ARTICLE **Communicated by Karl Friston**

Balancing New against Old Information: The Role of Puzzlement Surprise in Learning

Mohammadjavad Faraji
mjf.faraji@gmail.com
School of Computer and Communication Sciences and School of Life Sciences, Brain Mind Institute, École Polytechnique Fédéral de Lausanne, 1015 Lausanne EPFL, Switzerland

Kerstin Preuschoff
kerstin.preuschoff@unige.ch
Geneva Finance Research Institute and Center for Affective Sciences, University of Geneva, 1211 Geneva, Switzerland

Wulfram Gerstner
wulfram.gerstner@epfl.ch
School of Computer and Communication Sciences and School of Life Sciences, Brain Mind Institute, École Polytechnique Fédéral de Lausanne, 1015 Lausanne EPFL, Switzerland

Surprise describes a range of phenomena from unexpected events to behavioral responses. We propose a novel measure of surprise and use it for surprise-driven learning. Our surprise measure takes into account data likelihood as well as the degree of commitment to a belief via the entropy of the belief distribution. We find that surprise-minimizing learning dynamically adjusts the balance between new and old information without the need of knowledge about the temporal statistics of the environment. We apply our framework to a dynamic decision-making task and a maze exploration task. Our surprise-minimizing framework is suitable for learning in complex environments, even if the environment undergoes gradual or sudden changes, and it could eventually provide a framework to study the behavior of humans and animals as they encounter surprising events.

1 Introduction

To guide their behavior, humans and animals rely on previously learned knowledge about the world. Since the world is complex and models of the

K.P. and W.G. are co-senior authors.

Neural Computation **30, 34–83** (2018)
doi:10.1162/NECO_a_01025

world are never perfect, the question arises whether we should trust our internal world model that we have built from past data or readjust it when we receive a new data sample. Since a single data sample is not reliable in noisy environments, averaging over several data samples is normally a good strategy. However, when an unpredictable structural change occurs in the environment, the most recent data samples are the most informative ones, and we should put more weight on recent data samples than on older ones.

Indeed, both humans and animals adjust the relative contribution of old and newly acquired data on learning (Pearce & Hall, 1980; Behrens, Woolrich, Waltan, & Rushworth, 2007; Krugel, Biele, Mohr, Li, & Heekeren, 2009; Nassar et al., 2012) and rapidly adapt to changing environments (Pearce & Hall, 1980; Wilson, Boumphrey, & Pearce, 1992; Holland, 1997). To capture this behavior, two qualitatively different modeling approaches have been used. First, existing phenomenological models detect and respond to sudden changes using (absolute) reward prediction errors (Pearce & Hall, 1980; Hayden, Heilbronner, Pearson, & Platt, 2011; Roesch, Esber, Li, Daw, & Schoenbaumn, 2012), risk prediction errors (Preuschoff & Bossaerts, 2007; Preuschoff, Quartz, & Bossaerts, 2008), uncertainty-based jump detection (Nassar, Wilson, Heasly, & Gold, 2010; Payzan-Nestour & Bossaerts, 2011), or multi-timescale reward estimator comparison (Iigaya, 2016). Typically in these phenomenological models, a low-dimensional variable (related to the specific experiment) is used to trigger a rebalancing between new and old information.

Second, more principled statistical models involve either exact (Adams & MacKay, 2007; Behrens et al., 2007; Kolossa, Fingscheidt, Wessel, & Kopp, 2013; Kolossa, Kopp, & Finscheidt, 2015; Meyniel, Maheu, & Dehaene, 2016) or approximate (Yu & Dayan, 2005; Mathys, Daunizeau, Friston, & Stephan, 2011; Mathys et al., 2014; Gershman, Radulescu, Norman, & Niv, 2014) Bayesian updating of beliefs about the current model of the world. In the context of approximate Bayesian inference in a model with hidden variables, minimization of the variational free energy (Friston, 2010) reduces (a bound on) Shannon surprise (Shannon, 1948). These surprise-reducing, uncertainty-resolving assimilation schemes generally assume that the world is stationary and that random effects do not change with time. To accommodate fluctuations in the mean or in uncertainty, it is usually necessary to invoke hierarchical generative models such as the hierarchical gaussian filter model (Mathys et al., 2011, 2014). The disadvantage of complex models in a Bayesian framework is that model inversion, necessary for inference, typically involves nontrivial integrals for which it is not clear how the brain would compute these—even though Bayesian models of the brain have been highly successful in explaining data (Ernst & Banks, 2002; Körding & Wolpert, 2004; Knill & Pouget, 2004; Ma, Beck, Latham, & Pouget, 2006; Fiser, Berkes, Orban, & Lengyel, 2010; Berkes, Orban, Lengyel, & Fiser, 2011; Kolossa et al., 2013).

In this article, we introduce an alternative approach that does not call on inversion of deep hierarchical models yet still provides a flexible scheme of belief updating. In brief, the degree of belief updating is limited by an upper bound that is a monotonic function of surprise. This means that when subjects are surprised and uncertain, they pay more attention to new information and tip the balance in favor of new information relative to accumulated evidence in prior beliefs. We demonstrate our approach theoretically and in simulations.

Human surprise has attracted the attention of philosophers and experimental psychologists (Hurley, Dennett, & Adams, 2011). *Webster's Dictionary* defines *surprise* as "an unexpected event, piece of information" or "the feeling caused by something that is unexpected or unusual." Note that "unexpected" is different from "unlikely." An event can occur with very low probability without being unexpected: for example, you may park your car at the shopping mall next to a green BMW X3 with a license plate containing the number 5 without being surprised, even though the specific event is objectively very unlikely. But since you did not expect anything particular, this specific event was not unexpected. A pure likelihood-based definition of surprise, such as the Shannon information content (Shannon, 1948; Tribus, 1961), cannot capture this aspect. Note that something can be unexpected only if the subject is committed to a belief about what to expect (Schmidhuber, 2003). As Hurley et al. (2011) wrote, "What surprises us is . . . things we expected *not* to happen—because we expected something else to happen instead." In other words, surprise arises from a mismatch between a strong opinion and a novel event, but this notion needs a more precise mathematical formulation.

In practice, humans know when they are surprised (egocentric view), indicating that there are specific physiological brain states corresponding to surprise. Indeed, the state of surprise in other humans (observer view) is detectable as startle responses (Kalat, 2016) manifesting in pupil dilation (Hess & Polt, 1960; Preuschoff, t'Hart, & Einhauser, 2011) and tension in the muscles (Kalat, 2016). Neurally, the P300 component of the event-related potential (Squires, Wickens, Squires, & Donchin, 1976; Pineda, Westerfield, Kronenberg, & Kubrin, 1997; Missionier, Ragot, Derousne, Guez, & Renault, 1999) measured by electroencephalography (EEG) is associated with the violation of expectation (Squires et al., 1976; Verleger, Jaskowski, & Wauschkuhn, 1994; Kolossa et al., 2013, 2015; Meyniel et al., 2016). Furthermore, fMRI brain signals are correlated with surprise (Vossel et al., 2014; Iglesias et al., 2013). Finally, surprise drives attention (Itti & Baldi, 2009) and influences the development of sensory representations (Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001), as well as learning and memory formation (Hasselmo, 1999; Wallenstein, Hasselmo, & Eichenbaum, 1998; Ranganath & Rainer, 2003).

To implement our approach of surprise-based learning, we introduce a novel definition of confidence-corrected surprise that incorporates, and

extends, aspects of definitions of surprise that have been previously used in psychological theories of attention (Itti & Baldi, 2009), brain theories (Friston & Kiebel, 2009; Friston, 2010; Brea, Senn, & Pfister, 2013; Rezende & Gerstner, 2014), statistical models of information theory (Shannon, 1948), machine learning (Schmidhuber, 1991, 2010; Singh, Barto, & Chentanez, 2004; Sun, Gomez, & Schmidhuber, 2011; Frank, Leitner, Stollenga, Forster, & Schmidhuber, 2013), and active inference (Schmidhuber, 1991; Storck, Hochreiter, & Schmidhuber, 1995; Sun et al., 2011; Friston et al., 2015; Friston, Fitzgerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). In models of artificial curiosity, surprise is linked to sudden increases in learning progress (Schmidhuber, 1991), information gain (Storck et al., 1995), cumulative information gain (Sun et al., 2011), or algorithmic information gain measured by compression progress (Schmidhuber, 2006, 2010). Planning to be surprised so as to maximize information gain or epistemic value has been suggested as an optimal exploration technique in dynamic environments even in the absence of external reward (Storck et al., 1995; Sun et al., 2011; Little & Sommer, 2013; Frank et al., 2013; Joffily & Coricelli, 2013; Friston et al., 2015; Friston et al., 2017). In the framework of intrinsically motivated reinforcement learning (Singh et al., 2004; Oudeyer, Kaplan, & Hafner, 2007), researchers have defined ad hoc features (Singh et al., 2004; Sutton et al., 2011; Silver et al., 2016) or information-theoretic quantities (Mohamed & Rezende, 2015) that could replace the reward-prediction-error of classical reinforcement learning by a generalized model prediction error (Schmidhuber, 1991), which could be surprise related. However, surprise is important not just for active inference but also in the passive mode where a human subject receives auditory or visual stimuli (Squires et al., 1976; Kolossa et al., 2013, 2015; Meyniel et al., 2016). In this article, we do not consider active inference but limit our discussion to situations of passively perceiving a surprising event.

As the first aim of this article, we introduce a definition of confidence-corrected surprise that captures the notion of unexpectedness (as opposed to low probability) of an event. As a second aim, we study how surprise can influence learning. Similar to earlier theories (Schmidhuber, 2010; Friston et al., 2015, 2017) we study learning rules that minimize the surprise if the same data point appears a second time, but in contrast to these earlier theories, we start from our definition of confidence-corrected surprise as opposed to information gain measures (Schmidhuber, 2010) or a free-energy bound on the Shannon surprise (Friston et al., 2017). We demonstrate in two examples why surprising events increase the speed of learning and show that surprise can be used as a trigger to balance new information against old information. Before we turn to our definition of confidence-corrected surprise, we review existing theories.

2 Background: Theories of Surprise

A well-known saying states that when they listen to a joke, most people laugh twice: the first time when they hear the joke and the second time when they get it. This informal summary of observations (Hurley et al., 2011) suggests two different reasons for laughter linked to two different moments of surprise. The first moment of surprise occurs when we are puzzled: something seems to be wrong with our current interpretation of the story or, more generally, our current model of the world. We refer to this as "surprise in the sense of puzzlement" or, shorter, "puzzlement surprise." The second moment of surprise is when we understand the joke, when we have been enlightened, and have been able to give a different interpretation to the story or, more generally, an important update of our current model of the world. We refer to this as "surprise in the sense of enlightenment" or, shorter, "enlightenment surprise." Surprise in the sense of enlightenment also occurs when, after some hard work, we have understood an elegant proof of a mathematical theorem or hear a convincing explanation of a scientific discovery (Schmidhuber, 2010; Hurley et al., 2011; Friston et al., 2017).

With these notions in mind, we review some important existing theories of surprise (Shannon, 1948; Tribus, 1961; Baldi & Itti, 2010; Friston, 2010; Schmidhuber, 2010; Palm, 2012). Existing concepts can be roughly classified into two different categories.

First, the log likelihood of a single data point given a statistical model of the world has been called Shannon surprise or information content (Shannon, 1948; Tribus, 1961; MacKay, 2003; Palm, 2012). Formally, Shannon surprise is the information content of data point X calculated with the current world model of the subject. Let us introduce θ for the parameters of the (discrete or continuous) world model and θ^* for the specific parameter setting that has actually been used for generating the data. If the true causes of the data X (namely, θ^*) are known, the information content $-\ln p(X|\theta^*)$ for a specific outcome $X \in \mathcal{X}$ is the negative log likelihood of this data point (Tribus, 1961; Palm, 2012). In other words, the occurrence of a rare (i.e., unlikely to occur) data sample X is surprising. As the information content relates to the true probabilities $p(X|\theta^*)$ of samples in the real world, it is an objective, model-independent measure of unlikeliness.

However, for an application to human surprise, we will always work under the assumption that the true set of parameters θ^*, and thus the true probability $p(X|\theta^*)$, is not known to the observer, such that it is difficult to evaluate the exact information content of a data sample X. Let us denote by $\pi_n(\theta)$ our current belief (after having observed n data points) about the relevance of parameter value θ and introduce $p(X) = \int_\theta p(X|\theta)\pi_n(\theta)d\theta$ as the probability of data sample X after marginalizing over all possible model parameters. Note that the current belief $\pi_n(\theta)$ plays the role of a prior

in statistics. The Shannon surprise is defined as the negative-log-marginal-likelihood:

$$S = -\ln p(X) = -\ln \int_{\theta} p(X|\theta)\pi_n(\theta)d\theta. \tag{2.1}$$

Therefore, in the context of theories based on Shannon-surprise, an unlikely event becomes a surprising event. From our perspective, Shannon surprise and variants thereof incorporate an important aspect of what we have introduced as puzzlement surprise, but miss the influence of commitment to a belief. Empowerment in agents defined via mutual information (Mohamed & Rezende, 2015) or bounds on surprise in approximate Bayesian inference after minimization of the variational free energy (Friston, 2010; Joffily & Coricelli, 2013; Friston et al., 2017) fall roughly into the same class of models quantifying Shannon surprise.

Second, in the context of Bayesian models, surprise has been defined via the changes in model parameters during belief updating when assimilating a new data point $X = X_{n+1}$ (Storck et al., 1995; Itti & Baldi, 2009; Baldi & Itti, 2010). The belief after observation of n data points is given by the distribution $\pi_n(\theta)$ where θ are model parameters. Bayesian surprise introduced by Schmidhuber in the machine learning and active inference literature (Storck et al., 1995) and by Itti in the psychological literature (Itti & Baldi, 2009; Baldi & Itti, 2010) is defined as a KL divergence,

$$D_{KL}[\pi_n||\pi_{n+1}^{\text{Bayes}}], \tag{2.2}$$

between the prior belief $\pi_n(\theta)$ and the posterior belief $\pi_{n+1}^{\text{Bayes}}(\theta)$ that is calculated from the naive Bayes rule:

$$\pi_{n+1}^{\text{Bayes}}(\theta) = \frac{p(X|\theta)\pi_n(\theta)}{\int_{\theta} p(X|\theta)\pi_n(\theta)\, d\theta}. \tag{2.3}$$

Thus, in these theories, an event that causes a big change in the model of the world (characterized by the belief $\pi_n(\theta)$) becomes a surprising event (Schmidhuber, 1991). Optimization of (expected) free energy automatically contains Bayesian surprise in the belief update step (Friston et al., 2017). Surprise as successful progress in algorithmic compression of the agent's world model (Schmidhuber, 2006, 2010) is a non-Bayesian formulation of a related idea. Such a world model potentially includes the full history of all previous observations; making sense of observations then means finding a compressed representation of past experiences such that their storage take less memory space (Schmidhuber, 2006, 2010). In view of the paragraph at the beginning of this section, Bayesian surprise, or compression success, is a mathematical formulation of enlightenment surprise; it provides a

quantitative measure of how much our understanding of the world has improved after integrating data point X_{n+1} into our model.

Aspects of both Shannon surprise and Bayesian surprise can also be found in modern theories of approximate Bayesian inference. Variational methods in machine learning and Bayesian modeling use the free energy, which sets un upper bound on the Shannon surprise (MacKay, 2003; Friston, 2010; Joffily & Coricelli, 2013; Friston et al., 2017). To see where the bound comes from, let us consider the true posterior $\pi_n(\theta)$ after having seen the data sample X_n where θ refers to the hidden parameters, sometimes called sources or explanations of the data. More explicitly, the posterior can be written as $\pi_n(\theta) = p(\theta|X_n) = p(\theta, X_n)/p(X_n)$, and we will exploit these equivalent expressions in the next equation. Since the true sources θ are hidden, it is in general difficult to calculate the exact posterior $\pi_n(\theta)$ of complex Bayesian models. In variational methods, the true posterior is therefore approximated by another distribution $q_\mu(\theta)$, which is easier to manipulate mathematically. The index μ refers to parameters of this auxiliary distribution. To compare the auxiliary distribution with the (unknown) true distribution, we consider the Kullback-Leibler divergence,

$$\begin{aligned} D_{KL}[q_\mu||\pi_n] &= D_{KL}[q_\mu||p(\theta, X_n)/p(X_n)] \\ &= D_{KL}[q_\mu||p(\theta, X_n)] - [-\ln p(X_n)], \end{aligned} \tag{2.4}$$

where q_μ is a function of the hidden parameters θ. Since the Kullback-Leibler divergence involves an integration over θ, the result does not depend on θ. The last term on the right-hand side is the Shannon surprise of the nth data sample where $p(X_n) = \int p(X_n|\theta)\pi(\theta)d\theta$ is the likelihood of the data point. In the context of Bayesian modeling, the surprise is also called Bayesian model evidence (MacKay, 2003; Friston et al., 2015). The free energy of a single data point is defined as

$$F = D_{KL}[q_\mu||p(\theta, X_n)] \tag{2.5}$$

and can be interpreted in many different ways (MacKay, 2003; Friston, 2010; Joffily & Coricelli, 2013; Friston et al., 2015, 2017). Since the Kullback-Leibler divergence on the left-hand side of equation 2.4 is always positive, the free energy is an upper bound on the Shannon surprise. Therefore, the momentary value of the free energy for a data point X_n is, like the Shannon surprise, related to the puzzlement surprise.

