AARHUS UNIVERSITY

Hierarchical Gaussian Filters (HGFs) as a window into belief updates

Christoph Mathys

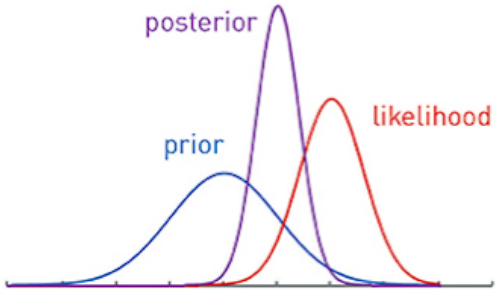
Psychology Faculty Colloquium
University of Marburg

22 June 2022

Computational modelling and the inferential brain

Bayes' Theorem

$$p(x|y,m) = \frac{p(y|x,m)p(x|m)}{p(y|m)}$$



Generative models as computational assays

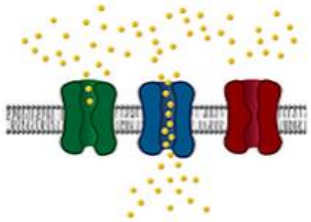
measurement y

forward model

$p(y|x,m)p(x|m)$

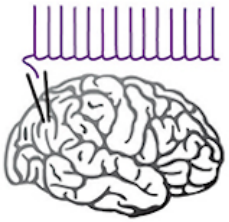
$p(x|y,m)$

inference



hidden system state x

Perception as the inversion of a generative model



neuronal activity y

forward model

$p(y|x,m)p(x|m)$

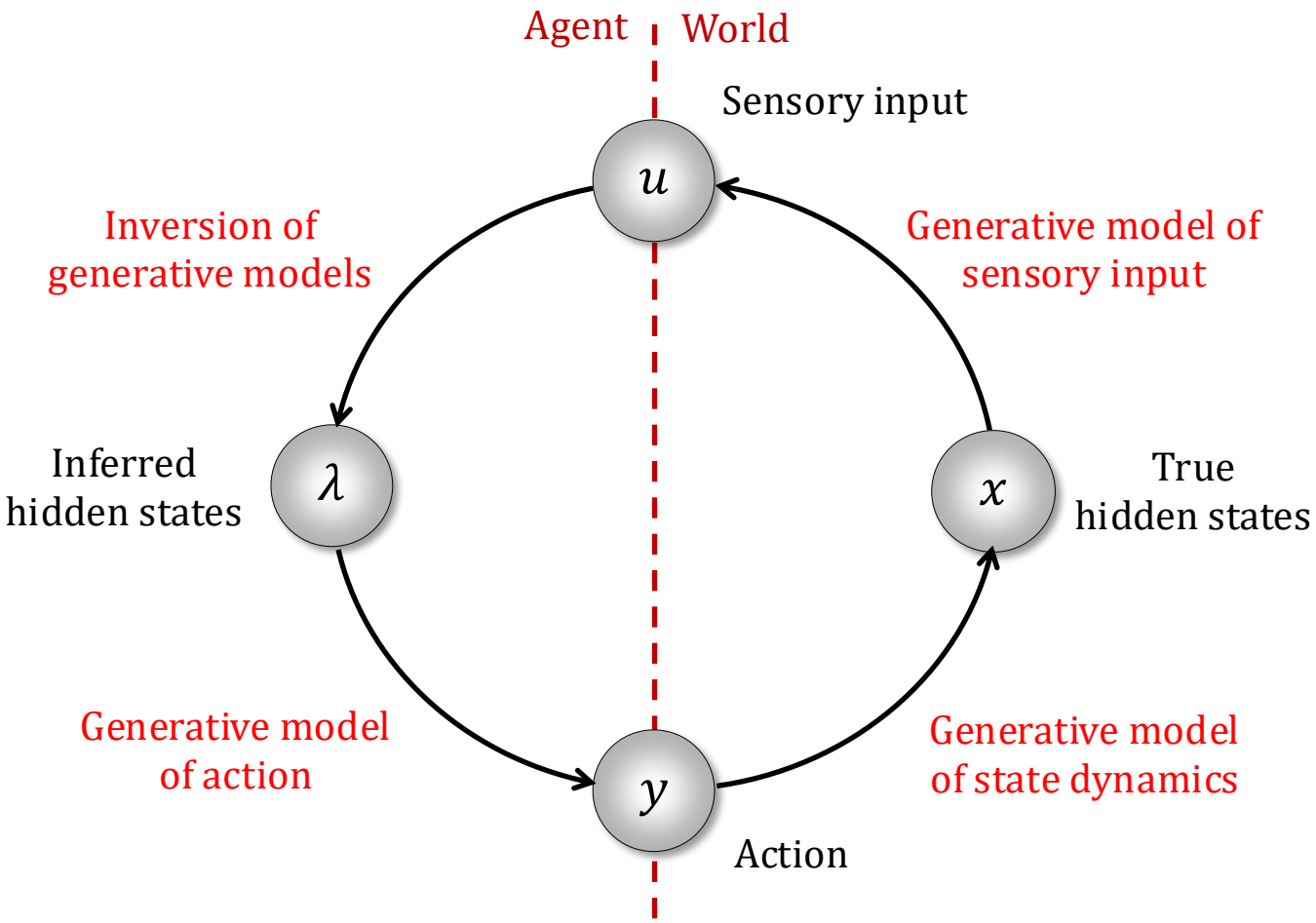
$p(x|y,m)$

perception

environmental state x

Stephan et al. (2016), Front. Hum. Neurosci., 10:550

Generative models of an agent's interaction with its environment



Neuron
Article

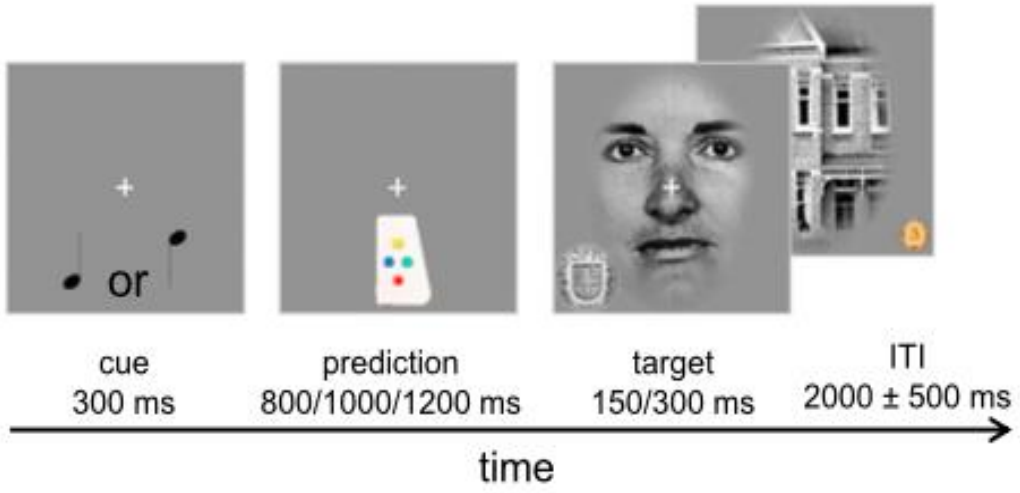
Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning

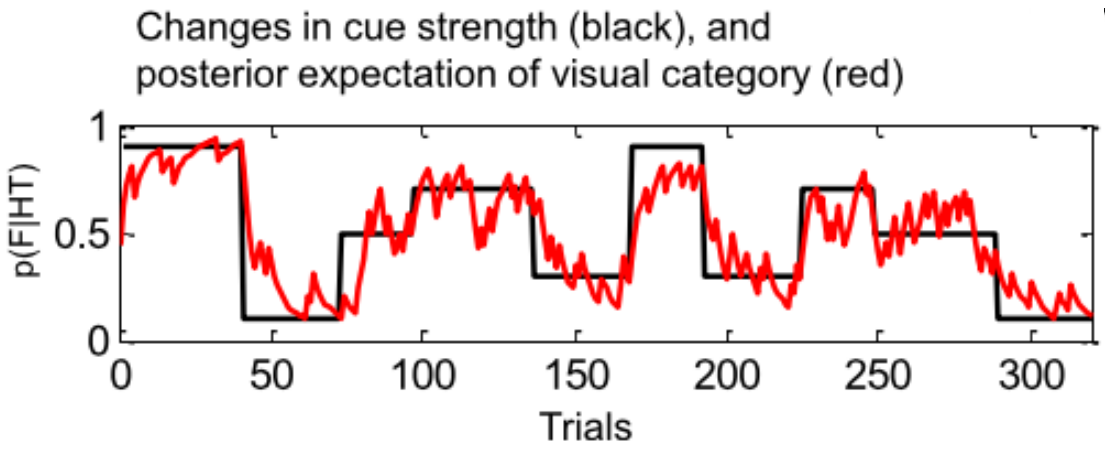
Sandra Iglesias,[1,2,*] Christoph Mathys,[1,2] Kay H. Brodersen,[1,2] Lars Kasper,[1,2] Marco Piccirelli,[2] Hanneke E.M. den Ouden,[3] and Klaas E. Stephan[1,2,4]

In Bayesian brain theories, hierarchically related prediction errors (PEs) play a central role for predicting sensory inputs and inferring their underlying causes, e.g., the probabilistic structure of the environment and its volatility. Notably, PEs at different hierarchical levels may be encoded by different neuromodulatory transmitters. Here, we tested this possibility in computational fMRI studies of audio-visual learning. Using a hierarchical Bayesian model, we found that low-level PEs about visual stimulus outcome were reflected by widespread activity in visual and supramodal areas but also in the midbrain. In contrast, high-level PEs about stimulus probabilities were encoded by the basal forebrain. These findings were replicated in two groups of healthy volunteers. While our fMRI measures do not reveal the exact neuron types activated in midbrain and basal forebrain, they suggest a dichotomy between neuromodulatory systems, linking dopamine to low-level PEs about stimulus outcome and acetylcholine to more abstract PEs about stimulus probabilities.

Neuron *80*, 519–530, October 16, 2013

Task


or
cue
300 ms
prediction
800/1000/1200 ms
target
150/300 ms
ITI
2000 ± 500 ms
time

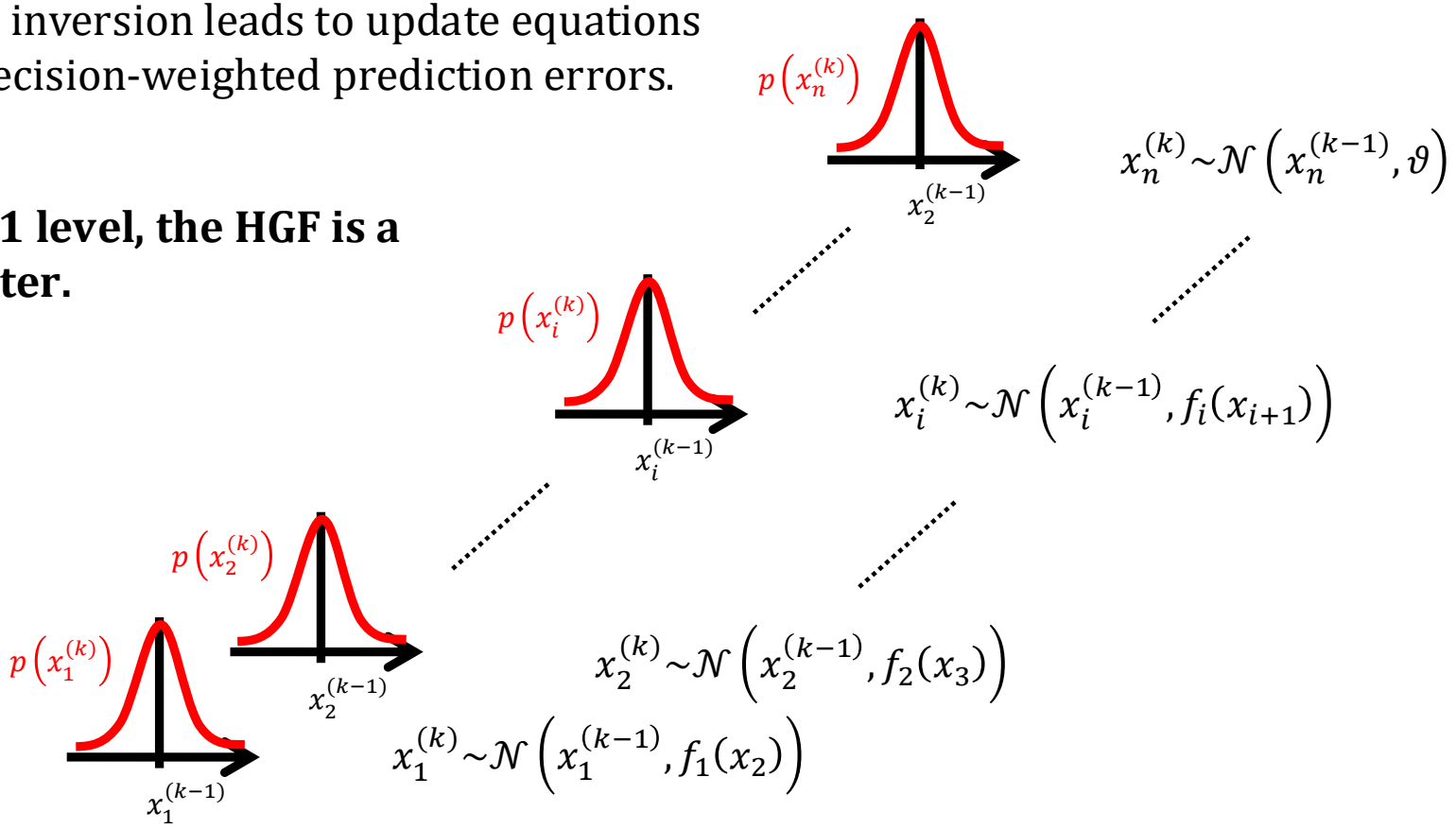

Changes in cue strength (black), and
posterior expectation of visual category (red)
p(F|HT)
1
0.5
0
0
50
100
150
200
250
300
Trials

Hierarchical Gaussian filters (HGF, Mathys et al., 2011; 2014)

HGFs provide a generic solution to the problem of adapting one's learning rate in a volatile environment.

