Computational and algorithmic approaches to understanding human minds:

Recent achievements and future goals

Christoph Mathys, PhD

Associate Professor of Cognitive Science

Aarhus University

Hearing concerning the chair „Interdisciplinary Digital Science for Healthy Human Development"

TU Dresden, 20 September 2022

Research portfolio

- My background
- Prediction updates in adaptive interconnected systems: hierarchical Gaussian filtering (HGF)
- Principles of HGF
- Experimental applications
- When pattern recognition goes wrong: delusional inference
- A generative framework for the study of delusions

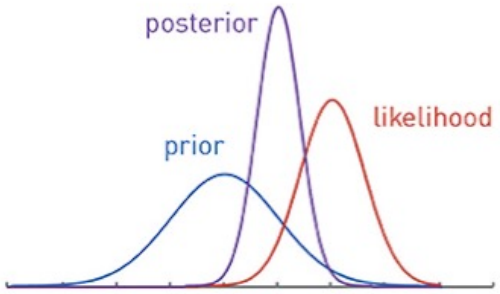
My background

- MSc in Interdisciplinary Sciences from ETH Zurich (major: theoretical physics)
- 4 years at a tech start-up sold to a major telecoms operator
- MSc in Psychology (major: neuropsychology) and Psychopathology from the University of Zurich
- Visiting researcher at Harvard Medical School
- Clinical psychologist intern at Charité Berlin psychiatry
- PhD in Translational Neuromodelling from the Department of Information Technology and Electrical Engineering (D-ITET) of ETH Zurich
- 4-year postdoctoral grant from the Max Planck Society
- Tenure-track position at SISSA, Trieste; tenure granted after 3 years
- Currently Associate Professor (tenured) at Aarhus University in Denmark

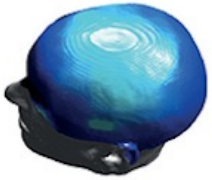
Algorithmic approaches to understanding inferential systems

Bayes' Theorem

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



Generative models as computational assays



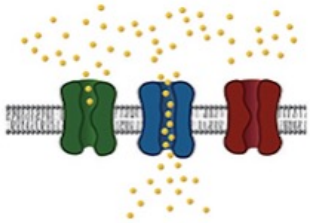
measurement y

forward model

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

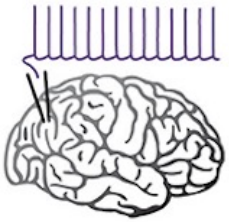
$p(x|y,m)$

inference



hidden system state x

Perception as the inversion of a generative model



neuronal activity y

forward model

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

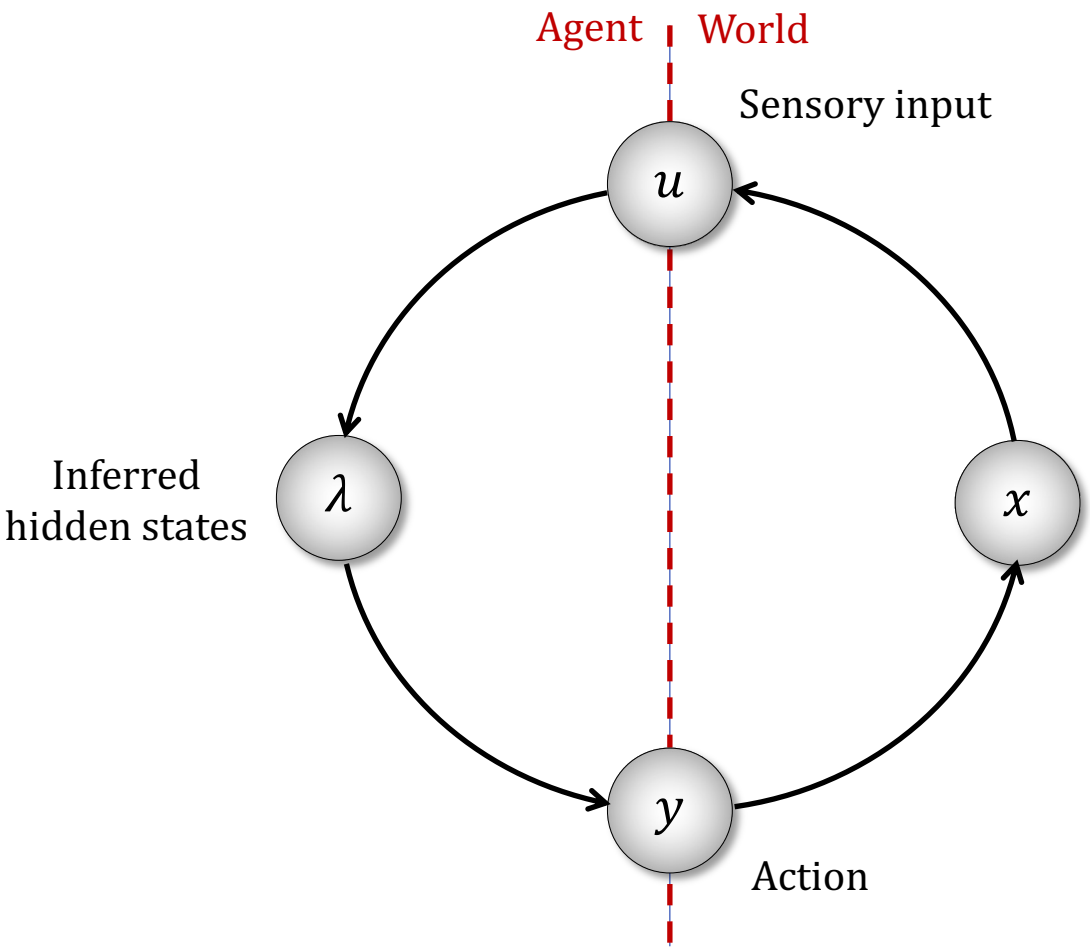
$p(x|y,m)$

perception

environmental state x

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

The next level: systems interacting with systems



The environment changes all the time: our agent is a system that interacts with other systems

Filtering is a challenge in dynamic environments

The fact that the environment is dynamic poses many challenges. I will focus on a particular one here: **filtering good information from noisy sequential observations.**

Put another way, the question is: how can we take into account that **old information becomes obsolete?**

If we do not, our learning rate becomes smaller and smaller

On the other hand, **we cannot simply rely on our latest observation** because observations are noisy and therefore using many of them allows for better predictions.

The challenge is that we need to give each observation just the right weight, and more recent observations need to count more than older ones.

The Kalman filter

- Two hierarchical equations – state equation and observation equation:

