The Influence of Stimulus Nature on Probabilistic Inference in Generative Models of Behavior and its Physiological Correlates

Moritz Gruber[1], Lukas Vogelsang[1], Jannis Born[1]

1 *Institute of Neuroinformatics, ETH Zürich and University of Zürich, Switzerland*

Abstract

The current state of psychiatric diagnostics calls for a transition from phenomenology to aetiology and from subjective to objective, physiologically interpretable classification criteria. Within the context of computational psychiatry non-invasively shedding light onto internal states by means of generative models, the presented study aims at devising a quantitative, potentially disease-relevant measure of behavioral states by modeling how aversive versus neutral stimuli bias behavior in an audio-visual learning task. In addition, we investigated possible peripheral physiological correlates of computational quantities delivered by the Hierarchical Gaussian Filter (HGF). Finally, we sketch a temporal perspective on ω_2, a key parameter in the HGF, and compare hierarchical Bayesian and non-hierarchical non-Bayesian models of learning (HGF and Rescorla-Wagner). In our experiment, behavioral data and skin conductance measurements from $n = 16$ subjects were acquired. Our findings suggest a bidirectional effect of aversive stimuli on behavior in two different subgroups and a promising hint at the existence of canonical physiological responses to prediction errors pertaining to the early levels of the generative hierarchy. Overall, our results, if replicated on a more significant — perhaps clinical — sample, represent a modest step towards the sorely needed, mechanistic psychiatric diagnostics by bridging the gap between adherents to the Bayesian brain hypothesis and adjacent neuroscientific fields.

All the code is available on https://github.com/tnm-finalproject.

KEYWORDS: computational psychiatry, generative model, hierarchical gaussian filter, galvanic skin response, skin conductance response, prediction error, Bayesian brain, temporal window of integration, Bayesian model selection

Introduction

Towards an objective and quantitative strategy for diagnostics in psychiatry

The status quo. Traditionally, psychiatric diagnostics are performed on the basis of a classification dictated by the Diagnostic and Statistical Manual (DSM), or similar, which comprises a comprehensive list of psychiatric disorders along with the signs and symptoms that make patients eligible for diagnosis. In the case of schizophrenia, one of the most prevalent psychiatric disorders (Saha et al., 2005), this list includes delusions, hallucinations, anhedonia, flat affect, asociality, disorganized behavior, and many more. According to the DSM, exhibiting at least two symptoms (at least one of the positive variety) over the course of at least one month is sufficient to obtain the label "schizophrenic". While this approach to diagnostics has proven effective in dealing with great numbers of returning soldiers post Wold War II, it inherently suffers from at least two major flaws. First, due to the diverse array of possible genetic predispositions and environmental risk factors, the pool of patients under the same diagnostic umbrella inevitably becomes highly heterogeneous in terms of pathologies. Secondly, there is no systematic and informed way of prescribing drugs, let alone predicting treatment courses: Drugs tend to be prescribed in a trial-and-error fashion, often based on what side effects the individual patient could tolerate best.

The computational approach. Deeming this status quo untenable – both for clinicians and patients, but also their relatives – the emerging field of computational psychiatry has set out to develop methods that allow for the objective and quantitative classification of patients. Moving away from subjective reports of symptoms, towards diagnostics that are grounded in aetiology rather than phenomenology, requires external sources of information in the form of neuroimaging or behavioral data. Prima facie, one could imagine that differential diagnosis may be feasible based on raw, anatomical data, such as magnetic resonance imaging (MRI) scans. While projects of this kind have been successfully undertaken (Liu et al., 2017), they bring us no closer to the bottom-up understanding of disease mechanisms that would enable us to prescribe targeted treatments.

Generative models of neuroimaging data

A more promising approach rests in the concept of generative modeling, in the context of which physiologically interpretable, computational models of the brain are constructed based on externally acquired data. More technically, those models mathematically describe the probabilistic links between hidden states of the brain and the noisy measurements we acquire using MRI or electroencephalography (EEG). The result of this approach is a set of subject-specific parameters that can be readily mapped onto physical entities (such as the average weight of excitatory long-range connections between two specific brain regions) and therefore exploited for a white-box classification into disease cohorts. The work of van Leeuwen et al. (2011) on synaesthetes serves as illustrious proof of concept of this approach: Based on an anatomically informed set

of prior models, two groups of different phenomenology, namely the "projectors" and the "associators", could be reliably distinguished, which has led to an understanding at the mechanistic level in terms of effective brain connectivity. Several similar studies within the clinical realm suggest that this approach can, in principle, be applied to assay specific neurophysiological parameters (*e.g.*, potassium channel (Gilbert et al., 2016) or NMDA receptor activity (Symmonds et al., 2018), marking a significant step towards objective, quantitative psychiatric diagnostics.

Generative models of behavior

Within the context of a theoretical framework he coined as the Free Energy Principle, Karl Friston argues that given the objective of minimizing energy expenditure, the brain's optimal policy is to hold and constitutively update a generative model of its inner and outer milieu, in an effort to avoid costly actions on the world or on itself (by changing its anatomy). The ensuing generative modeling approach of behavior rests on the Bayesian Brain hypothesis (Friston, 2012), which postulates a *hierarchical* generative model within the brain that is aimed at minimizing statistical *surprise*. At any point in time, an organism's objective to minimize surprise is accomplished in case it had perfectly predicted a prospective sensory input. The computational quantity describing the mismatch between a prediction and the actual sensory input has been postulated as **prediction error**(PE).

The Hierarchical Gaussian Filter (HGF) is the most established computational framework designed to mimic these requirements (Mathys et al., 2014). For the simple case of predicting a series of binary outcomes, the model would be structured as follows: At the lowest level, drawing from a Bernoulli distribution yields a binary outcome. To account for dynamic changes in the probability of the latter distribution, the second level represents it as a Gaussian random walk, the tendency, whose variance in turn depends on a random variable at the third level, termed the volatility. The parameters that govern the dependencies between different layers in the hierarchy can be fitted to individual subjects, given behavioral data from a suitable task. In addition to those parameters, individual model fits also yield interesting computational quantities for each trial, such as precision-weighted prediction errors (pwPEs), which can then be correlated to brain activity using conventional general linear models for fMRI. Following such an approach within the context of an audio-visual learning paradigm, Iglesias et al. (2013) found neurophysiological correlates of prediction errors at different levels of the generative hierarchy in different regions of the brain, e.g., low level PEs in the midbrain.

Motivation for presented work

This project is geared towards investigating the following questions:

A. Do computational quantities also have *peripheral* physiological correlates?

B. Does stimulus nature bias behavior in a sensory-motor learning task?

C. Can we derive a temporal interpretation of behavioral parameters that may serve as a diagnostic tool in computational psychiatry?

Based on Iglesias et al. (2013), we acquired behavioral data from 16 subjects in an audio-visual learning paradigm. Subjects were presented with a binary visual cue and subsequently asked to predict whether there will be an auditory stimulus or not. The difficulty of the task lies in the fact that the probabilities that govern the cue-stimulus contingencies change over time. In addition to behavioral data, we measured the participants' galvanic skin response over the course of the experiment. The details of the experimental setup are described in the next section.

In the following, we will sequentially motivate each of these questions and elaborate on how our experiment serves to address them.

A. Do computational quantities also have peripheral physiological correlates? It is intuitively sensible that surprise, in the colloquial sense, draws attention by creating arousal. For instance, when absent-mindedly attempting to cross a street, and noticing a car frantically breaking, you become aware of the danger *as a result of* an innate emotional response, triggered by the fact that your brain had not predicted this specific, highly salient input. Physiological responses such as increased heart rate, tense muscles and active sweat glands quickly ensue as a result of this near-fatal prediction error. Thus, it stands to reason that prediction errors— even those of the milder variety— or other abstract computational quantities, could in principle be assayed by reading out those peripheral states. Numerous studies have examined correlates of prediction errors on a variety of spatial scales ranging from the level of single-cell-recordings (O'Neill and Schultz, 2013) to widespread areas including millions of neurons (Brydevall et al., 2018). Prior work extends from cognitive to behavioral manifestations of different subtypes of prediction errors (Iglesias et al., 2017; Den Ouden et al., 2012). However, only a handful of studies have attempted to unravel possible relations between the Galvanic Skin Response (GSR, or skin conductance) and a PE signal (Bach et al., 2010; Spoormaker et al., 2011). Since the polarity of skin conductance prediction error signals to unconditioned stimuli has been controversially discussed (Spoormaker et al., 2012), we prefer using *surprise* (rather than error) as a directionally agnostic term. Hence, we present an effort to deepen the understanding of the role of skin conductance as a potential correlate of prediction errors. In a nutshell, a GSR device acquires a time series of skin conductances which vary as a function of autonomic modulation of perspiratory gland activity. It had therefore long been deemed to be one of the most robust markers of emotional arousal and surprise.