Typically the parameters μ of the approximate distribution q have been optimized such that, averaged over many data samples, the free energy is as small as possible, which makes the bound on surprise as tight as possible. If the value of μ is optimized iteratively by minimizing the free energy after each new data sample, the resulting learning step will move the approximate model of the world characterized by q_μ closer to the true posterior π_n.

This update step is therefore, just like Bayesian surprise, linked to the enlightenment surprise or epistemic value of the last data point (Friston et al., 2015, 2017).

In the first part of the results, we introduce a new quantification of puzzlement surprise that includes the commitment to an opinion. We then use this measure for surprise minimization but limit the update step by a bound on the enlightenment surprise formulated as Bayesian surprise.

3 Results

In section 3.1, we introduce our notion of confidence-corrected surprise as a measure of puzzlement surprise and apply it to a few examples. In section 3.2, we derive a learning rule from the principle of surprise minimization. The subsequent sections apply this learning rule to two scenarios, starting with a one-dimensional prediction task, followed by a maze exploration corresponding to a parameter space with more than 200 dimensions. In contrast to the framework of active inference (Joffily & Coricelli, 2013; Friston et al., 2015, 2017), both tasks are formulated in the framework of passive observers driven by a stream of inputs.

3.1 Definition of Confidence-Corrected Surprise. We aim for a measure of puzzlement surprise that captures the notion of a mismatch between an opinion (current world model) and a novel event (data point) and should have the following properties:

1. The puzzlement surprise associated with an event depends not only on the statistical probability of the event; it depends as well on the agent's commitment to her belief.
2. With the same level of commitment to a belief, surprise decreases with the probability of an event.
3. For an event of a low probability, surprise increases with commitment to the belief.
4. A surprising event will influence learning.

While the final point will be the topic of the section 3.2, we now present a definition of *puzzlement surprise* and check properties (1) to (3) by way of a few illustrative examples.

To mathematically formulate puzzlement surprise, we assume that a subject receives data samples X from an environment that is complex, potentially high-dimensional, only partially observable, stochastic, or changing over time. In contrast to an engineered environment where we might know the overall layout of the world (e.g., a hierarchical Markov decision process) and learn the unknown parameters from data, we do not want to assume that we have knowledge about the layout of the world. Our world model may therefore be conceptually insufficient to capture the intrinsic structure of the world and would therefore occasionally make

wrong predictions even when we have observed large amounts of data. In short, our model of the world is expected to be simplistic and wrong, but since we know this, we should be ready to readapt the world model when necessary.

In our framework, we construct the world model from many instances of simple models, each one characterized by a parameter $\theta \in \mathbb{R}^N$. The probability of a data point X under model θ is $p(X|\theta)$. In a neuronal implementation, we may imagine that different instantiations of the model (with different parameter values θ) are represented in parallel by different (potentially overlapping) neuronal networks in the brain. If a new data point X is provided as input to the sensory layer, a model with parameter θ responds with an activity $\hat{p}_X(\theta)$ that we define to be proportional to $p(X|\theta)$. The distribution $\hat{p}_X(\theta)$, evaluated for fixed input X as a function of θ, represents the naive response of the whole brain network (i.e., of all models) in a setting where all the models are equally likely. Formally, $\hat{p}_X(\theta)$ is the posterior probability under a flat prior (see section 6.1). We refer to $\hat{p}_X(\theta)$ as the scaled likelihood of a naive observer. The scaled likelihood of a naive observer will serve as a reference (null model) in our definition of puzzlement surprise.

A given subject, however, will not consider all models as equally likely. Based on the past observation of n data points, the subject has formed an opinion that assigns to each model θ its relevance $\pi_n(\theta)$ for explaining the world. The probability of the new data point X under the current opinion is $p(X) = \int_\theta p(X|\theta)\pi_n(\theta)d\theta$, where $\pi_n(\theta)$ summarizes the current opinion of the subject and the integral runs over all possible model instantiations, be it a finite number or a continuum. In Bayesian modeling, the current opinion $\pi_n(\theta)$ is taken as a prior for the interpretation of the next data point.

However, for our definition of puzzlement surprise, we are not primarily interested in the probability of a data point but rather in the degree of commitment of the subject to a specific opinion. The commitment is defined as the negative entropy of the current opinion:

$$\text{Commitment} = -H(\pi_n) = \int_\theta \pi_n(\theta) \ln \pi_n(\theta) d\theta. \tag{3.1}$$

In case the belief distribution is a gaussian with variance σ^2, the commitment is identical to the precision defined as $1/\sigma^2$ (MacKay, 2003; Mathys et al., 2014) which is in turn related to the confidence of a subject when reporting her belief (Meyniel, Sigman, & Mainen, 2015; Meyniel & Dehaene, 2017). A subject with a high commitment to her opinion (low entropy or high precision of belief distribution) will be viewed as a confident subject.

A definition of *puzzlement surprise* needs to measure the mismatch of a perceived data point X_{n+1} with the current opinion. The current opinion

(after observation of n data samples $X_1, \ldots, X_n$) is characterized by the distribution $\pi_n(\theta)$. On the other hand, the observed data point $X = X_{n+1}$ would lead in a naive observer to the scaled likelihood $\hat{p}_X(\theta)$ already introduced. We define the puzzlement surprise as the Kullback-Leibler divergence between these two distributions:

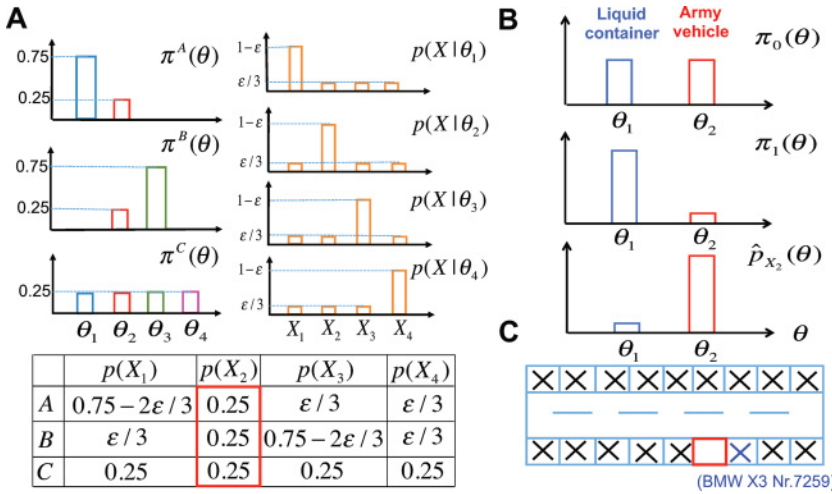
$$S_{cc}(X; \pi_n) = D_{KL}[\pi_n(\theta)||\hat{p}_X(\theta)] = \int_\theta \pi_n(\theta) \ln \frac{\pi_n(\theta)}{\hat{p}_X(\theta)}\, d\theta. \tag{3.2}$$

We call S_{cc} a confidence-corrected surprise because its definition includes the commitment to an opinion. Two aspects are important. First, confidence-corrected surprise is a measure of the puzzlement; therefore, it measures passive surprise and does not address the question of how the observer should update her model. Second, confidence-corrected surprise compares the current belief distribution always against the distribution of a naive observer; therefore, and in contrast to Bayesian surprise and related measures of enlightenment surprise, it does not compare the prior belief after n samples with the (posterior) belief after $n + 1$ samples, but rather the prior belief after n samples with the posterior belief of the naive observer (i.e., as if the new sample were the first one). Loosely speaking, whenever we receive a new data point, we compare the current model against the null model. To get acquainted with this unusual definition, let us look at a few examples.

First, imagine that three colleagues (A, B, and C) wait for the outcome of the selection of the next CEO. Four candidates are in the running. Suppose that we have four models, $\theta_1, \ldots, \theta_4$ where model θ_k means candidate $X = k$ wins with probability $(1 - \epsilon)$ (with small ϵ) and the remaining probability is equally distributed among the other candidates. Formally, the model (or basis function) with parameter θ_k predicts outcome probabilities $p(X = k|\theta_k) = 1 - \epsilon$, and $p(X = k'|\theta_k) = \epsilon/3$ for $k' \neq k$ (see Figure 1A, right).

The current opinion $\{\pi^A(\theta), \pi^B(\theta), \pi^C(\theta)\}$ of each colleague about the four possible models corresponds to the histogram in Figure 1A (left). Colleague A, who is usually well informed, has a weighting factor $\pi^A(\theta_1) = 0.75$ for the first model because he thinks the first candidate is likely to win. According to his opinion, the first candidate wins with probability $p^A(X = 1) = \sum_k p(X = 1|\theta_k)\pi^A(\theta_k)$, and he gives lower probabilities to the other candidates (see the Figure 1A table). Colleague B has heard rumors and favors the third candidate, while colleague C is uninformed as well as uninterested in the outcome and gives the same probabilities to each candidate. Note that colleagues A and B have the same commitment to their belief, $H(\pi^A) = H(\pi^B)$, but the likelihood of candidates differs. The commitment of colleague C is lower than that of A or B.

Evaluation of the confidence-corrected surprise measure indicates that (see appendix A.1 for the exact calculations):



	$p(X_1)$	$p(X_2)$	$p(X_3)$	$p(X_4)$
A	$0.75-2\varepsilon/3$	0.25	$\varepsilon/3$	$\varepsilon/3$
B	$\varepsilon/3$	0.25	$0.75-2\varepsilon/3$	$\varepsilon/3$
C	0.25	0.25	0.25	0.25

Figure 1: Examples of surprise. (A) A committed person is more surprised than an uncommitted one. Three colleagues A, B, and C have different beliefs (left) about the models $\theta_1, \ldots, \theta_4$ that determine the likelihood (right) that one of the four candidates is chosen as the CEO. Colleague A puts a lot of weight on the first model, indicating his preference for the first candidate (top left). The table (bottom) indicates the likelihood of each candidate (columns) being chosen as the CEO (after marginalization over all models) for each colleague (rows). If candidate 2 is selected (column highlighted by red color), subject C is less surprised than subjects A and B because C is not committed to a specific opinion, although the likelihood of candidate 2 being chosen is considered the same (0.25) for all three colleagues. (B) Surprise occurs only if a committed belief is disturbed. The first phrase of the goldfish joke transforms our belief about the meaning (θ) of *tank* from $\pi_0(\theta)$ (top) to $\pi_1(\theta)$ (middle). The second phrase then causes in a naive observer a distribution $\hat{p}_{X_2}(\theta)$ (bottom) that is very different from the last belief $\pi_1(\theta)$ (middle); thus, the listener is surprised. (C) For a driver who just cares about having a free parking place, finding a spot (empty red rectangle) next to a BMW X3 with the number plate 7259 (blue cross sign) is not very surprising, although it is very unlikely to occur.

1. If candidate 1 is selected, then A and B, despite having the same overall commitment to their belief, will be differently surprised due to different probabilities of candidate 1 in their models.
2. If candidate 2 is selected, then A is more surprised than if candidate 1 is selected because in his model, candidate 2 is less probable.
3. If candidate 2 is selected, then A will be more surprised than C. Although both colleagues assigned the same probability to this candidate, A's level of commitment to his belief is larger, which leads to a bigger surprise.

Note that if we take Shannon surprise as a measure of puzzlement, A and C will have the same amount of puzzlement surprise when candidate 2 is selected. This appears counterintuitive to us. Why should C be puzzled since he did not have a strong opinion in the first place? Note also that in order to compare the Bayesian surprise of A and C so as to find out that A is indeed more surprised than C, we would need to explicitly update both belief models *after* the election—which is tedious and potentially irrelevant since no future elections (with the same candidates) are on the horizon.

Second, we look at the theory of jokes developed by philosophers and cognitive psychologists (Hurley et al., 2011), which emphasizes that surprise in a joke can work only if the listener is committed to an opinion. Here is an example joke: "There are two goldfish in a tank. One turns to the other and says: 'You man the guns; I'll drive.'" The reason that some people find the joke funny is that "a perception of the world (manning the guns and driving the tank) suddenly corrects our mistaken preconception (tank as a liquid container)" (Hurley et al., 2011). Let us analyze the joke in the framework of our measure of confidence-corrected surprise. A naive English-speaking adult knows that *tank* can have two meanings: liquid container or a military vehicle (see Figure 1B, top). In the context of our theory, the two meanings correspond to two models—parameters θ_1 and θ_2, which have equal prior probability (opinion π_0). In the first sentence of the joke, the word *goldfish* (data point X_1) shifts the belief of the listener to a situation where he gives more weight to the liquid container. This becomes the opinion π_1 of the listener (see Figure 1B, middle). The opinion π_1 has low entropy, indicating a strong commitment. Now comes the second sentence, with the words *driving* and *guns*, which we may consider as data point X_2. These words trigger in a naive English-speaking adult a distribution $\hat{p}_{X_2}(\theta)$ (see Figure 1B, bottom) which favors the interpretation of tank as a military vehicle. Since the Kullback-Leibler divergence between the distributions in the second and third line is big, the listener is surprised. Similar conclusions could be drawn by theories that capture the enlightenment surprise by a change in belief (Storck et al., 1995; Itti & Baldi, 2009) or a progress in algorithmic compression (Schmidhuber, 2006, 2010). The difference is that our application of confidence-corrected surprise to a joke is meant to measure only the initial puzzlement-surprise. At this stage, the theory developed in this section does not make any statement about the moment when a listener gets the joke. The theory focuses on the moment when the listener realizes that something looks strange.

Third, let us return to the example of the green BMW X3 with a 5 in the license plate, mentioned in section 1. The probability of finding this type of car next to you in a shopping mall parking lot is extremely low (see Figure 1C), yet you are not surprised. If it were the parking lot of a company where every morning you see a little red car on this very same parking slot but today you see a green BMW, you might be surprised—quite

independent of the details of the green car. The difference arises from the degree of commitment.

The observations made in these examples can be mathematically formalized as follows:

- Our measure of surprise as defined in equation 3.2 is a linear combination of Shannon surprise and Bayesian surprise (and two further terms). Because it contains Shannon surprise as one of the terms, surprise decreases with increasing likelihood of the data under the current model (see section 6.2). This formal statement answers points 1 and 2 from the beginning of the section.
- Our measure of surprise as defined in equation 3.2 accounts for the differences in surprise between two subjects that reflect the differences in commitment to their opinion. In particular, a less confident individual (lower commitment to the current opinion) will generally be less surprised than a confident individual who is strongly committed to her opinion (see section 6.2). This formal statement answers point 3 from the beginning of the section.
- Our measure of surprise as defined in equation 3.2 can be computed rapidly because it uses only the scaled data likelihood (defined as the posterior of a naive observer given the new data point) and the degree of commitment to the current opinion. In particular, evaluation of surprise needs neither the lengthy evaluation of the posterior under the current model nor an update of the model parameters—in contrast to the Bayesian surprise model (Storck et al., 1995; Itti & Baldi, 2009; Baldi & Itti, 2010) with which our surprise measure otherwise shares important properties (see section 6.2). The question of how surprise relates to learning is the topic of the next section.

We emphasize that our measure of surprise is not restricted to discrete models but can also be formulated for models with continuous parameters θ (see section 6 and Figure 2A). Our proposed measure of surprise is consistent with formulations of Schopenhauer that link surprise to the "incongruity between representation of perception" (in our framework, the scaled likelihood response $\hat{p}_X(\theta)$ to a data X) and abstract representations (in our framework: the current opinion $\pi_n(\theta)$ formed from previous data points; freely cited after Hurley et al., 2011).

