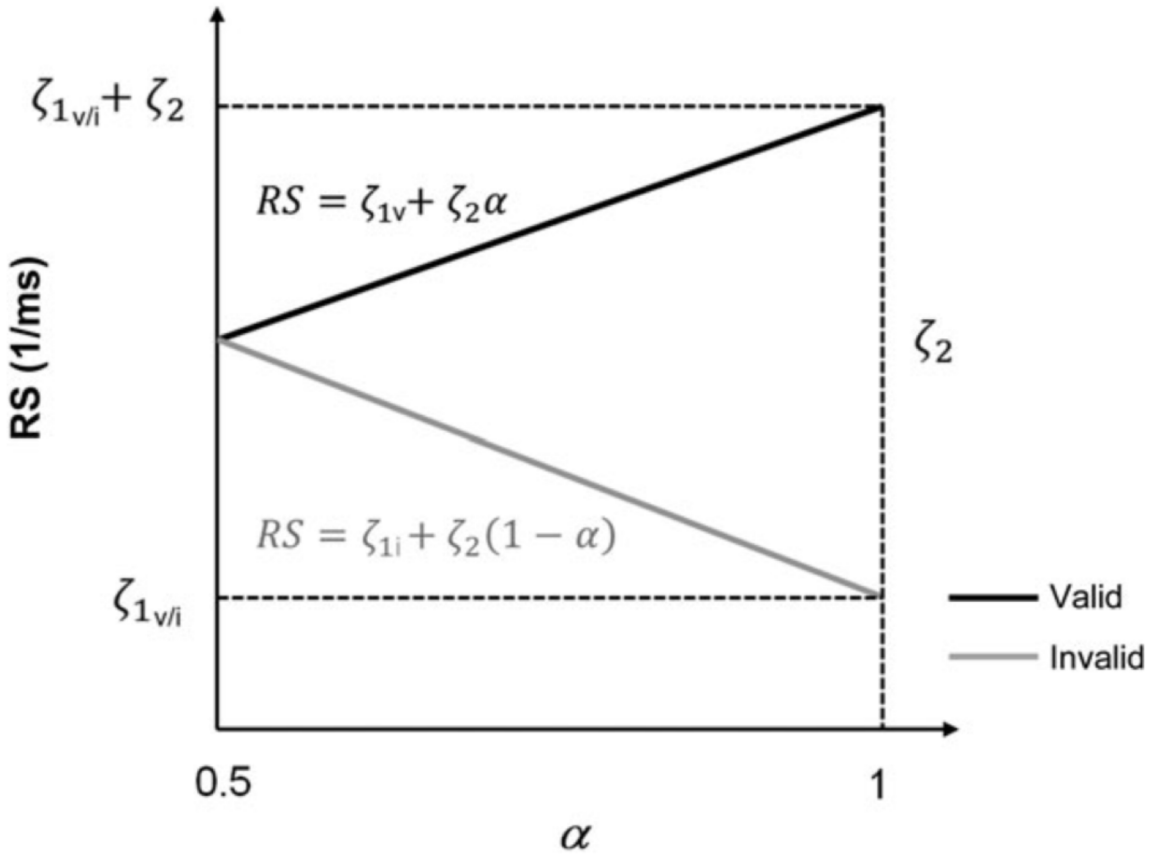



**Figure 3.** Illustration of the relationship between RS (inverse RT) and the quantity $\alpha$, representing the amount of attentional resources allocated to the cued location. For each response model, RS were assumed to be linearly related to $\alpha$ (which differs between the 3 models, see Appendix). Note the opposite behaviour of RS for increasing $\alpha$ on valid (black line and equation) and invalid (gray line and equation) trials (cf. eq. 5).

# Studies of attention and expectation

$$\mathrm{RS} = \begin{cases} \zeta_{1_{\mathrm{valid}}} + \zeta_2 & \text{for } x_1 = 1 \text{ (i.e., valid trial)}, \\ \zeta_{1_{\mathrm{invalid}}} + \zeta_2(1-\alpha) & \text{for } x_1 = 0 \text{ (i.e., invalid trial)}. \end{cases}$$