The reason we judge the first question as having substantial value is that unraveling GSR as a resilient byproduct of specific types of prediction errors would carry implications for the ethos of the Bayesian brain hypothesis as follows. Let us reformulate the first question to: Is it possible to infer the belief about the occurrence of a binary event in the *absence of a proper behavioral response*? Assuming the

stereotyped response for situations *with* a prediction error is orthogonal to the stereotyped response *without* a prediction error, a GSR device could in principle be used to trace back the belief about the occurrence of the binary event. In a behavioral paradigm that involves tracking of the predictive power of a cue about the occurrence of such an event, following that approach for every trial yields a binary trace of event predictions. The inferred belief trace is structurally identical to the online recorded behavioral response trace (e.g. collected via keyboard) and as such, amenable to all analyses performed on the behavioral trace (e.g. by generative models of behavior like the HGF) and, more particularly, for a qualitative comparison of the respectively inferred model parameters. Given this hypothesis is correct, a practical application could be present in case a motorically highly demanding procedure is used to acquire the behavioral responses (or in case a patient suffers from tremor, e.g. like in Parkinson's disease). In that scenario, the skin conductance response (SCR) inferred response trace could more closely reflect the participant's actual belief than the noise-distorted behavioral trace.

B. Does stimulus nature bias behavior in a sensory-motor learning task? Our experiment was designed to address a second question. In the first session of 150 trials, the auditory stimulus whose occurrence the participants were tasked to predict was a neutral, 200 Hz pure tone with a windowed onset. In the second session however, the tone was an aversive, sudden-onset white noise stimulus, at an amplitude calibrated to 90 % of the individual acoustic pain threshold. Comparing HGF parameter fits between neutral and aversive sessions would hopefully shed some light onto how the nature of the stimulus confounds behavior. One hypothesis is that participants whose ω_2, an additive term affecting the variance of the tendency (see Methods), increases in the aversive condition are more prone to erratic behavior, and vice versa. To investigate this hypothesis, subjects were asked to answer a short questionnaire where one of the questions was to rate how prone to anxiety and nervousness they judge themselves to be.

C. Can we derive a temporal interpretation of behavioral parameters that may serve as a diagnostic tool in computational psychiatry? We developed a complementary Pseudo-Bayesian framework to render the HGF differences in ω_2-differentials between our participants intuitively accessible in the time domain. Specifically, we show that, under certain constraints, ω_2 can be cast in terms of a temporal window of integration – a measure that may represent one dimension in the multidimensional continuum that the field of psychiatry is spanning; a measure that is interpretable by people both inside and outside of the Bayesian community.

Methods

Experimental Design

The behavioral experiment was based on Iglesias et al. (2013) and consisted of two sessions à 150 trials each. The setup of one trial is displayed in Figure 1. Cues and cue-stimulus contingencies were governed by the following probabilities:

$$\begin{aligned} p(\text{tone}|\text{square}) &= 1 - p(\text{no tone}|\text{square}) \\ &= 1 - p(\text{tone}|\text{circle}) = p(\text{no tone}|\text{circle}) \end{aligned} \tag{1}$$

$$p(\text{square}) = p(\text{circle}) \tag{2}$$

Every 30 trials, $p(\text{tone}|\text{square})$ changed in a discrete fashion in the following order: 0.9, 0.1, 0.5, 0.7, 0.3, adding up to a total of 5 blocks à 30 trials each. Importantly, the two sessions differed solely in the nature of the auditory stimulus: a neutral, windowed onset 200 Hz tone in the first session and an aversive, sudden-onset white noise tone in the second session.

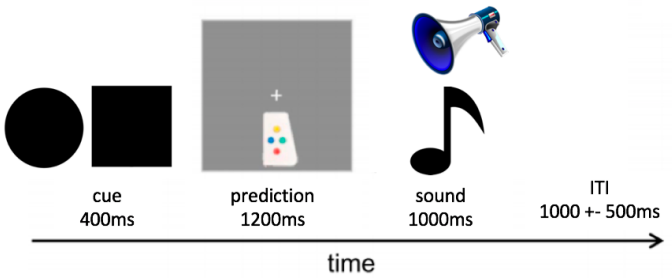

Figure 1: Each trial was compounded by a binary visual cue shown for 400 ms, followed by a 1200 ms response time, a 1000 ms auditory stimulus presentation and a variable inter-trial-interval of 1000 ms ± 500 ms.

Construction of Galvanic Skin Response device

In electrical terms, the skin can be modeled as a series of parallel resistors whose conductivity (or inverse resistance) increases with secretion of sweat. Because resistances are hard to measure using a microcontroller, the measuring device is based on a voltage divider, where R1 is a fixed resistor and R2 represents the skin. Custom-built electrodes were fabricated from 5-Rappen coins, which were selected because of their ratio of highly conductive copper (92 %). An `Arduino Uno` microcontroller was connected to the voltage divider, supplying 5 Volts to R2 and the node between R1 and R2 was connected to a 10-bit analog-to-digital converter (ADC) (see Figure 2). Thus, the measured values range from 1023 (short circuit) to 0 (open circuit). To balance the trade off between sensitivity and dynamic range, we selected R1 for each individual subject such that the baseline input value was in the vicinity of 600. Typically, resistors between 100 kΩ and 1 MΩ were used. The `Arduino Uno` was programmed to interface with `MATLAB`, providing samples at a rate of about 65 Hz.

Analysis of Galvanic Skin Response data

The measurements were acquired with the aforementioned microcontroller at a sampling rate that varied between 60 and 70 Hz. Prior to analysis, time stamps of the GSR signals were synchronized with the time stamps from the behavioral measurements. Next, data was downsampled to equidistant samples every 20 milliseconds (50 Hz) using `MATLAB`'s nearest neighbor interpolation (`interpn`). The `Arduino` backend limited the range of returned conductances to a 10-bit representation thus obliterating interpretation of the absolute values of the signal. For the following steps, `Ledalab`, an established `MATLAB`-based toolbox for

analysis of GSR data was used (Karenbach, 2005). By means of a continuous decomposition analysis, the phasic component of the GSR signal (to which we refer as skin conductance response, SCR) was extracted (Benedek and Kaernbach, 2010). The tonic component (skin conductance level) instead, was excluded from post-hoc analysis as it fluctuates in the range of seconds to minutes rather than 1-5 seconds post-stimulus as the SCR.

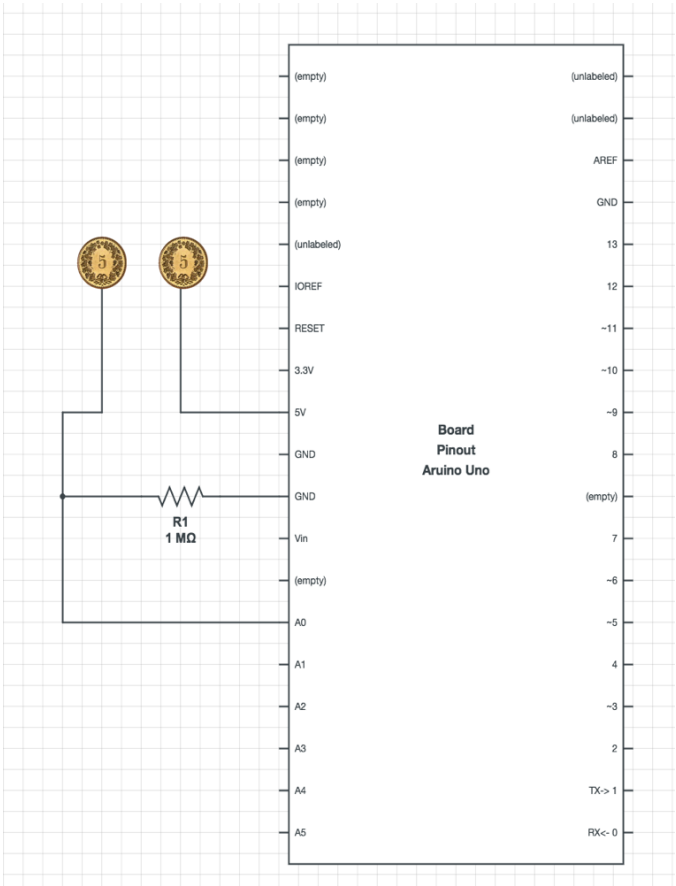

Figure 2: Circuit diagram of the GSR measuring device based on an Arduino microcontroller and 5-Rappen coins as electrodes. The voltage divider referenced in the text consists of R1, the fixed resistor and R2 (not annotated), which pertains to the combined resistance of electrodes and skin.

