Journal of Mathematical Psychology 110 (2022) 102712

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A taxonomy of surprise definitions

Alireza Modirshanechi *, Johanni Brea, Wulfram Gerstner

EPFL, School of Computer and Communication Sciences and School of Life Sciences, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 6 April 2022
Received in revised form 23 August 2022
Accepted 26 August 2022
Available online xxxx

Keywords:
Surprise
Prediction error
Probabilistic modeling
Predictive brain
Predictive coding
Bayesian brain

ABSTRACT

Surprising events trigger measurable brain activity and influence human behavior by affecting learning, memory, and decision-making. Currently there is, however, no consensus on the definition of surprise. Here we identify 18 mathematical definitions of surprise in a unifying framework. We first propose a technical classification of these definitions into three groups based on their dependence on an agent's belief, show how they relate to each other, and prove under what conditions they are indistinguishable. Going beyond this technical analysis, we propose a taxonomy of surprise definitions and classify them into four conceptual categories based on the quantity they measure: (i) 'prediction surprise' measures a mismatch between a prediction and an observation; (ii) 'change-point detection surprise' measures the probability of a change in the environment; (iii) 'confidence-corrected surprise' explicitly accounts for the effect of confidence; and (iv) 'information gain surprise' measures the belief-update upon a new observation. The taxonomy poses the foundation for principled studies of the functional roles and physiological signatures of surprise in the brain.

1. Introduction

Imagine you open the curtains one morning and find the street in front of your apartment covered by fresh snow. If you have expected a warm and sunny morning according to the weather forecast, you feel 'surprised' as you see the white streets; as a consequence of surprise, the activity of many neurons in your brain changes (Kolossa et al., 2015; Mars et al., 2008; Squires et al., 1976) and your pupils dilate (Antony et al., 2021; Nassar et al., 2012; Preuschoff et al., 2011). Surprise affects how we predict and perceive our future and how we remember our past. For example, some studies suggest that you would rely less on the weather forecast for your future plans after the snowy morning (Behrens et al., 2007; Nassar et al., 2010; Xu et al., 2021). Other studies predict that you would remember more vividly the face of the random stranger who walked past the street in that very moment you felt surprised (Rouhani & Niv, 2021; Rouhani et al., 2018), and some predict that this moment of surprise might have even modified your memory of another snowy morning in the past (Gershman et al., 2017; Sinclair & Barense, 2018). To understand and explain the computational role of surprise in different brain functions, one first needs to ask 'what does it really mean to be surprised?' and formalize how surprise is perceived by our brain. For instance, when you see the white street, do you feel 'surprised' because what you expected turned out to be wrong (Faraji et al., 2018; Gläscher et al., 2010; Meyniel et al., 2016) or because you need to change your trust in the weather forecast (Baldi, 2002; Liakoni et al., 2021; Schmidhuber, 2010)?

Computational models of perception, learning, memory, and decision-making often assume that humans implicitly perceive their sensory observations as probabilistic outcomes of a generative model with hidden variables (Findling et al., 2021; Fiser et al., 2010; Friston, 2010; Gershman et al., 2017; Liakoni et al., 2021; Soltani & Izquierdo, 2019; Yu & Dayan, 2005). In the example above, the observation is whether it snows or not and the hidden variables characterize how the probability of snowing depends on old observations and relevant context information (such as the current season, yesterday's weather, and the weather forecast). Different brain functions are then modeled as aspects of statistical inference and probabilistic control in such generative models (Behrens et al., 2007; Daw et al., 2011; Dubey & Griffiths, 2019; Findling et al., 2021; Friston et al., 2017; Gershman et al., 2017; Gläscher et al., 2010; Horvath et al., 2021; Liakoni et al., 2021; Meyniel et al., 2016; Nassar et al., 2012; Yu & Dayan, 2005). In these probabilistic settings, surprise of an observation depends on the relation between the observation and our expectation of what to observe.

In the past decades, different definitions and formal measures of surprise have been proposed and studied (Baldi, 2002; Barto et al., 2013; Faraji et al., 2018; Friston, 2010; Gläscher et al., 2010; Kolossa et al., 2015; Liakoni et al., 2021; Palm, 2012; Schmidhuber, 2010). These surprise measures have been successful both in explaining the role of surprise in different brain functions (Antony et al., 2021; Findling et al., 2021; Gershman et al., 2017; Itti & Baldi, 2006; Rouhani & Niv, 2021; Xu et al., 2021)

* Corresponding author.
E-mail address: alireza.modirshanechi@epfl.ch (A. Modirshanechi).