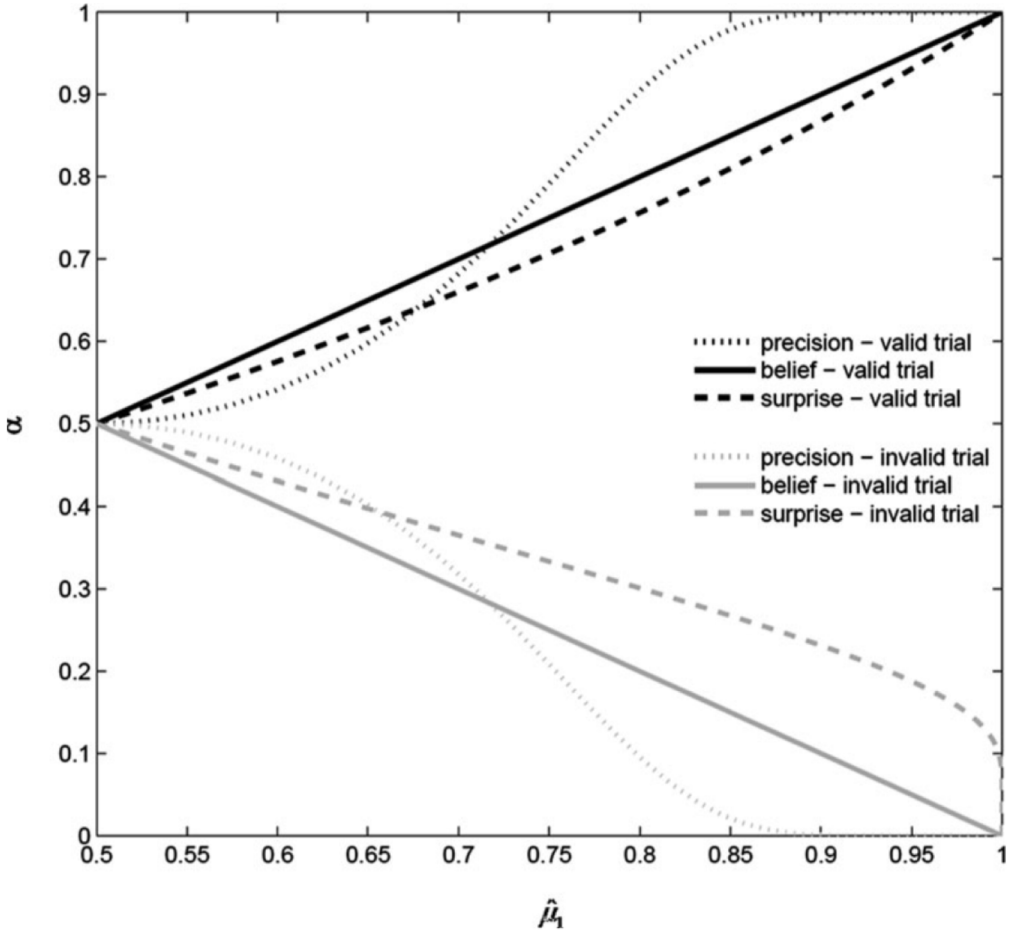
**Precision model:** $\alpha^{(t)} = s(\hat{\pi}_1^{(t)} - 4)$

**Belief model:** $\alpha^{(t)} = \hat{\mu}_1^{(t)} = s(\mu_2^{(t-1)})$

**Surprise model:** $\alpha^{(t)} = \dfrac{1}{(1 + \mathrm{surprise}(\hat{\mu}_1^{(t)}))}$

$$\mathrm{surprise}\,(\hat{\mu}_1^{(t)}) = -\log_2 p(x_1^{(t)} = 1 | \hat{\mu}_1^{(t)}) = -\log_2(\hat{\mu}_1^{(t)})$$