Experimental Procedure

Initialization. First, subjects were handed an individual sheet containing instructions (see GitHub repository) and a brief questionnaire, which they were instructed to start filling out. To guarantee their anonymity with respect to data analysis, subjects were first asked to pick an ID from an envelope and write it on their sheet. Next, any remaining questions as to the experimental procedure were answered, ensuring that no participant was told how or when the predictive strengths of the cues would change. Finally, to improve motivation, subjects were told that they would receive a performance-dependent compensation of 4 Rappen per correct trial, i.e., a maximum of 12 CHF.

GSR recording. Electrolyte gel was applied to two spots on the participants left hand and custom-made GSR electrodes were thoroughly mounted using medical tape. The data was recorded using a custom-written `MATLAB` interface specifically designed to sample GSR data at an effective sampling frequency of around 65 Hz.

Stimulus amplitude calibration. Before starting the experiment, the amplitude of the aversive stimulus was calibrated for each participant individually to maximize comparability. To this end, subjects were fitted with over-ear headphones, presented with a randomized sequence of the aversive white noise stimulus of 1000 ms duration at different amplitudes and were asked to rate each one on a painfulness scale from 0 (harmless) to 100 (pain threshold). Subsequently, the stimulus amplitude corresponding to 90 % of their individual pain threshold was extrapolated, presented to the participant for approval, and adjusted if necessary.

Experiment. Subjects sat in front of a screen with their left hand laying on the table (and being recorded from) and their right middle and index fingers on a keyboard to provide behavioral output. Presentations of cues and stimuli as well as recordings of behavioral and physiological data were carried out in `MATLAB`.

Questionnaire

During and after the experiment, subjects were asked to fill out a questionnaire (available on GitHub) consisting of three questions to be answered on a scale from 0 to 10.

1. How do you feel today?
2. In general, how easily do you become nervous or anxious in everyday life?
3. How well do you believe you performed the task?

A priori, the first question was designed to potentially eliminate significantly underperforming subjects with greater confidence, while the other questions would be plotted against experimental data and differences in model parameters (see Results).

Assessment of Performance Profile

For a first assessment of performance profiles, we analyzed subjects' responses in the five 30-trials blocks that are governed by different cue-stimulus-contingency probability distributions. This analysis was carried out for the neutral and aversive condition separately. A subject's condition-wise performance score is defined as the proportion of correct responses. A correct response is defined with respect to the underlying conditional probability distribution: in a block where cue-stimulus-contingency 1 will occur with a probability greater than 0.5, the correct response will be cue-stimulus-contingency 1; otherwise it will be 0. As such, the correct response is independent of the feedback received in any particular trial. Responses from the third block were disregarded for performance analysis as the conditional probability was at chance level.

Data Purge

The experiment was conducted with a cohort of $N = 16$ subjects. Due to high trial failure rates (i.e. no valid prediction inserted in $> 10\%$ of the trials), 4 subjects were discarded from all analyses. In addition, several hardware problems in the GSR measuring device, in particular (1)

high-frequency noise induced by concurrent charging and sampling, (2) long-range drifts caused by metal keyboards disturbing the electric circuit and (3) a lack of conductive paste (as used for EEG), forced us to purge all but the last 4 GSR recordings we had gathered.

Model-based Analysis: Hierarchical Gaussian Filters

Theory

In the context of the HGF (Mathys et al., 2014), a binary outcome can be *generated* by sampling from a Bernoulli distribution (5), whose parameter is a sigmoid of the tendency $\mathbf{x_2}$ (4), which evolves as a Gaussian random walk over time (k). The variance of the tendency depends on the third level of the hierarchy, the volatility $\mathbf{x_3}$, in an exponential parameterized by a multiplicative parameter κ_2 as well as an additive parameter ω_2. The volatility is also described by a Gaussian random walk with fixed parameter ω_3.

$$p\left(\mathbf{x_3}^{(k)} | \mathbf{x_3}^{(k-1)}\right) = \mathcal{N}\left(\mathbf{x_3}^{(k-1)}, \omega_3\right) \tag{3}$$

$$p\left(\mathbf{x_2}^{(k)} | \mathbf{x_2}^{(k-1)}, \mathbf{x_3}^{(k)}\right) = \mathcal{N}\left(\mathbf{x_2}^{(k-1)}, \exp(\kappa_2 \mathbf{x_3}^{(k)} + \omega_2)\right) \tag{4}$$

$$p\left(\mathbf{x_1}^{(k)} | \mathbf{x_2}^{(k)}\right) = \text{Bernoulli}\left(s(\mathbf{x_2}^{(k)})\right) \tag{5}$$

$$s(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

For every subject S, we fit the model and obtain parameter sets for *neutral* and *aversive* condition.

Fit

For conducting our analyses, we have worked with different variants of the HGF:

1. A 3-layer HGF with default parameter settings, the κ vector fixed to 1 and the unit square sigmoid observation model.

2. A 3-layer HGF with modified parameter settings: in order to allow ω_2 to explain more variance, we have reduced the variance of ω_3 to zero (thereby, effectively fixed ω_3 to -6) and increased the variance of ω_2 from 4^2 to 4^3.

3. A 2-layer HGF: The third layer was effectively eliminated by adjusting the respective update equations in the configuration file of the `TAPAS` toolbox. Given our experimental design, we are interested in comparing the quality of predictions emerging from a 2-layer and 3-layer HGF.

For all subjects and conditions, we fit an individual HGF and simulated traces based on that parameter estimation. All calculations were performed using the `TAPAS` toolbox and `MATLAB`.

Model-based Analysis: A Pseudo-Bayesian Model

Combining a Gaussian prior probability distribution (7) and a Gaussian likelihood function (8) via Bayesian Inference results in a Gaussian posterior distribution (9) whose update equations (10,11) have a simple analytical solution:

$$\text{Prior} : \mathcal{N}(\mu_\theta, \pi_\theta^{-1}) \tag{7}$$

$$\text{Likelihood} : \mathcal{N}(\text{theta}, \pi_\text{e}^{-1}) \tag{8}$$

$$\text{Posterior} : \mathcal{N}(\mu_{\theta|\text{x}}, \pi_{\theta|\text{x}}) \tag{9}$$

$$\text{Precision update} : \pi_{\theta|\text{x}} = \pi_\theta + \pi_\text{e} \tag{10}$$

$$\text{Mean update} : \mu_{\theta|\text{x}} = \mu_\theta + \frac{\pi_\text{e}}{\pi_{\theta|\text{x}}} \cdot (\text{x} - \mu_\theta) \tag{11}$$

Those sequential update equations result in a posterior precision that is increasing with the precision of each new sample. This translates into a uniform weighting of trials and a temporal window of integration ranging from the first to the last trial (see Figure 3).

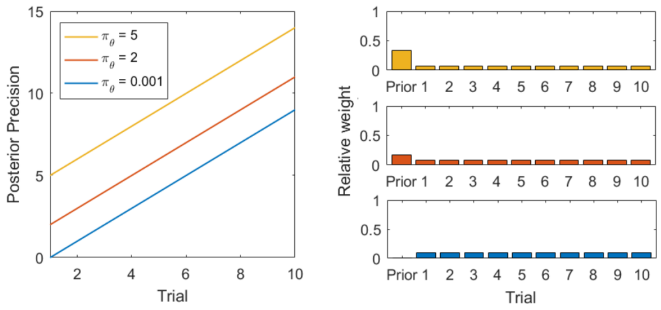

Figure 3: Temporal consequences of Bayes-optimality

This weighting is statistically optimal and the key characteristic of non-hierarchical Bayes-optimality. When having no prior, the posterior will simply be equal to the mean of all observations (assuming all observations have the same precision). With a prior, one could imagine adding $\frac{\pi_\theta}{\pi_e}$ pseudo observations "μ_θ" to the array of observations. Then, the posterior will be equal to the mean of that concatenated array.