https://doi.org/10.1016/j.jmp.2022.102712

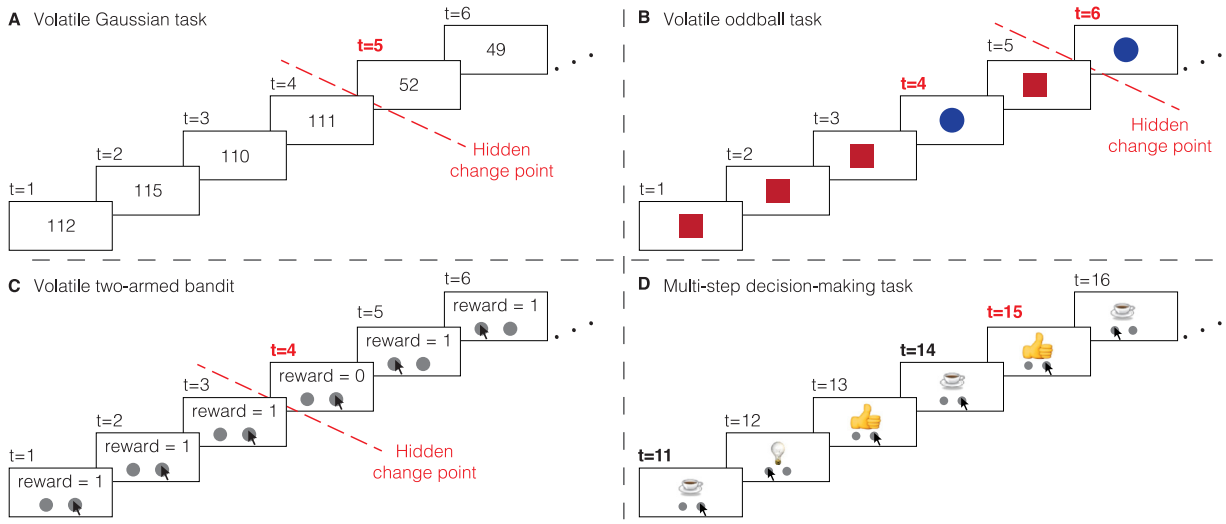

Fig. 1. Four typical experimental paradigms to study functional roles and physiological signatures of surprise in the brain. A. Volatile Gaussian task (Nassar et al., 2012, 2010): Participants see a sequence of numbers randomly sampled from a Gaussian distribution whose mean is piece-wise constant but abruptly changes at random points in time (change-points, e.g., $t = 5$ in the figure). The goal of participants is to predict the next observation; hence, the first few observations after a change-point are unexpected. Variants of this paradigm have been studied by O'Reilly et al. (2013) and Visalli et al. (2021). **B.** Volatile oddball task (Heilbron & Meyniel, 2019; Meyniel, 2020): Participants see a sequence of binary stimuli (e.g., a red square and a blue disk). The stimulus frequencies are piece-wise constant but abruptly change at random points in time (change-points, e.g., $t = 6$ in the figure). During the stationary periods between two consecutive change-points (before $t = 6$ in the figure), one stimulus (the blue disk, called 'deviant') is less frequent than the other (the red square, called 'standard') and hence more surprising than the other. Variants of the paradigm with more than 2 types of stimuli (Lieder et al., 2013; Mars et al., 2008) or without change-points (Huettel et al., 2002; Maheu et al., 2019; Modirshanechi et al., 2019; Squires et al., 1976) have also been studied. **C.** Volatile two-armed bandit task (Behrens et al., 2007; Horvath et al., 2021): Participants select one action (e.g., click on one of the gray disks in the figure) at a time and receive a reward value randomly sampled from a distribution specific to the selected action. The reward distributions are piece-wise stationary but switch at random change points (e.g., $t = 4$ in the figure). Participants optimize reward and have to adapt their strategy after a change-point. Variants of the paradigm include, e.g., multi-dimensional actions (Niv et al., 2015) or context-dependent reward distributions (Rouhani & Niv, 2021). **D.** Multi-step decision-making task (Gläscher et al., 2010; Liakoni et al., 2022; Xu et al., 2021): Participants move between states (e.g., images of different objects) by selecting one action (e.g., clicking on one of the disks in the figure) at a time. Assuming some transitions have been experienced before (e.g., the 'light bulb' state followed by selecting the right action in the 'cup' state), observing the 'light bulb' state at $t = 12$ is expected, whereas observing the 'thumb' state at $t = 15$ after the same stimulus-action sequence at $t = 14$ as at $t = 11$ is unexpected and hence surprising. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and in identifying signatures of surprise in behavioral and physiological measurements (Gijsen et al., 2021; Gläscher et al., 2010; Maheu et al., 2019; Mars et al., 2008; Modirshanechi et al., 2019; Rubin et al., 2016). However, there are still many open questions including, but not limited to: (i) Are the quantities that different definitions of surprise measure conceptually different? (ii) Can we identify mathematical relations between different surprise definitions? In particular, is one definition a special case of another one, completely distinct, or do they have some common ground?

In this work, we analyze and discuss 18 previously proposed measures of surprise in a unifying framework. We first present our framework, assumptions, and notation in Section 2. Then, in Section 3 to Section 6, we give definitions for each of the 18 surprise measures and show their similarities and differences. In particular, we identify conditions that make different surprise measures experimentally indistinguishable. Finally, in Section 7, we build upon our theoretical analyses and propose a taxonomy of surprise measures by classifying them into four conceptually different categories.

2. Subjective world-model: A unifying generative model

Our goal is to study the theoretical properties of different formal measures of surprise in a common mathematical framework. To do so, we need to make assumptions on how an agent (e.g., a human participant or an animal) thinks about its environment. We assume that an agent thinks of its observations as probabilistic outcomes of a generative model with hidden variables and, hence, consider a generative model that captures several key features of daily life and unifies many existing model environments in neuroscience and psychology (cf. Section 2.2). More specifically, we assume that the generative model describes the subjective interpretation of the environment from the point of view of the agent and, importantly, that the agent takes the possibility into account that the environment may undergo abrupt changes at unknown points in time (i.e., the environment is volatile), similar to the experimental paradigms studied by Behrens et al. (2007), Glaze et al. (2015), Heilbron and Meyniel (2019), Maheu et al. (2019), Nassar et al. (2010), Xu et al. (2021). See Fig. 1 for four typical experimental paradigms that are used to study behavioral and physiological signatures of surprise. Note that we do not assume that the environment has the same dynamics as those assumed by the agent.