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

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

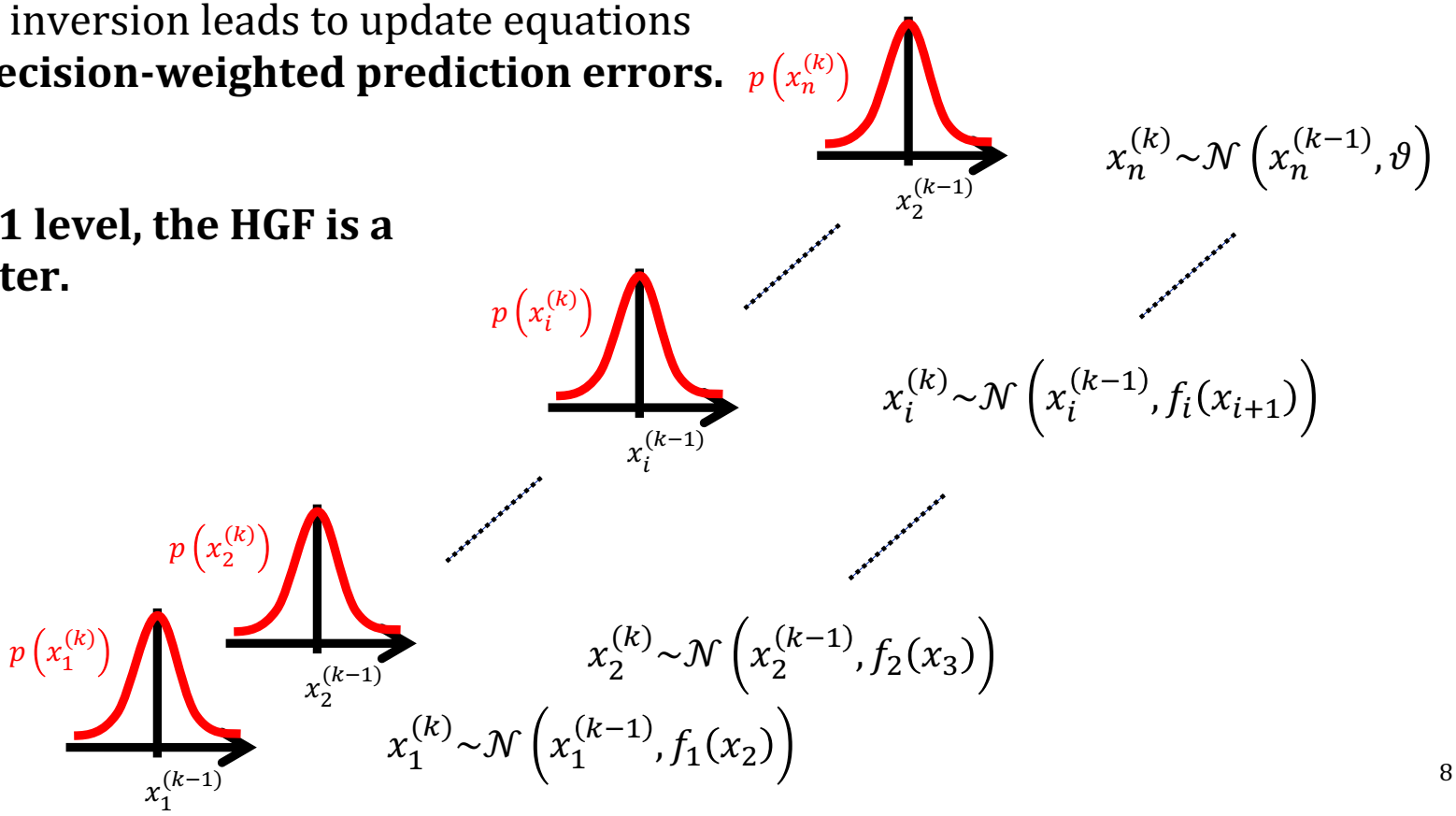
- The Kalman filter update equations can be derived exactly in closed form.
- They amount to adding a **precision-weighted prediction error** to the current prediction.
- The Kalman filter is **optimal for linear dynamic systems**.
- Unfortunately, except for simple physical systems, **the world is not linear**. Living organisms (and robots!) need to be able to filter **inputs whose rate of change changes**

A more adaptive filter: 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.**

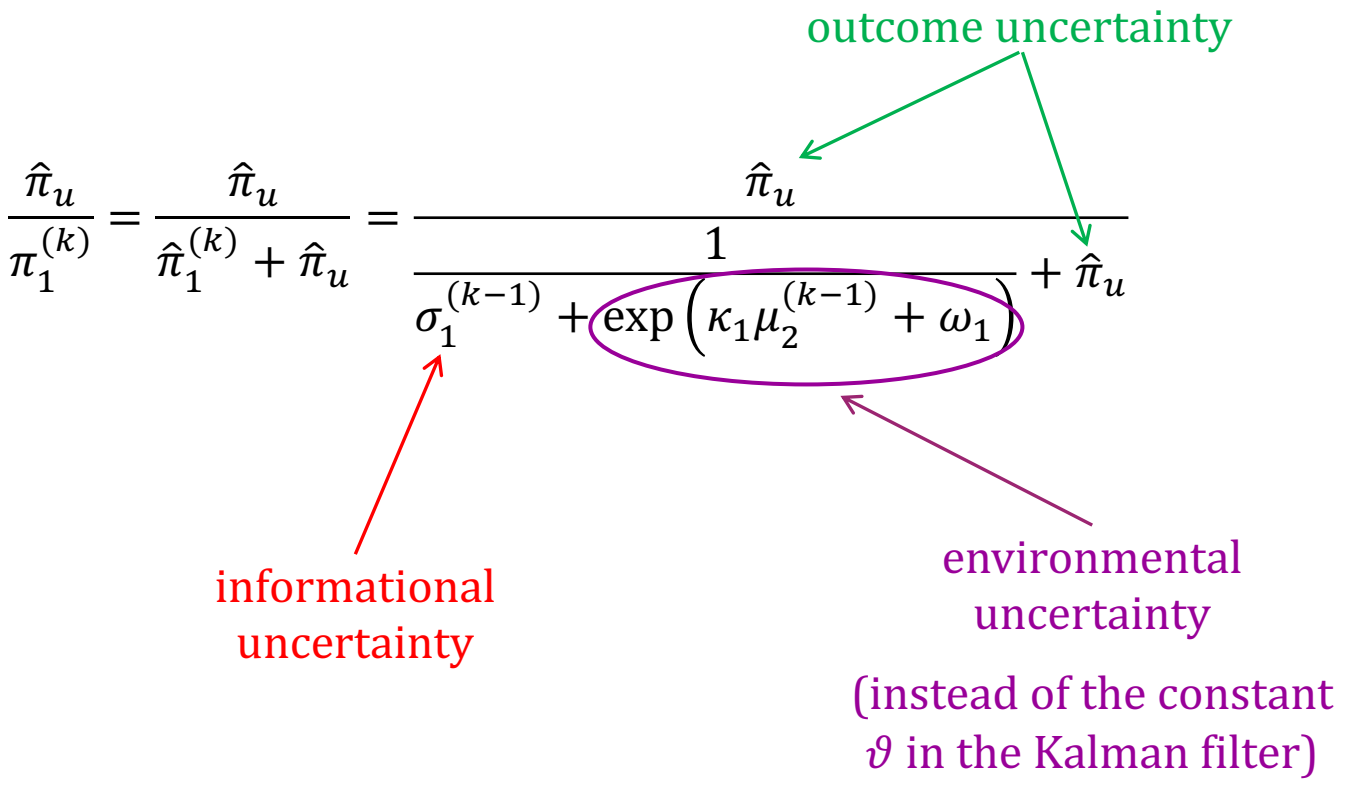
Variational inversion leads to update equations that are **precision-weighted prediction errors.**