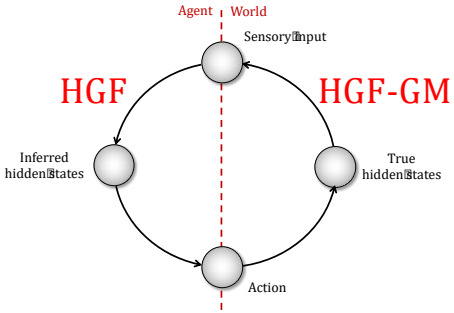
Variational inversion leads to update equations that are precision-weighted prediction errors.

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



Variational inversion and update equations

- Important distinction: **generative model** (HGF-GM) vs its inversion, **inference model** (HGF proper)



- Inversion of HGF-GM 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** (HGF proper) 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:

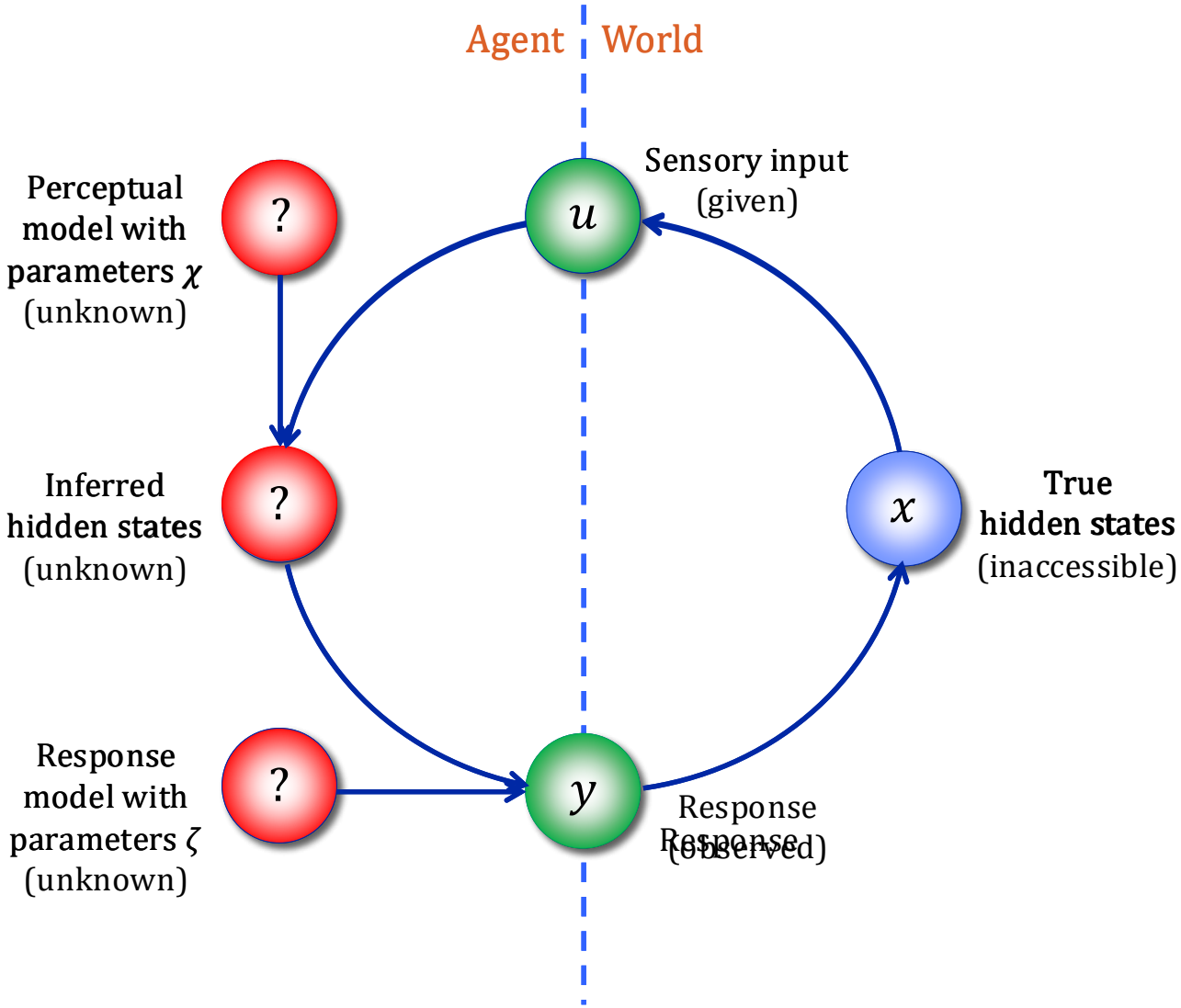
$$\Delta\mu_i \propto \frac{\hat{\pi}_{i-1}}{\pi_i}\delta_{i-1}$$

Prediction error

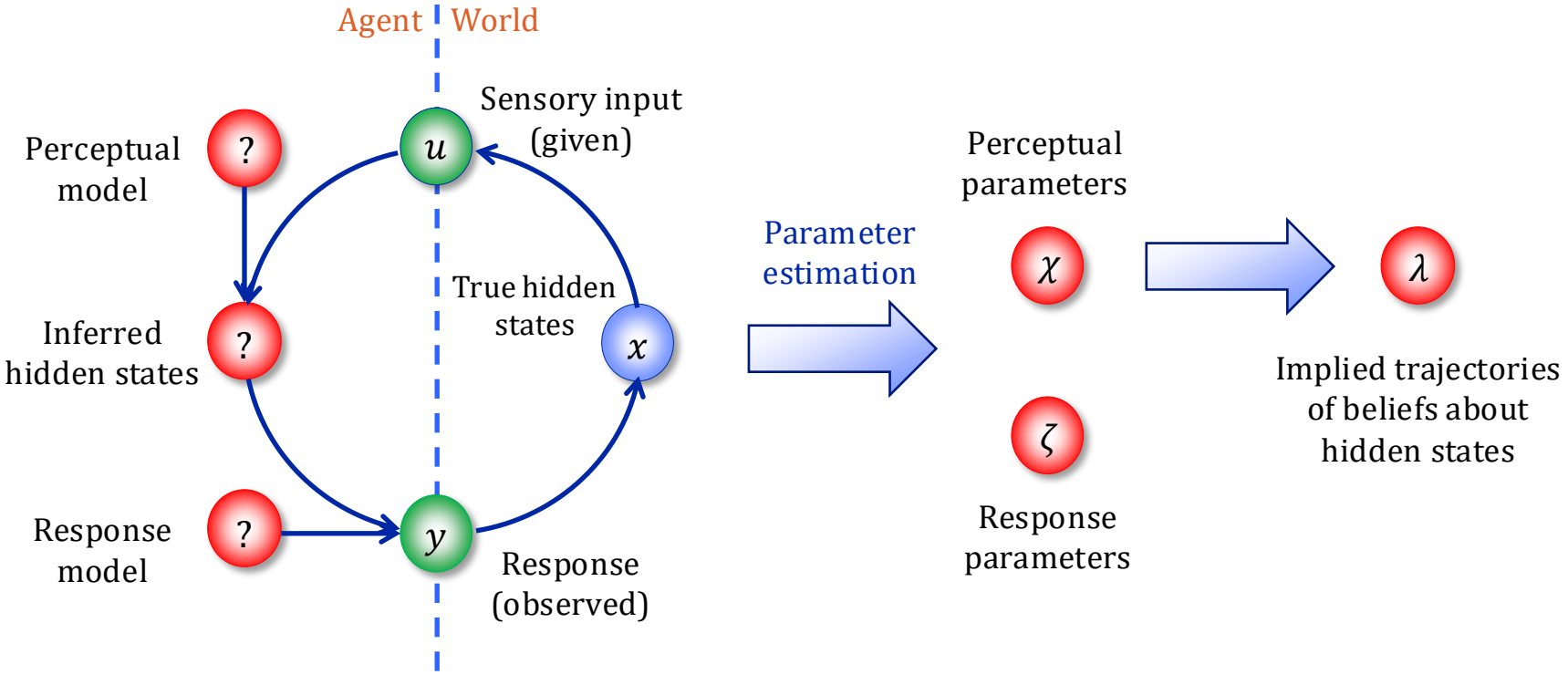
Precisions determine learning rate

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

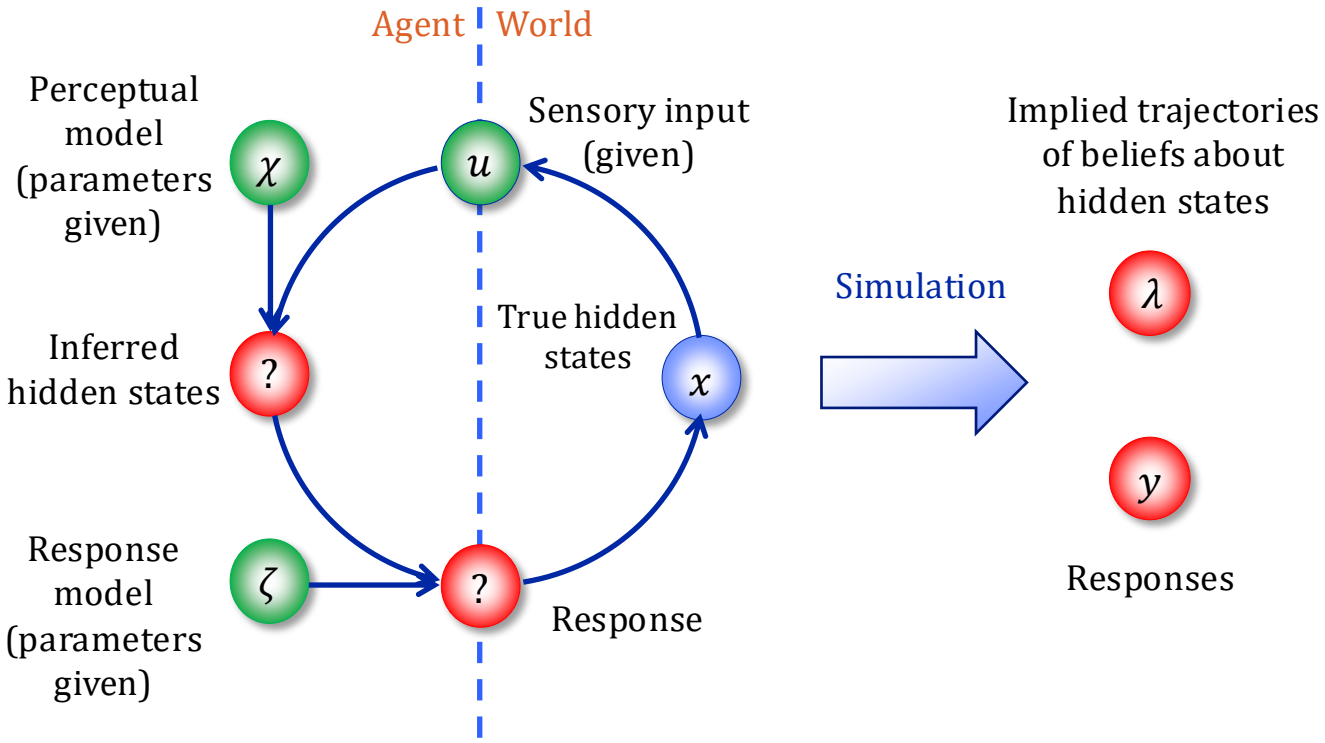
Application to experimental data



Application to experimental data: parameter estimation

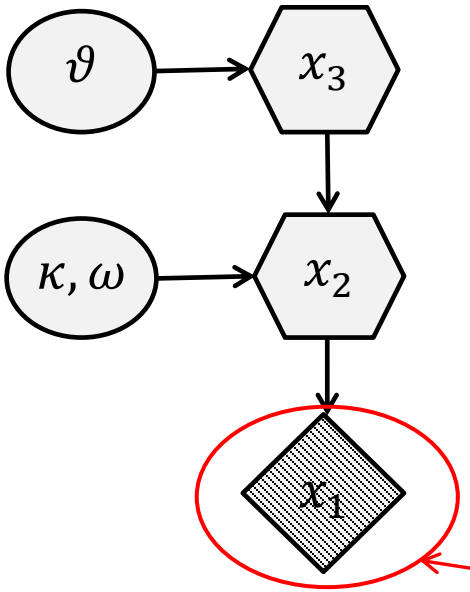


Model checking: simulation



Generative model for trial outcomes in the Iglesias et al. task:

3-level HGF for binary observations



$$x_3^{(k)} \sim \mathcal{N}\left(x_3^{(k-1)}, \vartheta\right)$$

$$x_2^{(k)} \sim \mathcal{N}\left(x_2^{(k-1)}, \exp\left(\kappa x_3^{(k)} + \omega\right)\right)$$

$$x_1^{(k)} \sim \text{Bern}\left(s\left(x_2^{(k)}\right)\right)$$

Outcome: 1 if high/low tone
followed by face/house, 0 otherwise

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

Updates at the first 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):

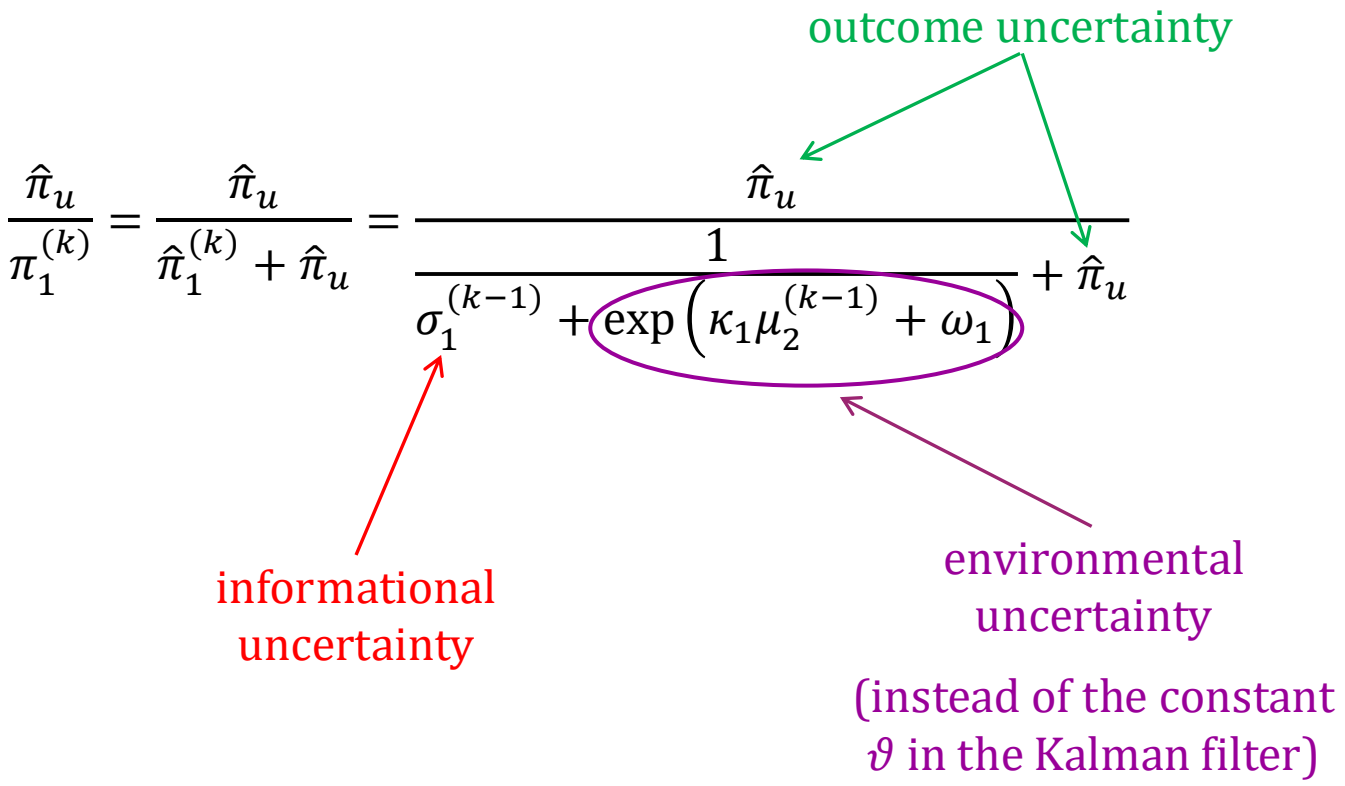
$$\pi_1^{(k)} = \hat{\pi}_1^{(k)} + \hat{\pi}_u$$

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

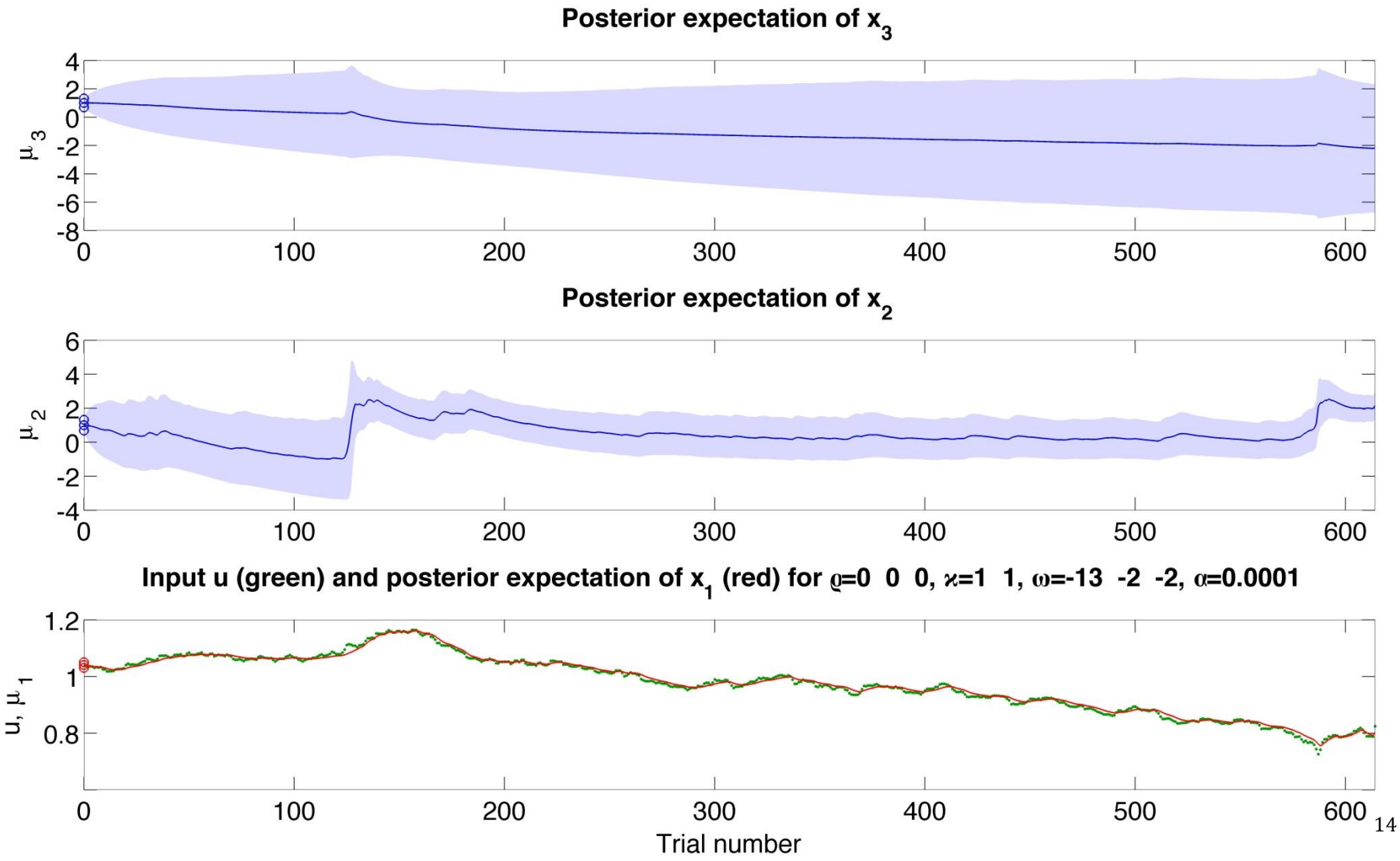
The familiar structure again – but now with a learning rate that is responsive to all kinds of uncertainty, including environmental (unexpected) uncertainty.

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

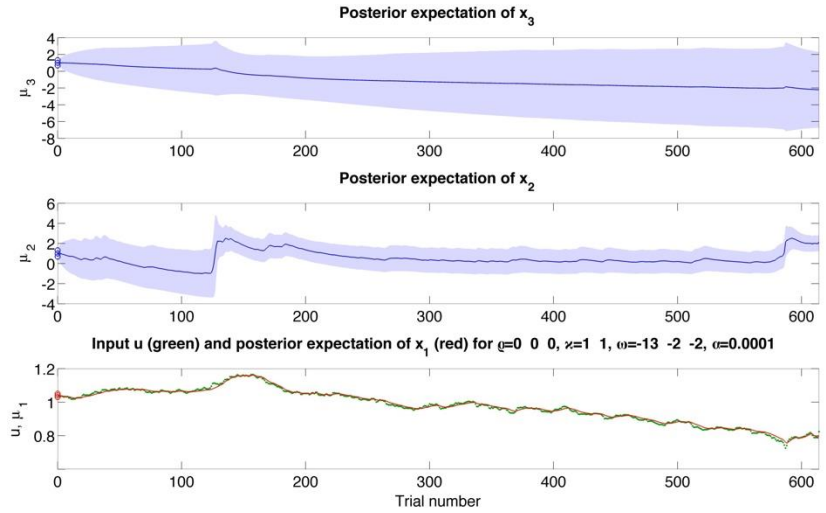
Unpacking the learning rate, we see:

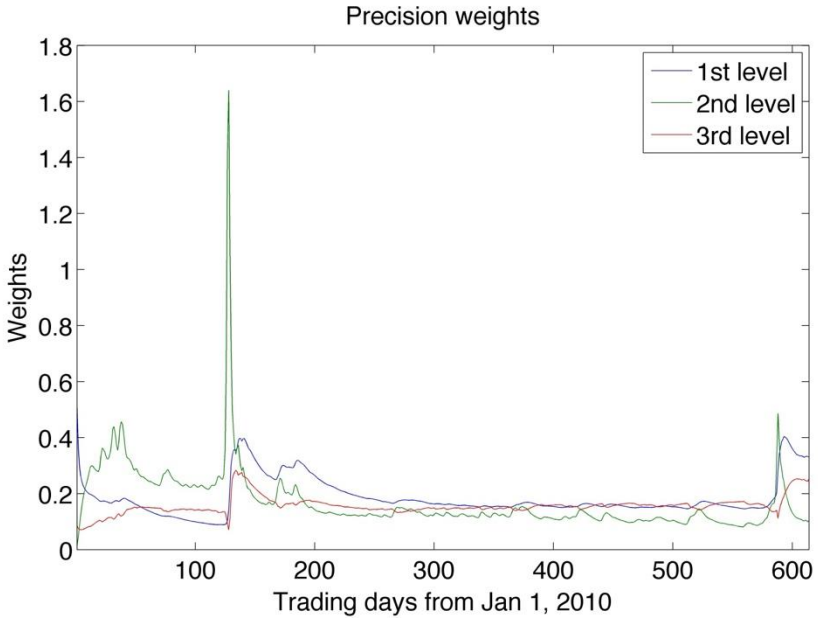


3-level HGF for continuous observations



Example of precision weight trajectory





Generative model, perceptual model, and decision model

- The *generative model* model describes (probabilistically) how the states x_i evolve in time: $x_i^{(k-1)} \longrightarrow x_i^{(k)}$
- Variationally inverting the generative gives us the *perceptual model* (also called *inference model* or *recognition model)*
- The perceptual model describes (deterministically, via update equations) how *beliefs* $\{\mu_i, \pi_i\}$ about states evolve in time: $\left\{\mu_i^{(k-1)}, \pi_i^{(k-1)}\right\} \longrightarrow \left\{\mu_i^{(k)}, \pi_i^{(k)}\right\}$
- The decision model (also called observation model or response model) describes (probabilistically) how beliefs are translated into observed actions: $\left\{\mu_i^{(k)}, \pi_i^{(k)}\right\}_{i=1,\ldots,l} \longrightarrow y^{(k)}$
- In the process of applying the HGF to data, we only need the perceptual and decision models. The generative model has already done its work: it has supplied the update equations of the perceptual model
- Both the perceptual and decision models have parameters that can be estimated individually for each dataset. Doing so requires defining priors for these parameters.