# Studies of attention and expectation



# Studies of attention and expectation

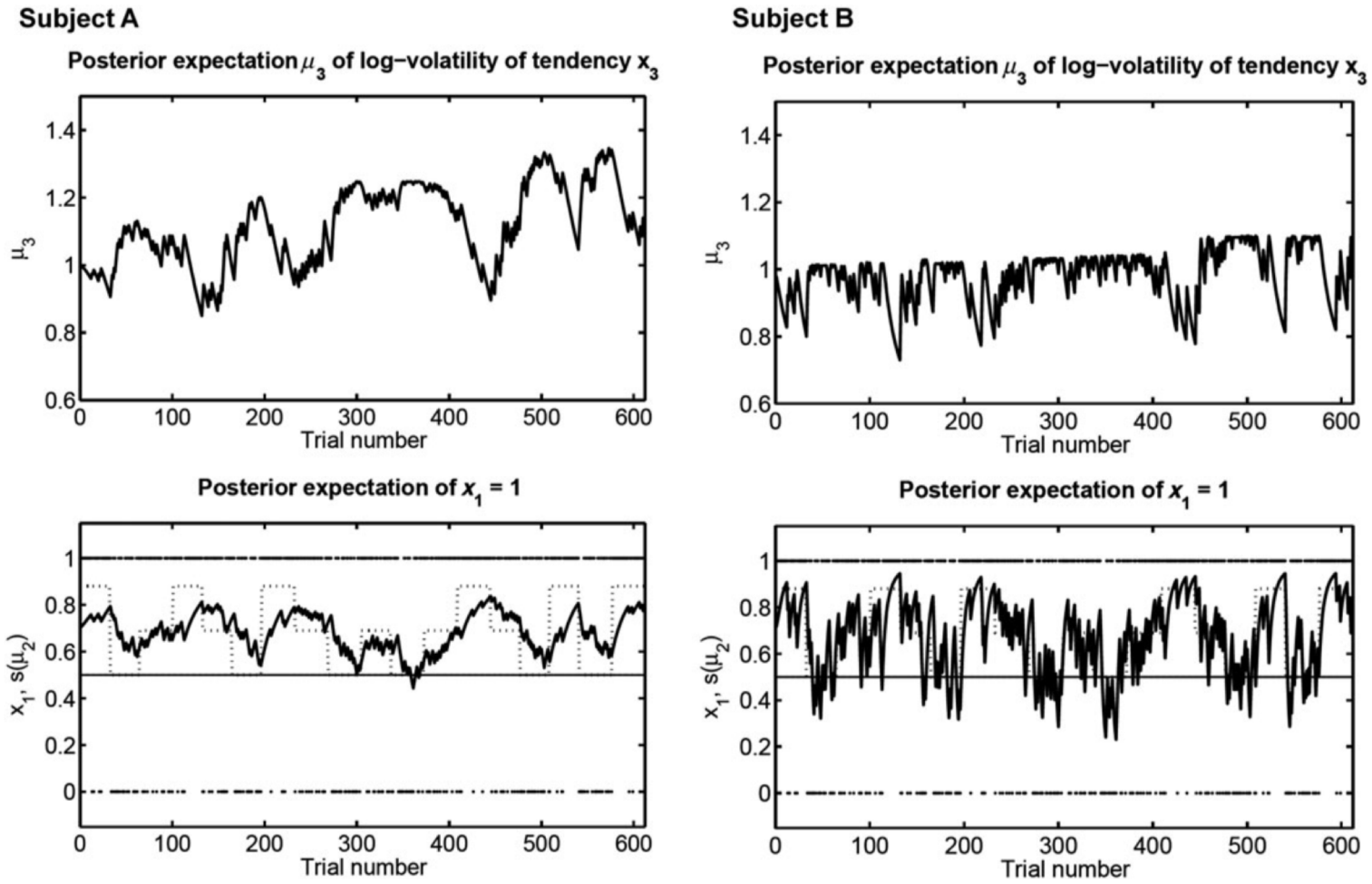
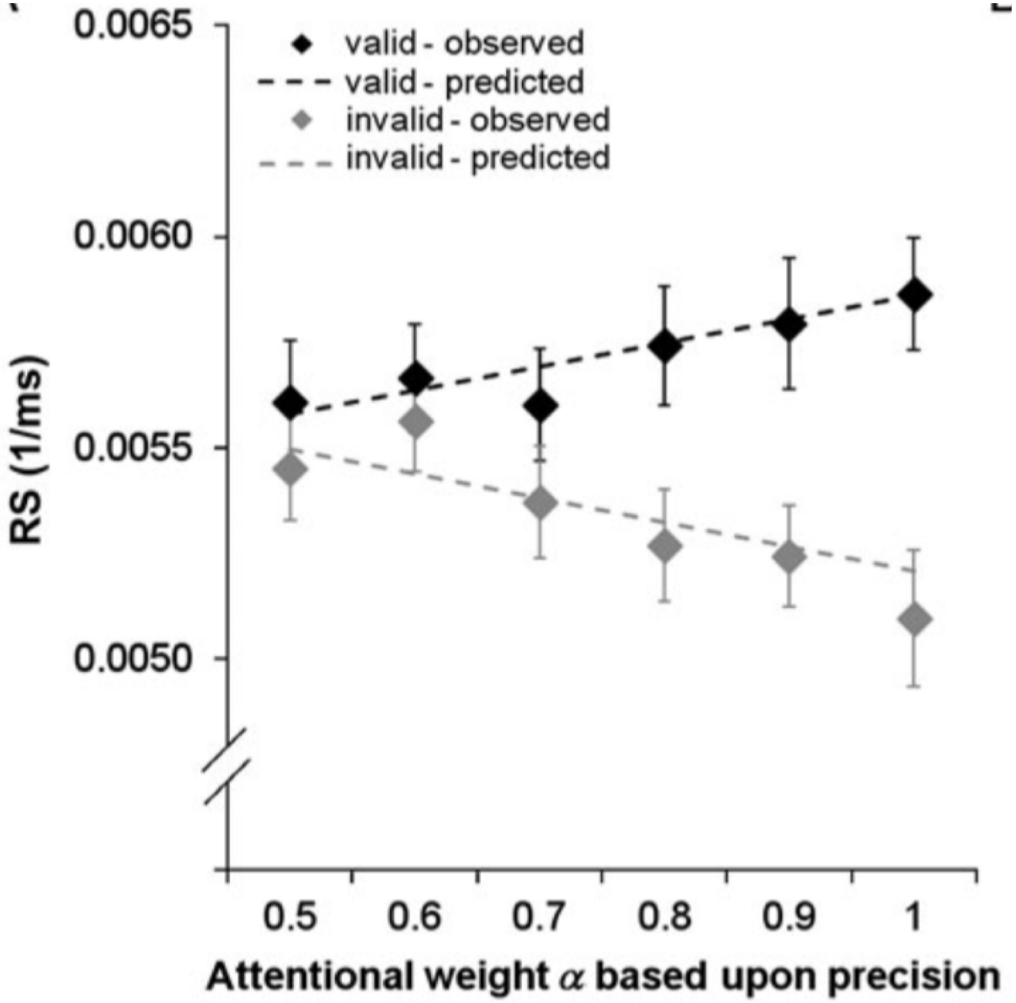



**Figure 8.** Illustration of the time course of $\mu_3$ (upper panels) and $s(\mu_2)$ (lower panels) during observation of $x_1$ (black diamonds) for 2 exemplary subjects with different parameters for $\omega$ and $\vartheta$. The true %CV is depicted as a dotted line. It can be seen that subject A ($\omega = -6.09$; $\vartheta = 0.97$) shows slower updating of the probability estimate that the target will appear at the cued location than subject B ($\omega = -2.78$; $\vartheta = 0.12$). This can be attributed to subject A's lower value of $\omega$ (reflecting the subject's belief in a less volatile environment).