While Bayes-optimality is an essential tool for statistical analysis, we may want to be able to model both sub-optimality in stable environments and optimality in volatile environments. Here we will illustrate a specific way of thinking about incorporating this non-hierarchical sub-optimality. If we modify the precision update equation by inserting a precision-leakage factor α, we can change the characteristic trajectories of non-hierarchical Bayesian Inference (12). For $\alpha \neq 1$, the temporal window of integration is smaller than n. For $\alpha < 1$, it is leaning towards more recent, for $\alpha > 1$ towards less recent events (Figure 4).

$$\text{Modified precision update: } \pi_{\theta|x} = \pi_\theta + \alpha \cdot \pi_e \tag{12}$$

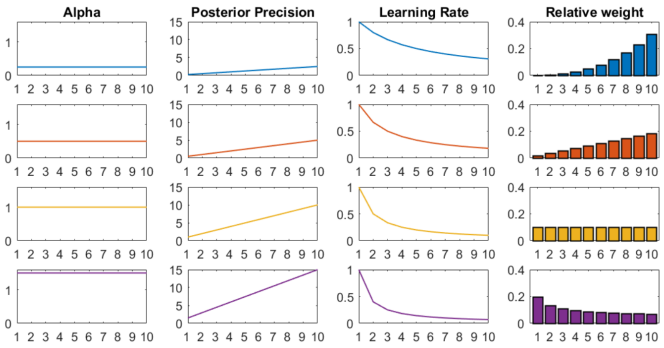

Figure 4: Relative weights for constant α

When allowing α to be a non-constant function of time, more complex trajectories can be described (Figure 5). However, in this report, we will focus on the case that α is constant.

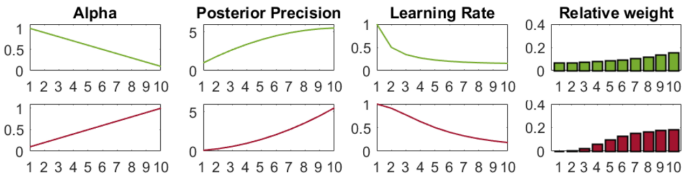

Figure 5: Relative weights for non-constant α

Given the behavioral recordings from the experiment described above, we can fit a subject- and condition-specific α and infer upon subject- and condition-specific temporal windows of integration. Furthermore, we can assess how this approach relates to the parameter estimation implicit in the HGF.

Hyperparameter Fitting in Pseudo-Bayesian Model

A crucial question for every generative model is that of model inversion: how to obtain the hyperparameter that best explain the observed data? Our Pseudo-Bayesian forward model (11),(12) is specified by prior mean, prior precision, likelihood precision and the precision-leakage parameter $(\mu_\theta, \pi_\theta, \pi_e, \alpha)$. Since the model dynamics are predominantly induced by α, the hyperparameter fitting focuses exclusively on α. In case one intends not to fit a continuous belief (e.g. the belief about the cue-stimulus contingency, i.e. $s(\mu_2)$ of the HGF), but a binary response (e.g. the trialwise prediction about the occurrence of an auditory event as in our paradigm), the forward model may optionally be extended by $\hat{y}_k = s(\mu_{\theta|x_k})$ where $\hat{y}_k$ is the binary response at time k and $s(\cdot)$ is a steep sigmoidal function approximating the discretization. In the former case, $s(\cdot)$ should be set to the identity function.

Our objective is to find an α that minimizes the sum of squared distances between the model predictions $\hat{\boldsymbol{y}}$ and the observations $\boldsymbol{y}$:

$$\min(\mathbf{e}^{\mathrm{T}}\mathbf{e}) = \min[(\boldsymbol{y} - \mathrm{s}(\boldsymbol{\mu}))^{\mathrm{T}}(\boldsymbol{y} - \mathrm{s}(\boldsymbol{\mu}))] \tag{13}$$

which we compute by deriving w.r.t. α:

$$\frac{\partial \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{\partial \alpha} = \frac{\partial}{\partial \alpha}[\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\mathrm{s}(\boldsymbol{\mu}(\alpha)) - \mathrm{s}(\boldsymbol{\mu}(\alpha))^{\mathrm{T}}\boldsymbol{y} + \mathrm{s}(\boldsymbol{\mu}(\alpha))^{\mathrm{T}}\mathrm{s}(\boldsymbol{\mu}(\alpha))] \tag{14}$$

$$= -2\frac{\partial \mathrm{s}(\boldsymbol{\mu}(\alpha))^{\mathrm{T}}}{\partial \alpha}\boldsymbol{y} + 2\frac{\partial \mathrm{s}(\boldsymbol{\mu}(\alpha))^{\mathrm{T}}}{\partial \alpha}\mathrm{s}(\boldsymbol{\mu}(\alpha)) \tag{15}$$

$$= 2\left[\left(\frac{\partial \mathrm{s}(\boldsymbol{\mu}(\alpha))}{\partial \boldsymbol{\mu}(\alpha)}\frac{\partial \boldsymbol{\mu}(\alpha)}{\partial \alpha}\right)^{\mathrm{T}}\mathrm{s}(\boldsymbol{\mu}(\alpha)) - \left(\frac{\partial \mathrm{s}(\boldsymbol{\mu}(\alpha))}{\partial \boldsymbol{\mu}(\alpha))}\frac{\partial \boldsymbol{\mu}(\alpha))}{\partial \alpha}\right)^{\mathrm{T}}\boldsymbol{y}\right] \tag{16}$$

The pitfall of (16) lies in the vectorized belief $\boldsymbol{\mu}$ which is subject to iterative updating by the forward model. Whilst it is still unknown to us, whether (16) may collapse to a closed-form analytical solution, we pragmatically decided to follow a rather unaesthetic approach by computing named expression iteratively as time progresses. Let us exemplary outline the value of (16) at an arbitrary time k. Whilst $\frac{\partial s(\mu_k(\alpha))}{\partial \mu_k(\alpha)}$ is obtained straightforwardly by deriving the chosen sigmoidal, $\frac{\partial \mu_k(\alpha)}{\partial \alpha}$ follows a recursive definition:

$$\frac{\partial \mu_k(\alpha)}{\partial \alpha} := \frac{\partial \mu_{k\text{-}1}(\alpha)}{\partial \alpha} - \frac{\partial \pi_k(\alpha)}{\partial \alpha}\frac{\pi_e(x_k - \mu_{k\text{-}1})}{\pi_k^2} - \frac{\partial \mu_{k\text{-}1}(\alpha)}{\partial \alpha}\frac{\pi_e}{\pi_k} \tag{17}$$

with the belief change in the first time step as a recursive anchor:

$$\frac{\partial \mu_1(\alpha)}{\partial \alpha} := -\frac{\pi_e^2}{\pi_\theta^2} \cdot (x_1 - \mu_\theta) \tag{18}$$

(17) is resolved by the precision's dependence on α:

$$\frac{\partial \pi_k(\alpha)}{\partial \alpha} := k \cdot \pi_e \tag{19}$$

To summarize, (18), i.e. the change in belief at any time k, depicts a precision weighted prediction error, whereas (19), i.e. the change in precision, increases linearly over time as in Figure 4 (2^{nd} to left column). The substrates (17)−(19) allowed in practice to infer upon those subject- and condition-specific alphas that best explain binary response traces or the HGF estimated belief trace about the hidden tendency $\boldsymbol{x_2}$. While we have quantitatively monitored that α governs the model fit compared to μ_θ, π_θ and π_e, the ansatz we sketch suffers from all the drawbacks undermining local optimization methods, in particular non-convex error surfaces[1]. Having verified convexity for the majority of our use cases, profound theoretical thoughts on desired conditions necessitating convexity are still lacking as well as a mathematically more solid derivation for hyperparameter fitting is.

Results

Performance profile

As can be seen from Figure 6, the assessment of averaged block-wise performances shows similar profiles in the neutral and aversive condition.