2.1. General definition

At each discrete time $t \in \{0, 1, 2, \ldots\}$, the agent's model of the environment is characterized by a tuple of 4 random variables $(X_t, Y_t, \Theta_t, C_t)$ (Fig. 2A). X_t and Y_t are observable, whereas Θ_t and C_t are unobservable (hidden). We refer to X_t as the cue and to Y_t as the observation at time t. Examples of an observation are an image on a computer screen (Kolossa et al., 2015; Mars et al., 2008) (e.g., Fig. 1), an auditory tone (Imada et al., 1993; Lieder et al., 2013), and an electrical stimulation (Ostwald et al., 2012). The cue variable X_t can be interpreted as a predictor of the next observation, since it summarizes the necessary information needed for predicting the observation Y_t. Examples of a cue variable are the previous observation Y_{t-1} (Meyniel et al., 2016; Modirshanechi et al., 2019), the last action of a participant (which we will denote by A_{t-1}) (Behrens et al., 2007; Horvath et al., 2021) (e.g., Fig. 1C-D), and a conditioned stimulus in Pavlovian conditioning tasks (Gershman et al., 2017).

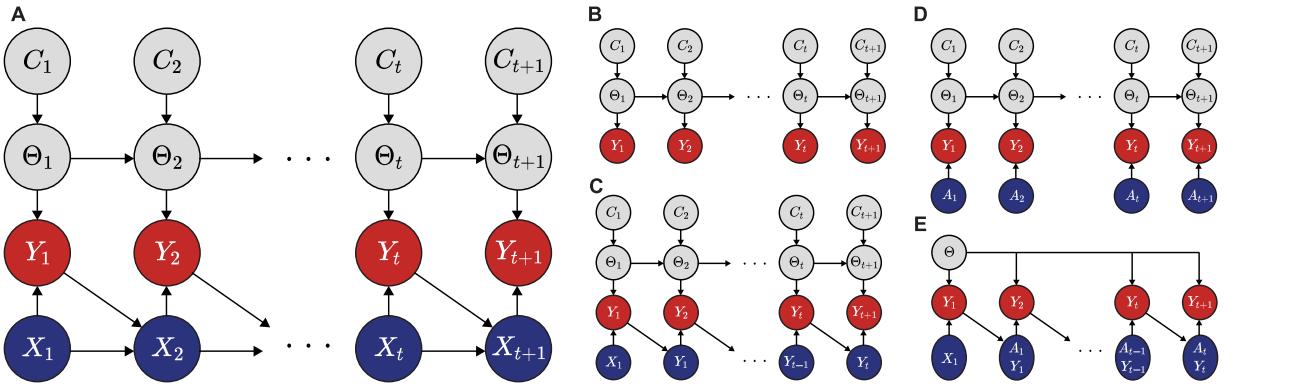

Fig. 2. Subjective model of the environment. A. The Bayesian network (Barber, 2012) corresponding to the most general case of our generative model in Eqs. (1) and (2). The arrows show conditional dependence, the gray nodes show the hidden variables ($C_{1:t+1}$ and $\Theta_{1:t+1}$), the red nodes show the observations ($Y_{1:t+1}$), and the blue nodes show the cue variables ($X_{1:t+1}$). A variety of tasks can be written in the form of a reduced version of our generative model. Specifically: **B.** Standard generative model for modeling and studying passive learning in experiments with volatile environments like the one in Fig. 1A (Adams & MacKay, 2007; Fearnhead & Liu, 2007; Liakoni et al., 2021; Nassar et al., 2012, 2010; Wilson et al., 2013), **C.** Generative model for modeling human inference about binary sequences in experiments like the one in Fig. 1B (Gijsen et al., 2021; Maheu et al., 2019; Meyniel et al., 2016; Modirshanechi et al., 2019; Mousavi et al., 2022), **D.** Generative model corresponding to variants of bandit and volatile bandit tasks like the one in Fig. 1C (Behrens et al., 2007; Findling et al., 2021; Horvath et al., 2021), where the cue variable $X_t = A_t$ is a participant's action, and **E.** Classic Markov Decision Processes (MDPs) to model experiments like the one in Fig. 1D (Daw et al., 2011; Gläscher et al., 2010; Huys et al., 2015; Lehmann et al., 2019; Schultz et al., 1997; Sutton & Barto, 2018), where the cue variable $X_t = (A_{t-1}, Y_{t-1})$ consists of previous action and observation. See Section 2.2 for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