# Studies of attention and expectation



# Studies of attention and expectation

The Journal of Neuroscience, November 19, 2014 • 34(47):15735–15742 • 15735

Behavioral/Cognitive

## Cholinergic Stimulation Enhances Bayesian Belief Updating in the Deployment of Spatial Attention

Simone Vossel,[1,2] Markus Bauer,[1,3] Christoph Mathys,[1,4,5] Rick A. Adams,[1] Raymond J. Dolan,[1] Klaas E. Stephan,[1,5,6] and Karl J. Friston[1]

# Studies of attention and expectation

The Journal of Neuroscience, November 19, 2014 • 34(47):15735–15742 • 15735

Behavioral/Cognitive

## Cholinergic Stimulation Enhances Bayesian Belief Updating in the Deployment of Spatial Attention

Simone Vossel,[1,2] Markus Bauer,[1,3] Christoph Mathys,[1,4,5] Rick A. Adams,[1] Raymond J. Dolan,[1] Klaas E. Stephan,[1,5,6] and Karl J. Friston[1]

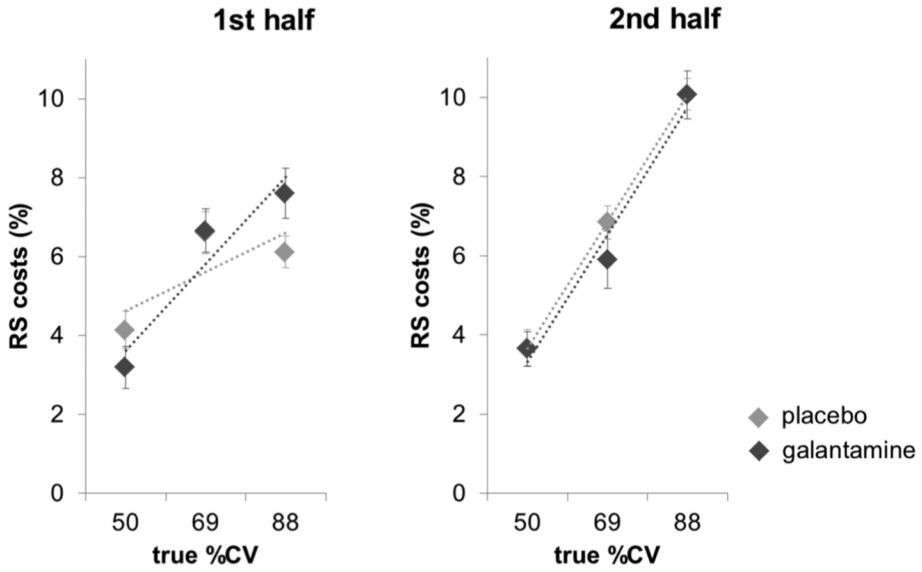



**Figure 3.** Illustration of the three-way interaction between cue, % CV, and drug for both halves of the experiment. For simplicity, relative RS differences between valid and invalid trials are shown. RS costs were related to the overall speed of responding and are expressed in percentage: (RS valid − RS invalid)/overall RS × 100. Error bars indicate SEM.