3.2 Surprise Minimization: the SMiLe-Rule. Successful learning implies an adaptation to the environment such that an event occurring for a second time is perceived as less surprising than the first time. In the following, *surprise minimization* refers to a learning strategy that modifies the internal model of the external world such that the unexpected observation becomes less surprising if it happens again in the near future. For example, successful compression of experiences after encountering a data sample X the first time may lead to enlightenment surprise via compression

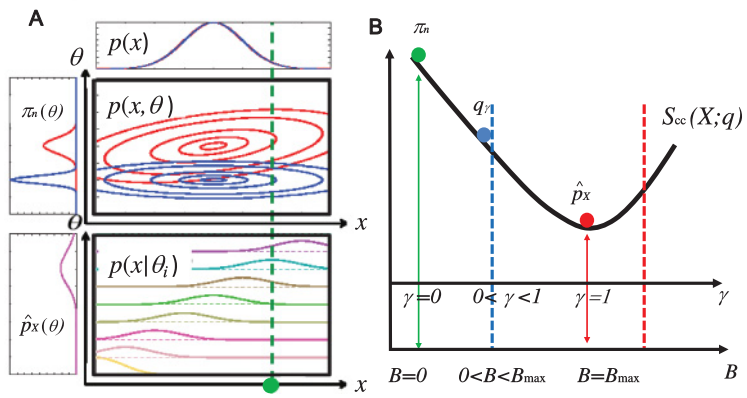


Figure 2: Confidence-corrected surprise and constraint surprise minimization. (A) Impact of confidence on surprise. Top: Two distinct internal models (red and blue), described by joint distributions $p(x, \theta)$ (contour plots) over observable data x and model parameters θ, may have the same marginal distribution $p(x) = \int_\theta p(x, \theta)d\theta$ (distributions along the x-axis coincide) but differ in the marginal distribution $\pi_n(\theta) = \int_x p(x, \theta)dx$ (distributions along the θ-axis). Surprise measures that are computed with respect to $p(x)$ neglect the uncertainty as measured by the entropy $H(\pi_n)$. Therefore, a given data sample X (green dot) may be equally surprising in terms of the raw surprise $S_{raw}(X)$ (see equation 6.3) but results in higher confidence-corrected surprise $S_{cc}(X)$ (see equation 3.2) for the blue as compared to the red model, because π_n in the red model is wider (corresponding to a larger entropy) than in the blue model. Bottom: The scaled likelihood $\hat{p}_X(\theta)$ (magenta) is calculated by evaluating the conditional probability distribution functions $p(x|\theta_i)$ (specified by different color for each θ_i) at $x = X$ (intersection of dashed green line with colored curves). The confidence-corrected surprise $S_{cc}(X)$ is the KL divergence between $\hat{p}_X(\theta)$ (bottom, magenta) and $\pi_n(\theta)$ (top, red). (B) Solutions to the (constraint) optimization problem in equation 3.5. The objective function, that is, the updated value of the surprise $S_{cc}(X; q)$ (black) for a given data sample X, is a parabolic landscape over γ where each γ corresponds to a unique belief distribution q_γ. Its global minimum is at $\gamma = 1$ (corresponding to $q_1 = \hat{p}_X$), which is equivalent to discarding all previously observed samples. The boundary B constrains the range of γ and thus the set of admissible belief distributions. At $B = 0$, no change is allowed, resulting in $\gamma = 0$ with an updated belief equal to the current belief π_n (green). $B \geq B_{\max} = D_{KL}[\hat{p}_X||\pi_n]$ (red dashed line) implies that we allow updating the belief to a distribution further away from the current belief than the sample itself so the optimal solution is the scaled likelihood $\hat{p}_X$ or $\gamma = 1$ as for the unconstrained problem. For $0 < B < B_{\max}$ (blue dashed line) the objective function is minimized by q_γ in equation 3.6 that fulfills the constraint $D_{KL}[q_\gamma||\pi_n] = B$ with $0 < \gamma < 1$.

progress, but for independent data samples, successful compression also implies that there will be less compression progress when seeing it a second time (Schmidhuber, 2006, 2010). Analogously, in variational Bayesian models, minimization of the free energies lowers an upper bound on the Shannon surprise so that the average surprise when encountering a data sample a second time is reduced (MacKay, 2003; Friston, 2010). Here we follow a similar surprise minimization strategy, except that we work with a confidence-corrected surprise instead of Shannon surprise or compression success and we set a bound on the maximally allowed change in belief. Surprise minimization is akin to, though more general than, reward prediction error learning. Reward-based learning modifies the reward expectation such that a recurring reward results in a smaller reward prediction error. Similarly, surprise-minimization learning results in a smaller surprise for recurring events (Schmidhuber, 2006; Friston et al., 2017).

To mathematically formulate learning through surprise minimization, we define a learning rule $L(X, \pi_n)$ as a mapping from the current belief $\pi_n(\theta)$ to a new belief $\pi_{n+1}(\theta) = q(\theta)$ after receiving data sample X: $q = L(X, \pi_n)$. The learning step after a single data sample will be called *belief update*.

We define the class $\mathcal{L}$ of plausible learning rules as the set of those learning rules L for which the surprise $\mathcal{S}(X; q)$ of a given data sample X under the new belief $q(\theta)$ is at most as surprising as the surprise $\mathcal{S}(X; \pi_n)$ of that data sample under the current belief $\pi_n(\theta)$:

$$\mathcal{L} = \{L : \mathcal{S}(X; q) \leq \mathcal{S}(X; \pi_n),\ q = L(X, \pi_n), \forall X \in \mathcal{X}\}. \tag{3.3}$$

In other words, if the same data sample X occurs a second time right after a belief update, it is perceived as less surprising than the first time.

After the belief update, we have a new belief $\pi_{n+1} = q$, and we may ask how much the data X have affected the internal model. To answer this question, we compare the surprise of data sample X under an arbitrary new belief q with that under the previous belief:

$$\Delta\mathcal{S}(q; \pi_n, X) = \mathcal{S}(X; \pi_n) - \mathcal{S}(X; q). \tag{3.4}$$

First, since the update under a reasonable learning rule cannot lead to an increase in surprise, we have $\Delta\mathcal{S}(q; \pi_n, X) \geq 0$. Second, for an update scheme with learning rule L, we can compare different data samples X and X'. A data sample X is considered more effective for a belief update than X', if $\Delta\mathcal{S}(q; \pi_n, X) > \Delta\mathcal{S}(q; \pi_n, X')$. Note that definitions in equations 3.3 and 3.4 do not depend on our specific choice of surprise measure $\mathcal{S}$. In the following we choose $\mathcal{S}$ to be the confidence-corrected surprise S_{cc} (see equation 3.2).

The impact function $\Delta S_{cc}(q; \pi_n, X)$; (see equation 3.4) for a given data sample X is maximal if the update makes the new belief distribution $q(\theta)$

equal to the scaled likelihood $\hat{p}_X(\theta)$. In this case, the confidence-corrected surprise vanishes after the update. However, as the new distribution $q = \hat{p}_X$ does not depend on the current belief π_n, it discards all previously learned information. Therefore, it amounts to overfitting the new data point.

To avoid overfitting to the last data sample, we need a regularizer. In order to limit our search to beliefs q that are not too different from the current opinion π_n, we consider only beliefs q that fulfill the constraint $D_{KL}[q||\pi_n] \leq B$ with an upper bound $B \geq 0$. The parameter B determines how much we allow our belief to change after receiving a data sample X. Note that the Kullback-Leibler divergence $D_{KL}[q||\pi_n]$ is closely related to the Bayes surprise in equation 2.2 except for a change of arguments. Thus, our regularizer limits the maximally allowed Bayes' surprise.

Maximizing the impact function $\Delta S_{cc}(q; \pi_n, X)$ under the above bound yields the following constraint optimization problem:

$$\min_{q: D_{KL}[q||\pi_n] \leq B} S_{cc}(X; q). \tag{3.5}$$

Using the method of Lagrange multipliers, we find (see section 6.3) the solution of the minimization problem in equation 3.5 to be

$$q_\gamma(\theta) = \frac{p(X|\theta)^\gamma \pi_n(\theta)^{1-\gamma}}{Z(X; \gamma)}, \tag{3.6}$$

where $Z(X; \gamma) = \int_\theta p(X|\theta)^\gamma \pi_n(\theta)^{1-\gamma}\, d\theta$ is a normalizing factor and the parameter γ with $0 \leq \gamma \leq 1$ is uniquely determined by the bound B. More precisely, γ is linked to B by a function $\gamma = F(B)$ that increases monotonously in the range for $0 \leq \gamma < 1$ (see appendix A.2 for the proof). Thus, once B has been chosen, γ is no longer a free parameter and vice versa. Learning is implemented by using the solution of equation 3.6 as the new opinion: $\pi_{n+1}(\theta) = q_\gamma(\theta)$.

Learning by updating according to equation 3.6 will be called the *surprise minimization learning (SMiLe)*, and we will refer to equation 3.6 as the SMiLe rule. The update step of the SMiLe rule is reminiscent of Bayes' rule except for the parameter γ, which modulates the relative contribution of the likelihood $p(X|\theta)$ and the current belief $\pi_n(\theta)$ to the new belief $\pi_{n+1}(\theta) = q_\gamma(\theta)$. Note that the SMiLe rule belongs to the class $\mathcal{L}$ of plausible learning rules for all $0 \leq \gamma \leq 1$.

Choosing γ in the range $0 \leq \gamma \leq 1$ is equivalent to choosing a bound $B \geq 0$. To understand how the optimal solution in equation 3.6, and thus γ, relates to the boundary B, we illustrate its limiting cases (see Figure 2B)

1. $B = 0$ yields $\gamma = 0$ and the new belief q is identical to the current belief π_n. In other words, the new information is discarded.

2. For $B \geq B_{\max} = D_{KL}[\hat{p}_X||\pi_n]$, the solution is always the scaled likelihood $\hat{p}_X$ (corresponding to $\gamma = 1$) because $q = \hat{p}_X$ fulfills the constraint $D_{KL}[q||\pi_n] \leq B$ for any $B \geq B_{\max}$ and minimizes $S_{cc}(X; q)$ among all possible belief distributions q. This is equivalent to the unconstrained case and implies that all previous information is discarded.
3. For $0 < B < B_{\max}$ the optimal solution is the new belief q_γ, equation (3.6), with $0 < \gamma < 1$ satisfying $D_{KL}[q_\gamma||\pi_n] = B$. Moreover, $B > B'$ implies $\gamma > \gamma'$ (see Figure 2B and appendix A.2 for the proof).

While the SMiLe rule, equation 3.6, depends on a parameter γ that is uniquely determined by the bound B, we have yet to indicate how to choose B. Insights of hierarchical Bayesian models linked to human behavior suggest that highly surprising data should result in larger belief shifts (Mathys et al., 2014; Meyniel et al., 2015; Meyniel & Dehaene, 2017). Therefore, the bound B should increase with the level of confidence-corrected surprise S_{cc}.

The definition of an optimal (nonlinear) mapping from S_{cc} to B (and thus to γ) would require further assumptions about how surprise is related to the bound, and we will therefore not search for optimality. However, it is instructive to study a few examples. For instance, if the nonlinear mapping were a step function, the system would make a binary choice between either keeping the old belief or relying on the last new data point. An extremely slow increase would amount to largely ignoring the surprise and sticking to the same old belief. Therefore, the sharpness of the transition in the mapping function matters. The exact link between the bound and surprise is, however, not crucial as long as B is monotonic in surprise in a reasonable way.

In the following, we choose a simple monotonic function to link the bound to the surprise. For each data sample X, we take

$$B(X) = \frac{mS_{cc}(X; \pi_n)}{1 + mS_{cc}(X; \pi_n)} B_{\max}(X), \tag{3.7}$$

where $B_{\max}(X) = D_{KL}[\hat{p}_X||\pi_n]$. Here, the monotonic function depends on a subject-specific parameter m that describes an organism's propensity toward changing its belief. Note that in equation 3.7, $m = 0$ indicates that the subject will never change her belief. As m increases, so does a subject's willingness to change her belief. We expect that differences in m from one subject to the next will eventually allow us to capture heterogeneity in belief update strategies when fitting human behavior. Although m is inserted in equation 3.7 to model subject dependence, one could also search for the best m algorithmically in a given simulated environment or other computational setting.

Note that biological correlates of surprise such as pupil dilation (Hess & Polt, 1960; Preuschoff et al., 2011) or the activity of a neuromodulator (Yu

Algorithm 1: Pseudo-Algorithm for Surprise-Modulated Belief Update (SMiLe).

$N \leftarrow$ number of data samples

$\text{Belief} \leftarrow \pi_0$ (the current belief)

$m \leftarrow 0.1$ (subject-dependent)

for n: 1 to N **do**

  $X_n \leftarrow$ a new data sample

  (i) evaluate the surprise $S_{cc}(X_n; \text{Belief})$, equation 3.2

  (ii-a) calculate $B_{max}(X_n) = D_{KL}[\hat{p}_{X_n}||\text{Belief}]$

  (ii-b) choose the bound $B(X_n) = \frac{mS_{cc}(X_n;\text{Belief})}{1+mS_{cc}(X_n;\text{Belief})} B_{max}(X_n)$

  (iii) find γ by solving $D_{KL}[q_\gamma||\text{Belief}] = B(X_n)$

  (iv) update using SMiLe, equation 3.6 $\text{Belief}(\theta) \leftarrow \frac{p(X_n|\theta)^\gamma \text{Belief}(\theta)^{1-\gamma}}{\int_\theta p(X_n|\theta)^\gamma \text{Belief}(\theta)^{1-\gamma}\, d\theta}$

Return Belief;

Note: In each iteration, we first calculate the surprise, step i, before the model is updated in step iv. Steps ii-a, ii-b, and iii can be merged and approximated by $\gamma = f(S_{cc}(X_n; \text{Belief}))$ where $f(.)$ increases with surprise.

& Dayan, 2005; Iglesias et al., 2013; Joffily & Coricelli, 2013; Vossel et al., 2014) will normally saturate at some maximal value, consistent with our choice of a saturating function in equation 3.7. The relation of our approach to existing data and theory in neurosciences is elaborated in section 4.

3.3 Surprise-Modulated Belief Update. The surprise-modulated belief update combines the confidence-corrected surprise, equation (3.2) and the SMiLe rule, equation 3.6, to dynamically update our belief: after receiving a new data point X, we evaluate the surprise $S_{cc}(X; \pi_n)$, which sets the bound B, equation 3.7, for our update and allows us to solve for γ. We then update the belief, using the SMiLe rule, equation 3.6, with parameter γ (see algorithm 1).

The parameter γ in the SMiLe rule controls the impact of a data sample X on belief update such that a bigger γ causes a larger impact. More precisely, the impact function $\Delta S_{cc}(q; \pi_n, X)$ in equation 3.4 with the SMiLe rule, equation 3.6, as the update scheme is an increasing function of γ (see appendix A.3 for the proof).

We note that in classical models of perception and attention (Itti & Baldi, 2009; Baldi & Itti, 2010), Bayesian surprise has been defined as a measure of belief change such as $D_{KL}[\pi_{n+1}||\pi_n]$ or its mirror form, $D_{KL}[\pi_n||\pi_{n+1}]$, where π_{n+1} is calculated by Bayes' formula, equation 2.3. The fact that Bayesian surprise is known only after the belief has been changed makes it a slow measure of surprise, likely to correspond to the moment of enlightenment surprise. We emphasize that our model of confidence-corrected surprise is "fast" in the sense that it can be evaluated before the beliefs are changed—as it should be as a measure of puzzlement surprise. The update step itself, however, is slow and can be linked to enlightenment surprise. Indeed, the impact function $\Delta S_{cc}(q; \pi_n, X)$ is given by (see appendix A.4 for derivation),

$$\Delta S_{cc}(q; \pi_n, X) = \frac{1}{\gamma} D_{KL}[\pi_n||q] + \left(\frac{1}{\gamma} - 1\right) D_{KL}[q||\pi_n] \geq 0, \tag{3.8}$$

where q is the new belief calculated with the SMiLe rule, equation (3.6). Thus, the impact function is closely linked to Bayesian surprise. Therefore, a larger reduction in the puzzlement surprise causes a bigger change in belief and therefore a larger enlightenment surprise. As an aside, we note that theories of active inference choose the next action so as to maximize the enlightenment surprise (Sun et al., 2011; Little & Sommer, 2013; Frank et al., 2013; Joffily & Coricelli, 2013; Friston et al., 2015, 2017).

3.4 Simulations. We now look at two examples to illustrate the functionality of our proposed surprise-modulated belief update algorithm 1. The first is a simple, one-dimensional dynamic decision-making task that has been used in behavioral studies (Behrens et al., 2007; Nassar et al., 2012) of learning under uncertainty. While somewhat artificial as a task, it is appealing as it nicely isolates different forms of uncertainty. This allows us to demonstrate the basic quantities and properties of our algorithm and show how its flexibility allows it to capture a wide range of behaviors. The second example is a multidimensional maze exploration task that we use to demonstrate how our algorithm extends to and performs in more complex and realistic experimental environments.

3.4.1 Gaussian Estimation. Task. In the one-dimensional dynamic decision-making task, subjects are asked to estimate the mean of a distribution based on consecutively and independently drawn samples. At each time step n, a data sample X_n is drawn from a normal distribution $\mathcal{N}(\mu_n, \sigma_x^2)$ and the subject is asked to provide her current estimate $\hat{\mu}_n$ of the mean of the distribution. Throughout the experiment, the mean may change without warning (see Figure 3A). Changes occur with a hazard rate of $H = 0.066$. In Figures 3C and 3D, the hazard rate H is varied. The variance σ_x^2 remains fixed.