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

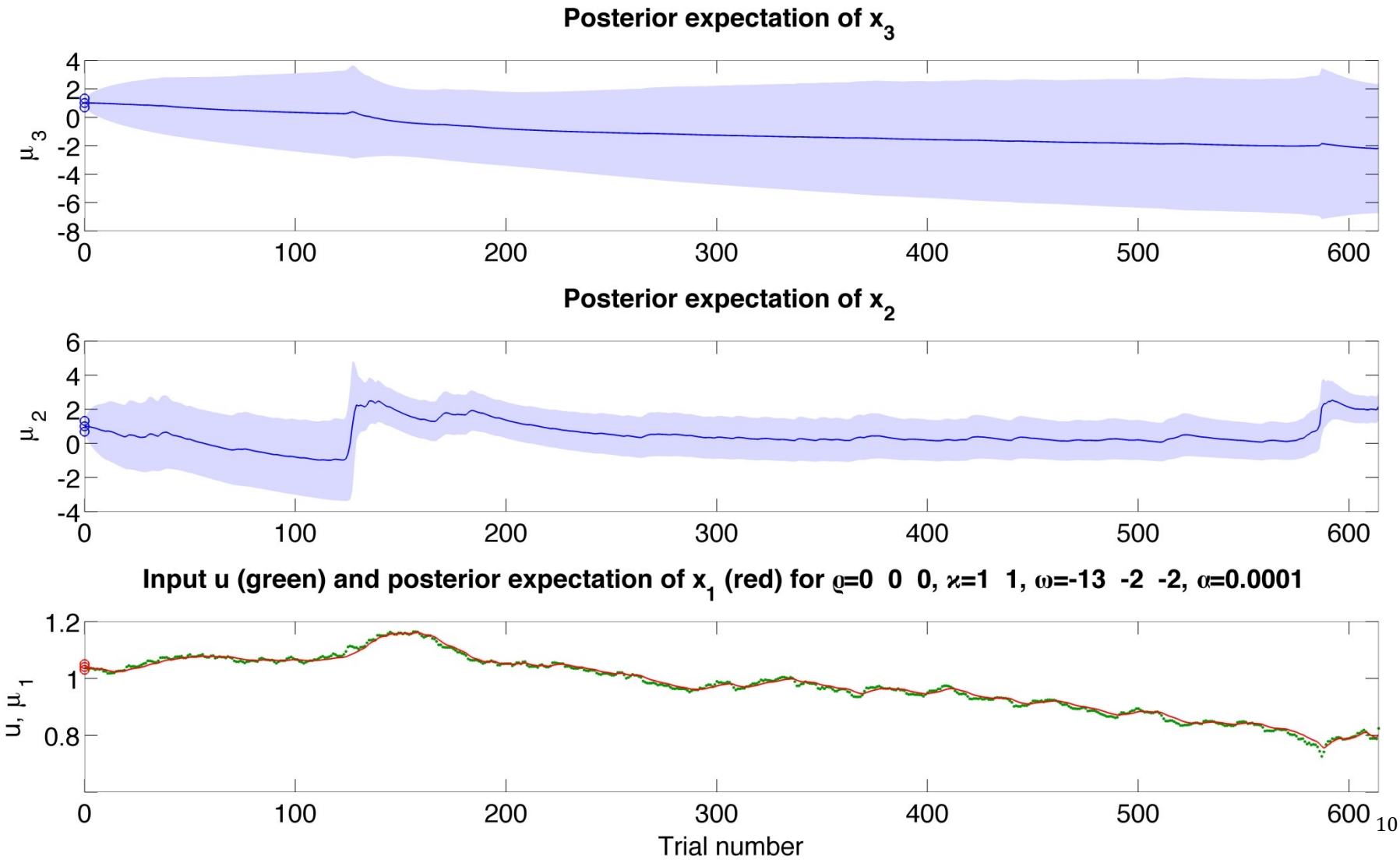


The learning rate ('Kalman gain') in the HGF

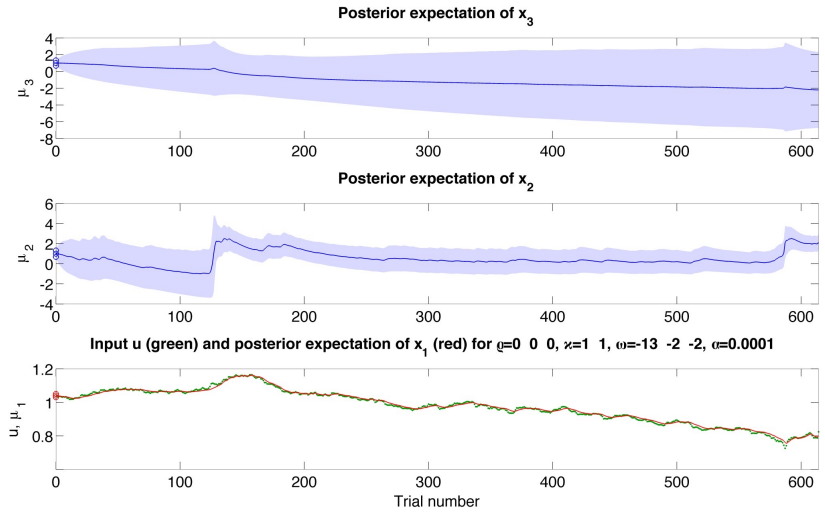
Unpacking the learning rate, we see:

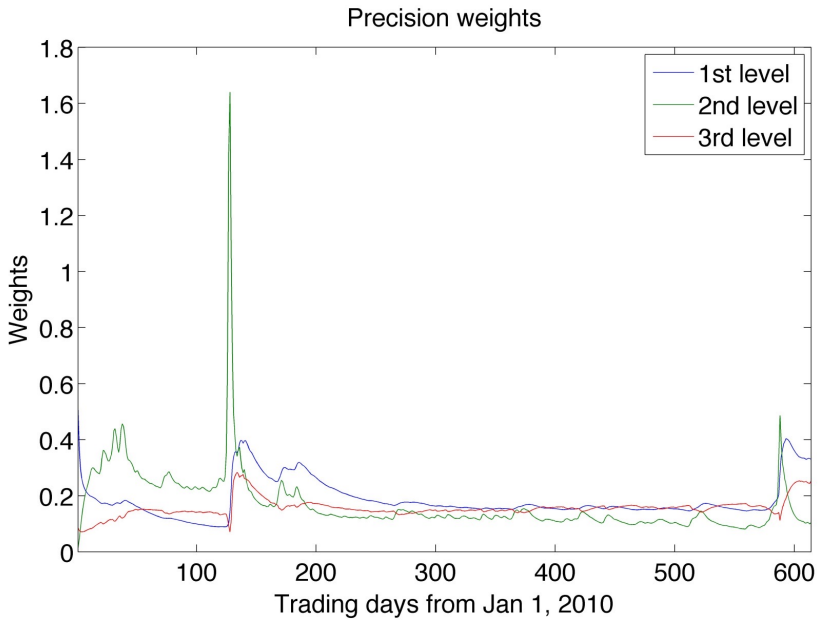


3-level HGF for continuous observations



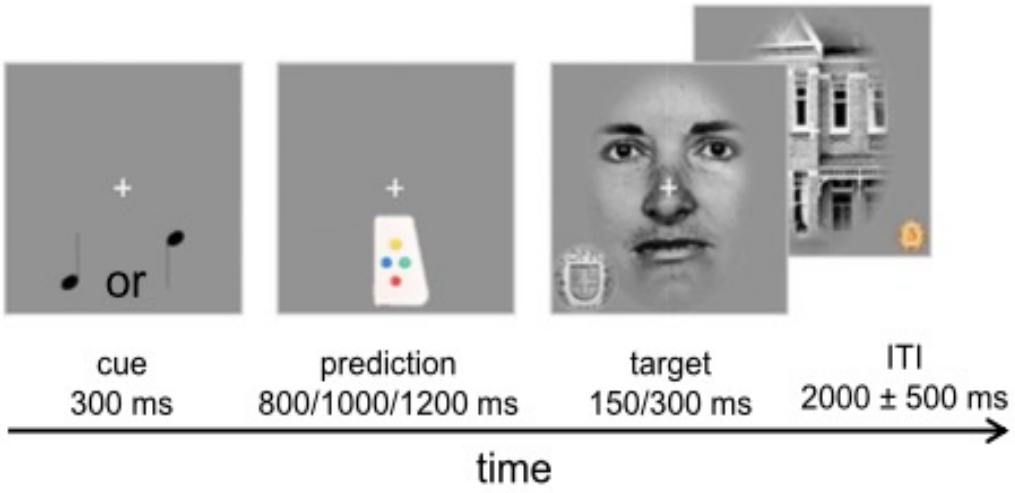
Example of precision weight trajectory

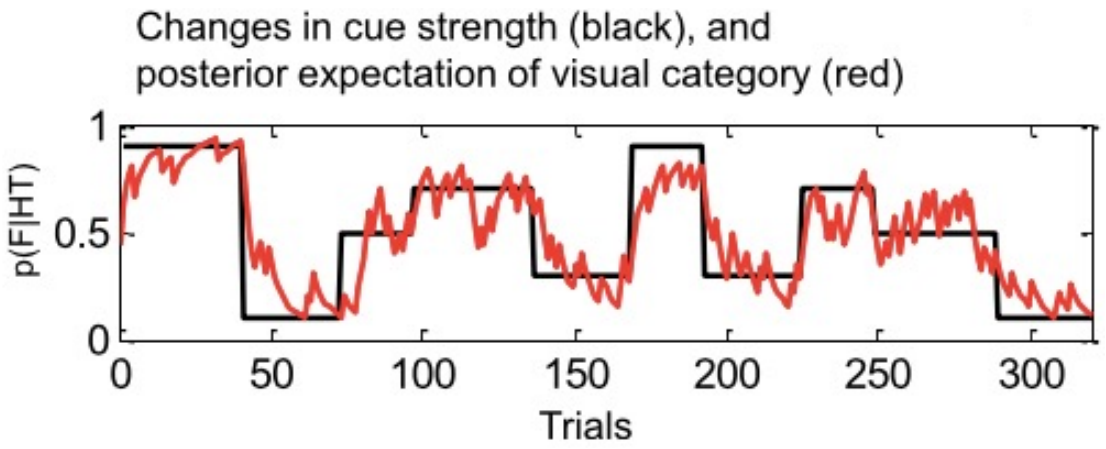




Where would we need a model with an adaptive learning rate?

Task of Iglesias et al., *Neuron*, (2013):