At time t, given the cue variable X_t, the agent assumes that the observation Y_t comes from a distribution that is conditioned on X_t and is parameterized by the hidden variable Θ_t. We do not put any constraints on the sets to which X_t, Y_t, and Θ_t belong. We refer to Θ_t as the environment parameter at time t. The sequence of variables $\Theta_{1:t} = (\Theta_1, \dots, \Theta_t)$ describe the temporal dynamics of the observations $Y_{1:t}$ given the cue variables $X_{1:t}$ in the agent's model of the environment. Similar to well-known models of volatile environments (Adams & MacKay, 2007; Behrens et al., 2007; Fearnhead & Liu, 2007; Findling et al., 2021; Glaze et al., 2015; Heilbron & Meyniel, 2019; Liakoni et al., 2021; Meyniel et al., 2016; Nassar et al., 2012, 2010; Wilson et al., 2013; Xu et al., 2021; Yu & Cohen, 2009; Yu & Dayan, 2005), the agent assumes that the environment undergoes abrupt changes at random points in time (e.g., Fig. 1A–C). An abrupt change at time t is specified by the event $C_t = 1$ and happens with a probability $p_c \in [0, 1)$; otherwise $C_t = 0$. If the environment abruptly changes at time t (i.e., $C_t = 1$), then the agent assumes that the environment parameter Θ_t is sampled from a prior distribution $\pi^{(0)}$ independently of Θ_{t-1}; if there is no change ($C_t = 0$), then Θ_t remains the same as Θ_{t-1}. We refer to p_c as the change-point probability.

We use $\mathbb{P}$ to refer to probability distributions: Given a random variable W and a value $w \in \mathbb{R}$, we use $\mathbb{P}(W = w)$ to refer to the probability of event $\{W = w\}$ for discrete random variables and, with a slight abuse of notation, to the probability density function of W at $W = w$ for continuous random variables. In general, we denote random variables by capital letters and their values by small letters. However, for any pair of arbitrary random variables W and V and their values w and v, whenever there is no risk of ambiguity, we either drop the capital- or the small-letter notation and, for example, write $\mathbb{P}(W = w|V = v)$ as $\mathbb{P}(w|v)$. When there is a risk of ambiguity, we keep the capital notation for the random variables, e.g., we write $\mathbb{P}(W = v, V = v)$ as $\mathbb{P}(W = v, v)$. Given this convention, the agent's model of the environment described above is formalized in Definition 1 (cf. Fig. 2A).

Definition 1 (*Subjective World-Model*)**.** An agent's model of the environment is defined for $t > 0$ as a joint probability distribution over $Y_{1:t}, X_{1:t}, \Theta_{1:t}$, and $C_{1:t}$ as

$$\mathbb{P}\big(y_{1:t}, x_{1:t}, \theta_{1:t}, c_{1:t}\big) := \mathbb{P}\big(c_1\big)\mathbb{P}\big(\theta_1\big)\mathbb{P}\big(x_1\big)\mathbb{P}\big(y_1|x_1, \theta_1\big)\times \prod_{\tau=2}^{t} \mathbb{P}\big(c_\tau\big)\mathbb{P}\big(\theta_\tau|\theta_{\tau-1}, c_\tau\big)\mathbb{P}\big(x_\tau|x_{\tau-1}, y_{\tau-1}\big)\mathbb{P}\big(y_\tau|x_\tau, \theta_\tau\big), \tag{1}$$

where c_1 is by definition equal to 1 (i.e., $\mathbb{P}(c_1) := \delta_{\{1\}}(c_1)$), $\mathbb{P}\big(\theta_1\big) := \pi^{(0)}(\theta_1)$ for an arbitrary distribution $\pi^{(0)}$, and

$$\begin{aligned} \mathbb{P}(c_\tau) &:= \text{Bernoulli}(c_\tau; p_c) \\ \mathbb{P}\big(\theta_\tau|\theta_{\tau-1}, c_\tau\big) &:= \pi^{(0)}(\theta_\tau)\delta_{\{1\}}(c_\tau) + \delta_{\{\theta_{\tau-1}\}}(\theta_\tau)\delta_{\{0\}}(c_\tau) \\ \mathbb{P}\big(y_\tau|x_\tau, \theta_\tau\big) &:= P_{Y|X}(y_\tau|x_\tau; \theta_\tau), \end{aligned} \tag{2}$$

where δ is the Dirac measure (cf. Table 1), and $P_{Y|X}$ is a time-invariant conditional distribution of observations given cues.[1] We do not make any assumption about $\mathbb{P}\big(x_1\big)$ and $\mathbb{P}\big(x_\tau|x_{\tau-1}, y_{\tau-1}\big)$.

See Table 1 for a summary of the notation.

2.2. Special cases and links to related works

Many of the commonly used experimental paradigms (e.g., see Fig. 1) can be formally described in our framework as special cases of Definition 1. The standard generative models for studying passive learning in volatile environments (Adams & MacKay, 2007; Liakoni et al., 2021; Nassar et al., 2012, 2010) are obtained if we remove the cue variables $X_{1:t}$ (Fig. 2B). For example, in the Gaussian experiment of Nassar et al. (2010) (Fig. 1A), Y_t is a sample from a Gaussian distribution with a mean equal to Θ_t and a known variance, and $\pi^{(0)}$ is a very broad uniform distribution.

The minimal model of human inference about binary sequences of Meyniel et al. (2016) (Fig. 2C) assumes that participants estimate probabilities of transitions between stimuli instead of stimulus frequencies, even when the stimuli are by design independent of each other. They show that such an assumption helps explaining many experimental phenomena. Their

[1] The last line of Eq. (2) implies that $\mathbb{P}\big(Y_\tau = y|X_\tau = x, \Theta_\tau = \theta\big) = \mathbb{P}\big(Y_{\tau'} = y|X_{\tau'} = x, \Theta_{\tau'} = \theta\big) = P_{Y|X}(y|x; \theta)$ for any τ and $\tau' \in \{0, 1, 2, \dots\}$.