1. which may partially be circumvented by computing the 2^{nd} derivative

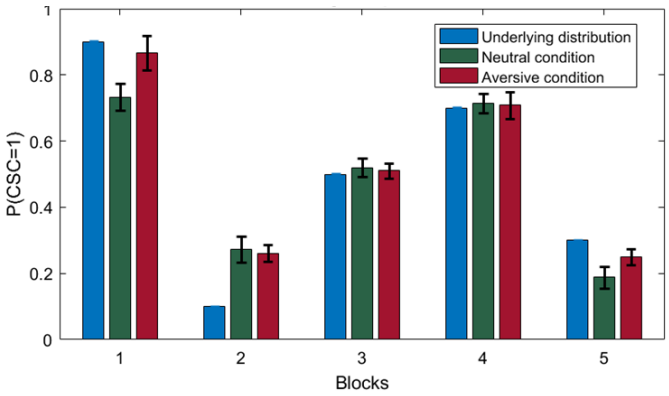

Figure 6: Block-wise performance profile.

Furthermore, participants traced the underlying cue-stimulus contingencies quite accurately.
Performance scores do not differ significantly between conditions (Figure 7). A paired-sample t-test does not reject the null hypothesis that the pairwise differences between performance scores of both conditions have a mean equal to zero (p=0.22).

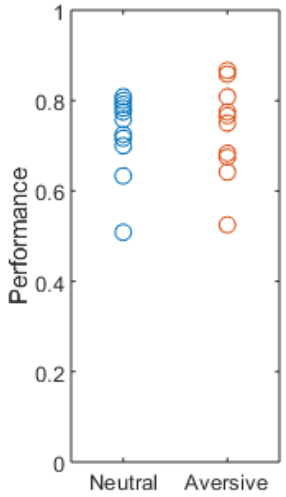

Figure 7: Performance scores

To conclude this subsection, Figure 8 displays the relation between subjective and objective measures of performance. The weak correlation speaks to the difficulty of assessing one's performance, suggesting that most decisions may not be based on a conscious strategy, but rather on unconscious, hierarchical processing, which is coherent with our modeling approach.

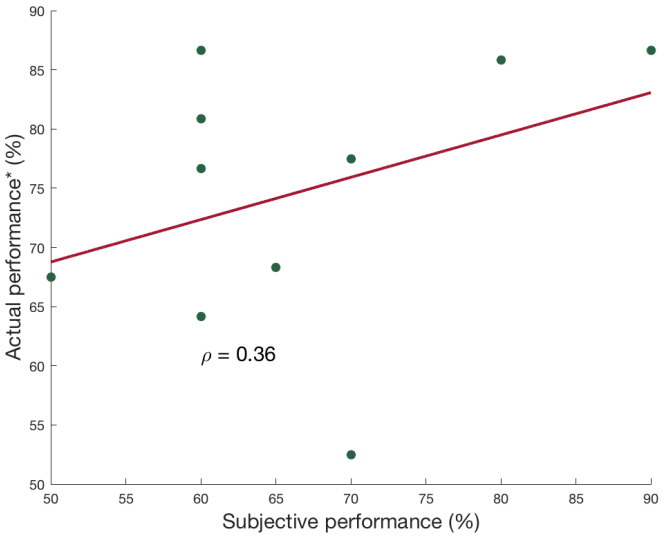

Figure 8: Subjective vs. actual performance. Actual performance refers to the objective performance, quantified as described above, subjective performance was assessed by the questionnaire.

Bayesian model selection

We fitted the aforementioned three HGF models to data from all subjects for neutral and aversive sessions, separately. The criterion we use to determine the best model in our family is the log model evidence (LME), which represents a trade-off between goodness of fit (likelihood) and complexity (deviance from prior). Given LMEs of all model fits, we can derive Bayes factors (BF), the ratio of evidences of two given models. At this stage, conventional assumptions are applied to assess significance: Typically, BFs north of 20 are considered strong evidence for one model over another.
For the sake of visual clarity it is important to emphasize that Figure 9 show the difference between LMEs, which can be viewed as an approximation to the log BF. ΔLMEs were averaged over neutral and aversive sessions for all 11 subjects that were analyzed (data from neutral and aversive sessions are virtually identical with respect to ΔLMEs).

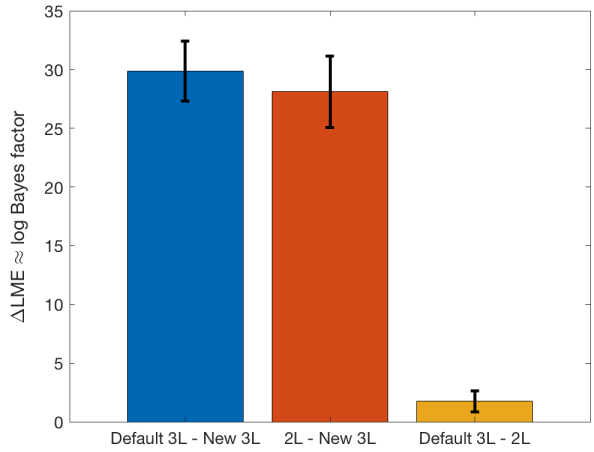

Figure 9: Bayesian model selection. The plots show the average ΔLMEs over neutral and aversive sessions for all 11 subjects that were analyzed ($n = 22$). **Default 3L** corresponds to Model 1 in Methods, **New 3L** to Model 2 and **2L** to Model 3. Error bars show standard error of the mean.

The results indicate that both the default 3-layer HGF as well was the modified 2-layer HGF clearly outperform the modified 3-layer HGFs. The difference between the default 3-layer HGF and the 2-layer HGF failed to reach significance (BF ¡ 20). We conclude that loosening the prior on ω_2 to allow for greater variance and fixing ω_3 (by setting its prior variance to 0) did not improve the model fits on average. Disentangling the individual effects of those two manipulations on model evidences was beyond the scope of this study.

HGF fit

Assessment of adaptation

For the sake of simplicity, the description of the results in the following sections is based on the 3-layer HGF fit with modified parameters. Figure 10 shows the distribution of ω_2 and ω_3 in the neutral and aversive condition for each subject. Instead of assessing statistics on the condition-wise distribution of parameter values, we analyzed the relative changes within each subject.

Figure 11 shows the distribution of condition differences in ω_2. 8 out of 12 subjects have a lower; 4 out of 12 a higher ω_2 in the aversive vs. the neutral condition. Assessing individual adaptation strategies may prove to be useful for extracting response properties that may be linked to psychologically relevant mechanisms.

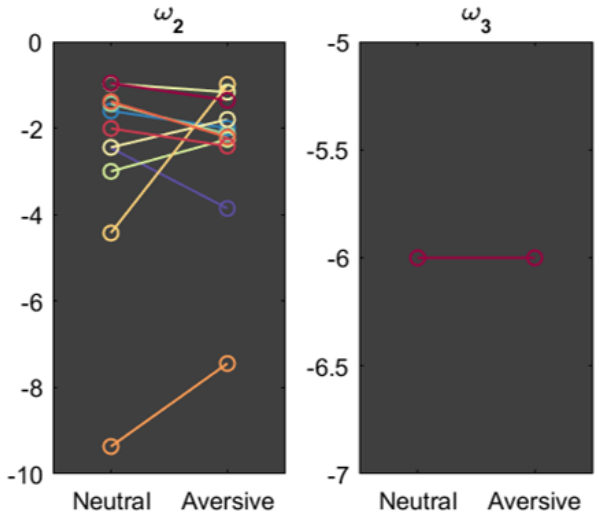

Figure 10: Distribution of HGF parameter values

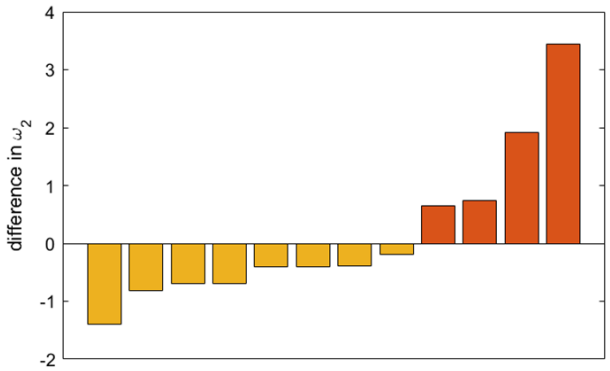

Figure 11: Distribution of individual ω_2 differences

Interpretation of $\Delta\omega_2$

To a layman's ears, the meaning of ω_2 within the HGF, i.e., "tonic component of the variance of the tendency", justly appears somewhat nebulous, that is, not readily understandable in cognitive or psychological terms. This situation not only impedes a clear interpretation and communication of our experimental results, but also the translational aspirations of this variant of generative modeling in a broader sense. Assuming our finding of two subgroups of different ω_2-differentials, $\Delta\omega_2$ is reproducible on a more significant scale, we will now propose a cognitive perspective as a complementary interpretation to the one pertaining to the classical HGF framework. Concretely, we propose an interpretation of ω_2 in terms of α, the auxiliary parameter describing a temporal window of integration.