Parameter estimation with the HGF Toolbox

- Available at

  www.translationalneuromodeling.org/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
```

Iglesias et al. (2013), Neuron

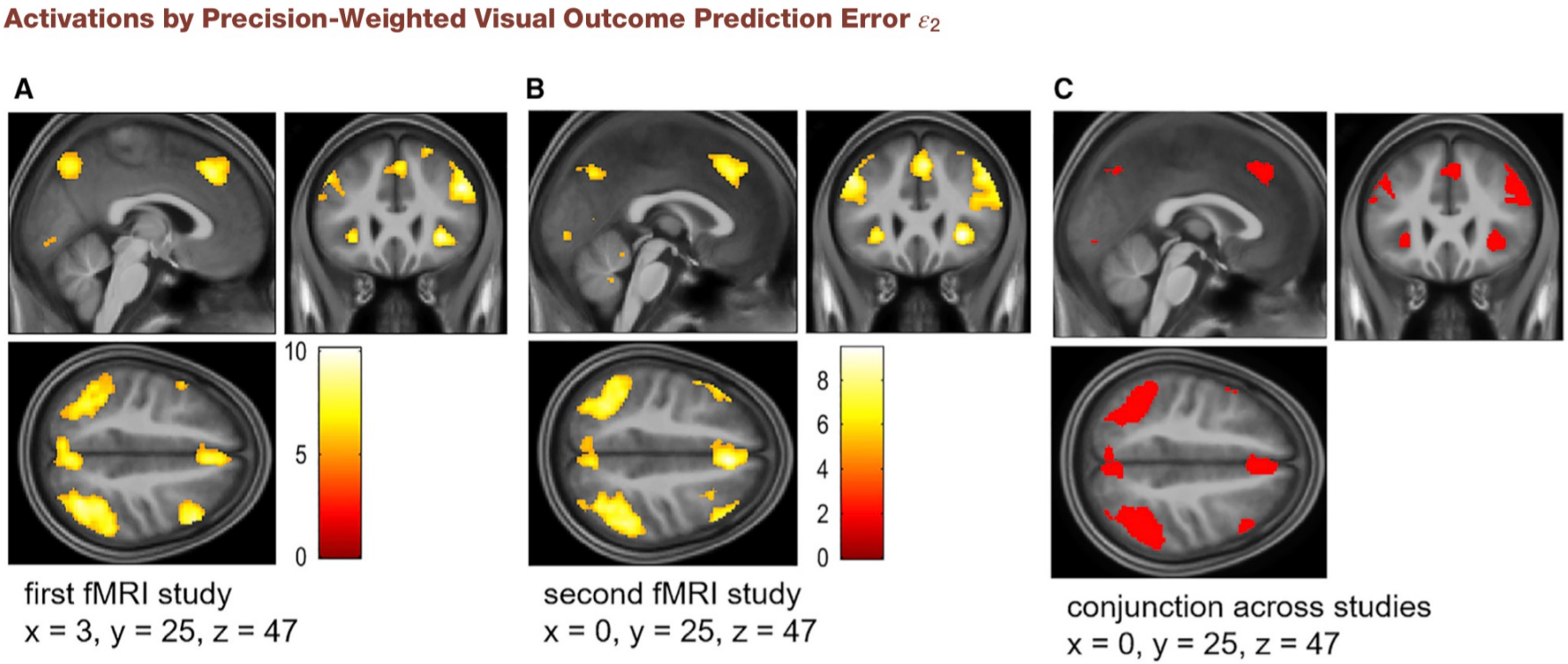


Figure 2. Whole-Brain Activations by ε_2

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)

Iglesias et al. (2013), Neuron

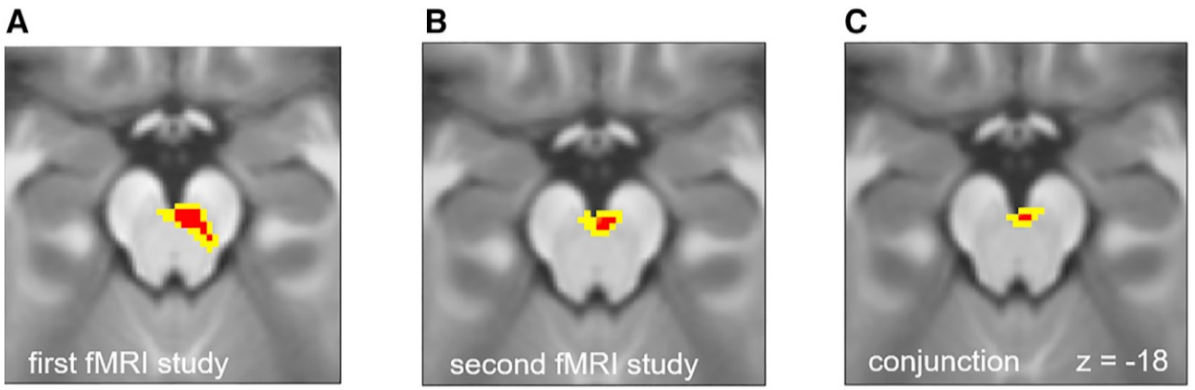


Figure 3. Midbrain Activation by ε_2

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)

Iglesias et al. (2013), Neuron

Activations by Precision-Weighted Prediction Error about Stimulus Probabilities ε_3

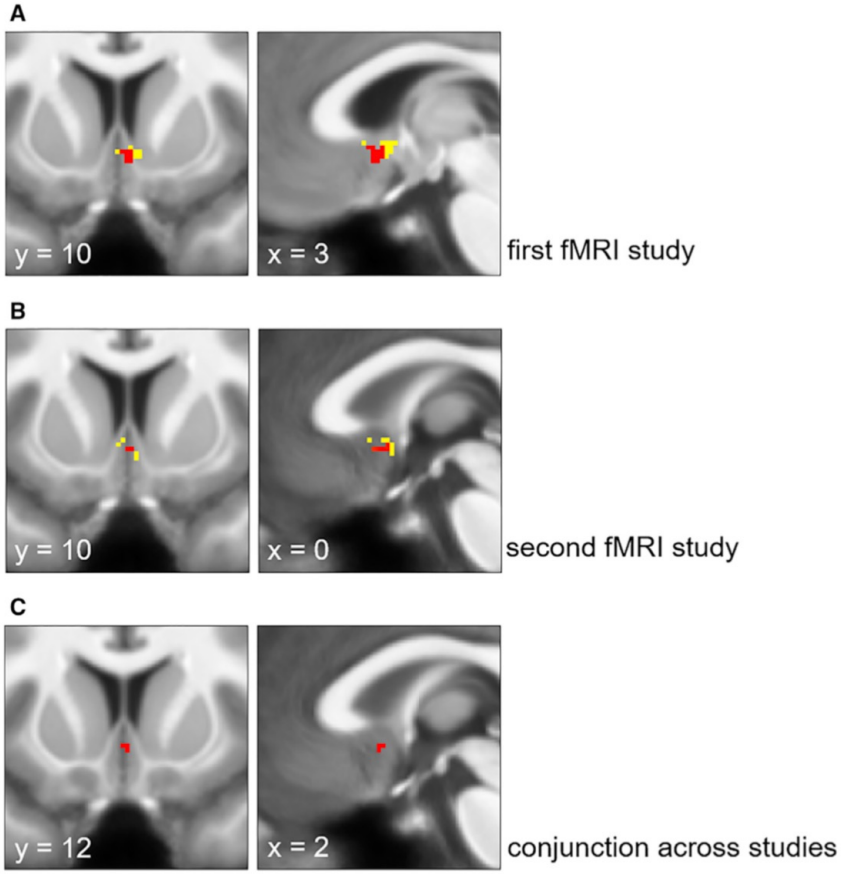
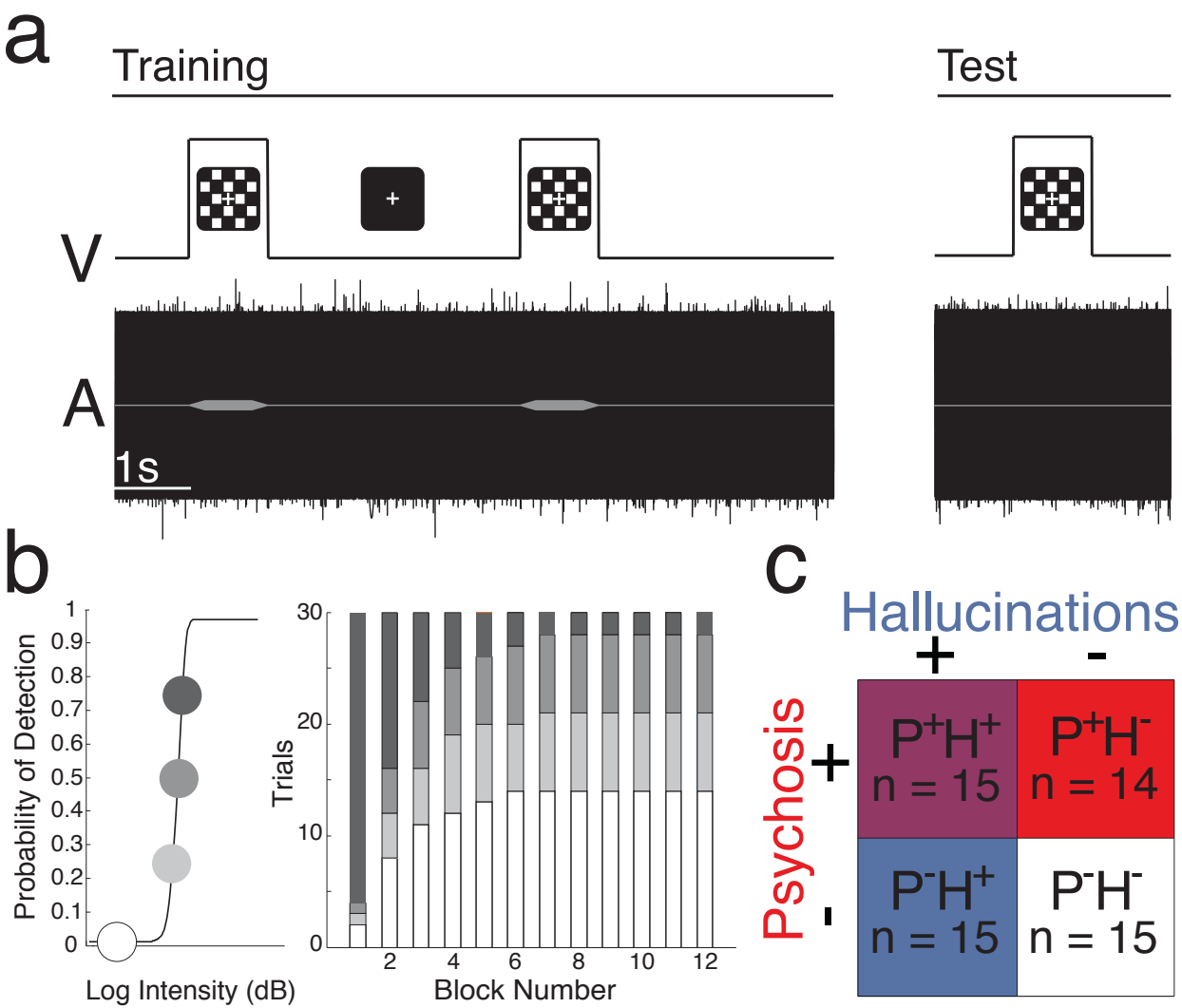


Figure 6. Basal Forebrain Activations by ε_3

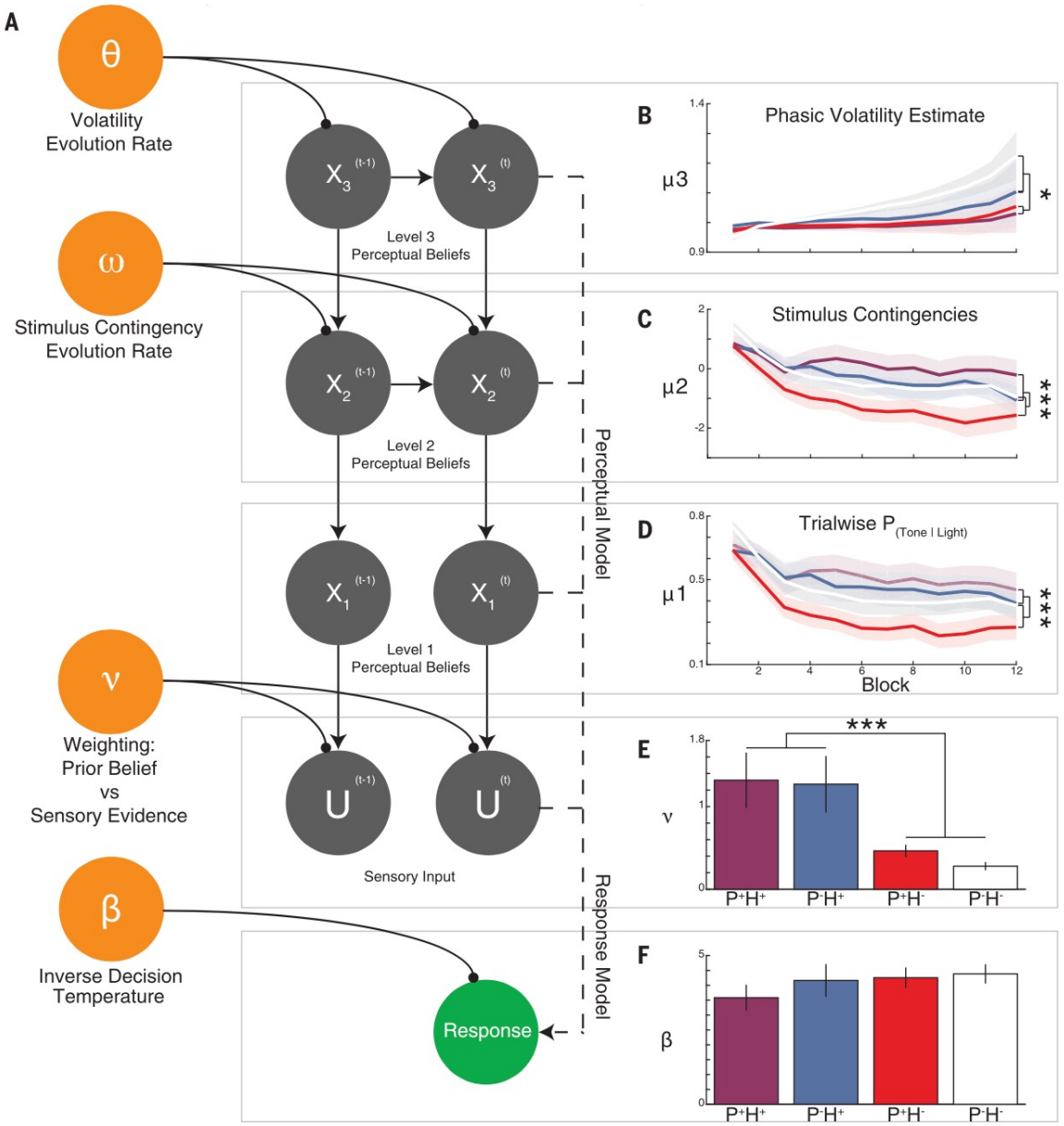
Activation of the basal forebrain by precision-weighted prediction error about stimulus probabilities ε_3 within the anatomically defined mask. For visualization of the activation area, we overlay the results thresholded at $p < 0.05$ FWE peak-level corrected for the entire anatomical mask (red) on the 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)

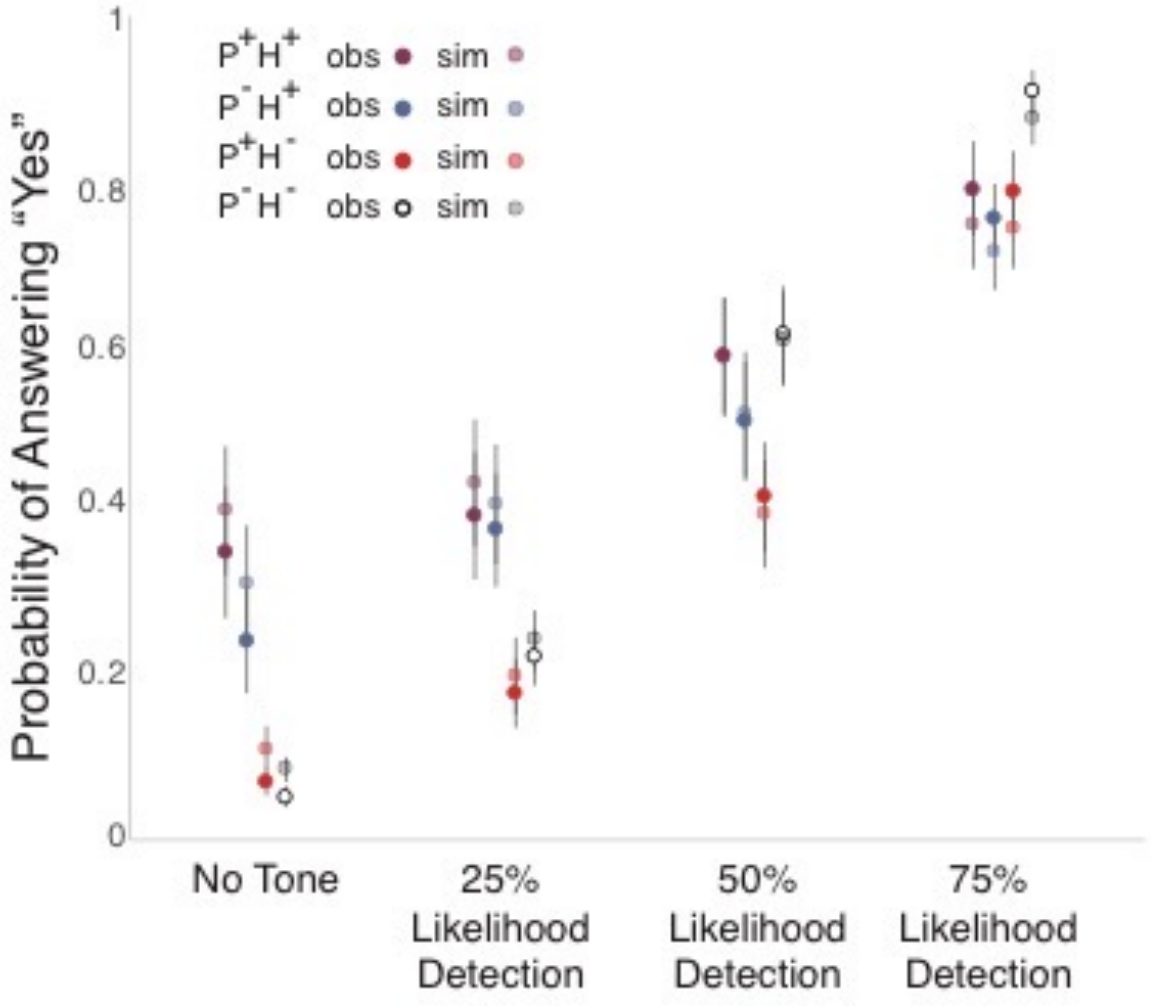
Conditioned hallucinations (Powers, Mathys, & Corlett, Science, 2017)



Conditioned hallucinations: HGF model and results



Conditioned hallucinations: posterior predictive simulations



Many other HGF applications

- Social learning in health and autism spectrum disorder, schizophrenia, major depressive disorder, and borderline personality disorder (e.g., Diaconescu et al., 2014; Henco et al., 2020b)
- Neural signatures of social learning (e.g., Henco et al. 2020a)
- Volatility learning in autism spectrum disorder (e.g., Lawson et al., 2017)
- Threat learning and psychopathic traits (Brazil et al., 2017)
- Different forms of uncertainty and their relation to stress (e.g., de Berker et al., 2016)
- …

Introducing HierarchicalGaussianFiltering.jl

- A Julia package (https://julialang.org) for the easy construction of arbitrarily large HGF models whose **nodes** are connnected via either **mean** or **variance coupling**