Table 1
Notation summary.

Notation	Meaning
X_t	Cue at time t
Y_t	Observation at time t
Θ_t	Environment parameter at time t
C_t	Change-point indicator at time t
p_c	Change-point probability, i.e., the probability of $C_t = 1$
$P_{Y\|X}(y\|x;\theta)$	Time invariant distribution of observation y given cue x, parameterized by θ
$\mathbb{P}$	The distribution corresponding to the subjective model of the environment; see Definition 1
$\mathbb{P}^{(t)}$	$\mathbb{P}$ conditioned on observations and cues until time t, i.e., $x_{1:t}$ and $y_{1:t}$
$\mathbb{P}^{(t)}_W$	An alternative notation for the distribution of random variable W conditioned on $x_{1:t}$ and $y_{1:t}$, i.e., $\mathbb{P}^{(t)}_W(w) := \mathbb{P}^{(t)}(W = w)$
$\pi^{(0)}$	Prior distribution over the environment parameter; equivalently, the distribution of Θ_t given $C_t = 1$
$\pi^{(t)}$	The belief about parameter Θ_t at time t, i.e., $\pi^{(t)}(\theta) := \mathbb{P}^{(t)}(\Theta_t = \theta)$
$P(y\|x;\pi^{(t)})$	The marginal probability of observation y given cue x and belief $\pi^{(t)}$; see Eq. (4)
$P(.\|x;\pi^{(t)})$	The full marginal distribution over the space of observations given cue x and belief $\pi^{(t)}$
$\|\|w\|\|_1$	ℓ_1-norm of the vector $w = (w_1, \dots, w_N) \in \mathbb{R}^N$ defined as $\|\|w\|\|_1 := \sum_{n=1}^N \|w_n\|$
$\|\|w\|\|_2$	ℓ_2-norm of the vector $w = (w_1, \dots, w_N) \in \mathbb{R}^N$ defined as $\|\|w\|\|_2 := \sqrt{\sum_{n=1}^N w_n^2}$
$\delta_{\{w^*\}}$	The Dirac measure at w^*, i.e., $\mathbb{P}(W = w) = \delta_{\{w^*\}}(w)$ implies that the probability of the event $\{W = w^*\}$ is one.

model is obtained as a special case of our generative model if the cue variable X_t is equal to the previous observation Y_{t-1}. There, Y_t, conditioned on Y_{t-1}, is a sample from a Bernoulli distribution with parameter Θ_t. In this setting, we have $\mathbb{P}(x_\tau|x_{\tau-1}, y_{\tau-1}) := \delta_{\{y_{\tau-1}\}}(x_\tau)$. This class of generative models has been used to study the neural signatures of surprise via encoding (Gijsen et al., 2021; Maheu et al., 2019) and decoding (Modirshanechi et al., 2019) models in oddball tasks (Fig. 1B).

Variants of bandit and reversal bandit tasks (Behrens et al., 2007; Findling et al., 2021; Horvath et al., 2021) can be modeled by considering the cue variables $X_{1:t}$ as actions $A_{1:t}$ (Fig. 2D). For example, in the experiment of Behrens et al. (2007) (Fig. 1C), $X_t = A_t$ is one of the two possible actions that participants can choose, Y_t is the indicator of whether they are rewarded or not, and Θ_t indicates which action is rewarded with higher probability. In this setting, $\mathbb{P}(x_\tau|x_{\tau-1}, y_{\tau-1}) = \mathbb{P}(x_\tau)$ is the probability that participants take action x_τ, independently of the dynamics of the environment.[2]

Classic Markov Decision Processes (MDPs) (Sutton & Barto, 2018) can also be written in the form of our generative model. To reduce our generative model to an MDP, we set $p_c = 0$, consider the observation Y_t as the pair of the current state and immediate reward value, and consider the cue variable X_t as the previous pair of action and observation (or state) (A_{t-1}, Y_{t-1}) (Fig. 2E). In this setting, we have $\mathbb{P}(X_\tau = (a_{\tau-1}, y)|x_{\tau-1}, y_{\tau-1}) := \delta_{\{y_{\tau-1}\}}(y)\mathbb{P}(a_{\tau-1}|y_{\tau-1})$, where $\mathbb{P}(a_{\tau-1}|y_{\tau-1})$ is called the action selection policy in Reinforcement Learning theory (Sutton & Barto, 2018) and is independent of the dynamics of the environment.[3] The theory of Reinforcement Learning for MDPs has been frequently used in neuroscience and psychology to model human reward-driven decision-making (Daw et al., 2011; Gläscher et al., 2010; Huys et al., 2015; Lehmann et al., 2019; Niv, 2009; Xu et al., 2021) (Fig. 1D).