Casting ω_2-differences in terms of a change in the temporal window of integration naturally leads to an alternative interpretation of our two subgroups (those with $\Delta\omega_2 > 0$ and those with $\Delta\omega_2 < 0$). Subjects whose ω_2 decreased in the aversive session now seem to behave more "Bayes-optimally" (non-hierarchically) by integrating more information from the past, and vice versa. In fact, one of our subjects has lucidly hinted at this perspective en passant, long before our theoretical considerations had matured to their current state.

Overall, we hypothesized that people whose temporal window of integration decreases with aversive stimuli would generally be more prone to erratic and anxious behavior in everyday life. Unfortunately however, correlating subjective measures of anxiety-proneness from our questionnaire with $\Delta\omega_2$ reveals only a weak relationship.

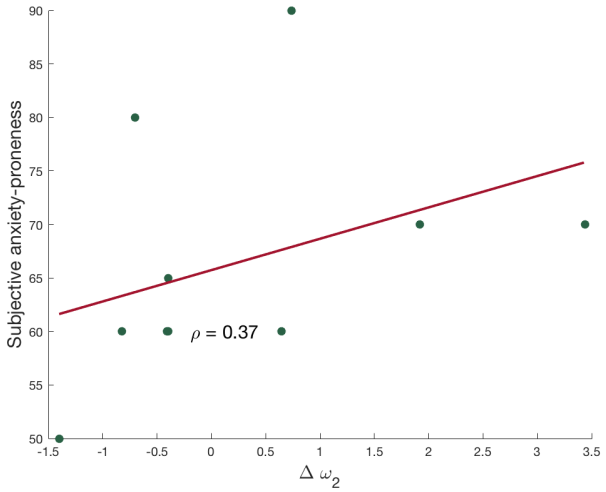

Figure 12: $\Delta\omega_2$ versus subjective measure of anxiety-proneness (%, see questionnaire).

An Engineer's perspective on ω_2

ω_2 influences the variance of the tendency towards a certain cue-stimulus-contingency. For a larger ω_2, the tendency is subject to more spontaneous changes; for a smaller ω_2, one would expect to observe a more stable belief trajectory. Taking an extremely small and an extremely large ω_2 from the distribution of parameter value fits, we can simulate belief trajectories and assess their shape(Figure 13,14).

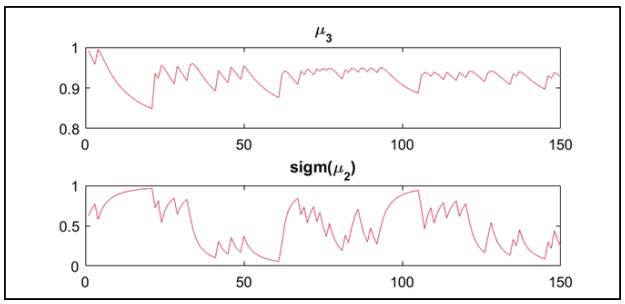

Figure 13: Belief trajectory with large ω_2

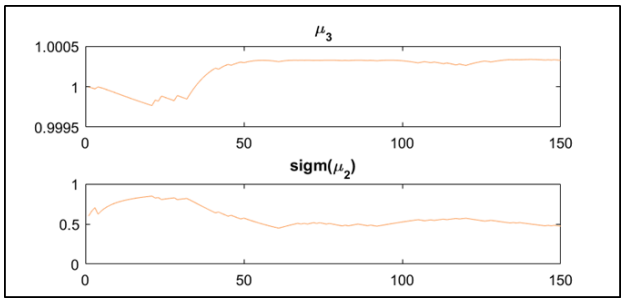

Figure 14: Belief trajectory with small ω_2

Applying the Engineering and Signal Processing view, one could interpret the smoothness of Figure 13 (small ω_2) as a result of filtering over more samples; in other words, applying a kernel with greater temporal extent. This, quite naturally, links to the aforementioned Pseudo-Bayesian model where α, the parameter determining the temporal window of integration, can be fit.

A Pseudo-Bayesian Model

Relation to the HGF

To test the flexibility of the precision-leakage Pseudo-Bayesian model proposed earlier, we compare its belief traces to the tendency trace from the HGF that was fitted before. We have then used our hyperparameter fitting algorithms to determine the α that would best describe the trajectory (representative samples are displayed in Figure 15,16).

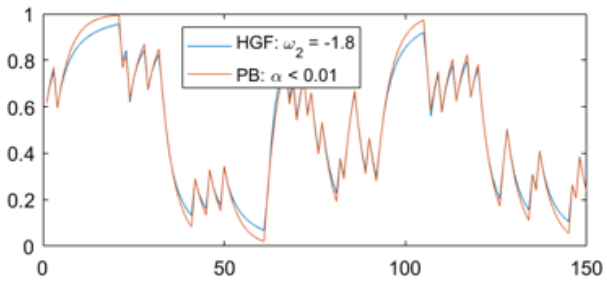

Figure 15: Pseudo-Bayes and HGF fit for large ω_2

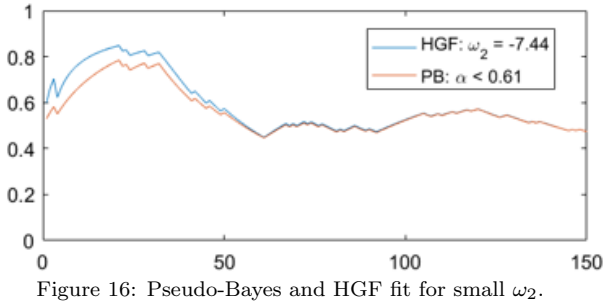

Figure 16: Pseudo-Bayes and HGF fit for small ω_2.

The small difference between both curves show that, in this example, we can, indeed, quite well describe the HGF trajectories in terms of our Pseudo-Bayesian model. After estimating α in this indirect fashion, we derived a relative weighting function that can be translated into a temporal window of integration (see Figure 17).

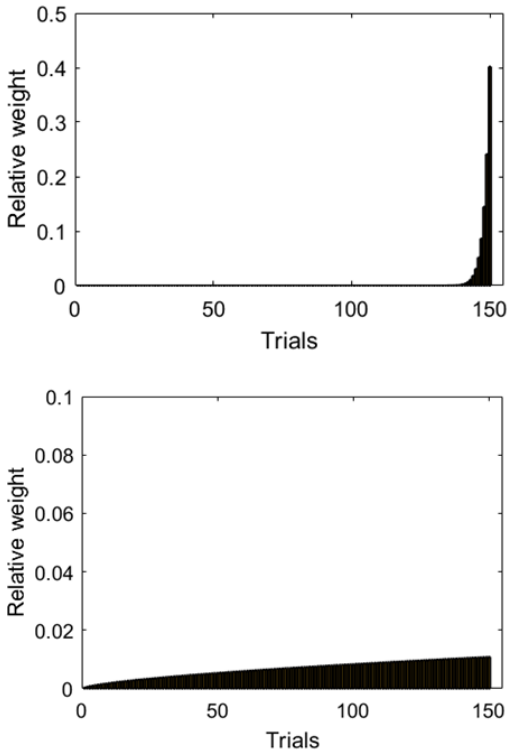

Figure 17: Extracting a temporal window of integration. The upper plot was generated from the belief trajectory in Figure 14, whereas the lower plot resulted from Figure 13.

Relation to the Rescorla-Wagner model

Part of the reason we chose Bayesian update equations as a starting point for developing our model is that it first enforces a Bayesian-inspired step-by-step idea derivation and secondly allows a comparison to hierarchical Bayesian models and non-hierarchical Bayesian models where prior or likelihood may be modified in order to account for certain phenomena. However, in terms of expressive power, there is a zoo of established models that allow a comparison, e.g. the Rescorla-Wagner (RW) model which culminates in:

$$\mu_k = \mu_{k-1} + \alpha \cdot (x_k - \mu_{k-1}) \tag{20}$$

When simulating pseudo-Bayesian Regression and the Rescorla-Wagner update rule with different values for α, we make a number of observations (see Figure 18).