# Studies of attention and expectation

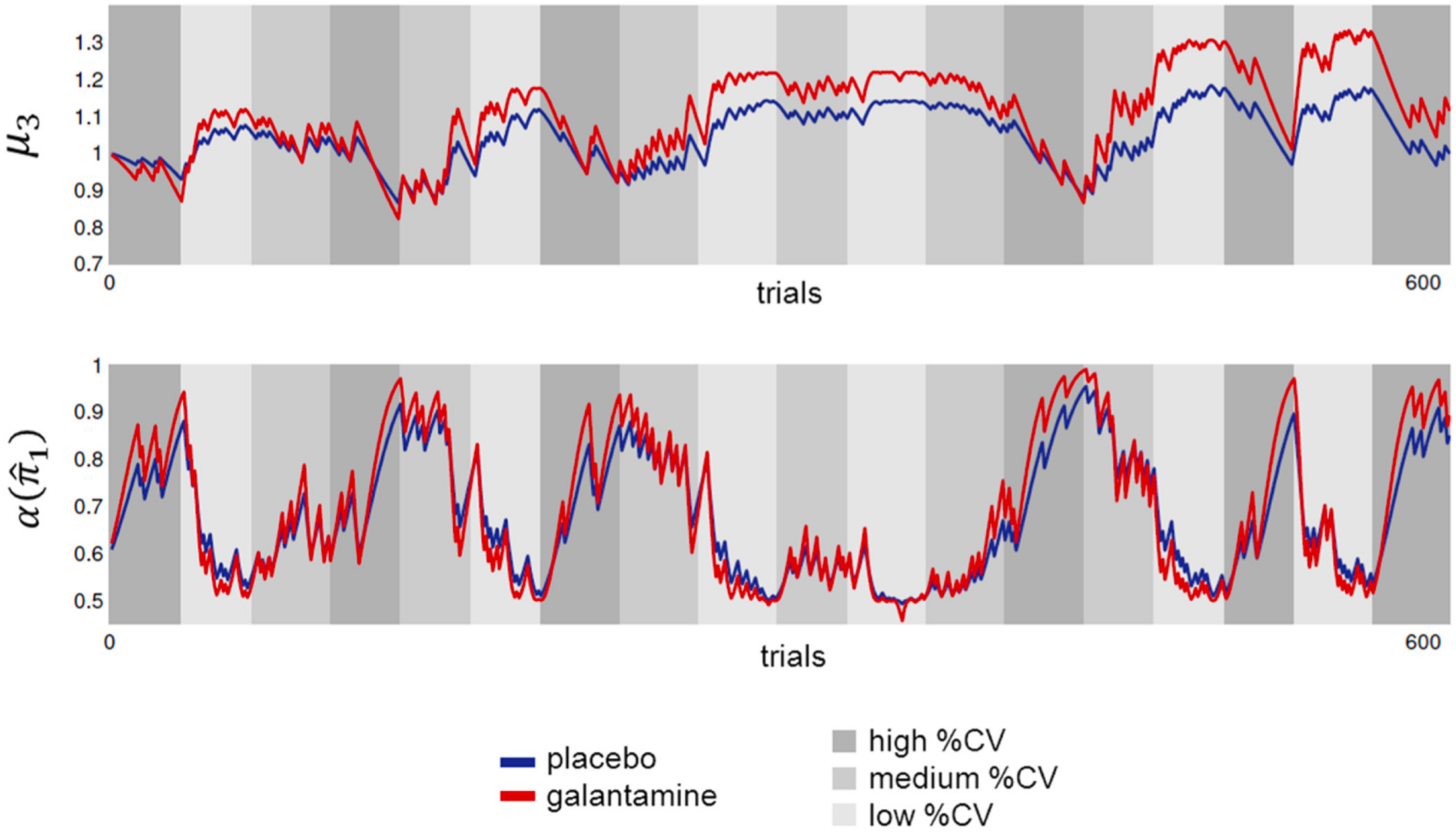



**Figure 4.** Illustration of the time course of the posterior expectations at the third level of the Bayesian model ($\mu_3$, top) and attentional gain allocated to the cued location $\alpha(\hat{\pi}_1)$ (bottom), based upon the group average values for $\omega$ and $\vartheta$. Because of the dependency of the galantamine-induced increase of $\omega$, we here exemplarily show the results for a normalized weight of 68 kg (placebo session: $\omega = -6.2$, $\vartheta = 0.60$; galantamine session: $\omega = -5.4$, $\vartheta = 0.61$). It can be seen that, under galantamine (red line, with a higher $\omega$), the changes in $\alpha(\hat{\pi}_1)$ in the different true % CV conditions are more pronounced than under placebo (blue line). The same sequence of valid and invalid trials was presented each subject and in each experimental session because the parameters of the learning process depend on the exact sequence of trials used.

# Studies of attention and expectation

11532 • The Journal of Neuroscience, August 19, 2015 • 35(33):11532–11542

Behavioral/Cognitive

## Cortical Coupling Reflects Bayesian Belief Updating in the Deployment of Spatial Attention

Simone Vossel,[1,2] Christoph Mathys,[1,3,4] Klaas E. Stephan,[1,4,5] and Karl J. Friston[1]