2.3. Additional notation, belief, and marginal probability

We define $\mathbb{P}^{(t)}$ as $\mathbb{P}$ conditioned on the sequences of observations $y_{1:t}$ and cue variables $x_{1:t}$. For example, for an arbitrary random variable W with value w, we write $\mathbb{P}^{(t)}(w) := \mathbb{P}(w|y_{1:t}, x_{1:t})$. Following this notation, we define an agent's belief about the parameter Θ_t at time t as

$$\pi^{(t)}(\theta) := \mathbb{P}^{(t)}(\Theta_t = \theta), \tag{3}$$

that is the posterior probability (or density, for continuous Θ_t) of $\Theta_t = \theta$ conditioned on $y_{1:t}$ and $x_{1:t}$. The belief plays a crucial role in the perception of surprise (cf. Section 3.1), and we assume that an agent constantly updates its belief, through either exact or approximate Bayesian inference, as it makes new observations — see Barber (2012) and Liakoni et al. (2021) for examples of inference algorithms in generative models similar to ours. According to exact Bayesian inference (Barber, 2012), the updated belief $\pi^{(t+1)}(\theta) = \mathbb{P}^{(t+1)}(\Theta_{t+1} = \theta)$ can be found by normalizing the product of the *prior* belief $\mathbb{P}^{(t)}(\Theta_{t+1} = \theta)$ about Θ_{t+1} and the *likelihood* $P_{Y|X}(y_{t+1}|x_{t+1};\theta)$. In Section 4.1, we give a simple and interpretable expression of the updated belief for the generative model of Definition 1 (cf. Proposition 1).

Another important quantity is the marginal probability of observing y given the cue x and a belief $\pi^{(t)}$:

$$\begin{aligned} P(y|x;\pi^{(t)}) &:= \mathbb{E}_{\pi^{(t)}}\Big[P_{Y|X}(y|x;\Theta)\Big] \\ &= \int P_{Y|X}(y|x;\theta)\pi^{(t)}(\theta)d\theta, \end{aligned} \tag{4}$$

where the integration is replaced by summation whenever θ is discrete.

3. Surprise measures and indistinguishability

Conditioned on the previous observations $y_{1:t}$ and cue variables $x_{1:t+1}$, how surprising is the next observation y_{t+1}? We address this question by examining previously proposed measures of surprise. In this section, we propose a technical classification of different surprise measures and a notion of indistinguishability between different measures and, in the next three sections, we define all surprise measures in the same mathematical framework and discuss their differences and similarities. We present the proofs of these results in Appendix.

3.1. A technical classification

Given θ_{t+1}, the observation y_{t+1} is independent of the previous observations $y_{1:t}$ and cue variables $x_{1:t}$ and only depends on x_{t+1} (Fig. 2A). Hence, the influence of $y_{1:t}$ and $x_{1:t}$ on the surprise of observing y_{t+1} is exclusively through the belief $\pi^{(t)}$, which indicates the importance of $\pi^{(t)}$ in surprise computation.

[2] We note that the action probability $\mathbb{P}(a_\tau)$ in bandit tasks often depends on the whole history of the agent, i.e., $a_{1:\tau-1}$ and $y_{1:\tau-1}$ (Sutton & Barto, 2018). In these situations, one can define x_τ as the concatenation of $a_{1:\tau}$ and $y_{1:\tau-1}$. In this case, the dynamics are described by $\mathbb{P}(X_\tau = (a'_{1:\tau}, y'_{1:\tau-1})|x_{\tau-1}, y_{\tau-1}) := \delta_{\{a_{1:\tau-1}\}}(a'_{1:\tau-1})\delta_{\{y_{1:\tau-1}\}}(y'_{1:\tau-1})\mathbb{P}(a'_\tau|a_{1:\tau-1}, y_{1:\tau-1})$ where $\mathbb{P}(a'_\tau|a_{1:\tau-1}, y_{1:\tau-1})$ is the non-stationary action selection policy — cf. Sutton and Barto (2018).

[3] Similar to the case of bandit tasks, action selection policies in reinforcement learning algorithms used for solving MDPs often depend on the sequence of previous actions $a_{1:\tau-1}$ and observations $y_{1:\tau-1}$, e.g., through estimation of action values (Sutton & Barto, 2018). In these situations, we can define x_τ as the concatenation of $a_{1:\tau}$ and $y_{1:\tau-1}$.