1. In a certain range, the same posterior estimates can be achieved with both models.
2. In order to do so, the parameter value matching must be chosen anti-proportionally.
3. While Bayes-optimality can be modeled with a constant $\alpha = 1$ in case of our Pseudo-Bayesian model, the Bayes-optimal α for the RW model depends on the length and structure of the trial and can as such not be determined for online learning paradigms.

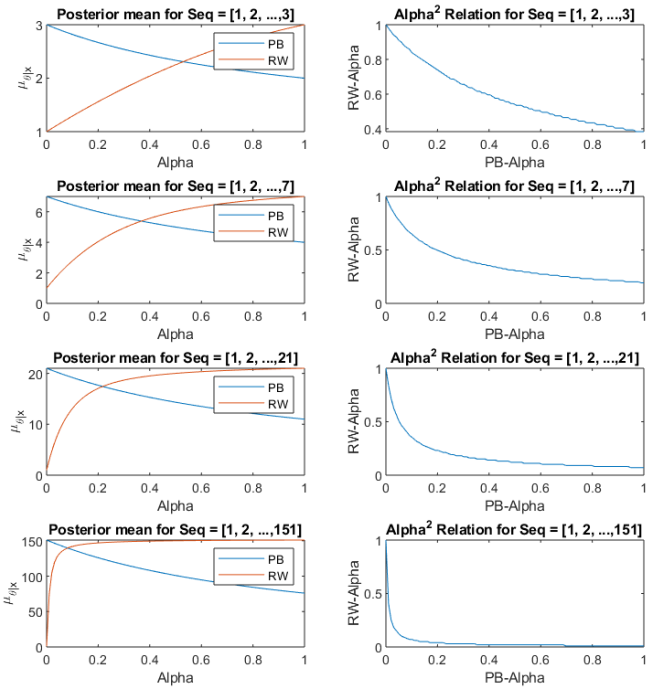

Figure 18: The effect of parameter value variation in the Pseudo-Bayesian and Rescorla-Wagner model.

Galvanic Skin Response

As outlined before, only the phasic component of the GSR signal (i.e. the skin conductance response, SCR) was analyzed, whereas the tonic component was discarded. For several variables of interest, the mean stimulus specific responses were inferred by cropping and averaging all time frames between 1 and 4.5 seconds post stimulus. The results of this process were averaged across the remaining 4 participants and are displayed in Figure 19.

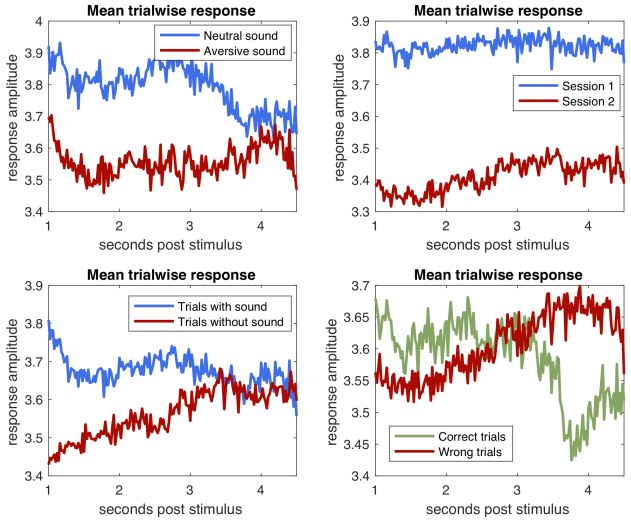

Figure 19: Grand average response of SCR signal depending on several variables of interest.

Surprisingly, their SCR were consistently higher for the first session compared to the second. More importantly, aversive sounds do not seem to have induced a higher response in general. The crucial subplot, however, is in the bottom right. It suggests a mediocre tendency towards an elevated signal for correct trials as well as a signal drop 3.5 sec post-stimulus for trials with a prediction error. While these tendencies are far from being indisputable, they yielded the opportunity to infer *canonical responses* for trials with and without prediction error, which were obtained via simple median filtering. This provided a pathway towards the first question we posed (see introduction). We verified to what extent inference on the beliefs is possible solely based on the physiological data obtained via the GSR device (in absence of any behavioral response). Based on the euclidean distance between the SCR of any trial with the two canonical responses, we generated a binary sequence of beliefs about the occurrence of the sound. However, examining the plain performance profiles of the inferred response traces, Figure 20 (left plot) reveals a significant drop for the physiologically inferred responses.

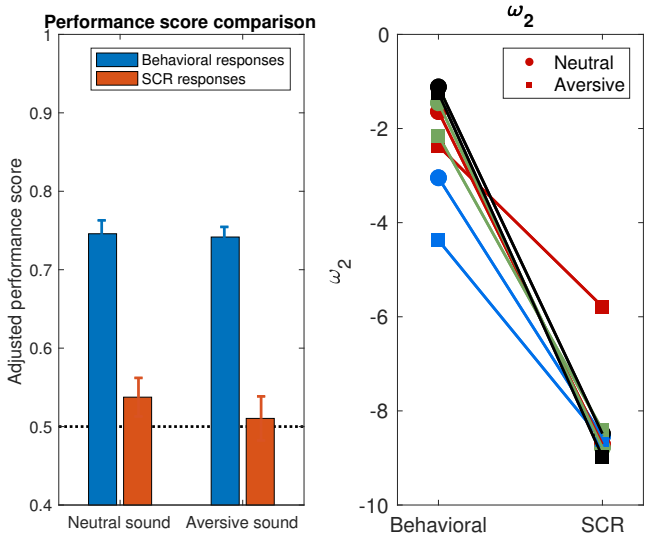

Figure 20: *Left:* Comparing the performances of the behaviorally recorded responses with the SCR inferred beliefs. Bars indicate confidence intervals, the dotted line indicates chance level. *Right:* Comparison of ω_2 parameter of HGF fits for behavioral and SCR data. Behavioral does refer to the default HGF fit (Model 1).

Indeed, the performance declines to a level just above chance. A binomial test could not reject the H_0 that the hypothetical responses derived from the SCR signal stem from the same distribution like random responses ($p = 0.11$), whereas the same test resulted in high significance ($p = 1e{-}66$) for the behavioral responses.

The right plot of Figure 20 shows a comparison of the ω_2 values of the HGF fitted to the behavioral and the SCR inferred responses. The default parameter settings were used for both models. For all participants and all sessions, a clear reduction in ω_2 can be observed. As ω_2 is inversely proportional to the volatility of the belief trajectory and the nature of our paradigm demands this volatility to be high in order to reach strong success rates, the drop in ω_2 may be precipitated by (1) a less (hierarchical) Bayes-optimal behavior or (2) inherent inconsistencies in the SCR belief traces and consequent failure of the HGF to fit them. Conclusively, this result does not provide evidence for our hypothesis that the beliefs about sound occurrence can be inferred from the skin conductance signal. Given that would be possible in this experiment, we had expected only marginal changes in both, the performance plot and the ω_2

comparison in Figure 20. In the discussion we outline several reasons for why this may be the case.
The HGF fits also supply ε_2, a precision-weighted PE between the outcome of each auditory event and its priorly estimated probability as well as ε_3, a precision weighted PE about the cue-outcome contingency. We compared the mean SCR for trials with low ε_2 (or ε_3) to the ones with high values (neglecting the error polarity). This did not unravel any significant insights.

Discussion

In the following, we will sequentially discuss the results of the 3 questions we raised in the introduction to motivate this project.

A. Do computational quantities also have *peripheral* physiological correlates?
Our analysis of the skin conductance response of 4 participants could not provide supporting evidence for the idea that a computational quantity like a prediction error about auditory stimulus outcome exhibits correlates encoded in peripheral physiological signals. This question was attempted by investigating to what extent inference on the belief of stimulus outcome can be made in the absence of behavioral responses. We can, however, report a difference in the SCR for correct and incorrect predictions about an auditory stimulus which may potentially be of qualitative nature. In addition, we have sketched a methodological pipeline on (1) how cognitive entities like beliefs about occurrences of sensory events may be inferred from physiological data and (2) how a quantitative comparison within established generative models of behavior can be conducted. Future projects should carefully consider amendments to the following aspects which may have caused the lack of supporting evidence:

1. The GSR device. Although our handcrafted setup was successfully tested prior to the experiment (it showed response increases due to physical exercising or suddenly getting scared), to our best knowledge, an `arduino` micro-controller has not been used previously in a scientific context to infer about SCR.