Parameter estimation with the HGF Toolbox

- Available at

  www.translationalneuromodeling.org/tapas or

  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
```

- ✨ Julia package HGF.jl will be released in September! 🎉

Parameter estimation with the HGF Toolbox

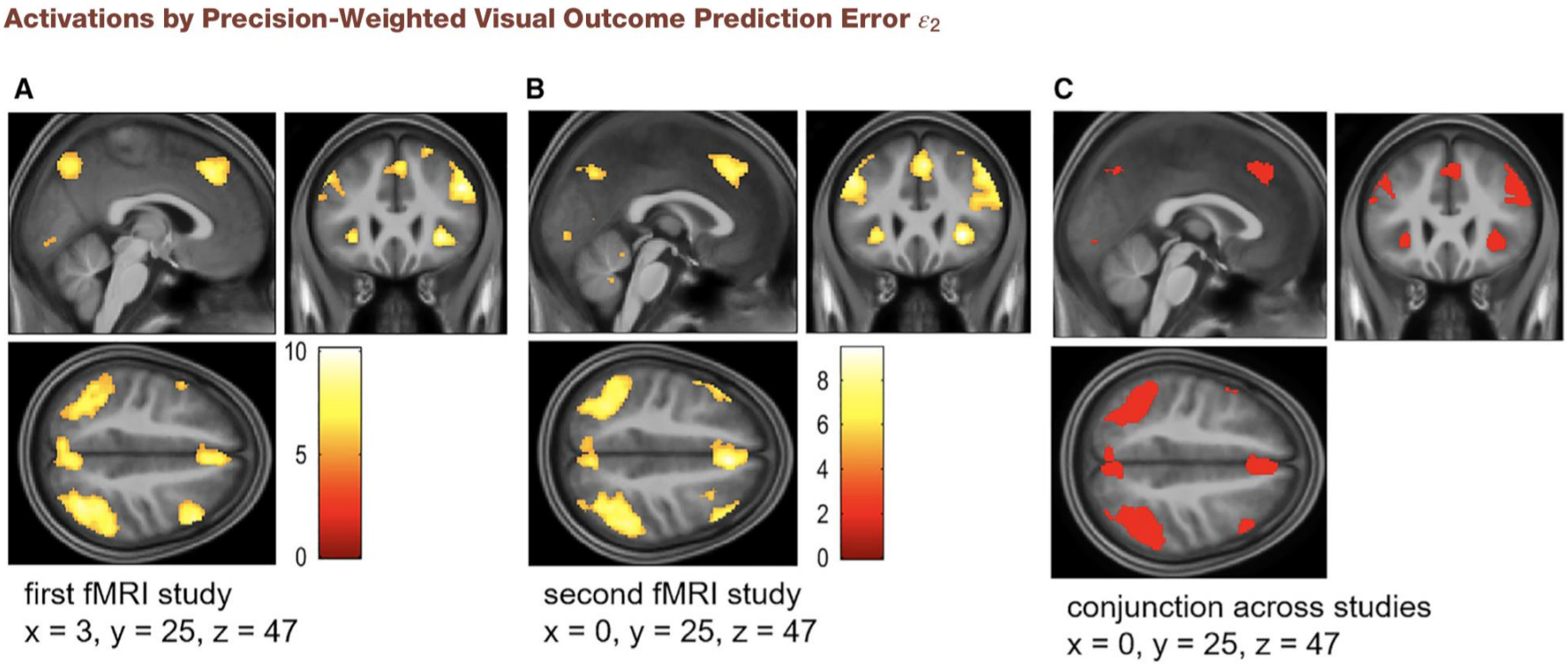


Figure 2. Whole-Brain Activations by ε_2

Activations by precision-weighted prediction errors 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 peak-level 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.

Iglesias et al, 2019 (Correction to Iglesias et al., 2013)

Parameter estimation with the HGF Toolbox

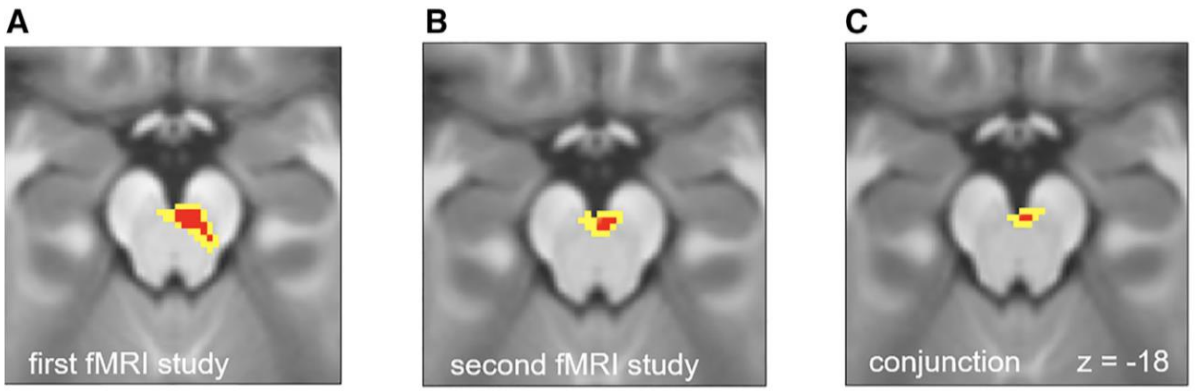


Figure 3. Midbrain Activation by ε_2

Activation of the dopaminergic VTA/SN by precision-weighted prediction errors about visual outcome, ε_2. The activation at $p < 0.05$ FWE peak-level corrected for the volume of our anatomical mask (comprising both dopaminergic and cholinergic brain structures: VTA/SN, PPT/LDT, and basal forebrain) is shown in red. The activation thresholded at $p < 0.001$ uncorrected is shown in yellow.

(A) Results from the first fMRI study. (B) Second fMRI study. (C) Conjunction (logical AND) across both studies.

Iglesias et al, 2019 (Correction to Iglesias et al., 2013)

Parameter estimation with the HGF Toolbox

Activations by Precision-Weighted Prediction Error about Stimulus Probabilities ε_3

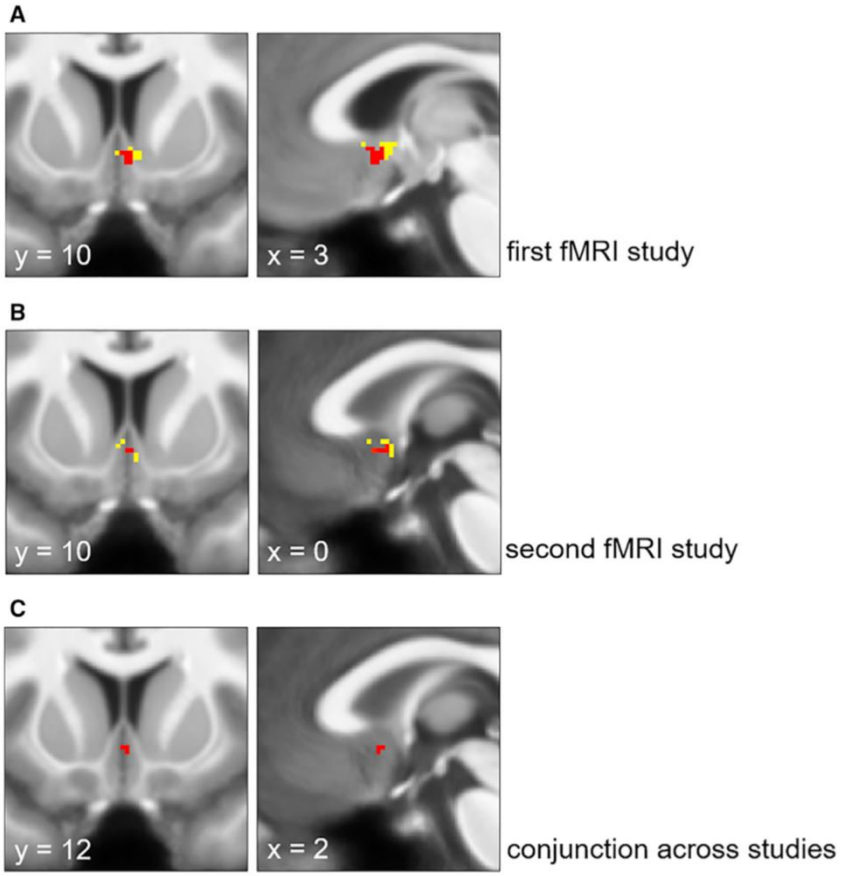


Figure 6. Basal Forebrain Activations by ε_3

Activation of the basal forebrain by 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 peak-level corrected for the entire anatomical mask (red) on the results thresholded at $p < 0.001$ (yellow; the yellow cluster also survives $p < 0.05$ FWE cluster-level correction for the entire anatomical mask). The anatomical mask comprised both dopaminergic and cholinergic brain structures: VTA/SN, PPT/LDT, and basal forebrain. (A) and (B) show results from the first (A: local maximum at x = 4, y = 12, z = –11, $t = 4.71$) and the second fMRI study (B: local maximum at x = 0, y = 10, z = –8, $t = 5.09$). (C) shows the conjunction analysis ("logical AND") across both studies. To ease visual comparison with Iglesias et al. (2013), the figure sections (x and y coordinates are indicated on each panel) are not located at the local maxima but correspond closely to those in Iglesias et al. (2013).