# Studies of attention and expectation

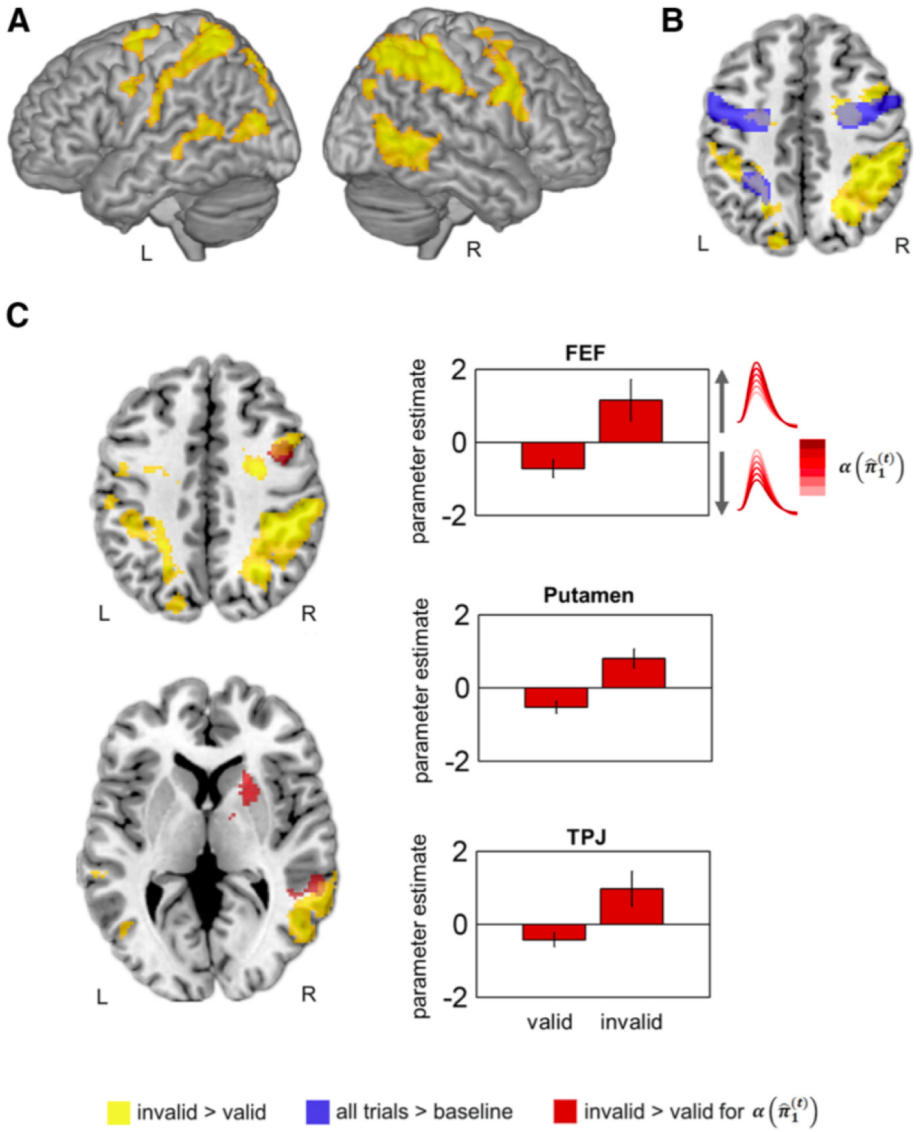
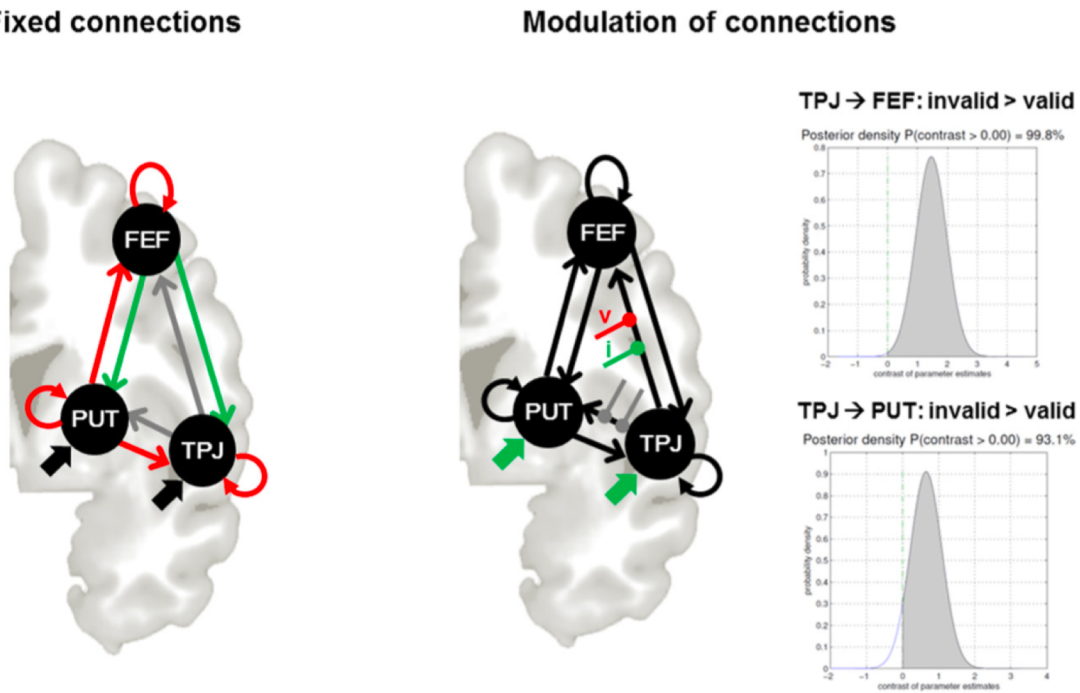



**Figure 3.** Results of the GLM analysis. ***A***, Reorienting network as revealed by a *t*-contrast (invalid > valid) for the main HRF regressor. ***B***, Overlap between the saccade network and the reorienting network in the FEF. ***C***, Precision-dependent differences between invalid and valid trials as revealed by a *t*-contrast (invalid > valid) for the parametric modulation with $\alpha(\hat{\pi}_1^{(t)})$. The bar charts depict the mean parameter ($\beta$) estimates ($\pm$ SEM) for the parametric regressor for valid and invalid trials.