- Sister package **ActionModels.jl** for constructing response models
- Both will be part of **TAPAS** (www.translationalneuromodeling.org/tapas)
- Parameter estimation via **Turing.jl** (https://turing.ml), a powerful general-purpose sampler which allows for **hierarchical parameter estimation**
- The most **common HGF models** (3-level binary, 2-level continuous, JGET, etc.) are **pre-implemented** for convenience

Schizophrenia Research 245 (2022) 42–49

Contents lists available at ScienceDirect

Schizophrenia Research

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

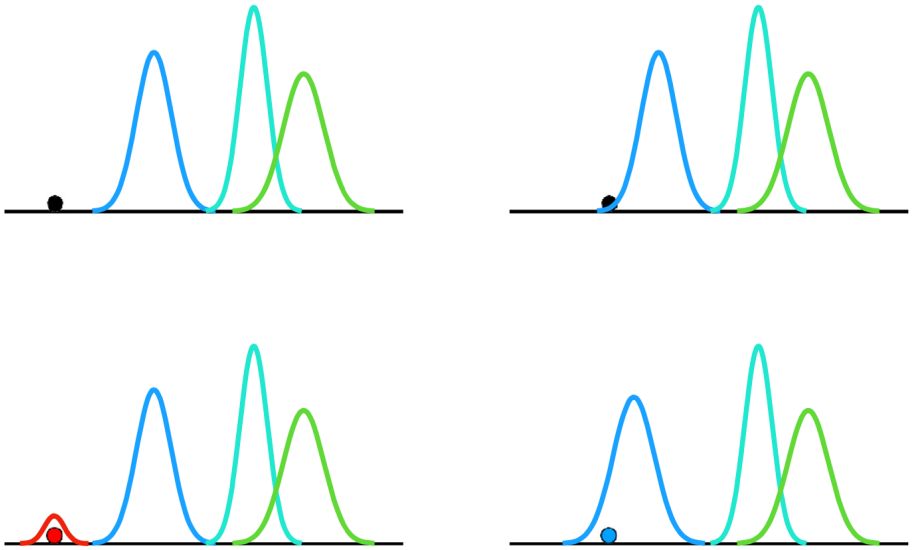
A generative framework for the study of delusions

Tore Erdmann [a], Christoph Mathys [b,c,*]

[a] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy
[b] Interacting Minds Centre, Aarhus University, Jens Chr. Skous Vej 4, 8000 Aarhus C, Denmark
[c] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Wilfriedstrasse 6, 8032 Zurich, Switzerland

A generative framework for the study of delusions

- Delusional inference is hard to model
- During delusion formation, the model has to learn too quickly
- During delusion maintenance, the model has to learn too slowly
- How to explain this contradiction?
- Our approach: hierarchical Dirichlet Process Mixture Models who are capable of insulating a central belief from countervailing evidence, which they explain away by creating ad-hoc causes for the offending observations



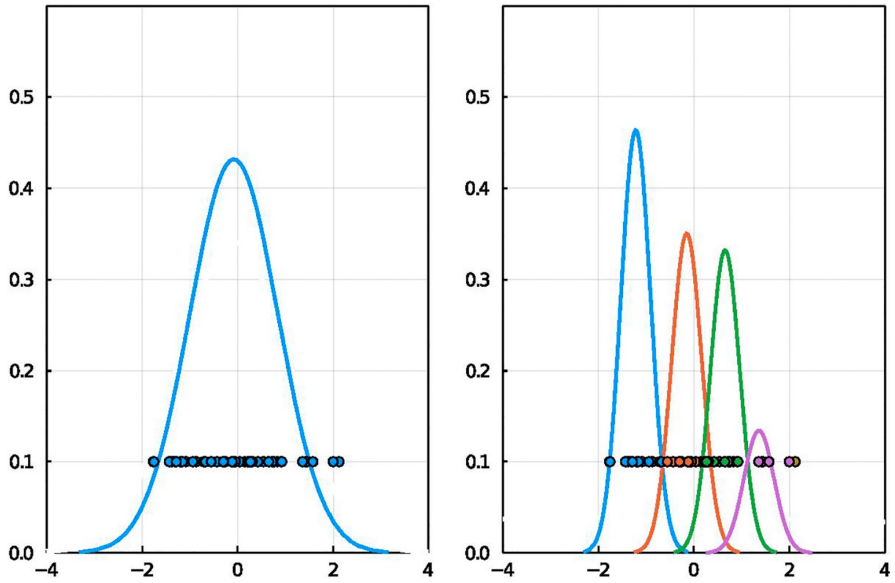
Thanks

SynoSys future research vision

- Studying delusional inference in the recent DPMM framework, augmented with explicitly causal models
- Large networks of HGF nodes
- Efficient concurrent filtering of rich and varied sensory modalities – combined with causal inference (adaptive weighting of inputs, exploiting correlation)
- Teaching and mentoring

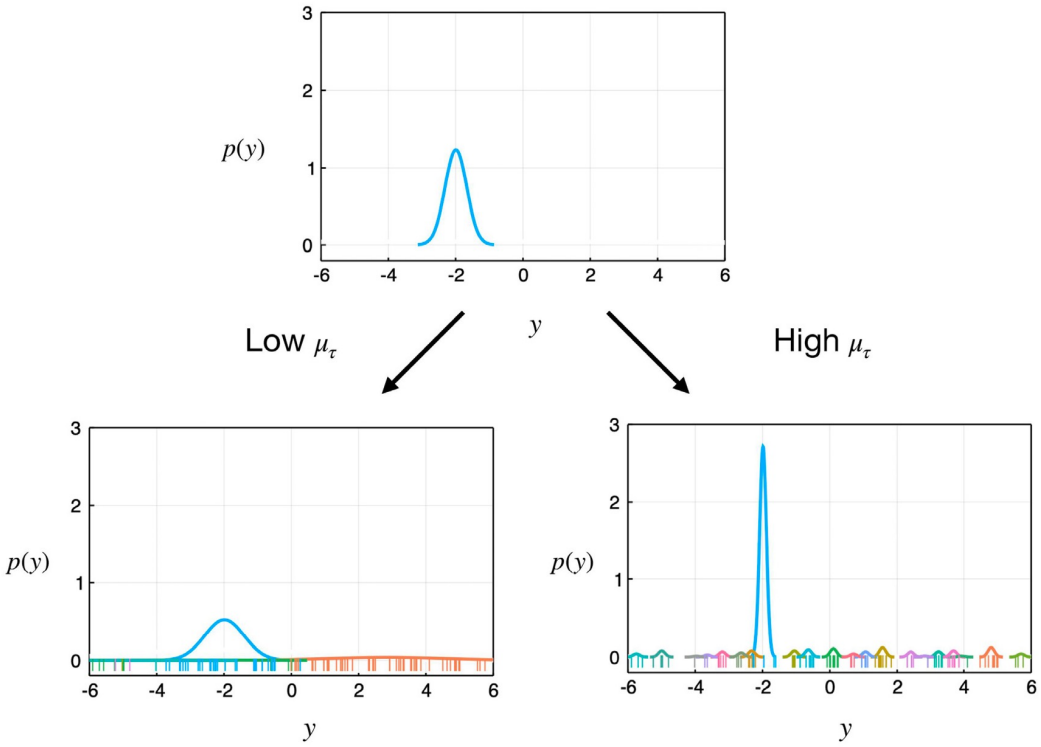
A generative framework for the study of delusions: capabilities waiting to be exploited