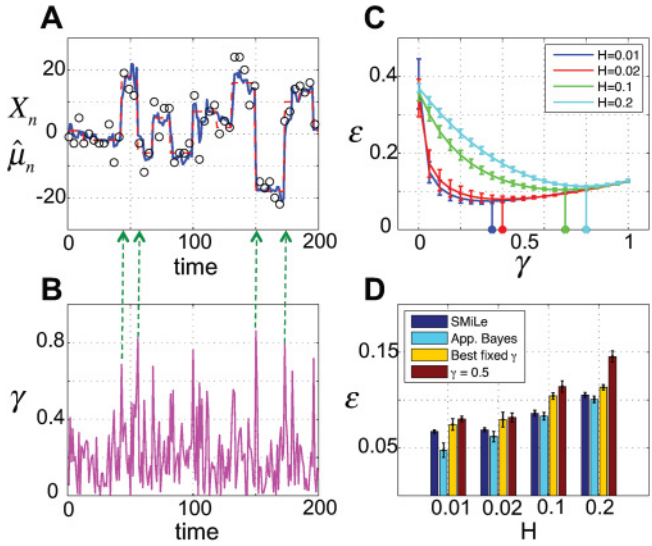


Figure 3: Gaussian mean estimation task. At each time step, a data sample X_n is independently drawn from a normal distribution whose underlying mean may change within the interval $[-20, 20]$ at unpredictable change points. On average, the underlying mean remains unchanged for 15 time steps corresponding to a hazard rate $H = 0.066$. The standard deviation of the distribution is fixed to 4 and is assumed to be known to the subject. (A) Using a surprise-modulated belief update (see algorithm 1), the estimated mean (blue) quickly approaches the true mean (dashed red) given observed samples (black circles). A few selected change points are indicated by green arrows. (B) The weight factor γ in equation 3.10 (magenta) increases at the change points, resulting in higher influence of newly acquired data samples on the new value of the mean. (C) The estimation error ϵ per time step versus the weight factor $0 \leq \gamma \leq 1$ in the delta rule method with constant γ for four different hazard rates. The minimum estimation error (for best fixed γ) is achieved by a γ (points on the horizontal axis) that decreases with the hazard rate, indicating that a bigger γ is preferred in volatile environments. Error bars indicate standard deviation over all trials and 50 episodes. (D) For all models, the average estimation error ϵ increases with the hazard rate. Moreover, surprise-modulated belief update (SMiLe, dark blue) outperforms the delta rule with the best fixed γ (best fixed γ, yellow). The best fixed γ for each hazard rate corresponds to the learning rate that has minimal estimation error (indicated by points on the horizontal axis in panel C). Although the surprise-modulated SMiLe rule performs worse than the approximate Bayesian delta rule (Nassar et al., 2010) (App. Bayes, light blue), the difference in the performance is not significant, except for the very small hazard rate of 0.01.

Model. We model the subject's belief before the nth sample X_n is observed, as the normal distribution $\mathcal{N}(\hat{\mu}_{n-1}, \sigma^2_{n-1})$ where $\hat{\mu}_{n-1}$ is the estimated mean and σ^2_{n-1} determines how uncertain the subject is about her estimation. In order to keep the scenario as simple as possible, we assume $\sigma^2_0 = \sigma^2_x$. Thus, in terms of the confidence-corrected surprise defined in equation 3.2, the unknown mean μ plays the role of the unknown parameter θ, and the belief distribution is described as a gaussian centered on $\hat{\mu}$, which we interpret as the most likely estimate of μ.

Results for the estimation task. We find that the updated value of the mean $\hat{\mu}_n$ resulting from the surprise-modulated belief update (see algorithm 1) is a weighted average of the current estimate of the mean $\hat{\mu}_{n-1}$ and the new sample X_n (see section 6.4 for derivation),

$$\hat{\mu}_n = \gamma X_n + (1 - \gamma)\hat{\mu}_{n-1}. \tag{3.9}$$

The weight factor, which determines to what extent a new sample X_n affects the new mean $\hat{\mu}_n$, is determined by γ, which increases with the surprise $S_{cc}(X_n)$ of that sample (see Figure 3B):

$$\gamma = \sqrt{\frac{mS_{cc}(X_n)}{1 + mS_{cc}(X_n)}}, \quad S_{cc}(X_n) = \frac{(X_n - \hat{\mu}_{n-1})^2}{2\sigma^2_x}. \tag{3.10}$$

Note that in this example, the confidence-corrected surprise measure is related to the normalized unsigned prediction error $|X_n - \hat{\mu}_{n-1}|/\sigma_x$. This outcome of our SMiLe update is consistent with recent approaches in reward learning that suggest rewarding prediction errors scaled by standard deviation or variance (Preuschoff & Bossaerts, 2007).

We can interpret the estimate $\hat{\mu}$ in equation 3.9 as an exponential filter over past data points with a filter constant related to γ, in agreement with Bayesian estimates of means in a world that switches at unknown times (Yu & Cohen, 2008). However, in our model, the filter parameter γ is modulated by the confidence-corrected surprise, which increases suddenly in response to the samples immediately after the change points. As a consequence, surprising samples increase the influence of a new data sample on the estimated mean (see Figure 3B). This result of the SMiLe rule for confidence-corrected surprise therefore leads to a result that is similar in spirit to updates in hierarchical gaussian filter models (Mathys et al., 2011, 2014) where the variance of the belief at the lower level is controlled by the estimated mean at the next higher level. In these models, the update step is variable and depends on the precision (inverse variance) of the beliefs (Mathys et al., 2011, 2014). Thus, both approaches, the hierarchical gaussian filter model and the SMiLe rule, lead to a delta rule for the estimated mean with an adaptive update rate. The advantage of hierarchical gaussian filter models is that they provide a systematic approach in a Bayesian modeling

framework so as to track both mean and precision of the belief (Mathys et al., 2011, 2014). The potential advantage of our confidence-corrected surprise measure is that it allows a rapid evaluation of surprise that directy modulates, via the SMiLe rule, the weighting factor γ, albeit in a less rigorous framework.

We compared our surprise-modulated belief update, equations 3.9 and 3.10, with a delta rule, equation 3.9, with constant weighting factor γ. To enable a fair comparison, we consider two situations: (1) we arbitrarily fix γ at 0.5 or (2) for a given hazard rate H, we first search for the optimal value of fixed γ so as to minimize the estimation error (see Figure 3C). We find that our surprise-modulated belief update outperforms the delta rule with any constant update rate (see Figure 3D). This clearly shows that an adaptive update rate is preferable to a fixed update rate.

We also compared our proposed algorithm with a delta rule that approximates the optimal Bayesian solution (Nassar et al., 2010). In the optimal model, the subject knows a priori that the mean will change at unknown points in time—that is, the subject makes use of a hierarchical statistical model of the world. The algorithm proposed in Nassar et al. (2010) provides an efficient approximate solution to estimate the parameters of the hierarchical model. In this algorithm, the subject increases the update rate as a function of the probability of encountering a change point at a given time step. This probability requires knowledge or online estimation of the hazard rate, which indicates how frequently change points occur. Although our surprise-modulated belief update does not outperform the approximate Bayesian delta rule, the difference in performance is, in most cases, not significant (see Figure 3D). In other words, our method, which does not require any information about the hazard rate, can almost reach the quality of the optimal Bayesian solution, with significantly reduced computational complexity. Note that the SMiLe rule is not designed for (almost) stationary environments where no fundamental change in context occurs. Therefore, in the case where the true mean is constant (low hazard rate), the SMiLe rule results in increased volatility in estimation. This is why the difference in performance of SMiLe and the optimized Bayesian delta rule becomes more evident for smaller hazard rates than bigger ones (see Figure 3D).

3.4.2 Maze Exploration. Task. The maze exploration task is similar to tasks used in behavioral neuroscience and robotics (Morris, Garrard, Rawlins, & O'Keefe, 1982; Gillner & Mallot, 1998; Nelson, Grant, Galeotti, & Rhody, 2004; Sun et al., 2011; Rezende & Gerstner, 2014). There are two environments $\mathcal{A}$ and $\mathcal{B}$, each composed of the same uniquely labeled (e.g., by colors or cue cards) rooms. $\mathcal{A}$ and $\mathcal{B}$ differ only in the spatial arrangement (topology) of rooms (see Figure 4). Neighboring rooms are connected and accessible through doors. Initially the agent is placed into either $\mathcal{A}$ or $\mathcal{B}$. At each time step, a door of the current room opens and the agent moves into the adjacent room, thus exploring the environment following a completely

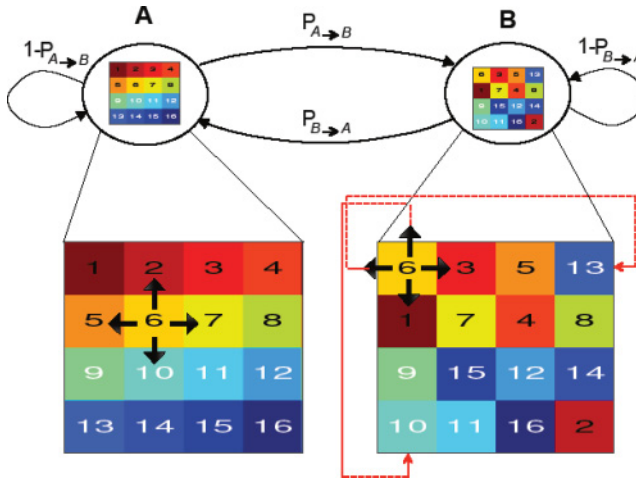


Figure 4: Maze exploration task. Environments $\mathcal{A}$ (left) and $\mathcal{B}$ (right) both consist of 16 rooms but differ in topology. At each time step, one of the four available doors (up, down, right, left) in the current room (e.g., $s = 6$) is randomly opened (with probability 0.25). While the learning agent is in environment $\mathcal{A}$, the environment may change to $\mathcal{B}$ with probability $P_{\mathcal{A}\to\mathcal{B}} \leq 0.1$ in the next time step of duration Δt. Similarly, $P_{\mathcal{B}\to\mathcal{A}}$ indicates the environment switches from $\mathcal{B}$ to $\mathcal{A}$. Therefore, as the agent starts moving out of state $s = 6$, depending on the current environment and switch probabilities $P_{\mathcal{A}\to\mathcal{B}}$ and $P_{\mathcal{B}\to\mathcal{A}}$, it will end up in environment $\mathcal{A}$ (i.e., $s' \in \{2, 10, 7, 5\}$) or $\mathcal{B}$ (i.e., $s' \in \{10, 1, 3, 13\}$). The duration of a stay in environment $\mathcal{A}$ is therefore exponentially distributed with mean $\tau_{\mathcal{A}} = \Delta t / P_{\mathcal{A}\to\mathcal{B}}$, where the parameter $\tau_{\mathcal{A}}$ determines the timescale of stability in environment $\mathcal{A}$; for larger $\tau_{\mathcal{A}}$, an agent has more time for adapting to $\mathcal{A}$ after a change point. The expected fraction of time spent in total within environment $\mathcal{A}$ is equal to $\psi_{\mathcal{A}} = P_{\mathcal{B}\to\mathcal{A}}/(P_{\mathcal{B}\to\mathcal{A}} + P_{\mathcal{A}\to\mathcal{B}})$. Note that $\tau_{\mathcal{A}}$ and $\psi_{\mathcal{A}}$ are two free parameters that we can change to study how the agent performs in different circumstances (e.g., see Figure 8).

random exploration strategy. After a random number of time steps, the environment is switched. The agent is not informed that a switch has occurred. Once the environment is changed, the agent must quickly adapt to the new environment. Note that this task differs from a reinforcement learning task because the task at hand consists of just the exploration phase. In particular, there is no reward involved in learning. Note that it also differs from active inference or active exploration because the agent is just randomly moving around to collect experiences. In particular, the agent does not choose an action so as to maximize information gain or surprise reduction.

Model. We model the knowledge of the environment by a learning agent that updates a set of parameters $\alpha(s, \check{s}) \geq 1$ used for describing its belief

about state transitions from $s \in \{1, 2, \ldots, 16\}$ to $\check{s} \in \{1, 2, \ldots, 16\} \backslash s$, where 16 is the number of rooms. More precisely, an agent's belief about how likely it is to visit $\check{s}$, given the current state s, is modeled by a Dirichlet distribution parameterized by a vector of parameters $\vec{\alpha}(s) \in \mathbb{R}^{15}$. The components of the vector $\vec{\alpha}(s)$ are denoted as $\alpha(s, \check{s})$. The 240 parameters $\alpha(s, \check{s})$ summarize the current belief of the agent about the model of the world. We emphasize that the agent has a structurally incomplete model of the world since it does not know that there are two different environments.

In order to see how well our proposed surprise-modulated belief update algorithm performs in this task, we compare it with a naive Bayesian learner and an online expectation-maximization (EM) algorithm in a hierarchical world model (Mongillo & Deneve, 2008). While the naive Bayesian model assumes that there is only a single stable, but stochastic, environment, the hierarchical Bayesian model exploits the true architecture of the hidden Markov model (HMM) corresponding to the task and approximates the optimal Bayesian solution using an EM algorithm (see appendix A.5).

Results for the maze task. The surprise-modulated belief update (see algorithm 1), with the Dirichlet distribution inserted, yields algorithm 2 for the maze exploration task (see section 6.5 for derivation). Immediately after a transition from the current state s to the next state s', the new belief q_γ obtained by the SMiLe rule, equation 3.6, is a Dirichlet distribution $\vec{\alpha}_{new}(s)$ with components $\alpha_{new}(s, \check{s}) = \gamma(1 + [\check{s} = s']) + (1 - \gamma)\alpha_{old}(s, \check{s})$, that can be written as a weighted average of the parameters of the current belief π_n (i.e., $\alpha_{old}(s, \check{s})$) and those of the scaled likelihood $\hat{p}_X$ (i.e., $1 + [\check{s} = s']$). Here, $[\check{s} = s']$ indicates a number that is 1 if the condition in brackets is satisfied and 0 otherwise. Similar update schemes for Dirichlet distributions during foraging also appear in a Bayesian framework (Sun et al., 2011; Friston et al., 2016, 2017).

Similar to the gaussian mean estimation task, surprise is initially high and slowly decreases as the agent learns the topology of the environment (see Figure 5A). When the environment is switched, the sudden increase in the surprise signal (see Figure 5A) causes the parameter γ to increase (see Figure 5B). This is equivalent to discounting previously learned information and results in a quick adaptation to the new environment. To quantify the adaptation to the new environment, we compare the state transition probabilities of the current model with the true transition probabilities of the two environments. We find that the estimation error of the state transition probabilities in the new environment is quickly reduced after the switch points (see Figure 5C). Following a change point, the model uncertainty U, measured as the entropy of the current belief about the state transition probabilities, increases, indicating that the current model of the topology is inaccurate (see Figure 5D). The commitment to a belief, defined in equation 3.1 as the negative entropy of the belief distribution, therefore decreases whenever a switch point occurs. A few time steps later, the uncertainty U slowly

Algorithm 2: Surprise-Modulated Belief Update for the Maze Exploration Task.

1: $N \leftarrow$ number of data samples

2: $\alpha(s, \check{s}) = 1, \quad \forall s \in \{1, 2, ..., 16\}, \check{s} \in \{1, 2, ..., 16\} \backslash \{s\}$ (a uniform prior belief)

3: $m \leftarrow 0.1$ (subject-dependent)

4: Start in state s

5: **for** n: 1 to N **do**

at this time step we only update the parameters that describe state transitions from the current state s to all possible next states $\check{s} \in \{1, 2, ..., 16\} \backslash \{s\}$. The current belief, for the state s, is $\pi_{n-1} \sim Dir(\mathbf{a}),\ \mathbf{a} \in \mathbb{R}^{15},\ \mathbf{a}(\check{s}) = \alpha(s, \check{s})$.

6: $X_n : s \rightarrow s'$ (a new transition is observed)

the scaled likelihood is $\hat{p}_X \sim Dir(\mathbf{b}), \mathbf{b} \in \mathbb{R}^{15},\ \mathbf{b}(\check{s}) = 1 + [\check{s} = s']$

7: (i) $S_{cc}(X_n; \pi_{n-1}) = D_{KL}[Dir(\mathbf{a})||Dir(\mathbf{b})]$

8: (ii-a) $B_{max}(X_n) = D_{KL}[Dir(\mathbf{b})||Dir(\mathbf{a})]$

9: (ii-b) $B(X_n) = \frac{mS_{cc}(X_n;\pi_{n-1})}{1+mS_{cc}(X_n;\pi_{n-1})} B_{max}(X_n)$

10: (iii) find γ by solving $D_{KL}[Dir(\gamma\mathbf{b} + (1-\gamma)\mathbf{a})||Dir(\mathbf{a})] = B(X_n)$

11: (iv) $\alpha(s, \check{s}) \leftarrow (1-\gamma)\alpha(s, \check{s}) + \gamma(1 + [\check{s} = s'])$

12: **Return** $\alpha(s, \check{s}), \forall s, \check{s}$;

Note: $D_{KL}[Dir(\mathbf{m})||Dir(\mathbf{n})] = \ln\Gamma(\sum_{\check{s}} \mathbf{m}(\check{s})) - \ln\Gamma(\sum_{\check{s}} \mathbf{n}(\check{s})) - \sum_{\check{s}} \ln\Gamma(\mathbf{m}(\check{s})) + \sum_{\check{s}} \ln\Gamma(\mathbf{n}(\check{s})) + \sum_{\check{s}} (\mathbf{m}(\check{s}) - \mathbf{n}(\check{s}))(\Psi(\mathbf{m}(\check{s})) - \Psi(\sum_{\check{s}} \mathbf{m}(\check{s})))$. $\Gamma(.)$ and $\Psi(.)$ denote the gamma and digamma functions, respectively. $[\check{s} = s']$ denotes the Iverson bracket, a number that is 1 if the condition in brackets is satisfied, and 0 otherwise.

decreases, indicating an increased confidence in what is learned in the new environment and thus an increased commitment to the current model.