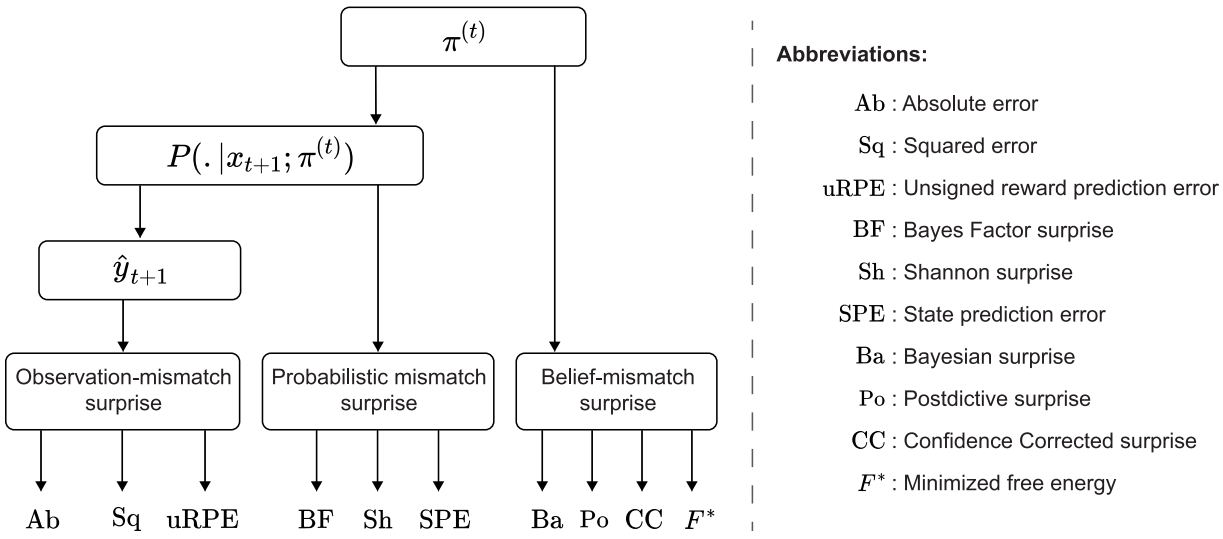

Fig. 3. Technical classification of surprise measures based on the form of their dependence upon the agent's belief. Surprise depends on expectations. Therefore, all surprise measures depend on the belief $\pi^{(t)}$. However, the specific form of the dependence changes between one measure and another. 'Observation-mismatch' surprise measures use the marginal distribution $P(.|x_{t+1};\pi^{(t)})$ (cf. Table 1) to calculate an estimate $\hat{y}_{t+1}$ of the next observation, which is then compared with the real observation y_{t+1} by an error function such as $\|\hat{y}_{t+1} - y_{t+1}\|_1$ (cf. Table 1). 'Probabilistic mismatch' surprise measures use the marginal probability $P(y_{t+1}|x_{t+1};\pi^{(t)})$ directly, without extracting a specific estimate. 'Belief-mismatch' surprise measures use the belief $\pi^{(t)}$ directly, without extracting the marginal probability $P(y_{t+1}|x_{t+1};\pi^{(t)})$. See Section 3 for details.

More precisely, a surprise measure is a function $\mathcal{S} : \mathcal{Y} \times \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ that takes an observation $y_{t+1} \in \mathcal{Y}$, a cue $x_{t+1} \in \mathcal{X}$, and a belief $\pi^{(t)} \in \mathcal{P}$ as arguments and gives the value $\mathcal{S}(y_{t+1}|x_{t+1};\pi^{(t)}) \in \mathbb{R}$ as the corresponding surprise value. However, the specific form of how $\pi^{(t)}$ influences surprise computation changes between one measure and another. Based on how they depend on $\pi^{(t)}$, we divide existing surprise measures into three categories: (i) probabilistic mismatch, (ii) observation-mismatch, and (iii) belief-mismatch surprise measures (Fig. 3). *Probabilistic mismatch* surprise measures depend on the belief $\pi^{(t)}$ only through the marginal probability $P(y_{t+1}|x_{t+1};\pi^{(t)})$; an example is the Shannon surprise (Barto et al., 2013; Tribus, 1961). In other words, probabilistic mismatch surprise depends only on the integral $P(y_{t+1}|x_{t+1};\pi^{(t)}) = \int P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi^{(t)}(\theta)d\theta$ (Eq. (4)) and is independent of other characteristics of the belief $\pi^{(t)}$. *Observation-mismatch* surprise measures depend on $\pi^{(t)}$ only through some estimate $\hat{y}_{t+1}$ of the next observation according to the marginal distribution $P(.|x_{t+1};\pi^{(t)})$ (cf. Table 1); an example is the absolute difference between y_{t+1} and $\hat{y}_{t+1}$ (Nassar et al., 2010; Prat-Carrabin et al., 2021). In other words, observation-mismatch surprise depends only on some statistics (e.g., average or mode) of $P(.|x_{t+1};\pi^{(t)})$ that is used as the estimate $\hat{y}_{t+1}$ and is independent of the other characteristics of $\pi^{(t)}$ and $P(.|x_{t+1};\pi^{(t)})$. To compute the *belief-mismatch* surprise measures, however, we need to have the whole distribution $\pi^{(t)}$; an example is the Bayesian surprise (Baldi, 2002; Schmidhuber, 2010). In other words, neither the marginal distribution $P(.|x_{t+1};\pi^{(t)})$ nor the estimate $\hat{y}_{t+1}$ can solely determine the value of a belief-mismatch surprise measure.

3.2. Notion of indistinguishability

Surprise measures are commonly used in experiments to study whether a behavioral or physiological variable Z (e.g., the amplitude of the EEG P300 component (Kolossa et al., 2015)) is sensitive to or representative of surprise. Given two measures of surprise $\mathcal{S}$ and $\mathcal{S}'$, a typical experimental question is which one of them (if any) more accurately explains the variations of the variable Z (Gijsen et al., 2021; Kolossa et al., 2015; Ostwald et al., 2012; Visalli et al., 2021); see Fig. 4A1. However, if there exists a strictly increasing mapping between $\mathcal{S}$ and $\mathcal{S}'$ (e.g., as in Fig. 4A2), then the two surprise measures have the same explanatory power with respect to Z — because any function of $\mathcal{S}$ can be written in terms of $\mathcal{S}'$ and vice-versa. For example, assume that $\mathcal{S} = f(\mathcal{S}')$ for a strictly increasing function f. If an estimator of the variable Z is found using the measure $\mathcal{S}$ as $\hat{Z} = g(\mathcal{S})$, then we can rewrite the same estimator in terms of $\mathcal{S}'$ as $\hat{Z} = \tilde{g}(\mathcal{S}') = g(f(\mathcal{S}'))$. Because $g(\mathcal{S})$ and $\tilde{g}(\mathcal{S}')$ have the same explanatory power given any function g and any measure of performance, the two surprise measures $\mathcal{S}$ and $\mathcal{S}'$ are equally informative about the variable Z in this regard.[4] We formalize this idea in Definition 2.