Iglesias et al, 2019
(Correction to Iglesias et al., 2013)

Parameter recovery

- *Parameter recovery* means the ability to estimate the 'ground truth' parameter values put into a simulation
- This is never possible to do exactly because simulation is stochastic
- The stochastic nature of the simulation means that the parameter value best able to explain the simulated data is not the same as that used in the simulation. By nature, the simulation only noisily represents the 'ground truth' parameter values.
- It is therefore somewhat misleading to call the value used in the simulation 'true'. One could just as well argue that the estimated value is the 'true' one since it best explains the data.
- In order to get an idea of how exactly a parameter can be recovered, run many (on the order of tens to hundreds) simulations and look at the distribution of parameter estimates.
- **Estimating a parameter that isn't well recovered from simulations can still make sense because adjusting its value can improve the model by leading to a better explanation of the data**

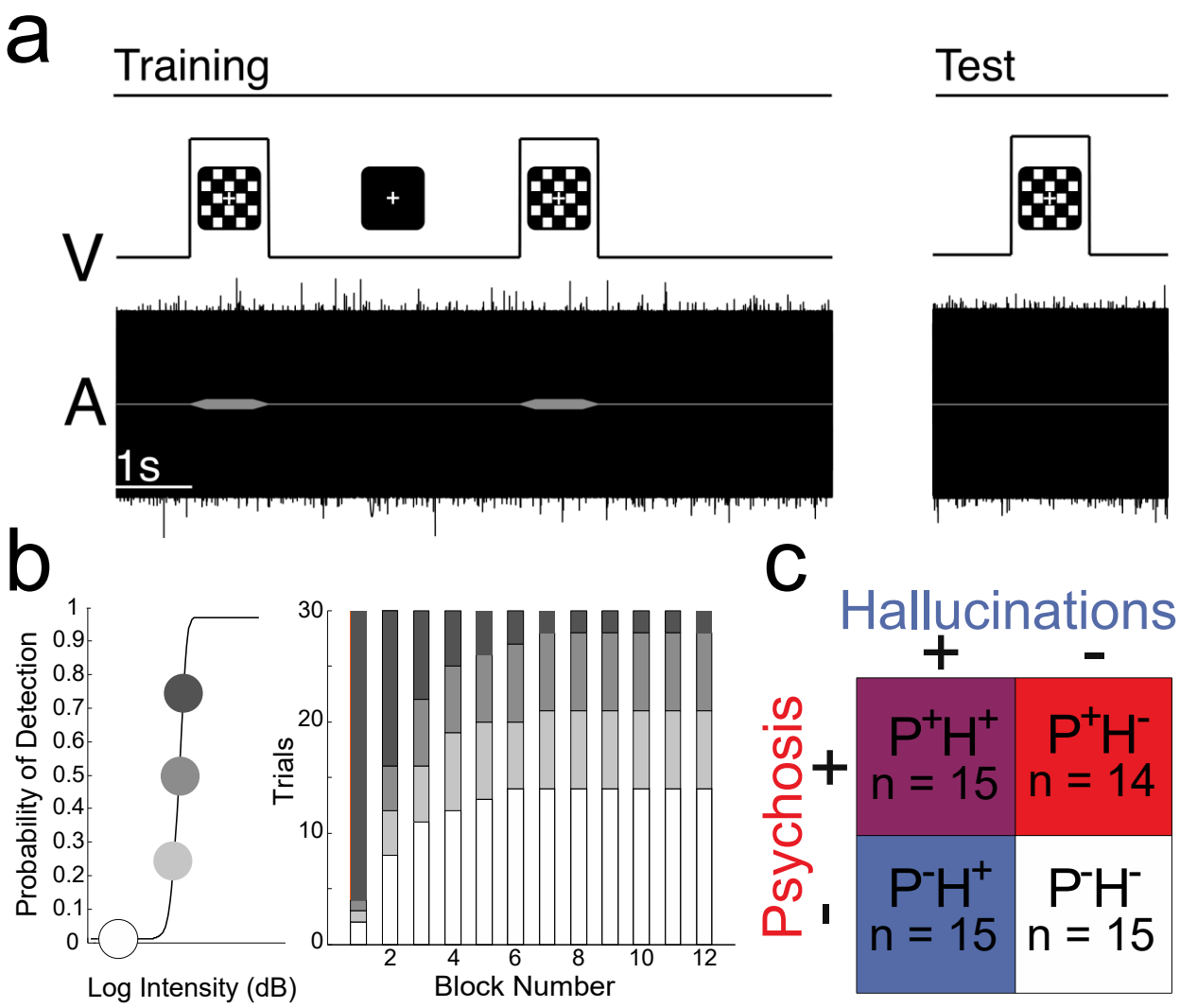
Model comparison

- There is a range of scores that help in choosing a well-performing model: AIC (Akaike information criterion), BIC (Bayesian information criterion), Bayes factors, LME (log-model evidence), free energy, etc.
- Each model gets a particular score (which is on its own uninterpretable!)
- The difference in score between models is what counts
- However, model selection is not straightforward. AIC and BIC penalize complexity based on simple heuristics, which may not reflect complexity accurately. LME is better on that count, but is very sensitive to the modeller's choice of priors.
- **Three important considerations:**

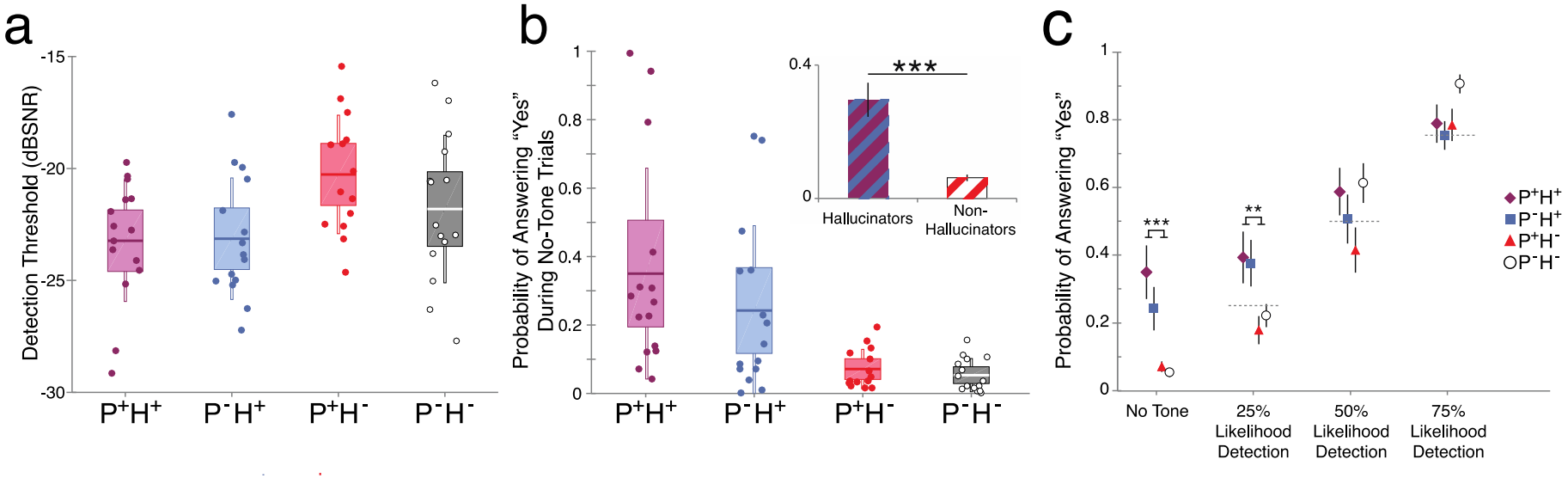
  1. **Does the model allow me to answer my question of interest?**
  2. **Does the *prior predictive* distribution of observations make sense?**
  3. **Does the *posterior predictive* distribution of observations make sense?**

  When the answer to all three is yes, the model is fine.