If we look more closely at the model parameters, we find that the surprise-modulated belief update (see algorithm 2) enables the agent to adjust the estimated state transition probabilities. In Figure 6 we compare the

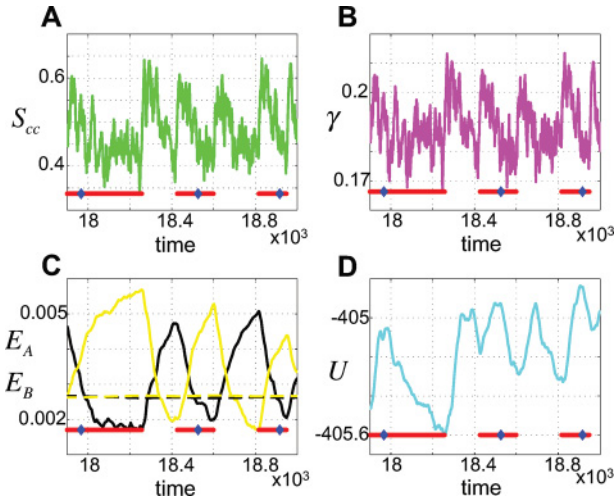


Figure 5: Time series of relevant signals in the surprise-modulated belief update, algorithm 2, applied to the maze exploration task. All curves have been smoothed with an exponential moving average (EMA) with a decay constant 0.1. The plots are shown for 1100 time steps (horizontal axis) toward the end of a simulation with 20,000 time steps. The agent visits environments $\mathcal{A}$ and $\mathcal{B}$ equally often and spends on average 200 time steps in each environment before a switch occurs. Red bars indicate the time that the agent explores environment $\mathcal{A}$. Blue diamonds indicate 100 time steps after a change point from $\mathcal{B}$ to $\mathcal{A}$. (A) Confidence-corrected surprise S_{cc} (see equation 3.2) (green) increases at switch points and decreases (with fluctuations) until the next change point. (B) The parameter γ (magenta) increases with the surprise at the change points and causes the next data samples to be more effective on belief update than the samples before the change point. (C) The estimation errors for the transition matrix $\hat{T}$, $E_{\mathcal{A}}[t] = ||\hat{T}[t] - T_{\mathcal{A}}||_2 = 256^{-1} \sum_{s,s'} [\hat{T}[t](s, s') - T_{\mathcal{A}}(s, s')]^2$ (solid black) and $E_{\mathcal{B}}[t] = ||\hat{T}[t] - T_{\mathcal{B}}||_2$ (solid yellow) while in environment $\mathcal{A}$ and $\mathcal{B}$, respectively, indicate a rapid adaptation to the new environment after the change points. The dashed black and yellow lines correspond to the estimation errors $E_{\mathcal{A}}$ and $E_{\mathcal{B}}$, respectively, when the naive Baye's rule (as a control experiment) is used for belief update. The naive Bayes' rule converges to a stationary solution (no significant change in the estimation error after a switch of environment). (D) The model uncertainty (light blue) increases for a few time steps following a change in the environment, an alert that the current model might be wrong. It then starts decreasing as the agent becomes more certain in the new environment.

estimated and the true transition probabilities 100 time steps after a switch. Given that the environment is characterized by 64 different transitions (in a space of $16 \times 15 = 240$ potential transitions), 100 time steps allow an agent

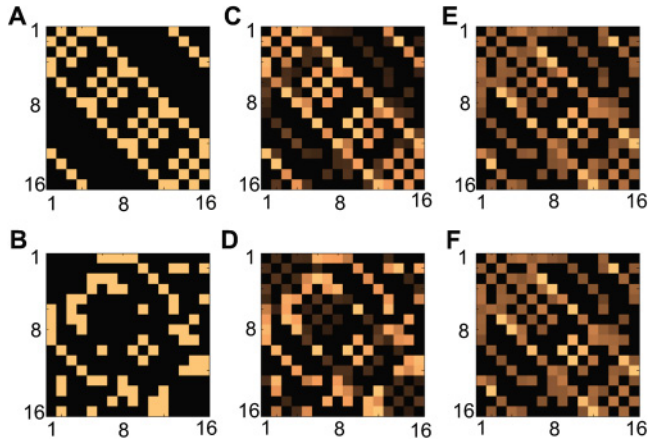


Figure 6: True and estimated state transition probabilities in the maze exploration task. The color intensity for each entry (s, s') represents the probability of transition from a current state s (row) to a next state s' (column). (A) The true state transition probability matrix $T_{\mathcal{A}}(s, s')$ in environment $\mathcal{A}$. Each row $T_{\mathcal{A}}(s, :)$ has only four nonzero entries (small light brown squares) whose positions indicate the neighboring rooms of state s in environment $\mathcal{A}$. Note that $\sum_{\check{s}} T_{\mathcal{A}}(s, \check{s}) = 1\,\forall s$. (B) The true state transition probability matrix $T_{\mathcal{B}}(s, s')$ for the environment $\mathcal{B}$, which has a different topology compared to $\mathcal{A}$. (C) The estimated state transition probability matrix $\hat{\mathcal{T}}_{\mathcal{A}}$ when the surprise-modulated algorithm 2 is used for belief update. $\hat{\mathcal{T}}_{\mathcal{A}} = K^{-1} \sum_{k=1}^{K} \hat{T}[t^k_{\mathcal{B}\to\mathcal{A}} + 100]$ is calculated by averaging the estimated transition matrix $\hat{T}[t]$ at 100 time steps after each of K change points $t^k_{\mathcal{B}\to\mathcal{A}}$. Here, $t^k_{\mathcal{B}\to\mathcal{A}}$ denotes the kth time that the environment is changed from $\mathcal{B}$ to $\mathcal{A}$ and has remained unchanged for at least the next 100 time steps (relevant time points are indicated by blue diamonds in Figure 5). The similarity between $\hat{\mathcal{T}}_{\mathcal{A}}$ and $T_{\mathcal{A}}$ indicates that algorithm 2 enables the agent to quickly adapt to environment $\mathcal{A}$ once a switch from $\mathcal{B}$ to $\mathcal{A}$ occurs. (D) The estimated transition matrix $\hat{\mathcal{T}}_{\mathcal{B}}$ (similarly defined as $\hat{\mathcal{T}}_{\mathcal{A}}$ but for environment $\mathcal{B}$) when algorithm 2 is used for belief update. Note its similarity to the true matrix $T_{\mathcal{B}}$. (E, F) The estimated state transition probability matrices $\hat{\mathcal{T}}_{\mathcal{A}}$ (top) and $\hat{\mathcal{T}}_{\mathcal{B}}$ (bottom) when the naive Bayesian method (as a control experiment) is used for belief update. A Bayesian agent does not adapt well to the new environment after a switch occurs because it learns a weighted average of true transition matrices $T_{\mathcal{A}}$ and $T_{\mathcal{B}}$, where the weight is proportional to the fraction of time spent in each environment. Since both environments are visited equally in this experiment, the estimated quantities approach $(T_{\mathcal{A}} + T_{\mathcal{B}})/2$.

to explore only a fraction of the potential transitions. Nevertheless, 100 time steps after a switch, the matrix of transition probabilities already resembles that of the present environment (see Figures 6C and 6D).

The surprise-modulated belief update is a method of quick learning. How well does our SMiLe update rule perform relative to other existing models? We compared it with two well-known models. First, we compared it to a naive Bayesian learner that tries to estimate the 240 state transition probabilities using Bayes' rule. Note that by construction, the naive Bayesian learner is not aware of the switches between the environments. Second, we compared it to a hierarchical statistical model that reflects the architecture of the true world as in Figure 4. The task is to estimate the 2×240 state transitions in the two environments, as well as transition probabilities between the environments $p_{\mathcal{A}\to\mathcal{B}}$ and $p_{\mathcal{B}\to\mathcal{A}}$ by an online EM algorithm.

For the naive Bayesian learner, we find that its behavior indicates a steady increase in certainty, regardless of how surprising the samples are. In other words, it is incapable of changing its belief after it has sufficiently explored the environments (see Figure 5C). The state transition probabilities are estimated by averaging over the true parameters of both environments, where the weight of averaging is determined by the fraction of time spent in the corresponding environment (see Figures 6E and 6F).

The comparison of our surprise-modulated belief update with the online EM algorithm (Mongillo & Deneve, 2008) for the hierarchical Bayesian model associated with the changing environments provides several insights (see Figure 7). First, after fewer than 1000 time steps, the estimation error for environment $\mathcal{A}$ during short episodes in environment $\mathcal{A}$ drops below $E_{\mathcal{A}} = 0.002$. The online EM algorithm takes 10 times longer to achieve the same level of accuracy. While the solution of the SMiLe rule in the long run is not as good as that of the online EM algorithm, our algorithm benefits from a reduced computational complexity and simpler implementation.

To further investigate the ability of an agent to adapt to the new environment after a switch, we analyzed performance as a function of two free parameters that control the setting of the task: the fraction of time spent in environment $\mathcal{A}$ and the average time spent in environment $\mathcal{A}$ before a switch to $\mathcal{B}$ occurs. To do so, we calculate the average estimation error in state transition probabilities 64 time steps after a switch occurs. We consider only those switches after which the agent stays in that environment for at least 64 time steps. Note that 64 is the minimum number of time steps that is required to ensure that all possible transitions from 16 rooms to their 4 neighbors could occur. A smaller estimation error for a given pair of free parameters indicates a faster adaptation to the new environment for that setting.

We found that the surprise-modulated belief enables an agent to quickly readjust its estimation of model parameters, even if the fraction of time spent in an environment is relatively short. In that sense, it behaves similarly to the approximate hierarchical Bayesian approach (online EM algorithm). This is not, however, the case for a naive Bayesian learner whose estimation error in each environment depends on the fraction of time spent in the corresponding environment (see Figure 8).

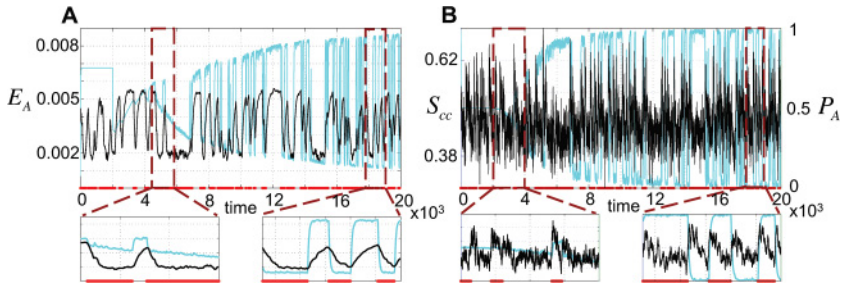


Figure 7: Comparison of surprise-modulated belief update with an online EM algorithm for the hierarchical Bayesian model. (A) The estimation error $E_{\mathcal{A}}$ (vertical axis) of state transition probabilities within environment $\mathcal{A}$ versus time (horizontal axis), for surprise-modulated belief update (black) and online EM learner (blue). Bottom plots depict zooms during the early (left) and late (right) phases of a simulation of 20,000 time steps. In the early phase of learning (bottom left), the surprise-modulated belief update enables the agent to quickly learn model parameters after a switch to environment $\mathcal{A}$ (indicated by red bars). In the late phase of learning (right), however, the online EM algorithm adapts to the new environment faster and more accurately than the surprise-modulated belief update. (B) The inferred probability $P_{\mathcal{A}}$ of being in environment $\mathcal{A}$ (blue, right vertical axis) used in the online EM algorithm and the confidence-corrected surprise S_{cc} (black, left vertical axis) used in the surprise-modulated belief update.

The naive Bayesian learner suffers from low accuracy in estimation and cannot adapt to environmental changes. A full hierarchical Bayesian model, however, requires prior information about the task and is computationally demanding. For example, the computational load of the hierarchical Bayesian model increases with the number N of environments between which switching occurs. The surprise-modulated belief update, however, balances accuracy and computational complexity: computational complexity remains, by construction, independent of the number of switched environments. In other words, since we accept from the beginning that our model of the world will be approximate and structurally incomplete, the model can perform reasonably well after having seen a small number of data samples.

4 Discussion

For many decades, surprise has been an influential concept for research in information theory (Shannon, 1948; Tribus, 1961), experimental neuroscience of EEG signals (Squires et al., 1976) or pupil dilation (Hess & Polt, 1960), psychophysics of covert attention (Itti & Baldi, 2009; Baldi & Itti,

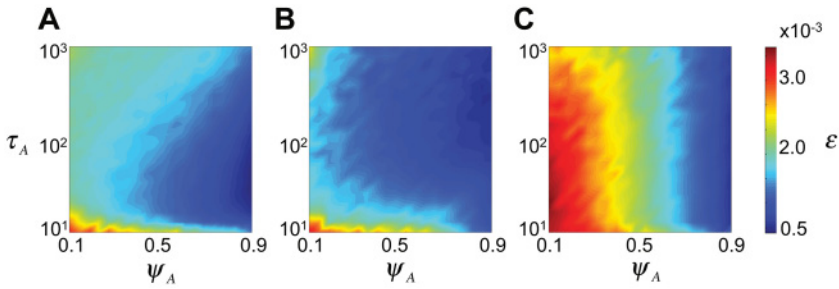


Figure 8: The estimation error ϵ in the maze exploration task, as a function of (1) the average time spent in environment $\mathcal{A}$ before a switch to environment $\mathcal{B}$ ($\tau_{\mathcal{A}} = \Delta t / p_{\mathcal{A}\to\mathcal{B}}$, vertical axis) and (2) the fraction of time spent in environment $\mathcal{A}$ ($\psi_{\mathcal{A}} = P_{\mathcal{B}\to\mathcal{A}}/(P_{\mathcal{B}\to\mathcal{A}} + P_{\mathcal{A}\to\mathcal{B}})$, horizontal axis). (A) The average estimation error (of state transition probabilities), 64 time steps after a switch from $\mathcal{B}$ to $\mathcal{A}$, when surprise-modulated belief update (see algorithm 2) is used for learning. The spread of blue (lower estimation error) illustrates that the surprise-modulated belief update enables an agent to quickly adapt to the environment visited after a switch. For each pair $(\tau_{\mathcal{A}}, \psi_{\mathcal{A}})$, the simulation is repeated for 20 episodes, each consisting of 20,000 time steps. In each episode, a different rearrangement of rooms for building environment $\mathcal{B}$ is used to make sure that the result is not biased by a specific choice of this environment. (B) The average estimation error when the online EM algorithm is used for learning the hierarchical statistical model. (C) The average estimation error when the naive Bayesian learner is used for belief update. The estimation error for this model is mainly determined by the fraction of time spent in environment $\mathcal{A}$ (i.e., $\psi_{\mathcal{A}}$). The estimation error decreases with the time spent in environment $\mathcal{A}$ regardless of the timescale of stability determined by $\tau_{\mathcal{A}}$.

2010), fMRI (Iglesias et al., 2013; Meyniel & Dehaene, 2017), machine learning (Storck et al., 1995; Schmidhuber, 1991, 2006, 2010), neuromodulation (Yu & Dayan, 2005; Iglesias et al., 2013; Joffily & Coricelli, 2013; Vossel et al., 2014), and Bayesian variational learning (Friston, 2010), to name just a few examples. Moreover, surprise minimization is a well-known learning principle for active inference in learning agents (Storck et al., 1995; Schmidhuber, 2006; Friston et al., 2016, 2017), and linked to intrinsically generated reward signals, analogous to reinforcement learning (Storck et al., 1995; Frank et al., 2013; Schmidhuber, 2010; Singh et al., 2004; Sun et al., 2011; Schultz, 2015). On the background of this rich tradition, the contributions of this article are two-fold.

First, we introduced the notion of confidence-corrected surprise as a quantification of immediate (i.e., low reaction time) puzzlement surprise. Earlier quantifications of puzzlement surprise have measured the Shannon surprise (also called information content or model evidence) (Tribus, 1961;

MacKay, 2003) or approximated Shannon surprise by a free-energy bound (MacKay, 2003; Friston, 2010). However, in our opinion, Shannon surprise neglects the importance of the commitment to an opinion (also called precision in case of a gaussian belief distribution). In our framework, two agents A and B with different world models may assign the same low probability to a data point X—yet if A is more committed to his opinion than B, he will be more surprised.

Second, for models that do not capture the complexity of the world, we proposed a surprise-modulated update scheme for learning in nonstationary environments derived from a heuristic approach toward surprise minimization. In this update scheme, sudden changes are identified by high surprise and result in placing more weight on new information. Our surprise-modulated learning rule is closely related to well-known precision-based update schemes used in Kalman filters (Kalman, 1960) or in iterative solutions of hierarchical Bayesian models (Mathys et al., 2011, 2014). But while the class of solvable Bayesian models is limited and typically involves nontrivial integrals, our approach avoids the specification and inversion of complicated or hierarchical Bayesian models. In fact, despite the obvious success of Bayesian models for understanding the brain (Ernst & Banks, 2002; Körding & Wolpert, 2004; Knill & Pouget, 2004; Ma et al., 2006; Beck et al., 2008; Fiser et al., 2010; Berkes et al., 2011; Kolossa et al., 2013), we believe that the brain normally does not use the correct hypothesis class (e.g., a multilevel hierarchical partially observable Markovian process) and Bayes-optimal updates rules to capture dependencies in the outside world or in a specific experimental design; rather, the brain may develop incomplete models using an imperfect hypothesis class.

We discuss some of the above insights in more detail.

4.1 Puzzlement Surprise Is Rapid. We found it useful to distinguish puzzlement surprise from enlightenment surprise. We conjecture that the P300 component of the EEG, which is visible within less than 400 ms after the surprising stimulus (Squires et al., 1976), indicates puzzlement surprise. Typically the P300 component has been correlated with violation of expectation, with Shannon surprise, or with its free-energy bound (Squires et al., 1976; Friston, 2005; Meyniel et al., 2016). One of the questions for the future is whether confidence-corrected surprise correlates better with EEG and fMRI data than Shannon surprise or its free-energy bound.