# Studies of attention and expectation



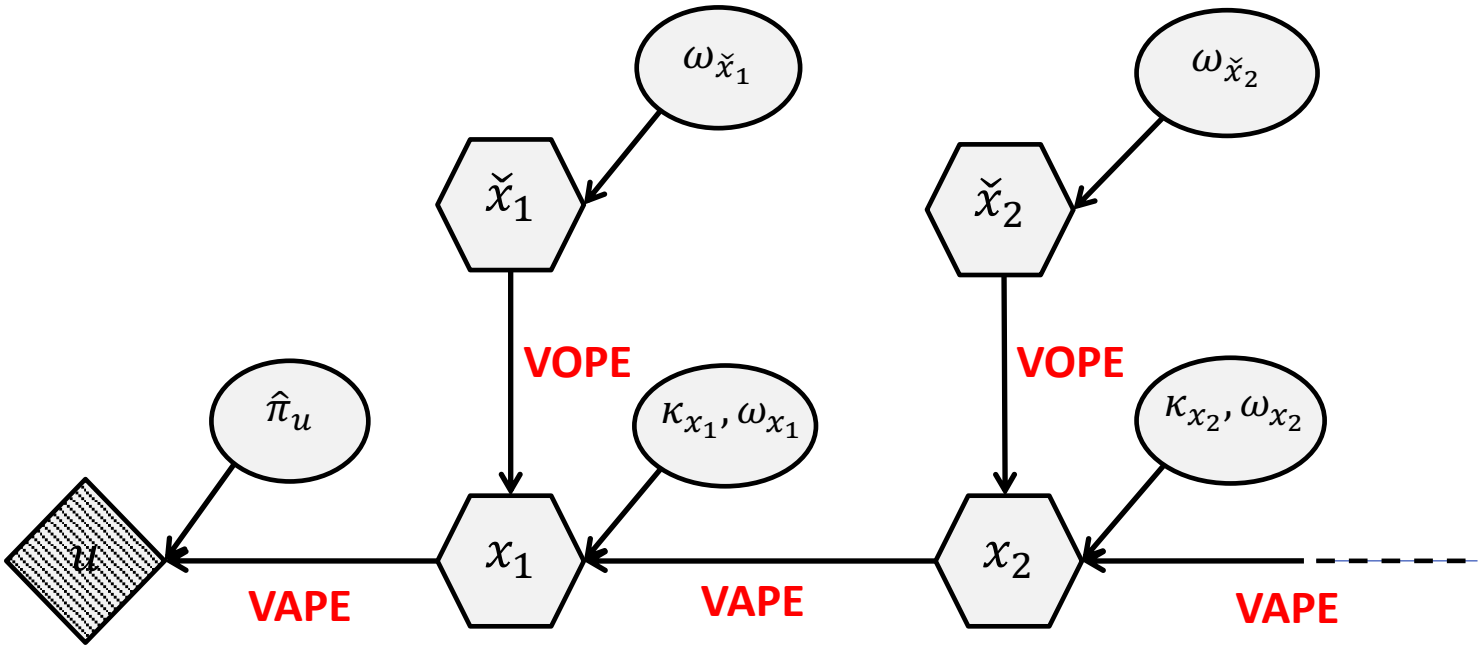

| | | From FEF | From TPJ | From PUT |
|---|---|---|---|---|
| To | FEF | -0.57 (1.00) | -0.01 (0.78) | -0.15 (1.00) |
| To | TPJ | 0.75 (1.00) | -0.56 (1.00) | -0.16 (1.00) |
| To | PUT | 0.27 (1.00) | -0.001 (0.62) | -0.42 (1.00) |

| | Modulation of connectivity by $\alpha\left(\hat{\pi}_1^{(t)}\right)$: valid trials | Modulation of connectivity by $\alpha\left(\hat{\pi}_1^{(t)}\right)$: invalid trials |
|---|---|---|
| TPJ→FEF | -0.55 (0.97) | 0.91 (0.99) |
| TPJ→PUT | -0.21 (0.81) | 0.44 (0.89) |