- We are only just starting to exploit the capabilities of our new framework. A quick taste of this in what follows.
- Capgras delusion: someone close to me has been replaced by an impostor
- Case study (Hirstein & Ramachandran, 1997): Capgras patient was presented with a sequence of photos of the same person from different directions. Patient identified the photos as “different women who look just like each other”.
- A simple model based on our framework produces such an effect with the adjustment of a single parameter: the expected precision of explanations.



A generative framework for the study of delusions: capabilities waiting to be exploited

- Adjustment of the same single parameter decides between **broadening and elaborating** a central belief **versus protecting and narrowing** it.
- A side effect of protecting the central belief is the need to generate many disconnected ad-hoc explanations



Vision for a SynoSys department

- My potential contribution to the field of Collective Adaptive Systems: **novel algorithms and experimental methods based on HGFs and our delusion framework**, especially when augmented with explicitly causal models.
- Synergies **with robotics, artificial intelligence, agent-based modelling, and cognitive modelling** in humans and animals (concurrent filtering of varied sensory modalities; adaptive weighting of inputs, exploiting correlation)
- Synergies with **clinical psychology** and (computational as well as clinical) **psychiatry**
- Synergies with **virtual/augmented reality researchers** regarding the design of delusion-inducing virtual environments (considerable experimental challenge!)
- Synergies with **neuroscientists** on the testing of HGF-based models of cortical message passing (large networks of HGF nodes)

Teaching and mentoring (1)

- I have taught at university since my PhD student days
- E.g., I have for many years been an invited lecturer at Statistical Parametric Mapping (SPM) courses and at Computational Psychiatry courses in London and Zurich