Our working hypothesis is that the brain evaluates puzzlement surprise even before recognition, inference, or learning occurs. We thus need to evaluate surprise before we update our belief so that puzzlement surprise may control update rates. For the confidence-corrected surprise measure introduced in this article, a rapid evaluation (before an update of the model occurs) is indeed possible. While an evaluation of enlightenment surprise using measures of Bayesian surprise (Storck et al., 1995; Itti & Baldi, 2009) might also be rapid, it is, in our view, more difficult to imagine how

Bayesian surprise could modulate update rates because Bayesian surprise or compression surprise (Schmidhuber, 2006) are quantities that are available only after the update step.

In our view, the update step happens after the evaluation of puzzlement surprise. In our model and completely analogous to free-energy models (Friston, 2010), the update step in turn leads to minimization of future puzzlement surprise. The amount of belief change during one update step of the SMiLe rule measures enlightenment surprise and can be quantified by Bayesian surprise (Storck et al., 1995; Itti & Baldi, 2009; Baldi & Itti, 2010) which is in turn linked to the reduction in (expected) free energy or epistemic value (Friston, 2010; Friston et al., 2017) and the progress in compression (Schmidhuber, 2006, 2010).

4.2 New versus Old Information. The performance of the proposed belief update algorithm is primarily achieved by two features of the update rule: (1) the algorithm adaptively increases the influence of new data on the belief update as a function of how surprising the data were and (2) the algorithm increases model uncertainty in the face of surprising data, thus increasing the influence of new data on current and future belief updates. The importance of the first point has been recognized and incorporated previously (Nassar et al., 2012; Pearce & Hall, 1980). The second point is automatically incorporated in hierarchical gaussian filter models (Mathys et al., 2011, 2014) but has often been omitted in previous simpler models. Essentially, a surprising sample not only signals a potential change; it also signals that our current model may be wrong, so that we should be less certain about it. This increase in model uncertainty (decrease in commitment or precision) implies discounting the influence of past information in current and future belief updates.

Both humans and animals adaptively adjust the relative contribution of old and newly acquired data on learning (Behrens et al., 2007; Nassar et al., 2012; Krugel et al., 2009; Pearce & Hall, 1980) and rapidly adapt to changing environments (Pearce & Hall, 1980; Wilson et al., 1992; Holland, 1997). A full (hierarchical) Bayesian approach is possible only if the subject is aware of the correct (hierarchical) architecture of the generative model of the task, (e.g., the timescale of change in the environment or the number of environments between which switches occur). Calculating the probability of a change point in a gaussian estimation task (Nassar et al., 2010), estimating the volatility of the environment in a reversal learning task (Behrens et al., 2007), and dynamically forgetting the past information with a controlled time constant (Ruter, Marcille, Sprekeler, Gerstner, & Herzog, 2012) are all examples of addressing learning in changing environments without explicit knowledge of the full Bayesian model of the real world. While our proposed surprise-based algorithm may not be theoretically optimal, it approximates the optimal (hierarchical) Bayesian solution with minimal knowledge regarding the task or the environment.

A corollary of point (2) is that the SMiLe rule guarantees that a small model uncertainty remains even after a long stationary period. This remaining uncertainty ensures that an organism can still detect a change even after having spent an extensive amount of time in a given environment (see Figure 5C). It also means that the learning rate never falls to zero, so that recent data samples are more important than old ones, even during a stationary fully stochastic sequence, consistent with behavioral experiments (Squires et al., 1976; Meyniel et al., 2016).

4.3 Potential Applications. Surprise as a violation of expectation can be considered as a model prediction error (Squires et al., 1976; Yu & Dayan, 2005; Friston, 2005) that is more general than prediction errors used in reward-based learning paradigms (O'Doherty et al., 2003, 2004; Schultz, 2010). Analogous to existing neuroscience studies with traditional surprise measures (Squires et al., 1976; Friston, 2005; Yu & Dayan, 2005; Iglesias et al., 2013; Joffily & Coricelli, 2013; Vossel et al., 2014; Meyniel et al., 2016; Meyniel & Dehaene, 2017), confidence-corrected surprise can be used in EEG and fMRI experiments to identify and model surprise-based responses within and across subjects. Similar to prediction error estimation in hierarchical gaussian filter models (Mathys et al., 2011; Iglesias et al., 2013), confidence-corrected surprise can be estimated from behavioral data. At the group level, individual subjects are characterized by different values of the parameter m while the effective learning rate for updating the model provides a sample-by-sample measure to model responses to confidence-corrected surprise across time and within subjects.

Moreover, it is in principle possible to fit γ to behavioral data without computing surprise or to control γ by something other than surprise. In general, replacing the full hierarchical Bayesian model update during a learning task in a changing environment with the SMiLe rule simplifies calculations, which should make the SMiLe-framework an attractive candidate for fitting relevant parameters to behavioral data.

Furthermore, both confidence-corrected surprise and the SMiLe rule have wide-reaching implications outside the framework presented here. On the one hand, our surprise measure can not only modulate learning but can be used as a trigger signal for an algorithm that needs to choose between several uncertain states or actions, as is the case in change point detection (Nassar et al., 2010; Ruter et al., 2012; Wilson, Nassar, & Gold, 2013), memory and cluster formation (Gershman & Niv, 2015), exploration-exploitation trade-off (Cohen, McClure, & Yu, 2007; Jepman & Nieuwenhuis, 2011), novelty detection (Bishop, 1994; Knight, 1996), and network reset (Bouret & Sara, 2005).

4.4 Neuromodulatory Signals and Surprise. There is ample evidence for a neural substrate of surprise. Existing measures of expectation violations such as absolute and variance-scaled reward prediction errors

(Schultz, 2015, 2016), unexpected uncertainty (Yu & Dayan, 2005), risk prediction errors (Preuschoff et al., 2008), model prediction error in hierarchical models (Iglesias et al., 2013), emotional valance (Joffily & Coricelli, 2013), and Bayesian belief updating have been linked to different neuromodulatory systems. Among those, the noradrenergic system has emerged as a prime candidate for signaling unexpected uncertainty: noradrenergic neurons respond to unexpected changes such as the presence of a novel stimulus, unexpected pairing of stimulus with a reinforcement during conditioning, and reversal of the contingencies (Sara, Vankov, & Herve, 1994; Vankov, Minvielle, & Sara, 1995; Aston-Jones, Rajkowski, & Kubiak, 1997; Sara, 2009). The P300 component of the event-related potential (Squires et al., 1976; Pineda et al., 1997; Missionier et al., 1999), which is associated with novelty (Donchin, Ritter, & McCallum, 1978) and surprise (Verleger et al., 1994), is modulated by noradrenaline. It also modulates pupil size (Costa & Rudebeck, 2016) as a physiological response to surprise. The dynamics of the noradrenergic system are fast enough to quickly respond to unexpected events (Rajkowski, Kubiak, & Aston-Jones, 1994; Clauton, Rajkowski, Cohen, & Aston-Jones, 2004; Bouret & Sara, 2004), a functional requirement for surprise to control learning; see (Sara, 2009; Bouret & Sara, 2005; Aston-Jones & Cohen, 2005) for a review. We predict that in experiments with changing environments, the activity of noradrenaline should exhibit a high correlation with the confidence-corrected surprise signal. It is an empirical question to check whether the correlation with confidence-corrected surprise is better or worse than that with competing surprise measures (Shannon, 1948; Storck et al., 1995; Schmidhuber, 2006; Itti & Baldi, 2009; Friston, 2010).

Furthermore acetylcholine (ACh) is a candidate neuromodulator for encoding expected uncertainty (Yu & Dayan, 2005) and thus is linked to the model uncertainty (although it might also be linked to other forms of uncertainty such as environmental stochasticity).

A variety of experimental findings are consistent with and can be explained by our definition of confidence-corrected surprise and the SMiLe rule. It has been shown both theoretically (Yu & Dayan, 2005) and empirically (Gu, 2002) that noradrenaline and ACh interact such that ACh sets a threshold for noradrenaline to indicate fundamental changes in the environment (Yu & Dayan, 2005). This is consistent with our hypothesis that if an agent is uncertain about its current model of the world, unexpected events are perceived as less surprising than when the agent is almost certain about the model (which is the key idea behind the confidence-corrected surprise). The impairment of adaptation to contextual changes due to noradrenaline depletion (Sara, 1998) can be explained by the incapability of subjects to respond to surprising events signaled by noradrenaline. The absence or suppression of ACh (low model uncertainty) implies little or no variability of the environment so that even small

prediction error signals are perceived as surprising (Jones & Higgins, 1995), consistent with the excessive activation of noradrenergic system in such situations.

Moreover, there is empirical evidence that noradrenaline and ACh both affect synaptic plasticity in the cortex and the hippocampus (Bear & Singer, 1986; Gu, 2002), suppress cortical processing (Kimura, Fukuda, & Tsumoto, 1999; Kobayashi et al., 2000), and facilitate information processing from thalamus to the cerebral cortex (Hasselmo, Wyble, & Wallenstein, 1996; Gil, Connors, & Amitai, 1997; Hsie, Cruikshank, & Metherate, 2000). This is consistent with our theory that surprise balances the influence of newly acquired data (thalamocortical pathway) and old information (corticocortical pathway) during belief update.

In our view, the activity of a single neuromodulator, or a combination of several neuromodulators, conveys the amount of puzzlement surprise to many areas of the brain, where it regulates the learning rate of those synapses that have been involved in representing the recent stimuli (Frémaux & Gerstner, 2016). This view has been illustrated, for example, for networks of stochastically spiking neurons in the context of free-energy minimization, where the learning rate was modulated by an amount proportional to free-energy minus expected free energy (Rezende & Gerstner, 2014; Brea et al., 2013). In this article, we showed the modulation of learning rate by confidence-corrected surprise, albeit in a more abstract, nonspiking model.

5 Conclusion

In summary, we have proposed confidence-corrected surprise as a measure of puzzlement surprise and a surprise-modulated belief update algorithm that can be used for modeling how humans and animals learn in changing environments. The belief updating step gives, similar to compression progress (Schmidhuber, 2006, 2010) or Bayesian surprise (Storck et al., 1995; Itti & Baldi, 2009) or parameter update in free-energy minimization (Friston, 2010), an interpretation of enlightenment surprise.

Our results suggest that the SMiLe rule in an imperfect world model can approximate an optimal hierarchical Bayesian learner (e.g., Mathys et al., 2011, 2014; Mongillo & Deneve, 2008), with significantly reduced computational complexity. Our model of confidence-corrected surprise provides a framework for future work, including computational studies, so as to find out how the proposed model can be neurally implemented; neurobiological studies, so as to unravel the interaction between different neural circuits that are functionally involved in learning under surprise; and behavioral studies with human subjects, so as to correlate confidence-corrected surprise with EEG or pupil size.

6 Mathematical Methods

In this section we provide mathematical explanations for statements made in the section 3. Long and detailed calculations have been moved to the appendix.

6.1 The Scaled Likelihood is the Posterior Belief under a Flat Prior. Assume that all model parameters θ stay in some bounded convex interval of volume A. The volume A can be arbitrarily large. Given a data sample X, the posterior belief $p^{flat}(\theta|X)$ about the model parameters θ (derived by Bayes' rule) under the assumption of a flat prior $\hat{\pi}_0(\theta) = 1/A$ is

$$p^{flat}(\theta|X) = \frac{p(X|\theta)\hat{\pi}_0(\theta)}{\int_\theta p(X|\theta)\hat{\pi}_0(\theta)\, d\theta} = \frac{p(X|\theta)}{\int_\theta p(X|\theta)\, d\theta} = \frac{p(X|\theta)}{||p_X||} = \hat{p}_X(\theta), \tag{6.1}$$

where $||p_X|| = \int_\theta p(X|\theta)d\theta$ is a data-dependent constant. Therefore, the scaled likelihood $\hat{p}_X(\theta)$ is the posterior under a flat prior. Note that the result is independent of the volume of A so that we can take the limit of $A \to \infty$.

6.2 Confidence-Corrected Surprise Increases with Shannon Surprise and Bayesian Surprise. In this section, we show that confidence-corrected surprise can be written as a sum of terms that include Shannon surprise, Bayesian surprise, and commitment. Two corollaries will be that a committed subject is more surprised than an uncommited one (if the data have the same probability for both subjects) and that (for the same level of commitment) the surprise decreases with the probability of the data.

The confidence-corrected surprise in equation 3.2 can be expressed as

$$S_{cc}(X; \pi_n) = -\int_\theta \pi_n(\theta) \ln p(X|\theta) d\theta + \ln ||p_X|| - H(\pi_n), \tag{6.2}$$

where $||p_X||$ is a data-dependent constant defined in equation 6.1 and $H(\pi_n)$ denotes the entropy of the current belief (cf. equation 3.1). Let us call the first term $-\int_\theta \pi_n(\theta) \ln p(X|\theta)$ in equation 6.2 the raw surprise $S_{raw}(X; \pi_n)$ of a data sample X:

$$S_{raw}(X; \pi_n) = -\int_\theta \pi_n(\theta)\ \ln p(X|\theta)\, d\theta. \tag{6.3}$$

We now show that the raw surprise $S_{raw}(X; \pi_n)$ in equation 6.3 increases with the Shannon surprise and the Bayesian surprise.

In the following, small numbers above an equals sign refer to equations in the text. The raw surprise $S_{raw}(X; \pi_n)$ in equation 6.3 can be rewritten

as a linear combination of both Bayesian surprise and Shannon surprise because

$$\begin{aligned} S_{raw}(X;\pi_n) &\overset{(6.3)}{=} -\int_\theta \pi_n(\theta)\ln p(X|\theta)\,d\theta \\ &\overset{(2.3)}{=} -\int_\theta \pi_n(\theta)\ln\left[\frac{\pi_{n+1}^{\text{Bayes}}(\theta)\left(\int_\theta p(X|\theta)\pi_n(\theta)\,d\theta\right)}{\pi_n(\theta)}\right]d\theta \\ &= D_{KL}[\pi_n||\pi_{n+1}^{\text{Bayes}}] - \ln\left[\int_\theta p(X|\theta)\pi_n(\theta)\,d\theta\right], \end{aligned} \tag{6.4}$$

where the first term $D_{KL}[\pi_n||\pi_{n+1}^{\text{Bayes}}]$ stands for the Bayesian surprise and the second term $-\ln\left[\int_\theta p(X|\theta)\pi_n(\theta)\,d\theta\right]$ stands for the Shannon surprise.

Note that for the calculation, we introduced a hypothetical belief π_{n+1}^{Bayes} that would correspond to an update step under Bayes' rule—even though we do not actually perform such an update step. This notation was introduced just to highlight that the raw surprise $S_{raw}(X;\pi_n)$ in equation 6.4 combines the puzzlement surprise of Shannon (information content) and the enlightenment surprise defined as Bayesian surprise after Bayesian belief updating. We emphasize again that π_{n+1}^{Bayes} refers to the update step under a naive Bayes' rule, but this is *not* the rule that we apply in our surprise-based learning scheme.

As an aside we note the formal similarity of equation 6.4 with the formalism of free energy in equation 2.4. In both cases, the Shannon surprise (model evidence) is additively combined with a Kullback-Leibler divergence.

Corollary 1. *Less probable data lead to a larger surprise S_{cc}.* Our proposed confidence-corrected surprise measure $S_{cc}(X;\pi_n)$ in equation 3.2 inherits the property of the Shannon surprise from the raw surprise $S_{raw}(X;\pi_n)$ in equation 6.4. In particular, for a fixed opinion π_n and for the same value of $||p_X||$, a data point of lower probability leads to a larger surprise than one of higher probability.

Corollary 2. *Committed subjects are more surprised than uncommitted ones.* The value of the confidence-corrected surprise, equation 6.2, depends on a subject's commitment to her belief. The commitment to the current model of the world is represented by the negative entropy—$H(\pi_n) = \int_\theta \pi_n(\theta)\ln\pi_n(\theta)d\theta$. Equation 6.2 shows that the confidence-corrected surprise decreases with entropy, which is equivalent to an increase with commitment. Therefore, given the same probability of the data point under two different world models, the subject with a stronger commitment (smaller entropy) is more surprised than the subject with a weaker commitment (higher entropy) (see the example of Figure 1). Intuitively, if we are uncertain about what to expect (because we have not yet learned the structure of the world), receiving a data sample that occurs with low probability

under the present model is less surprising than a low-probability sample in a situation when we are almost certain about the world (see Figure 2A).

6.3 Derivation of the SMiLe Rule. We note that the KL divergence $D_{KL}[a||b]$ is convex with respect to the first argument a. Therefore, both the objective function $S_{cc}(X; q)$ in equation 3.2 and the constraint $D_{KL}[q||\pi_n] \leq B$ in the optimization problem in equation 3.5 are convex with respect to q, which ensures the existence of the optimal solution.