Conditioned hallucinations (Powers, Mathys, & Corlett, Science, 2017)



Conditioned hallucinations



Conditioned hallucinations

- Belief/percept formation model specifically created for this task
- Probability that the subject will respond "yes" to detection on a given trial:

$$P(\text{"yes"}|belief) = sigmoid(belief)$$

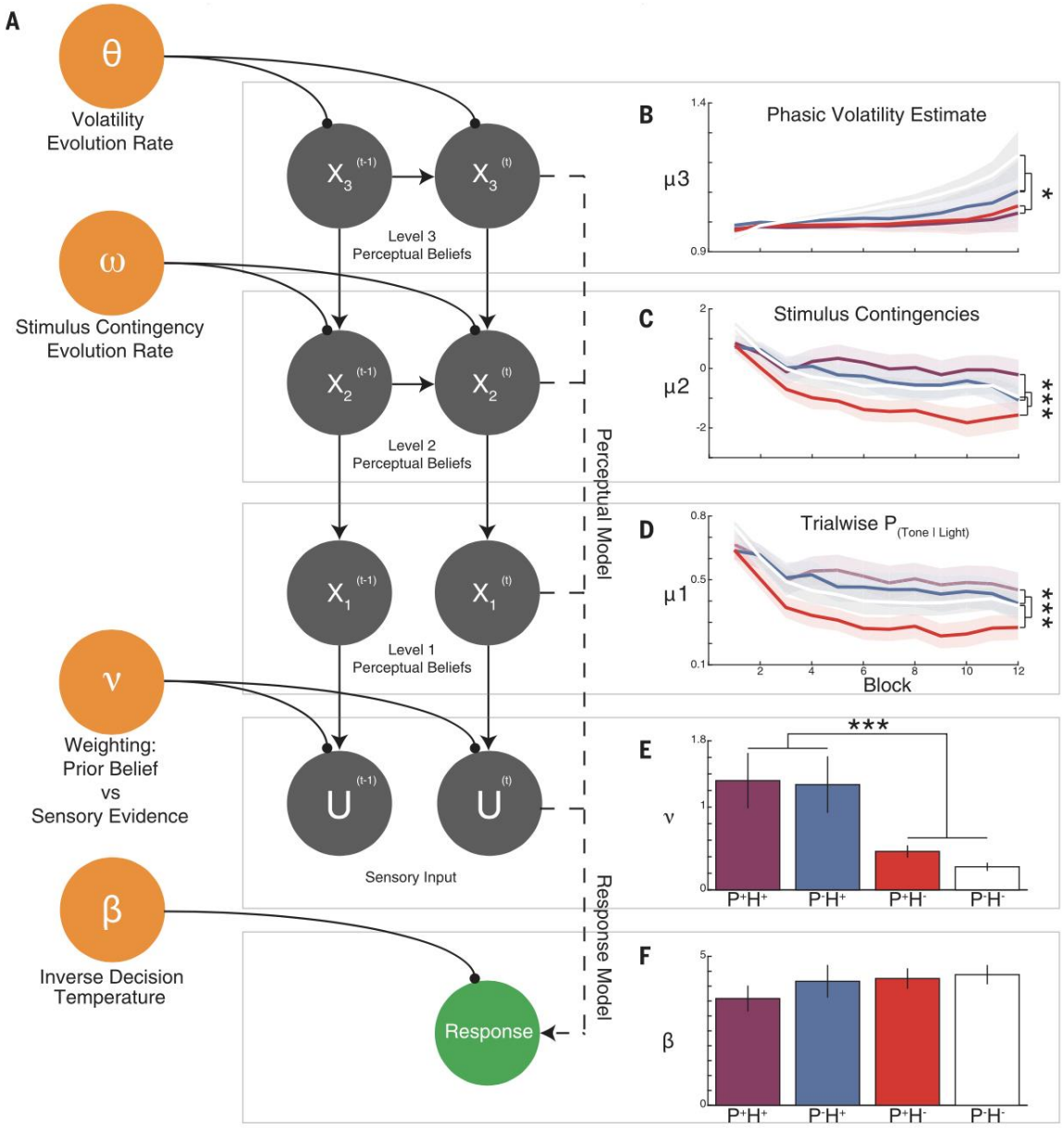
- $belief$ is formalized as the Bayesian posterior mean of a beta distribution

$$belief = prior + \frac{1}{1+\nu}(input - prior)$$

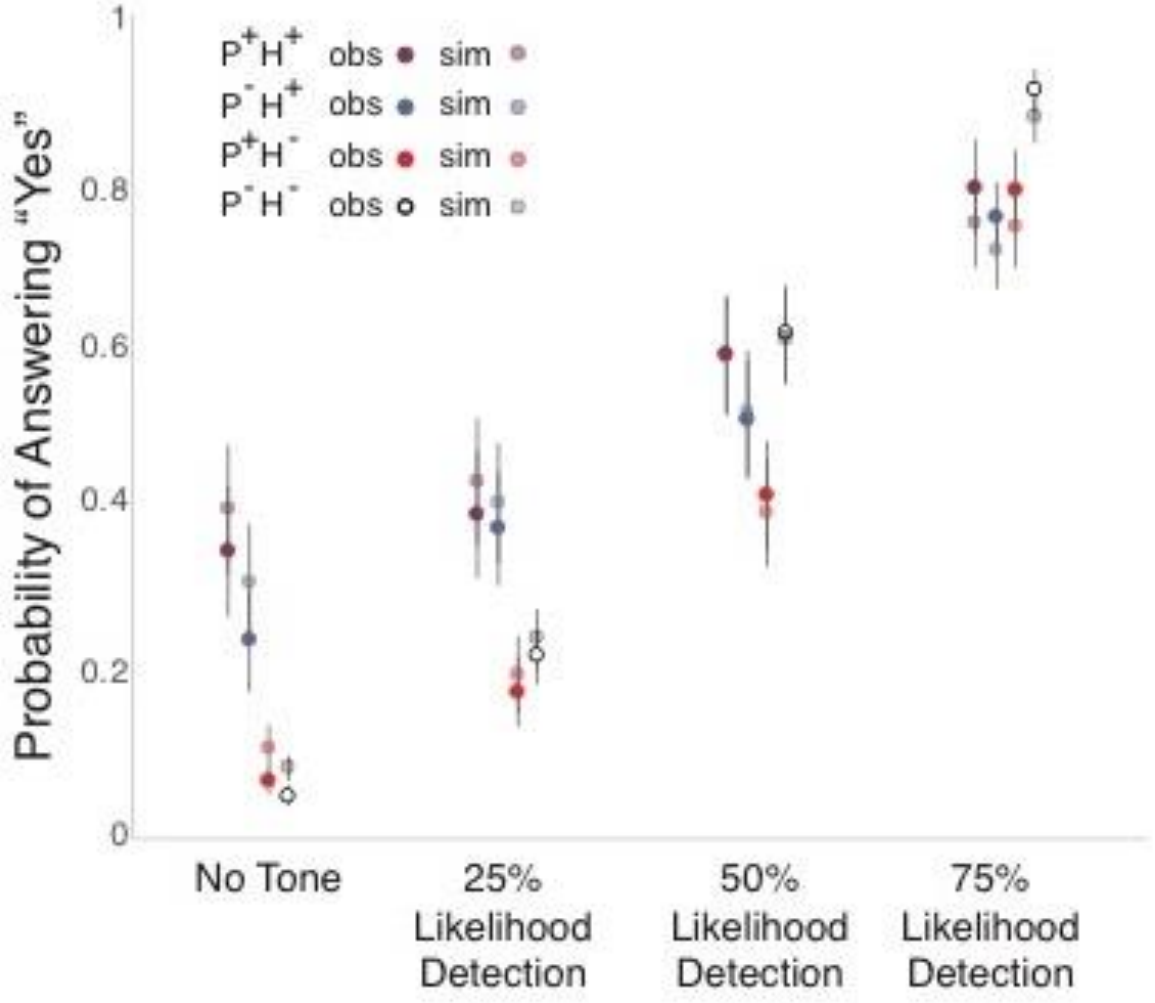
where

- $input$ is given by experimental design: the true positive rate of the tone presented without light at each trial: 25%, 50%, or 75%
- $prior$ is the prior from learning using the HGF: $\hat{\mu}_1$
- ν is a subject-specific parameter indicating the relative weight of the prior compared to the input.
- For $\nu = 1$, prior and input have equal weight; for $\nu > 1$ the prior has more weight than the input; and for $\nu < 1$ the input has more weight than the prior.

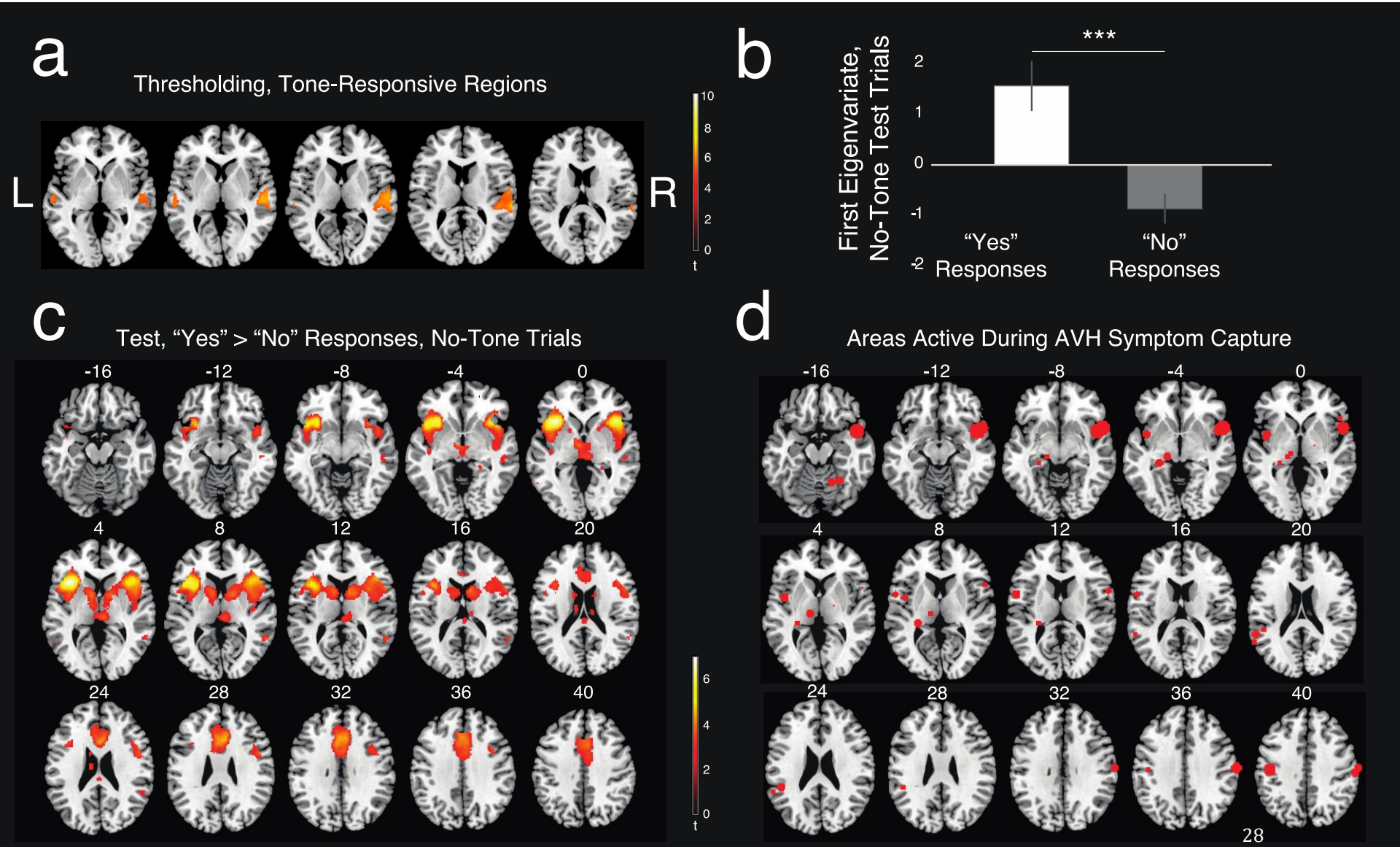
Conditioned hallucinations



Conditioned hallucinations



Conditioned hallucinations



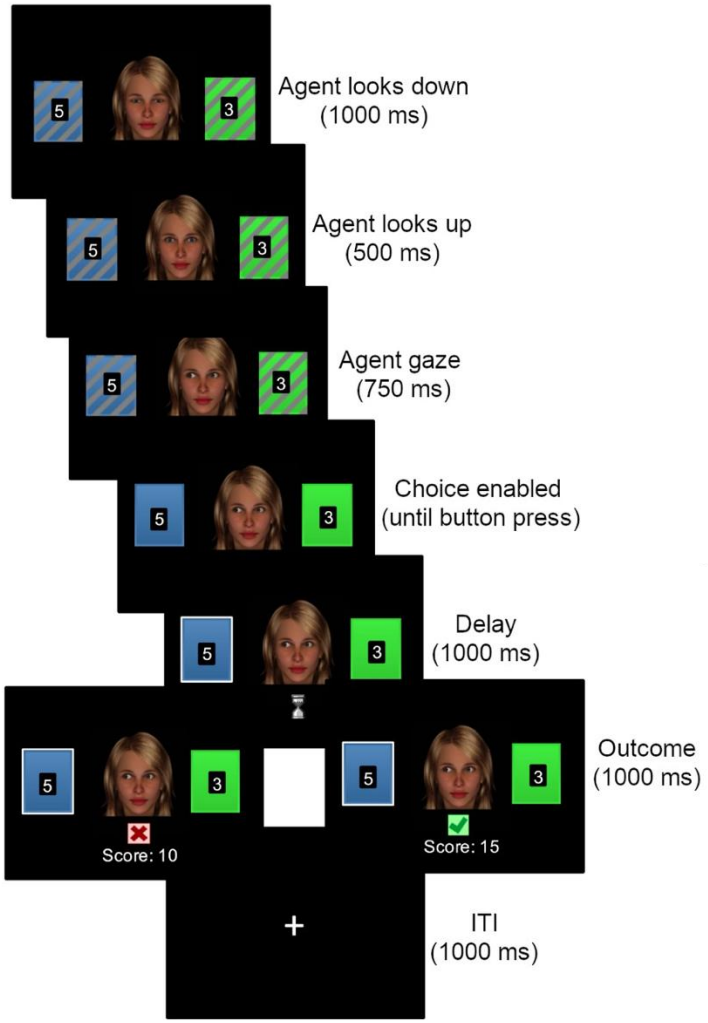
PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder

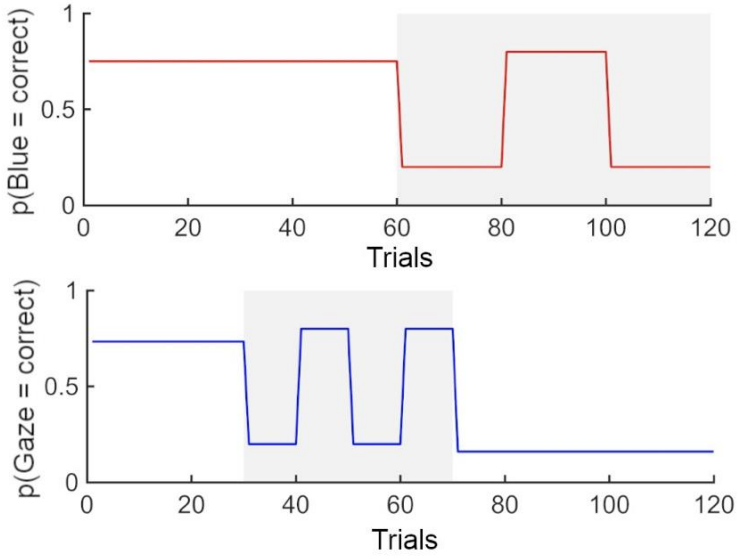
Lara Henco[1,2]*, Andreea O. Diaconescu[3,4,5], Juha M. Lahnakoski[1,6,7], Marie-Luise Brandi[1], Sophia Hörmann[1], Johannes Hennings[8], Alkomiet Hasan[9,10], Irina Papazova[9], Wolfgang Strube[9], Dimitris Bolis[1,11], Leonhard Schilbach[1,2,11,12☯], Christoph Mathys[4,13,14☯]

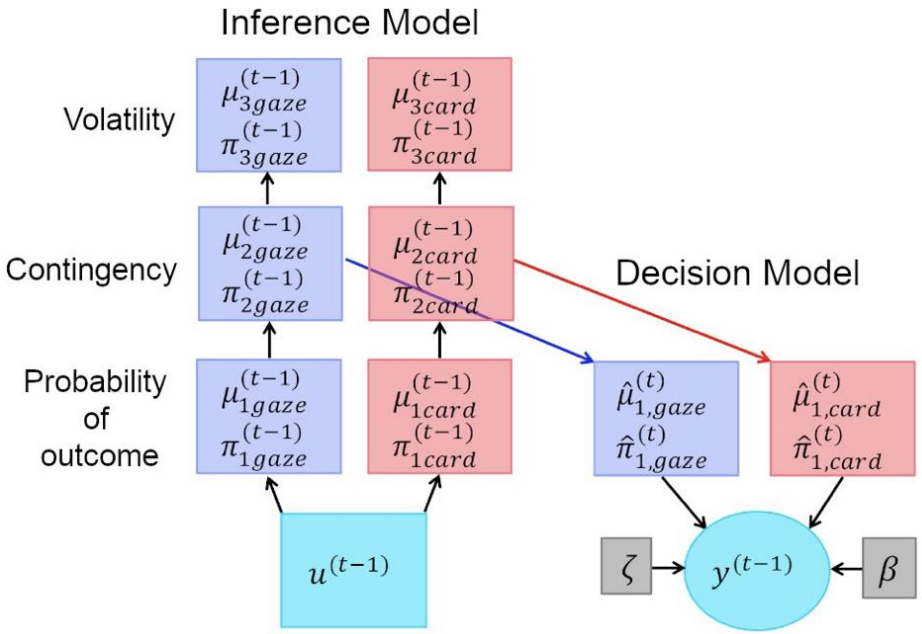
Social inference in borderline personality disorder, schizophrenia, and major depression.



Henco et al, PLOS Comput Biol, 2020

Social inference in borderline personality disorder, schizophrenia, and major depression.





Henco et al, PLOS Comput Biol, 2020

Social inference in borderline personality disorder, schizophrenia, and major depression.

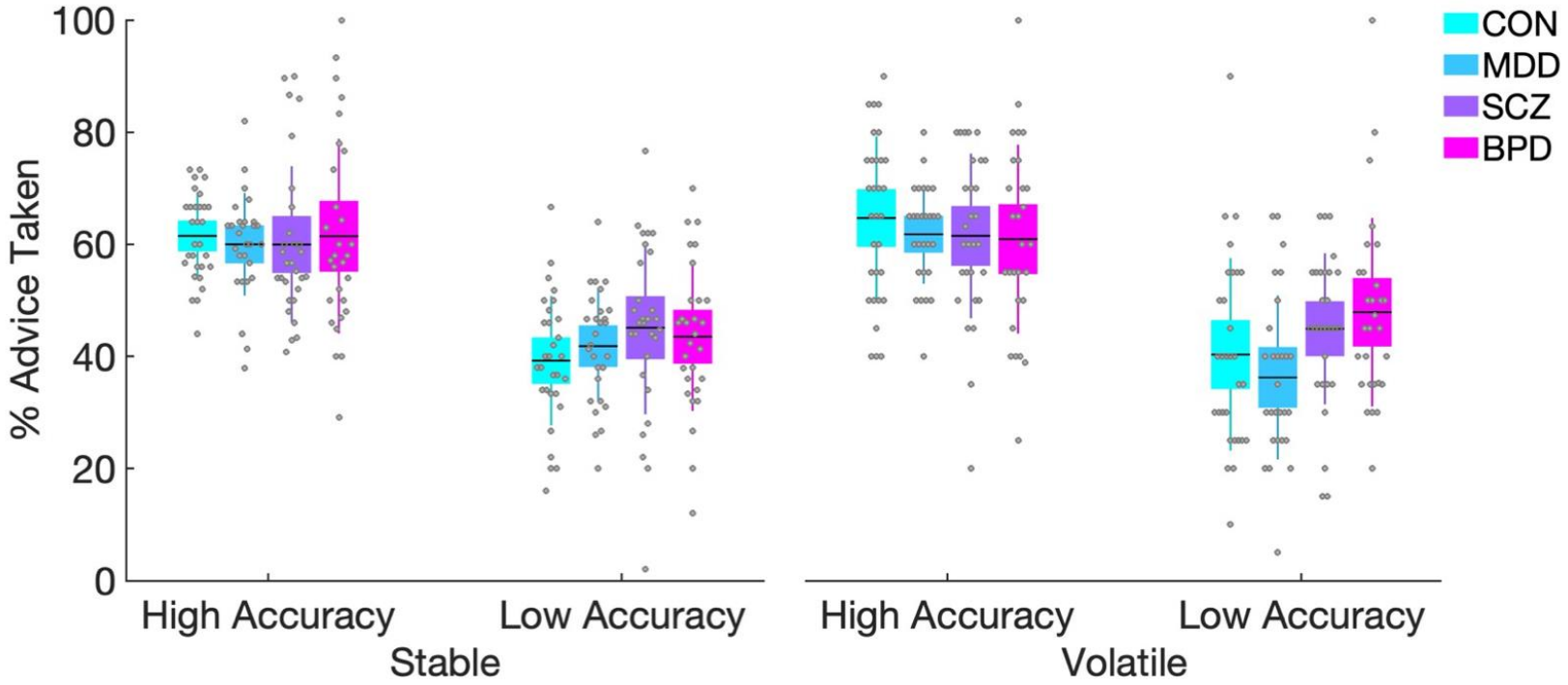