- I taught various research-based courses during my time in Trieste, both at SISSA and at the International Centre for Theoretical Physics (ICTP)
- Organizer of the 2018 ICTP Winter School on Learning and Artificial Intelligence, aimed primarily at students from developing countries
- Currently teaching three courses (2 BSc, 1 MSc; 10 ECTS each) in Aarhus University's Cognitive Science programmes
- Methods 2 (BSc): The General Linear Model (including linear algebra and multivariate calculus)
- Methods 4 (BSc): Bayesian Computational Modelling
- Data Science (MSc)
- Previously also taught Decision Making (MSc, 10 ECTS)

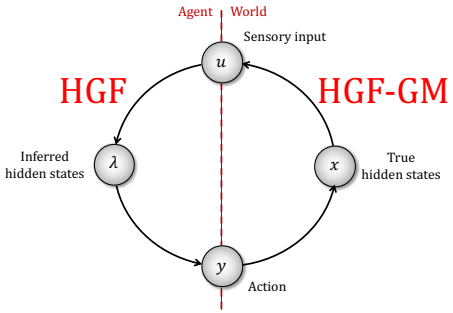
Teaching and mentoring (2)

- I make extensive use of the ‘flipped classroom’ model:
- Students read texts and/or watch lecture videos prepared by me and/or external sources
- Lectures and classes are used to deepen and expand the watched/read material and to familiarize oneself with practical applications of the methods introduced
- Successfully supervised and co-supervised many PhD, MSc, and BSc students.
- Most recent PhD graduate moved on to a postdoc position at University College London.
- Postdoc alumnus recently moved to another postdoc position at the University of Bonn.
- Mentored several PhD students whose primary supervisor was someone else.

Thanks

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**.

Updates at the first level

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

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

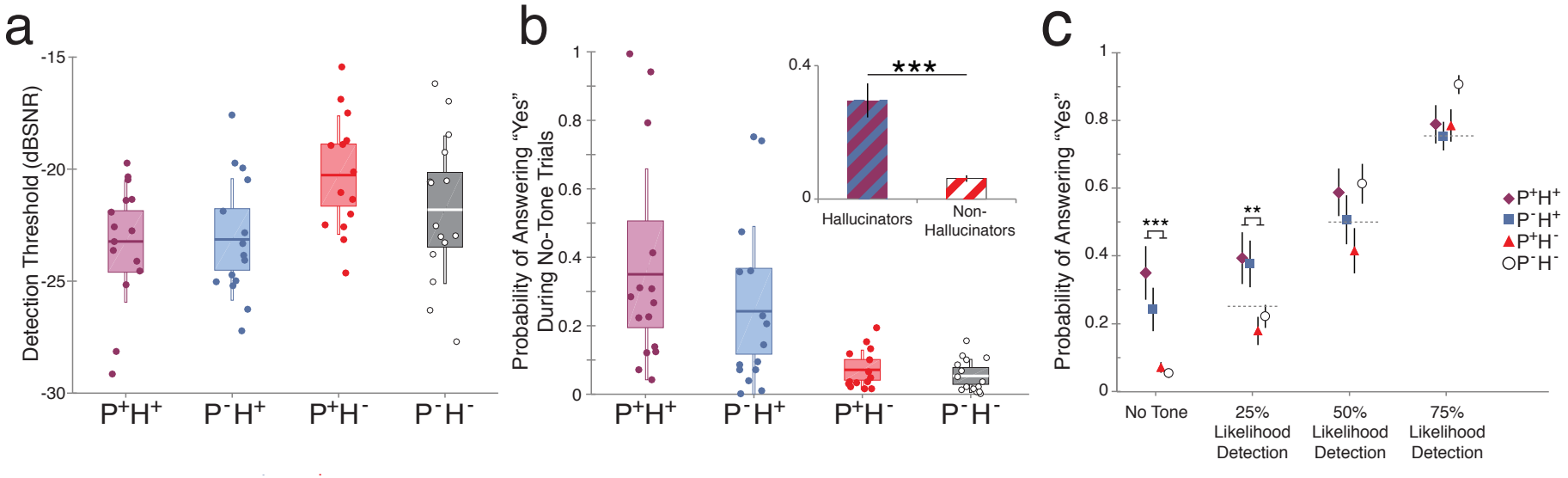
This gives us the following update for our belief on x_1 (our quantity of interest):

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

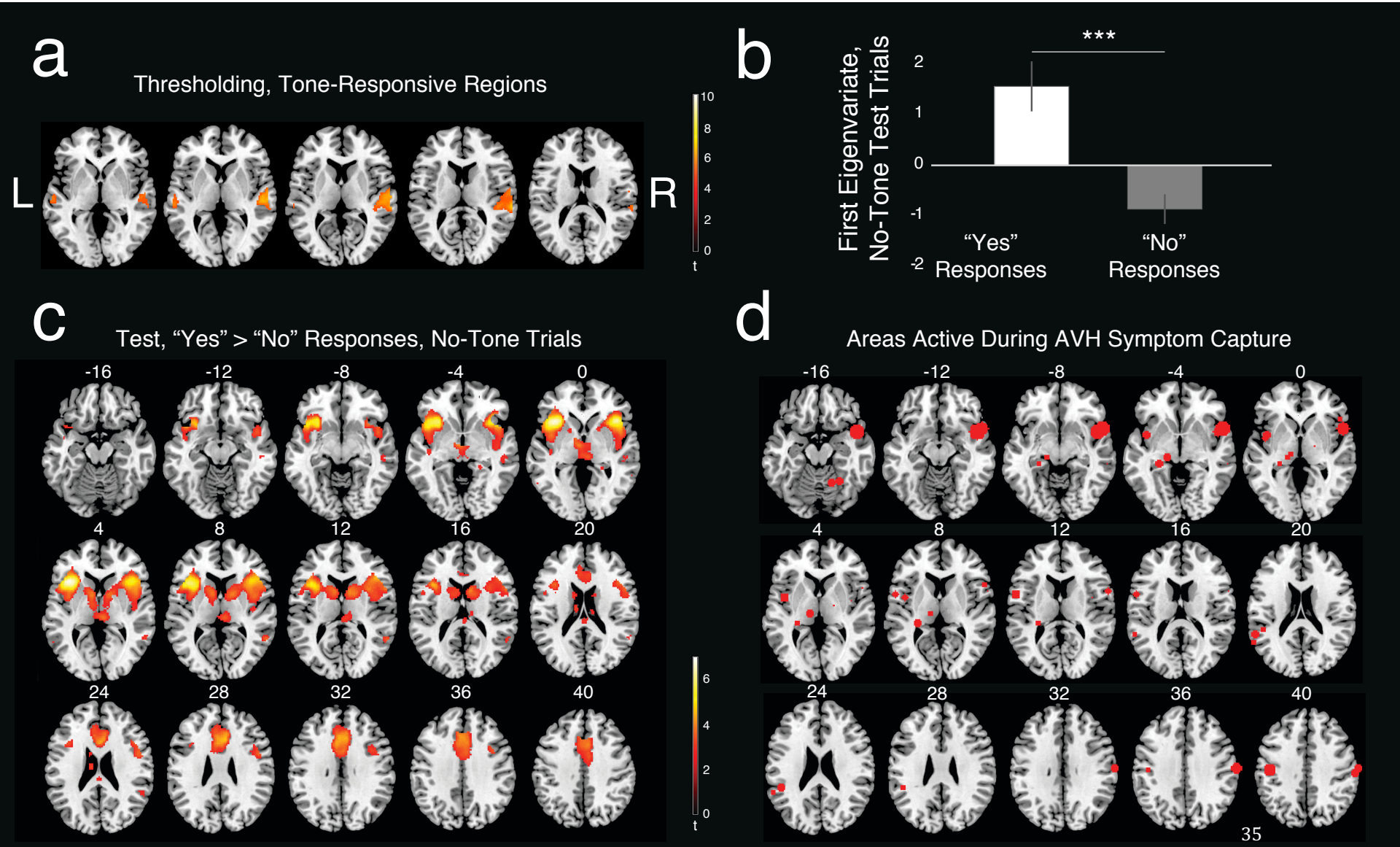
$$\mu_1^{(k)} = \mu_1^{(k-1)} + \frac{\hat{\pi}_u}{\pi_1^{(k)}}\left(u^{(k)} - \mu_1^{(k-1)}\right)$$

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

Conditioned hallucinations: sanity checks



Conditioned hallucinations: were participants really hallucinating?



Remarks on model comparison / model selection

- 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.