We solve the constraint optimization by introducing a nonnegative Lagrange multiplier $\lambda^{-1} \geq 0$ and a Lagrangian

$$\begin{aligned}\mathbb{L}(q, \lambda) &= S_{cc}(X; q) - \frac{1}{\lambda}(B - D_{KL}[q||\pi_n]) \\ &\overset{(6.2)}{=} \left\langle -\ln p(X|\theta) + \ln q(\theta) + \frac{1}{\lambda} \ln \frac{q(\theta)}{\pi_n(\theta)} \right\rangle_q - \frac{B}{\lambda} + \ln ||p||, \qquad (6.5)\end{aligned}$$

where $\langle . \rangle_q$ denotes the average with respect to q. Similar to standard approaches used in support vector machines (Schölkopf & Smola, 2002), the Lagrangian $\mathbb{L}$ defined in equation 6.5 must be minimized with respect to the primal variable q and maximized with respect to the dual variable λ (i.e., a saddle point must be found). Therefore the constraint problem in equation 3.5 can be expressed as

$$\arg \min_q \max_{\lambda \geq 0} \mathbb{L}(q, \lambda). \qquad (6.6)$$

By taking the derivative of $\mathbb{L}$ with respect to q and setting it equal to zero,

$$\frac{\partial \mathbb{L}}{\partial q} = -\ln p(X|\theta) + \left[1 + \ln q(\theta)\right] + \frac{1}{\lambda}\left[1 + \ln \frac{q(\theta)}{\pi_n(\theta)}\right] = 0, \qquad (6.7)$$

we find that the Lagrangian in equation 6.5 is minimized by the SMiLe rule, equation 3.6, that is, $q(\theta) \propto p(X|\theta)^{\gamma} \pi_n(\theta)^{1-\gamma}$, where γ is determined by the Lagrange multiplier λ:

$$0 \leq \gamma = \frac{\lambda}{\lambda + 1} \leq 1. \qquad (6.8)$$

Note that the constant $Z(X; \gamma)$ in equation 3.6 follows from normalization of $q(\theta)$ to integral one.

6.4 The SMiLe Rule for Beliefs Described by a Gaussian Distribution. Suppose we have drawn $n - 1$ samples $X_1, \ldots, X_{n-1}$ from a gaussian distribution of known variance σ_x^2 but unknown mean. Our empirical estimate of the current mean after $n - 1$ samples is denoted by $\hat{\mu}_{n-1}$.

Assume that the current belief about the mean μ is a normal distribution: $\pi_{n-1}(\mu) \sim \mathcal{N}(\hat{\mu}_{n-1}, \sigma^2_{n-1})$. Since the likelihood of receiving a new sample X_n is also normal, $p(X_n|\mu) \sim \mathcal{N}(\mu, \sigma^2_x)$, the updated belief obtained by the SMiLe rule, equation 3.6, is

$$q_\gamma(\mu) \propto \exp\left(-\frac{(X_n-\mu)^2}{2(\sigma'_x)^2}\right)\exp\left(-\frac{(\mu-\hat{\mu}_{n-1})^2}{2(\sigma'_{n-1})^2}\right), \tag{6.9}$$

where $(\sigma'_x)^2 = \sigma^2_x/\gamma$ and $(\sigma'_{n-1})^2 = \sigma^2_{n-1}/(1-\gamma)$. Because the product of two gaussians is a gaussian, we arrive at a distribution $q_\gamma \sim \mathcal{N}(\hat{\mu}_n, \sigma^2_n)$ with the mean $\hat{\mu}_n = w_n X_n + (1-w_n)\hat{\mu}_{n-1}$ $\left(\text{with } w_n = \frac{(\sigma'_{n-1})^2}{(\sigma'_x)^2+(\sigma'_{n-1})^2}\right)$, and the variance $\sigma^2_n = \left(\frac{1}{(\sigma'_x)^2} + \frac{1}{(\sigma'_{n-1})^2}\right)^{-1}$ (MacKay, 2003). Assuming $\sigma^2_{n-1} = \sigma^2_x$, we find $\sigma^2_n = \sigma^2_x$, and $w_n = \gamma$. Thus, if the variance of the belief distribution is initialized at $\sigma^2_0 = \sigma^2_x$, it will always stay at this value. Moreover, we can evaluate the confidence-corrected surprise to be

$$S_{cc}(X_n; \pi_{n-1}) = D_{KL}[\mathcal{N}(\hat{\mu}_{n-1}, \sigma^2_{n-1})||\mathcal{N}(X_n, \sigma^2_x)] = \frac{(X_n-\hat{\mu}_{n-1})^2}{2\sigma^2_x}, \tag{6.10}$$

where we have used (assuming again $\sigma^2_x = \sigma^2_{n-1}$),

$$D_{KL}[\mathcal{N}(a_1, b^2_1)||\mathcal{N}(a_2, b^2_2)] = \frac{(a_1-a_2)^2}{2b^2_2} + \frac{1}{2}\left(\frac{b^2_1}{b^2_2} - 1 - \ln\frac{b^2_1}{b^2_2}\right). \tag{6.11}$$

6.5 The SMiLe Rule for Beliefs Described by a Dirichlet Distribution. Assume that the current belief about the probability of transition from state $s \in \{1, 2, \ldots, D\}$ to all $D-1$ possible next states $\check{s} \in \{1, 2, \ldots, D\}\backslash s$ is described by a Dirichlet distribution $\pi_n(\theta_s) \propto \Pi_{\check{s}}\, \theta(s, \check{s})^{\alpha(s,\check{s})-1}$ parameterized by $\alpha_s = \alpha(s, :)$. Here, $\theta_s = \theta(s, :)$ denotes a vector of random variable $\theta(s, \check{s})$ that determines the probability of transition from s to $\check{s}$ with properties $0 \le \theta(s, \check{s}) \le 1$ and $\sum_{\check{s}} \theta(s, \check{s}) = 1$. The likelihood function for an occurred transition $X : s \to s'$ is $p(X|\theta_s) = \theta(s, s') = \Pi_{\check{s}}\, \theta(s, \check{s})^{[\check{s}=s']}$, where [.] denotes the Iverson bracket (that is equal to 1 if the condition inside the bracket is correct and 0 otherwise). Therefore, the updated belief $q_\gamma(\theta_s)$ obtained by the SMiLe rule, equation 3.6,

$$q_\gamma(\theta_s) \propto \left(\Pi_{\check{s}}\, \theta(s, \check{s})^{[\check{s}=s']}\right)^{\gamma} . \left(\Pi_{\check{s}}\, \theta(s, \check{s})^{\alpha(s,\check{s})-1}\right)^{1-\gamma} \propto \Pi_{\check{s}}\, \theta(s, \check{s})^{\beta(s,\check{s})-1}, \tag{6.12}$$

is again a Dirichlet distribution parameterized by $\beta(s, \check{s}) = (1 - \gamma)\alpha(s, \check{s}) + \gamma(1 + [\check{s} = s'])$.

The probability $\hat{T}[t](s, s')$ of transition from s to s' at time step t is estimated by $\hat{T}[t](s, s') = \frac{\alpha[t](s,s')-1+\epsilon}{\sum_{\check{s}}(\alpha[t](s,\check{s})-1+\epsilon)}$, where $\alpha[t](s, \check{s})$ denotes the updated model parameter at time step t. Here, $\epsilon > 0$ is a very small number, which prevents the denominator from being zero.

Appendix

A.1 Calculation of Surprise for the Example of CEO Election. If candidate 1 is selected, the surprise of colleague B, $S_{cc}(X = 1; \pi^B)$ is bigger than the surprise of colleague A, $S_{cc}(X = 1; \pi^A)$. Both colleagues are equally committed to their beliefs, but the outcome "candidate 1" is less likely for colleague B than A. The evaluation of confidence-corrected surprise yields

$$\begin{aligned} & S_{cc}(X = 1; \pi^B) - S_{cc}(X = 1; \pi^A) \\ &= \sum_k \pi^B(\theta_k) \ln \frac{\pi^B(\theta_k)}{\hat{p}_{X=1}(\theta_k)} - \sum_k \pi^A(\theta_k) \ln \frac{\pi^A(\theta_k)}{\hat{p}_{X=1}(\theta_k)} \\ &= \sum_k \left(\pi^A(\theta_k) - \pi^B(\theta_k)\right) \ln \hat{p}_{X=1}(\theta_k), \end{aligned} \tag{A.1}$$

which yields $0.75 \ln \frac{(1-\epsilon)}{\epsilon/3} > 0$ once we insert the numbers. Therefore, B is more surprised than A.

For colleague A, surprise of the outcome "candidate 2," $S_{cc}(X = 2; \pi^A)$, is bigger than the surprise of outcome "candidate 1" (his favorite), $S_{cc}(X = 1; \pi^A)$ because the second candidate is less likely to win in his opinion (see point 2 at the beginning of section 3.1):

$$S_{cc}(X = 2; \pi^A) - S_{cc}(X = 1; \pi^A) = \sum_k \pi^A(\theta_k) \ln \frac{\hat{p}_{X=1}(\theta_k)}{\hat{p}_{X=2}(\theta_k)}, \tag{A.2}$$

which yields $0.5 \ln \frac{1-\epsilon}{\epsilon/3} > 0$.

More important, however, if the second candidate wins, the surprise of colleague A is bigger than that of colleague C, even though both have assigned the same low probability to the second candidate. The evaluation of surprise yields

$$\begin{aligned} S_{cc}(X = 2; \pi^A) - S_{cc}(X = 2; \pi^C) = & \sum_k \left(\pi^C(\theta_k) - \pi^A(\theta_k)\right) \ln \hat{p}_{X=2}(\theta_k) \\ & - H(\pi^A) + H(\pi^C). \end{aligned} \tag{A.3}$$

The terms with the ln in equation A.3 add up to zero, so that we just need to evaluate the entropies, which yields a difference of $0.75 \ln 3 > 0$. In other words, since colleague A is more committed to his opinion than colleague C, that is, $H(\pi^C) > H(\pi^A)$, colleague A will be more surprised (see point 3 in section 3.1).

A.2 A Bound $B > B'$ Implies $\gamma > \gamma'$ in the SMiLe Rule. For $0 < B < B_{\max}$ the solution of the optimization problem in equation 3.5 is the updated belief q_γ, equation 3.6, with $0 < \gamma < 1$ satisfying $D_{KL}[q_\gamma || \pi_n] = B$. In order to prove that $B > B'$ implies $\gamma > \gamma'$, we need to show that $D_{KL}[q_\gamma || \pi_n]$ is an increasing function of γ. We therefore evaluate its derivative with respect to γ.

As a first step, we calculate the derivative of $q_\gamma(\theta)$, equation 3.6, with respect to γ:

$$\frac{\partial}{\partial \gamma} q_\gamma(\theta) = q_\gamma(\theta) \left(\ln \frac{p(X|\theta)}{\pi_n(\theta)} - \left\langle \ln \frac{p(X|\theta)}{\pi_n(\theta)} \right\rangle_{q_\gamma} \right). \tag{A.4}$$

We use this result together with

$$\int_\theta \frac{\partial}{\partial \gamma} q_\gamma(\theta)\, d\theta = 0. \tag{A.5}$$

to calculate the derivative of $D_{KL}[q_\gamma || \pi_n]$ with respect to γ:

$$\begin{aligned}
&\frac{\partial}{\partial \gamma} D_{KL}[q_\gamma || \pi_n] \\
&= \int_\theta \left(\ln \frac{q_\gamma(\theta)}{\pi_n(\theta)} + 1 \right) \frac{\partial}{\partial \gamma} [q_\gamma(\theta)]\, d\theta \\
&\overset{(A.4)}{=} \gamma \int_\theta \left(\ln \frac{p(X|\theta)}{\pi_n(\theta)} \right) \left(\ln \frac{p(X|\theta)}{\pi_n(\theta)} - \left\langle \ln \frac{p(X|\theta)}{\pi_n(\theta)} \right\rangle_{q_\gamma} \right) q_\gamma(\theta)\, d\theta \\
&= \gamma \left(\left\langle \left(\ln \frac{p(X|\theta)}{\pi_n(\theta)} \right)^2 \right\rangle_{q_\gamma} - \left(\left\langle \ln \frac{p(X|\theta)}{\pi_n(\theta)} \right\rangle_{q_\gamma} \right)^2 \right) \geq 0.
\end{aligned} \tag{A.6}$$

This finishes the proof of our claim.

A.3 The Impact Function $\Delta S_{cc}(q; \pi_n, X)$ Increases with the Parameter γ in the SMiLe Rule. To prove the statement above, we need to show that the impact function $\Delta S_{cc}(q; \pi_n, X)$ in equation 3.4, where the SMiLe rule, equation 3.6, is used for belief update (i.e., when $q = q_\gamma$), increases with the

parameter γ. We therefore consider the derivative

$$\frac{\partial}{\partial\gamma}\Delta S_{cc}(q_\gamma;\pi_n,X)=\int_\theta\left(\ln\frac{p(X|\theta)}{q_\gamma(\theta)}-1\right)\frac{\partial}{\partial\gamma}\big[q_\gamma(\theta)\big]\,d\theta. \tag{A.7}$$

We now use equations 3.6, A.5, and A.4 and find

$$\frac{\partial}{\partial\gamma}\Delta S_{cc}(q_\gamma;\pi_n,X)=(1-\gamma)\left[\left\langle\left(\ln\frac{p(X|\theta)}{\pi_n(\theta)}\right)^2\right\rangle_{q_\gamma}-\left(\left\langle\ln\frac{p(X|\theta)}{\pi_n(\theta)}\right\rangle_{q_\gamma}\right)^2\right], \tag{A.8}$$

which is always nonnegative.

A.4 A Larger Reduction in the Surprise Implies a Bigger Change in Belief. The minimal value of the Lagrangian $\mathbb{L}(q,\lambda)$ in equation 6.5 that is achieved by the updated belief q_γ in equation 3.6, obtained by the SMiLe rule, is equal to

$$\begin{aligned}\mathbb{L}(q_\gamma,\lambda)&\overset{(6.5)}{=}\left\langle-\ln p(X|\theta)+\ln q_\gamma(\theta)+\frac{1}{\lambda}\ln\frac{q_\gamma(\theta)}{\pi_n(\theta)}\right\rangle_{q_\gamma}\overbrace{-\frac{B}{\lambda}+\ln||p||}^{=C}\\&=-\frac{1}{\gamma}\ln Z(X;\gamma)+C,\end{aligned} \tag{A.9}$$

where we used the SMiLe rule, equation 3.6, and the equality $\frac{1}{\gamma}=1+\frac{1}{\lambda}$, from equation 6.7. If the optimal solution q_γ is approximated by any other potential next belief q, then its corresponding functional value $\mathbb{L}(q,\lambda)$ differs from its minimal value $\mathbb{L}(q_\gamma,\lambda)$ in proportion to the KL divergence $D_{KL}[q||q_\gamma]$. This is because

$$\begin{aligned}\mathbb{L}(q,\lambda)-\mathbb{L}(q_\gamma,\lambda)&\overset{(6.5),(A.3)}{=}\left\langle-\ln p(X|\theta)+\ln q(\theta)+\frac{1}{\lambda}\ln\frac{q(\theta)}{\pi_n(\theta)}\right\rangle_q\\&\quad+\frac{1}{\gamma}\ln Z(X;\gamma)\\&=\frac{1}{\gamma}\left\langle\ln\frac{q(\theta)^{\gamma(1+\frac{1}{\lambda})}Z(X;\gamma)}{p(X|\theta)^\gamma\pi_n(\theta)^{\frac{\gamma}{\lambda}}}\right\rangle_q\\&=\frac{1}{\gamma}D_{KL}[q||q_\gamma].\end{aligned} \tag{A.10}$$

Replacing q with π_n in equation yields the impact function $\Delta S_{cc}(q; \pi_n, X)$ in equation 3.4:

$$\Delta S_{cc}(q; \pi_n, X) \overset{(6.5)}{=} \mathbb{L}(\pi_n, \lambda) + \frac{1}{\lambda}B - \mathbb{L}(q_\gamma, \lambda) - \frac{1}{\lambda}(B - D_{KL}[q_\gamma||\pi_n])$$
$$\overset{(A.10)}{=} \frac{1}{\gamma}D_{KL}[\pi_n||q_\gamma] + \frac{1}{\lambda}D_{KL}[q_\gamma||\pi_n] \geq 0, \tag{A.11}$$

where $1/\lambda = 1 - 1/\gamma$ and q_γ is the updated belief under the SMiLe rule: $\pi_{n+1} = q_\gamma$. Therefore, the reduction in the surprise upon a second exposure to the same data sample is related to the belief changes $D_{KL}[\pi_n||\pi_{n+1}]$ and $D_{KL}[\pi_{n+1}||\pi_n]$. The equality in equation A.11 holds if and only if there is no change in the current belief, that is, if $q_\gamma = \pi_{n+1} = \pi_n$. This happens only if $\gamma = 0$, which is equivalent to neglecting the new data point when updating the belief.

A.5 The Online EM Algorithm for the Maze-Exploration Task. The online EM algorithm, presented in Mongillo and Deneve (2008), is an estimation algorithm for the unknown parameters of a hidden Markov model (HMM). For the maze exploration task we adapted the method presented in Mongillo and Deneve (2008) such that the transition probability to a new room also depends on the previously visited room (and not just the current environment). The HMM of the maze exploration task consists of two sets of unknown parameters: (1) a set $\mathbf{P} = [P_{ij}]_{2\times 2}$ of (unknown) switch probabilities from environment i to j (where we use 1 for environment $\mathcal{A}$ and 2 for environment $\mathcal{B}$), and (2) a set $\mathbf{T} = [T_{jss'}]_{2\times 16\times 16}$ of state transition probabilities, where $T_{jss'}$ denotes the probability of transition from state s to state s' within environment j. The set of all unknown parameters is denoted by $\theta \equiv (\mathbf{P}, \mathbf{T})$.