Fig 3. Behavioral results with regard to advice taking. All participants followed the advice significantly more during phases of high compared to low accuracy. The descriptive data shows a trend of BPD patients to follow the advice more during volatile phases of low accuracy (right most plot) but the interaction was not significant. Means are plotted with boxes marking 95% confidence intervals and vertical lines showing standard deviations.

Henco et al, PLOS Comput Biol, 2020

Social inference in borderline personality disorder, schizophrenia, and major depression.

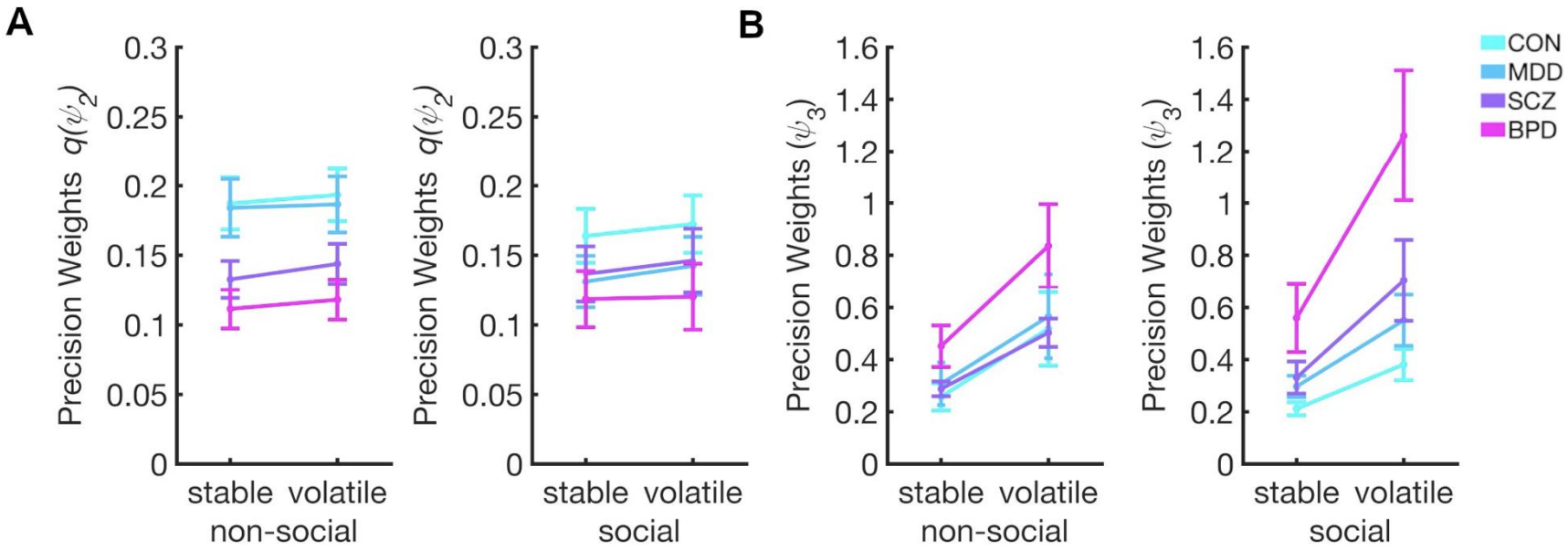


Fig 4. Results for mixed ANOVA using precision weights for updating beliefs about social and non-social contingency and volatility. (A) Precision weights $q(\psi_2)$ and (B) precision weights ψ_3. Overall, $q(\psi_2)$ and ψ_3 increase when transitioning from stable to volatile phase. Patients with BPD show reduced overall $q(\psi_2)$. At same time, patients with BPD show higher ψ_3 compared to the other groups and a more pronounced increase in response to volatility. Bars indicate SEM. See also S1 Fig, S7 Table and S8 Table for all results.

Henco et al, PLOS Comput Biol, 2020

Henco et al. (2020): Summary

- Patients diagnosed with major depressive disorder (MDD, N = 29), schizophrenia (SCZ, N = 31), and borderline personality disorder (BPD, N = 31) as well as healthy controls (CTL, N = 34) performed a probabilistic reward learning task in which participants could learn from social and non-social information.
- SCZ and BPD performed more poorly on the task than CTL and MDD
- BPD performed better in the social compared to the non-social domain. In contrast, CTL and MDD showed the opposite pattern, and SCZ showed no difference between domains.
- In effect, BPD gave up a possible overall performance advantage by concentrating their learning in the social at the expense of the non-social domain.
- BPD showed slower learning from social and non-social information and an exaggerated sensitivity to changes in environmental volatility.
- Compared to CTL and MDD, BPD and SCZ showed a stronger reliance on social relative to non-social information when making choices.

HGF and active inference

Inferring in Circles: Active Inference in Continuous State Space Using Hierarchical Gaussian Filtering of Sufficient Statistics

Peter Thestrup Waade[1,2(✉)], Nace Mikus[1,3], and Christoph Mathys[1,4,5]