| | input |
|---|---|
| PUT | 0.01 (0.90) |
| TPJ | 0.03 (1.00) |

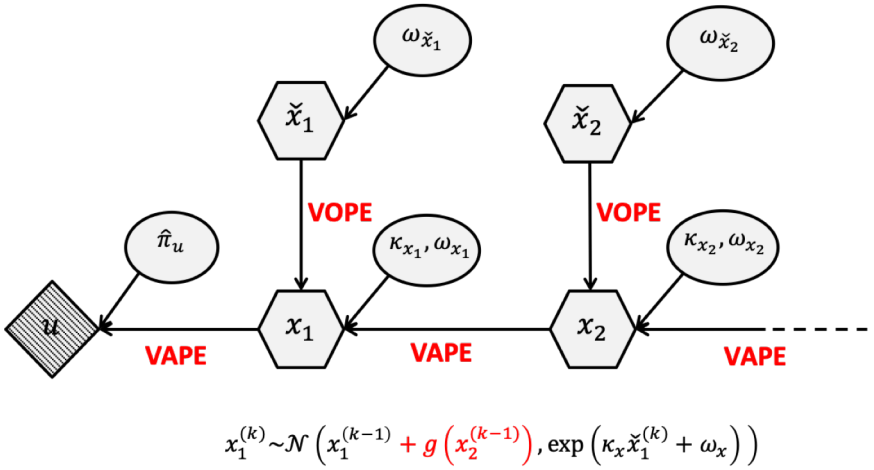
**Figure 6.** Results of the Bayesian parameter averaging across subjects under the winning model (model 10 of Fig. 5 with inputs into putamen und TPJ). Illustration of fixed connections and modulations of connections by $\alpha(\hat{\pi}_1^{(t)})$ in valid and invalid trials: parameter estimates for valid and invalid trials were contrasted against each other (cf. graphs on the right). Tables depict the Bayesian parameter averages with posterior probabilities in parentheses.

# ghgf: The basic idea




$$x_1^{(k)} \sim \mathcal{N}\left(x_1^{(k-1)} + g\left(x_2^{(k-1)}\right), \exp\left(\kappa_x \breve{x}_1^{(k)} + \omega_x\right)\right)$$

# ghgf: Overview




- Large networks of nodes coupled by either VAPEs or VOPEs
- VAPE: value prediction error
- VOPE: volatility prediction error
- Nonlinear (e.g. ReLU) VAPE coupling
- Continuous time
- Python toolbox
- Predictive coding: interpretation of neuronal activity in terms of HGF message passing

# ghgf: Predictive coding

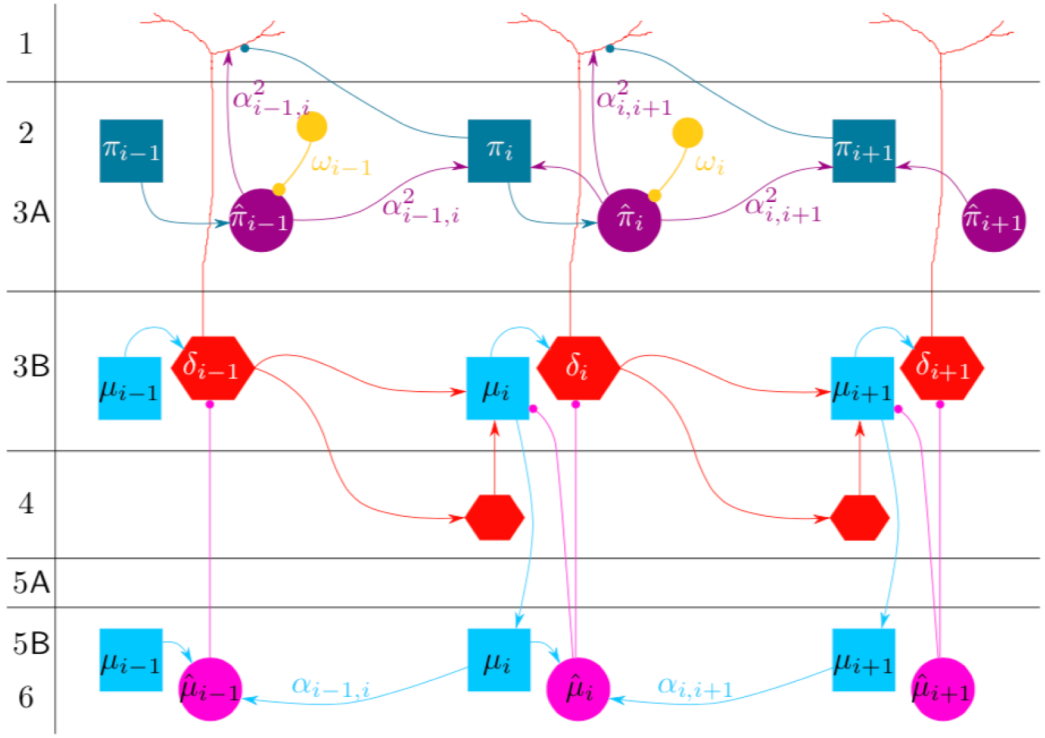



Figure 4: Overview of possible layer-specific message passing in VAPE coupling. Assignment of nodes to layers is loosely based on figure 3 in [4], reproduced in this document in figure 3. Connections with arrowheads indicate positive influences, or excitatory connections, while connections with circular heads indicate inhibitory influences.

# Thanks

Rick Adams

Archie de Berker

Sven Bestmann

Kay Brodersen

Jean Daunizeau

Andreea Diaconescu

Peter Dayan

Ray Dolan

Karl Friston

Sandra Iglesias

Lars Kasper

Rebecca Lawson

Ekaterina Lomakina

Read Montague

Tobias Nolte

Geraint Rees

Robb Rutledge

Klaas Enno Stephan

Philipp Schwartenbeck

Simone Vossel

Lilian Weber