At each time step t, we estimate the probability $q_l^t = P(E_t = l|s_{0\to t})$ of being in environment $E_t = l \in \{1, 2\}$, given all previous state transitions $s_{0\to t} = \{s_0, s_1, \ldots, s_t\}$. The probability q_l^t can be recursively calculated by

$$\hat{q}_l^t = \sum_m \hat{q}_m^{t-1}\gamma_{ml}^t, \tag{A.12}$$

where $\gamma_{ml}^t = \frac{P(s'=s_t|s=s_{t-1},E_t=l)P(E_t=l|E_{t-1}=m)}{P(s'=s_t|s_{0\to(t-1)})}$ belongs to a set of auxiliary variables $\mathbf{\Gamma} = [\gamma_{lh}]_{2\times 2}$ that are calculated by the last estimate $\hat{\theta}^{t-1}$ of the model parameters:

$$\gamma_{lh}^t = \frac{\hat{P}_{lh}^{t-1}\hat{T}_{hs_{t-1}s_t}^{t-1}}{\sum_{m,n}\hat{q}_m^{t-1}\hat{P}_{mn}^{t-1}\hat{T}_{ns_{t-1}s_t}^{t-1}}. \tag{A.13}$$

Then, using these auxiliary variables γ_{lh}, a set $\mathbf{\Phi} = [\hat{\phi}_{i,j,s,s',h}]_{2\times2\times16\times16\times2}$ of parameters is recursively updated:

$$\hat{\phi}^t_{i,j,s,s',h} = \sum_l \gamma^t_{lh}\Big[(1-\eta)\hat{\phi}^{t-1}_{i,j,s,s',l} + \eta\hat{q}^{t-1}_l \Delta^{lhs_{t-1}s_t}_{ijss'}\Big], \tag{A.14}$$

where $\Delta^{lhs_{t-1}s_t}_{ijss'} = \delta(i-l)\delta(j-h)\delta(s-s_{t-1})\delta(s'-s_t)$, $\delta(.)$ is the Kronecker delta (i.e., 1 when its argument is zero and 0 otherwise), and η is the learning rate.

Finally, the model parameters are updated by

$$\hat{P}^t_{ij} = \frac{\sum_{s,s',h}\hat{\phi}^t_{ijss'h}}{\sum_{j,s,s',h}\hat{\phi}^t_{ijss'h}}; \quad \hat{T}^t_{jss'} = \frac{\sum_{i,h}\hat{\phi}^t_{ijss'h}}{\sum_{i,s',h}\hat{\phi}^t_{ijss'h}}. \tag{A.15}$$

We emphasize that in order for the online EM algorithm to work properly, some technical considerations must be respected. For instance, at the beginning of learning, only the online estimation of $\mathbf{\Phi}$ must be updated (without updating the model parameters θ), so that the estimation error for the first 2000 time steps of our simulation (see Figure 7A, blue) remains fixed. Moreover, we found that the online EM algorithm works well only if it is correctly initialized. To make our comparison fair, we assumed the agent "believes in" frequent transitions between environments by initializing the probabilities $\hat{P}^0_{ij}$ that describe the switch between environment $\mathcal{A}$ and $\mathcal{B}$ to be very close to true ones. Without such an assumption, the online EM takes even more time than what we reported here to learn the maze exploration task. The actual initialization values were $\hat{P}^0_{12} = \hat{P}^0_{21} = 0.1$, while the true values were $P_{12} = P_{21} = 0.005$.

Acknowledgments

This project has been funded by the European Research Council under grant agreement 268689 and by the European Union's Horizon 2020 research and innovation program under grant agreement 720270.

References

Adams, R., & MacKay, D. (2007). *Bayesian online changepoint detection*. arXiv: 0710.3742.

Aston-Jones, G., & Cohen, J. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

Aston-Jones, G., Rajkowski, J., & Kubiak, P. (1997). Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience*, *80*, 697–715.

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, *23*, 649–666.

Bear, M., & Singer, W. (1986). Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*, *320*, 172–176.

Beck, J., Ma, W., Kiani, R., Hanks, T., Churchland, A. K., Roitma, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*, 1142–1152.

Behrens, T., Woolrich, M., Waltan, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nat. Neuroci.*, *10*, 1214–1221.

Berkes, P., Orban, G., Lengyel, M., & Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, *331*, 83–87.

Bishop, C. (1994). Novelty detection and neural network validation. *IEEE Proceedings in Vision, Image and Signal Processing*, *141*, 217–222.

Bouret, S., & Sara, S. (2004). Reward expectation, orientation of attention and locus coeruleus–medial frontal cortex interplay during learning. *Europ. J. Neurosci.*, *20*, 791–802.

Bouret, S., & Sara, S. (2005). Network reset: A simplified overarching theory of locus coeruleus. *Trends in Neurosciences*, *28*, 574–582.

Brea, J., Senn, W., & Pfister, J.-P. (2013). Matching recall and storage in sequence learning with spiking neural networks. *J. Neuroscience*, *33*, 9565–9575.

Clauton, E., Rajkowski, J., Cohen, J., & Aston-Jones, G. (2004). Phasic activation of monkey locus coeruleus neurons by simple decisions in a forced-choice task. *J. Neurosci.*, *24*, 9914–9920.

Cohen, J., McClure, S., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. Roy. Soc. London B*, *362*, 933–942.

Costa, V., & Rudebeck, P. (2016). More than meets the eye: The relationship between pupil size and locus coeruleus activity. *Neuron*, *89*, 8–10.

Donchin, E., Ritter, W., & McCallum, W. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. H. Kroslov (Eds.), *Event-related potentials in man*. New York: Academic Press.

Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fairhall, A., Lewen, G., Bialek, W., & de Ruyter van Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412* (6849), 787–792.

Fiser, J., Berkes, P., Orban, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cogn. Sci.*, *14*, 119–130.

Frank, M., Leitner, J., Stollenga, M., Forster, A., & Schmidhuber, J. (2013). Curiosity driven reinforcement learning for motion planning on humanoids. *Front. Neurorobotics*, *7*, 25.

Frémaux, N., & Gerstner, W. (2016). Neuromodulated spike-timing dependent plasticity and theory of three-factor learning rules. *Front. Neural Circuits*, *9*, 85.

Friston, K. (2005). A theory of cortical responses. *Phil. Trans. R. Soc. B*, *360*, 815–830.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.*, *11*, 127–138.
Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neurosci. and Behav. Rev.*, *68*, 862–879.
Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*, 1–49.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Phil. Trans. Roy. Soc. London B*, *364*, 1211–1221.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.*, *6*, 187–214.
Gershman, S., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cogn. Sci.*, *7*, 391–415.
Gershman, S., Radulescu, A., Norman, K., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Comput. Biol.*, *10*, e1003939.
Gil, Z., Connors, B., & Amitai, Y. (1997). Differential regulation of neocortical synapses by neuromodulators and activity. *Neuron*, *19*, 679–686.
Gillner, S., & Mallot, H. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *J. Cognitive Neuroscience*, *10*, 445–463.
Gu, Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience*, *111*, 815–835.
Hasselmo, M. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, *3*, 351–359.
Hasselmo, M., Wyble, B., & Wallenstein, G. (1996). Encoding and retrieval of episodic memories: Role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus*, *6*, 693–708.
Hayden, B., Heilbronner, S., Pearson, J., & Platt, M. (2011). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *J. Neurosci.*, *31*, 4178–4187.
Hess, E., & Polt, J. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.
Holland, P. (1997). Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. *Animal Learning and Behavior*, *25*, 373–399.
Hsie, C., Cruikshank, S., & Metherate, R. (2000). Differential modulation of auditory thalamocortical and intracortical synaptic transmission by cholinergic agonist. *Brain Research*, *880*, 51–64.
Hurley, M., Dennett, D., & Adams, R. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Cambridge, MA: MIT Press.
Iglesias, S., Mathys, C., Brodersen, K., Kasper, L., Piccirelli, M., den Ouden, H., & Stephan, K. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, *80*, 519–530.
Iigaya, K. (2016). Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *eLife*, *5*, e18073.
Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295–1306.

Jepman, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *J. Cogn. Neurosci.*, *23*, 1587–1596.

Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Comput. Biology*, *9*, e1003094.

Jones, J., & Higgins, G. (1995). Effect of scopolamine on visual attention in rats. *Psychopharmacology*, *120*, 142–149.

Kalat, J. (2016). *Biological psychology*. Boston: Cengage Learning.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.*, *82*, 35–45.

Kimura, F., Fukuda, M., & Tsumoto, T. (1999). Acetylcholine suppresses the spread of excitation in the visual cortex revealed by optical recording: Possible differential effect depending on the source of input. *Europ. J. Neurosci.*, *11*, 3597–3609.

Knight, R. T. (1996). Contribution of human hippocampal region to novelty detection. *Nature*, *383*, 256–259.

Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*, 712–719.

Kobayashi, M., Imamura, K., Sugai, T., Onoda, N., Yamamoto, M., Komai, S., & Watanabe, Y. (2000). Selective suppression of horizontal propagation in rat visual cortex by norepinephrine. *Europ. J. Neurosci.*, *12*, 264–272.

Kolossa, A., Fingscheidt, T., Wessel, K., & Kopp, B. (2013). A model-based approach to trial-by-trial P300 amplitude fluctuations. *Front. Hum. Neurosci.*, *6*, 359.

Kolossa, A., Kopp, B., & Finscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *NeuroImage*, *106*, 222–237.

Körding, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Krugel, L., Biele, G., Mohr, P., Li, S.-C., & Heekeren, H. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Natl. Acad. Sci. USA*, *106*, 17951–17956.

Little, D., & Sommer, F. (2013). Learning and exploration in action-perception loops. *Front. Neural Circuits*, *7*, 37.

Ma, W., Beck, J., Latham, P., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, *9*, 1432–1438.

MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. (2011). A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.*, *5*, 39.

Mathys, C. D., Lomakina, E. I., Daunizeau, K., Iglesias, S., Brodersen, K., Friston, K., & Stephan, K. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Front. Hum. Neurosci.*, *8*, 825.

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci. USA*, *114*, 3859–3868.

Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human inferences about sequences: A minimal transition probability model. *PLoS Comput. Biol.*, *12*, e1005260.

Meyniel, F., Sigman, M., & Mainen, Z. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*, 78–92.

Missionier, P., Ragot, R., Derousne, C., Guez, D., & Renault, B. (1999). Automatic attentional shifts induced by a noradrenergic drug in Alzheimers disease: Evidence from evoked potentials. *Int. J. Psychophysiology*, *33*, 243–251.

Mohamed, S., & Rezende, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 2125–2133). Red Hook, NY: Curran.

Mongillo, G., & Deneve, S. (2008). Online learning with hidden Markov models. *Neural Computation*, *20*, 1706–1716.

Morris, R., Garrard, P., Rawlins, J., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*, 681–683.

Nassar, M., Rumsey, K., Wilson, R., Parikh, K., Heasly, B., & Gold, J. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nat. Neurosci.*, *15*, 1040–1046.

Nassar, M., Wilson, R., Heasly, B., & Gold, J. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.*, *30*, 12366–12378.

Nelson, A., Grant, E., Galeotti, J., & Rhody, S. (2004). Maze exploration behaviors using an integrated evolutionary robotics environment. *Robotics and Autonomous Systems*, *46*, 159–173.

O'Doherty, J., Dayan, P., Friston, K., Critchley, H., & Dolan, R. (2003). Temporal difference learning model accounts for responses in human ventral striatum and orbitofrontal cortex during Pavlovian appetitive learning. *Neuron*, *38*, 329–337.

O'Doherty, J., Dayan, P., Schultz, J., Deischmann, R., Friston, K., & Dolan, R. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

Oudeyer, P., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.*, *11*, 265–286.

Palm, G. (2012). *Novelty, information and surprise*. Heidelberg: Springer.

Payzan-Nestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput. Biol.*, *7*, e1001048.

Pearce, J., & Hall, G. (1980). A model of Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.*, *87*, 532–552.

Pineda, J., Westerfield, M., Kronenberg, B., & Kubrin, J. (1997). Human and monkey P3-like responses in a mixed modality paradigm: Effects of context and context-dependent noradrenergic influences. *Int. J. Psychophysiol.*, *27*, 223–240.

Preuschoff, K., & Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. *Annals of New York Acad. Sci.*, *1104*, 135–146.

Preuschoff, K., t'Hart, B., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenalines role in decision making. *Front. Neurosci.*, *5*, 115.

Preuschoff, K., Quartz, S., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *J. Neurosci.*, *28*, 2745–2752.

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Res. Bull.*, *35*, 607–616.

Ranganath, C., & Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat. Rev. Neurosci.*, *4*, 193–202.

Rezende, D., & Gerstner, W. (2014). Stochastic variational learning in recurrent spiking networks. *Frontiers in Computational Neuroscience*, *8*, 38.

Roesch, M., Esber, G., Li, J., Daw, N., & Schoenbaumn, G. (2012). Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *Europ. J. Neurosci.*, *35*, 1190–1200.

Ruter, J., Marcille, N., Sprekeler, H., Gerstner, W., & Herzog, M. (2012). Paradoxical evidence integration in rapid decision processes. *PLoS Comput. Biol.*, *8*, e1002382.

Sara, S. (1998). Learning by neurones: Role of attention, reinforcement and behaviour. *Comptes Rendus de l'Académie des Sciences—Séries III—Sciences de la Vie*, *321*, 193–198.

Sara, S. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nat. Rev. Neurosci.*, *10*, 211–223.

Sara, S., Vankov, A., & Herve, A. (1994). Locus coeruleus-evoked responses in behaving rats: A clue to the role of noradrenaline in memory. *Brain Res. Bull.*, *35*, 457–465.

Schmidhuber, J. (1991). Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks, Singapore* (vol. 2, pp. 1458–1463). Piscataway, NJ: IEEE Press.

Schmidhuber, J. (2003). Exploring the predictable. In A. Ghosh & S. Tsutsui (Eds.), *Advances in evolutionary computing* (pp. 579–612). New York: Springer.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, *18*, 173–187.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Autonom. Mental Developm.*, *2*, 230–247.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Schultz, W. (2010). Dopamine signals for reward value and risk: Basic and recent data. *Behavioral and Brain Functions*, *6*, 24.

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiol. Rev.*, *95*, 853–951.

Schultz, W. (2016). Dopamine reward prediction-error signalling: A two-component response. *Nat. Rev. Neurosci.*, *17*, 183–195.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 37–423.

Silver, D., van Haselt, H., Hessel, M., Schaul, T., Guez, A., Harley, T., . . . Degris, T. (2016). *The predictron: End-to-end learning and planning*. arXiv:1612.08810.

Singh, S., Barto, A., & Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In L. Bottou, L. K. Saul, & Y. Weiss (Eds.), *Advances in neural information processing systems*, *17* (pp. 1281–1288). Cambridge, MA: MIT Press.

Squires, K., Wickens, C., Squires, N., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, *193*, 1141–1146.

Storck, J., Hochreiter, S., & Schmidhuber, J. (1995). Reinforcement-driven information acquisition in non-deterministic environments. In *Proc. International Conference on Artificial Networks* (vol. 2, pp. 159–164). Piscataway, NJ: IEEE.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal bayesian exploration in dynamic environments. In J. Schmidhuber, K. R. Thorisson, & M. Looks (Eds.), *Artificial general intelligence* (pp. 41–51). New York: Springer.

Sutton, R., Modayil, J., Delp, M., Degris, T., Pilarski, P., White, A., & Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems* (vol. 2, pp. 761–768). International Foundation of Autonomous Agents.

Tribus, M. (1961). Information theory as the basis for thermostatics and thermodynamics. *J. Appl. Mechanics*, *28*, 1–8.

Vankov, A., Minvielle, H., & Sara, S. (1995). Response to novelty and its rapid habituation in locus coeruleus neurons of the freely exploring rat. *Europ. J. Neurosci.*, *7*, 1180–1187.

Verleger, R., Jaskowski, P., & Wauschkuhn, B. (1994). Suspense and surprise: On the relationship between expectancies and P3. *Psychophysiology*, *31*, 359–369.

Vossel, S., Bauer, M., Mathys, C., Adams, R., Dolan, R., Stephan, K., & Friston, K. (2014). Cholinergic stimulation enhances Bayesian belief updating in the deployment of spatial attention. *J. Neurosci.*, *34*, 15735–15742.

Wallenstein, G., Hasselmo, M., & Eichenbaum, H. (1998). The hippocampus as an associator of discontiguous events. *Trends in Neurosci.*, *21*, 317–323.

Wilson, P., Boumphrey, P., & Pearce, J. (1992). Restoration of the orienting response to a light by a change in its predictive accuracy. *Quart. J. Exp. Psych., B*, *44*, 17–36.

Wilson, R., Nassar, M., & Gold, J. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. *PLoS Comput. Biol.*, *9*, e1003150.

Yu, A., & Cohen, J. (2008). Sequential effects: Superstition or rational behavior? In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*, *21* (pp. 1873–1880). Cambridge, MA: MIT Press.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*, 681–692.

Received March 1, 2017; accepted July 7, 2017.