https://doi.org/10.1007/978-3-030-93736-2_57

https://github.com/ilabcode/hgf-active-inference

HGF and active inference

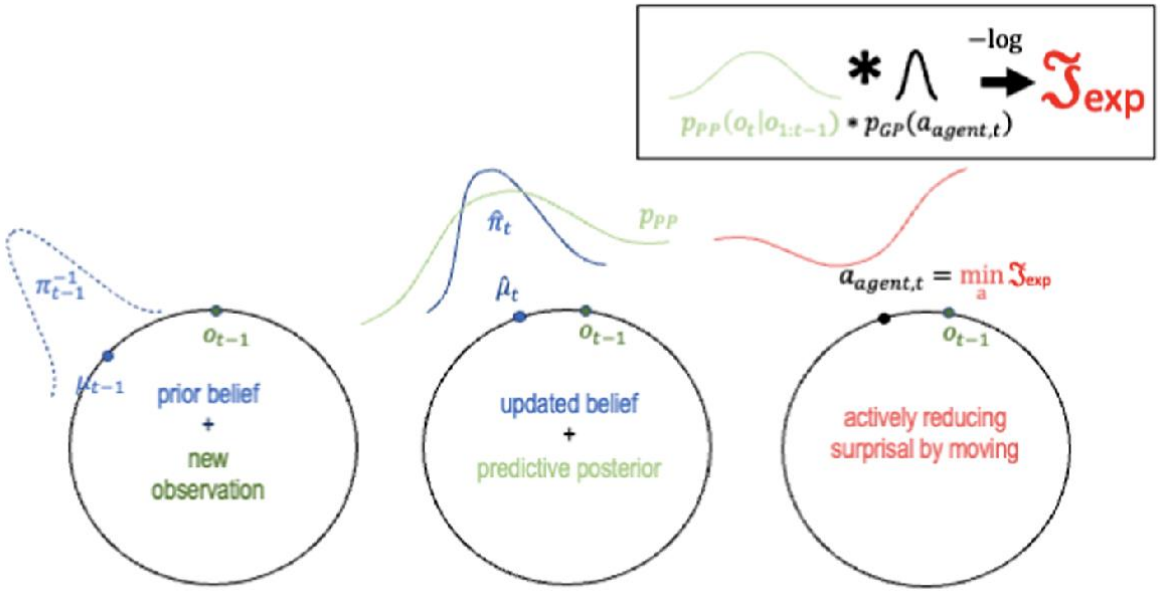
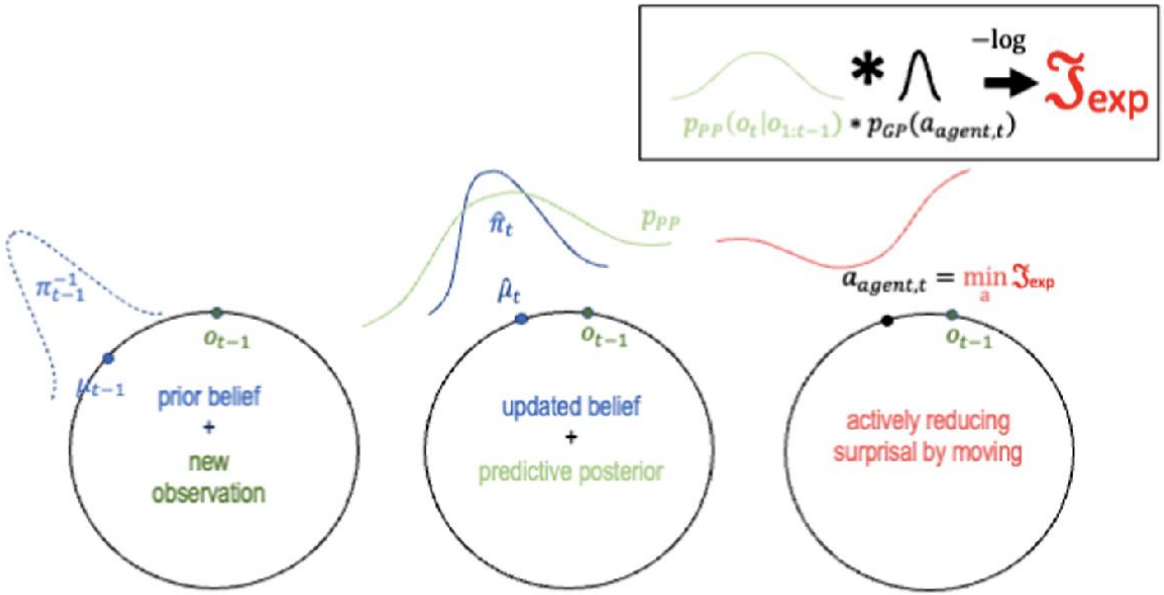
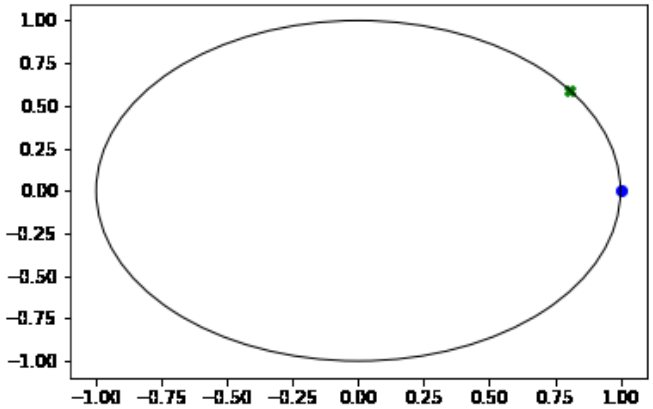


Fig. 1. Graphical sketch of the agent's action process, visualized on the circle. The agent starts with a Gaussian prior belief about the target's position with mean μ_{t-1} and precision π_{t-1}, and makes a new observation o_t. From that a new belief is computed with the HGF (also taking into consideration the drift ρ), and a predictive posterior t-distribution p_{PP} can be calculated. Finally, the predictive posterior is convolved with the 'goal posterior' $p_{GP}(a_{agent,t})$ i.e. the goal prior over agent positions given that the target is observed at its expectation (see Eq. 9). The negative log of the resulting probability distribution is the expected surprisal associated with the agent moving to different positions, of which the lowest is selected. If the static goal prior $p_{GP}(o_1 - o_{agent})$ is symmetric and centered around 0, μ_t, the mean of p_{PP} and the agent's action $a_{agent,t}$ coincide.

HGF and active inference





HGF and active inference

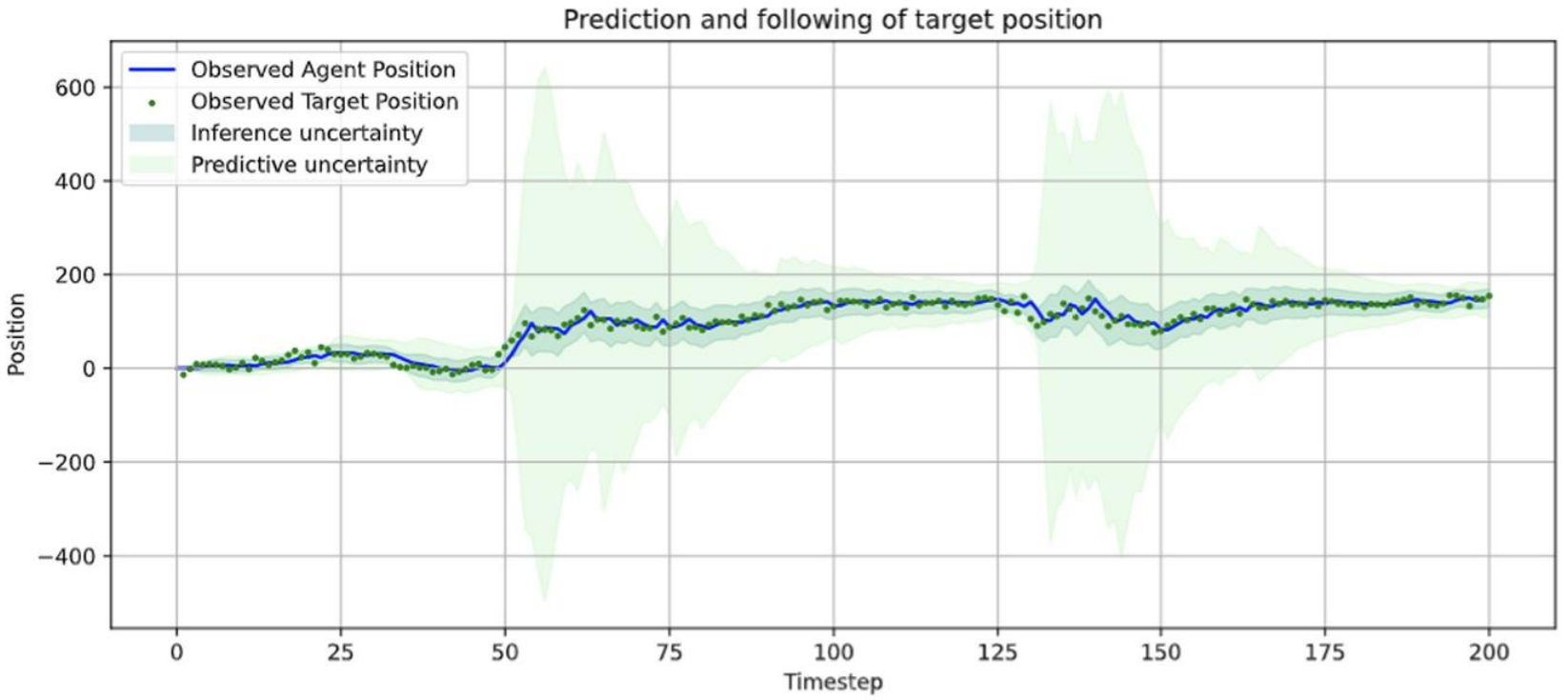


Fig. 2. Relative positions of the agent and the target. Inner shaded area is the inverse precision of the inference of the target position. Outer shaded area is the 68% confidence interval of the predictive posterior

HGF and active inference

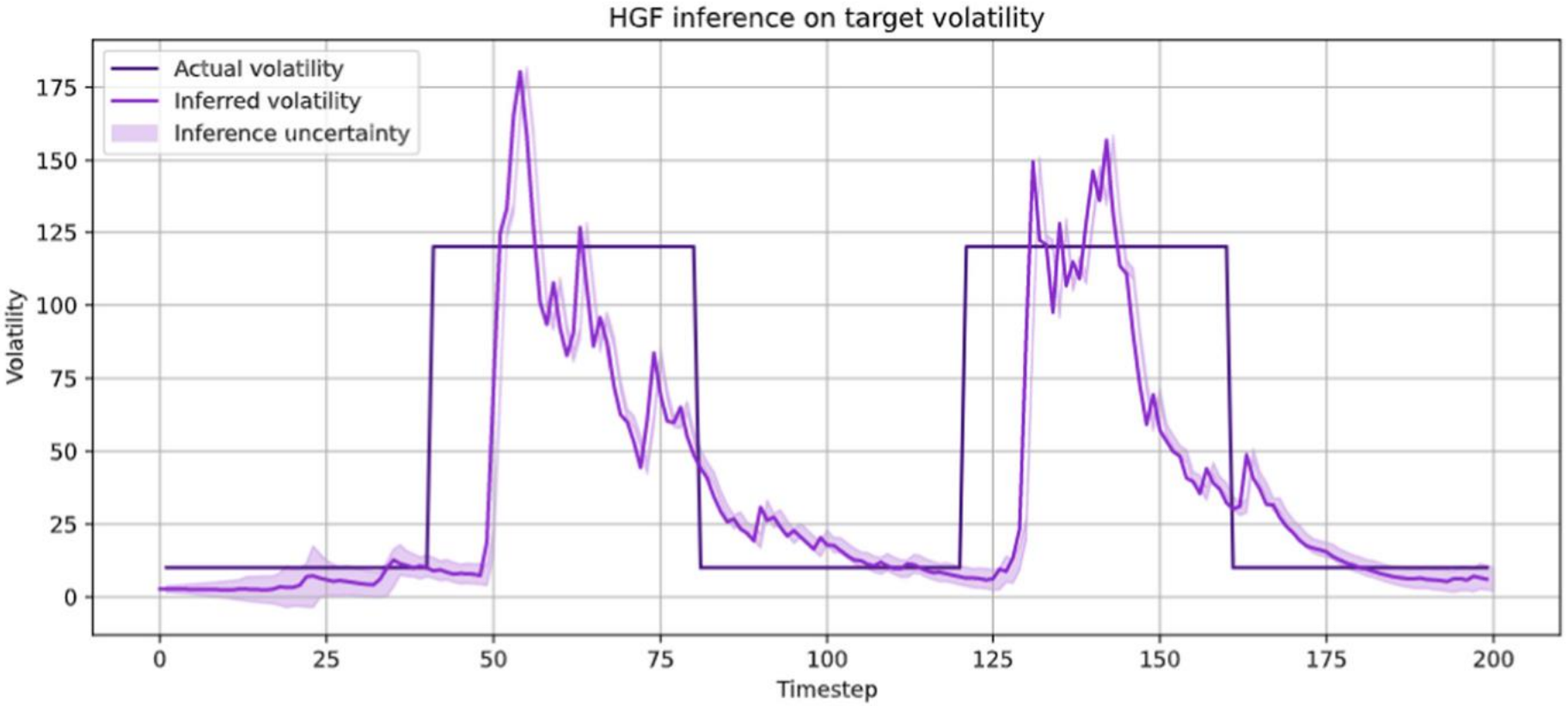


Fig. 3. The agent's HGF-based inference of the volatility (standard deviation) of the target's Gaussian random walk. Shaded area is the inverse precision of the inference.

Outlook:
Predictive coding in cortex with generalized HGF networks

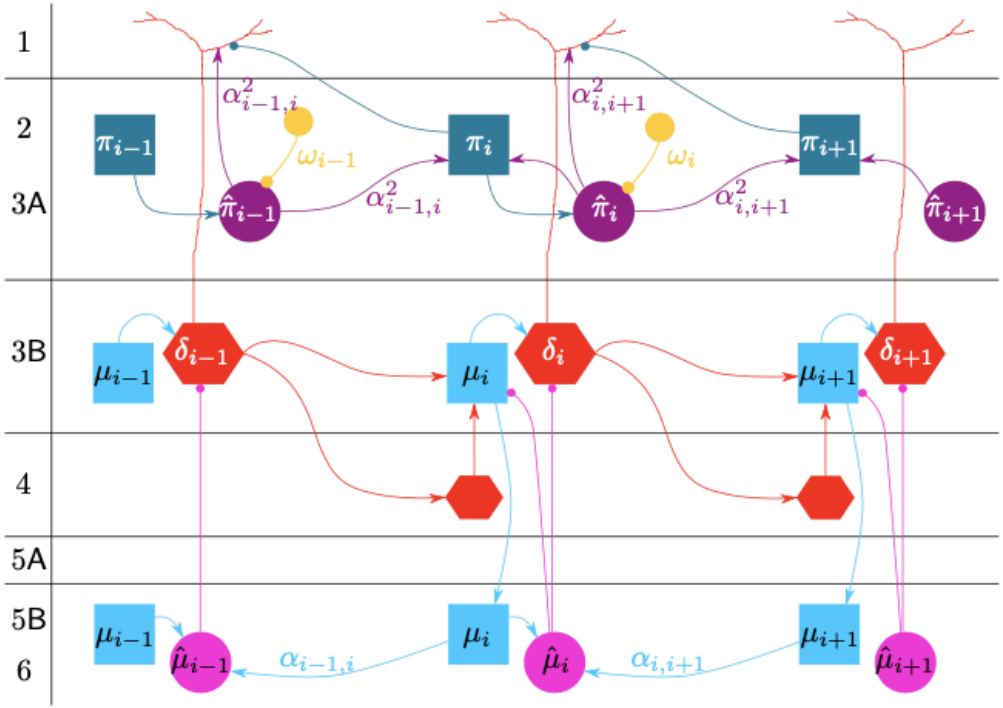


Figure 4: Overview of possible layer-specific message passing in VAPE coupling. Assignment of nodes to layers is loosely based on figure 3 in [4], reproduced in this document in figure 3. Connections with arrowheads indicate positive influences, or excitatory connections, while connections with circular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

Weber et al, in preparation

Thanks