2. The amount of participants. Due to the aforementioned difficulties, our sample size ($N = 4$) was insufficient to allow strong conclusions in any direction.

3. Our expectation that the SCR inferred belief traces should yield similar performance scores and ω_2 distributions like the behavioral responses assumes that the behavioral responses indeed closely reflect the belief about cue-stimulus contingency (compare Figure 20).

4. The "canonical" skin conductance response for the prediction error was inferred from the same trials that were used to generate the SCR belief trace. This induced an inherent bias which could have been avoided by (1) cross validation or (2) a widely accepted (actually canonical) SCR signal, similar to the hemodynamic response function in fMRI. To our best knowledge, this canonical response does not yet exist in literature, but may be a worthwhile consideration for future work in the field.

B. Does stimulus nature bias behavior in a sensory-motor learning task?
Fitting the modified HGFs to data from neutral and aversive sessions individually and comparing the parameters revealed two subgroups in the examined cohort — with the discriminating factor being ω_2: In 4 subjects, ω_2 increased, whereas in the remaining 8 subjects, it decreased. With respect to the meaningfulness of this observation, we emphatically acknowledge our awareness of the fact that drawing a dozen numbers from a normal distribution twice and subtracting them pairwise would in all likelihood yield precariously similar "$\Delta\omega_2$". Only further investigations could reveal whether this observation indeed reveals itself as a red herring.
Cautionary statements aside, we further hypothesized that subjects whose ω_2 increased with aversive stimuli would generally tend to behave more erratically, and vice versa. In an attempt to link $\Delta\omega_2$ to a colloquially more tangible trait, namely general anxiety-proneness, we evaluated the relationship between subjective anxiety-proneness and $\Delta\omega_2$— unfortunately however, to little avail: the correlation of 0.37 prohibits us from drawing strong conclusions. In short, while our results certainly hint at its existence, the degree to which aversive stimuli bias behavior has yet to be convincingly disentangled from experimental noise.
Finally, we propose a reinterpretation of ω_2 in terms of a temporal window of integration, where an increase in ω_2 can be understood as a decrease in the extent of the temporal window with which past information is integrated into the present belief.

C. Can we derive a temporal interpretation of behavioral parameters that may serve as a diagnostic tool in computational psychiatry?
We developed a Pseudo-Bayesian framework to render the individual differences in ω_2 and ω_2-differentials intuitively accessible in the time domain. Specifically, we show that, under certain constraints, ω_2 can be understood in terms of a temporal window of integration. This measure may prove to be useful in understanding inter-individual differences in information processing; specifically, in the dynamic integration of sensory information over time. As such, it may represent one out of the many dimensions that are essential for understanding psychiatric diagnosis. Furthermore, it is a measure that is interpretable both by researchers and clinicians inside and outside of the Bayesian community. This may allow setting up joint research projects with, e.g., experimental neuroscientists interested in the accumulation and integration of sensory evidence over time.
A Pseudo-Bayesian follow-up to this project will be addressing a set of questions and concerns:

1. The similarity analysis was carried out on data that was recorded in an experiment with little to no variance of volatility. The effect of ω_3 and the relation to a Pseudo-Bayesian model with non-constant α needs to be studied more carefully.

2. A larger study will be carried out to compare ω_2 to the α of the Pseudo-Bayesian fit based on the HGF trajectory and to the α of the Pseudo-Bayesian fit based on the binary input sequence. This will be repeated for a diverse set of behavioral responses. So far, we have primarily worked with representative samples and fits based on HGF-trajectories to define the scope and power of the models.

3. The temporal quantities described above will be assessed in clinical and non-clinical subpopulations based on experimental recordings.

Acknowledgements

First and foremost, we would like to acknowledge Prof. Klaas Enno Stephan, for passionately and eloquently planting a very valuable seed in our minds, a seed that quickly found fertile ground and grew into a strong interest and personal involvement in the field of translational neuromodeling. All TNU members and guest lecturers who gave equally insightful, entertaining and inspiring lectures or comments deserve our utmost gratitude. Specifically, we owe our thanks to Lillian Weber, for providing invaluable, lucid feedback, disclosing some of the tricks of the trade (adjusting priors!), and assisting us in coding up the two-layer HGF configuration. Next, we acknowledge Jakob Heinzle for kindly providing the electrolyte paste that rendered our SCR measurements usable to begin with. Finally, Ethan Palmiere carefully injected some enlightening thoughts and Seraina Steiger kindly assisted us with LaTex advice.

References

Dominik R Bach, Jean Daunizeau, Karl J Friston, and Raymond J Dolan. Dynamic causal modelling of anticipatory skin conductance responses. *Biological psychology*, 85(1): 163–170, 2010.

Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

Maja Brydevall, Daniel Bennett, Carsten Murawski, and Stefan Bode. The neural encoding of information prediction errors during non-instrumental information seeking. *Scientific reports*, 8(1):6134, 2018.

Hanneke EM Den Ouden, Peter Kok, and Floris P De Lange. How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, 3:548, 2012.

Karl Friston. The history of the future of the Bayesian brain. *NeuroImage*, 62(2):1230–1233, 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2011.10.004. URL http://dx.doi.org/10.1016/j.neuroimage.2011.10.004.

Jessica R. Gilbert, Mkael Symmonds, Michael G. Hanna, Raymond J. Dolan, Karl J. Friston, and Rosalyn J. Moran. Profiling neuronal ion channelopathies with non-invasive brain imaging and dynamic causal models: Case studies of single gene mutations. *NeuroImage*, 124:43–53, 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.08.057. URL http://dx.doi.org/10.1016/j.neuroimage.2015.08.057.

Sandra Iglesias, Christoph Mathys, Kay H Brodersen, Lars Kasper, Marco Piccirelli, Hanneke EM den Ouden, and Klaas E Stephan. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, 80(2):519–530, 2013.

Sandra Iglesias, Sara Tomiello, Maya Schneebeli, and Klaas E Stephan. Models of neuromodulation for computational psychiatry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(3), 2017.

C Karenbach. Ledalab-a software package for the analysis of phasic electrodermal activity. *Internal report, Allgemeine Psychologie, Institut für Psychologie, Institut ftir Psychologie, Tech. Rep.*, 2005.

Jin Liu, Min Li, Yi Pan, Fang Xiang Wu, Xiaogang Chen, and Jianxin Wang. Classification of Schizophrenia Based on Individual Hierarchical Brain Networks Constructed from Structural MRI Images. *IEEE Transactions on Nanobioscience*, 16(7):600–608, 2017. ISSN 15361241. doi: 10.1109/TNB.2017.2751074.

Christoph D Mathys, Ekaterina I Lomakina, Jean Daunizeau, Sandra Iglesias, Kay H Brodersen, Karl J Friston, and Klaas E Stephan. Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in human neuroscience*, 8:825, 2014.

Martin O'Neill and Wolfram Schultz. Risk prediction error coding in orbitofrontal neurons. *Journal of Neuroscience*, 33(40):15810–15814, 2013.

Sukanta Saha, David Chant, Joy Welham, and John McGrath. A systematic review of the prevalence of schizophrenia. *PLoS Medicine*, 2(5):0413–0433, 2005. ISSN 15491277. doi: 10.1371/journal.pmed.0020141.

Victor I Spoormaker, KC Andrade, Manuel S Schröter, A Sturm, Roberto Goya-Maldonado, Philipp G Sämann, and Michael Czisch. The neural correlates of negative prediction error signaling in human fear conditioning. *Neuroimage*, 54(3):2250–2256, 2011.

Victor I Spoormaker, Jens Blechert, Roberto Goya-Maldonado, Philipp G Sämann, Frank H Wilhelm, and Michael Czisch. Additional support for the existence of skin conductance responses at unconditioned stimulus omission. *Neuroimage*, 63(3):1404–1407, 2012.

Mkael Symmonds, Catherine H Moran, M Isabel Leite, Camilla Buckley, Sarosh R Irani, Klaas Enno Stephan, Karl J Friston, and Rosalyn J Moran. Ion channels in EEG : isolating channel dysfunction in NMDA receptor antibody encephalitis. (May):1–12, 2018. ISSN 0006-8950. doi: 10.1093/brain/awy107.

T. M. van Leeuwen, H. E. M. den Ouden, and P. Hagoort. Effective Connectivity Determines the Nature of Subjective Experience in Grapheme-Color Synesthesia. *Journal of Neuroscience*, 31(27):9879–9884, 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0569-11.2011. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0569-11.2011.