Definition 2 (*Indistinguishability*)**.** For the generative model of Definition 1, we say $\mathcal{S}$ and $\mathcal{S}'$ are indistinguishable if there exists a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{S} = f(\mathcal{S}')$ for all choices of belief $\pi^{(t)}$, cue x_t, and observation y_t.

One of our goals in the next three sections is to determine under what conditions different surprise measures are indistinguishable (Fig. 4B and Table 2).

4. Probabilistic mismatch surprise measures

4.1. Bayes Factor surprise

An abrupt change in the parameters of the environment influences the sequence of observations. Therefore, a sensible way to define the surprise of an observation is that 'surprise' measures the probability of an abrupt change in the eye of the agent, given the present observation. To detect an abrupt change, it is not enough to measure how unexpected the observation is

[4] This statement is not necessarily true if one restricts the estimators to a particular class of functions — e.g., if the estimators are constrained to be linear with respect to surprise measures while f is nonlinear. Such limitations can be avoided by using non-parametric statistical methods like Spearman or Kendall correlations (Corder & Foreman, 2014). For example, the Spearman correlation (a measure of monotonic relationship between two random variables) between $\mathcal{S}'$ and Z is the same as the Spearman correlation between $\mathcal{S} = f(\mathcal{S}')$ and Z, but this is not the case for Pearson correlation (a measure of linear relationship between two random variables) if f is nonlinear.

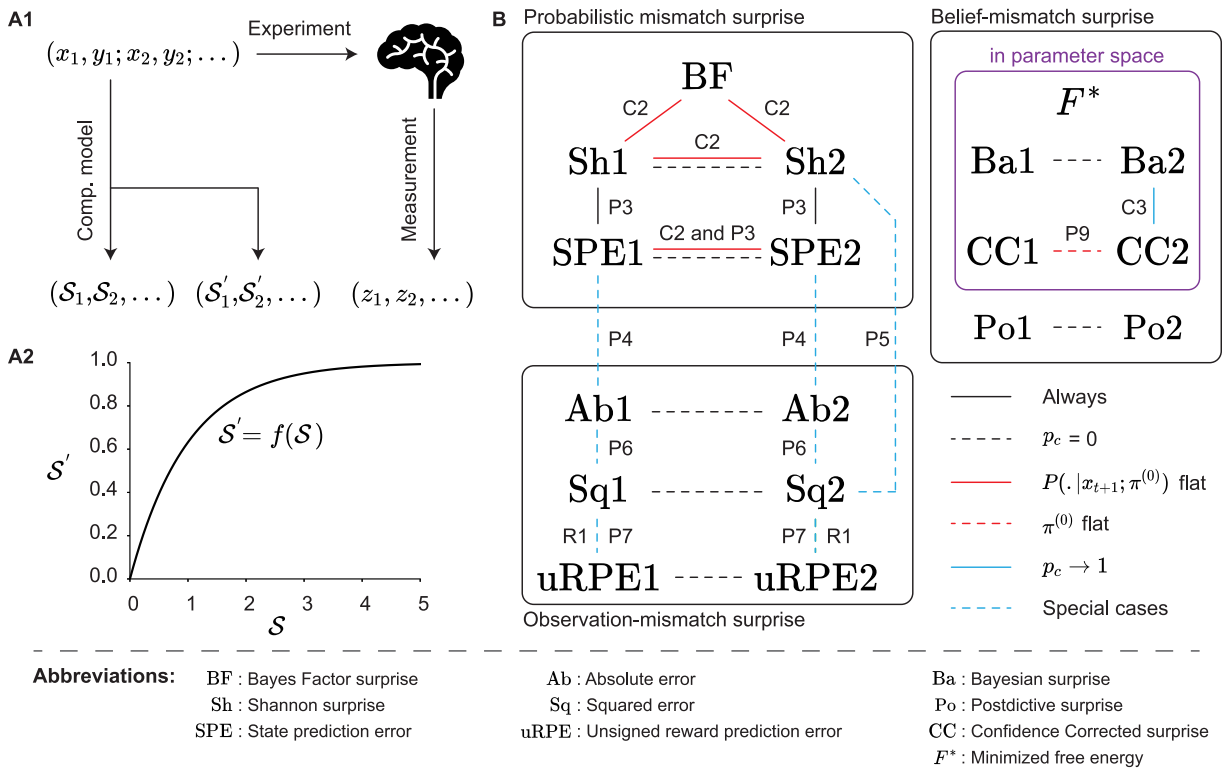

Fig. 4. Indistinguishable surprise measures. A. A typical question in human and animal experiments is whether a surprise measure $\mathcal{S}$ explains the variations of a behavioral or physiological variable Z better than an alternative surprise measure $\mathcal{S}'$. **A1.** A common experimental paradigm: A sequence of cues $x_{1:t}$ and observations $y_{1:t}$ is presented to participants, the sequence $z_{1:t}$ is measured, and the sequence of surprise values $\mathcal{S}_{1:t}$ or $\mathcal{S}'_{1:t}$ is predicted by computational modeling. Then statistical tools are used to study whether the sequence $\mathcal{S}_{1:t}$ or $\mathcal{S}'_{1:t}$ is more informative about the sequence of measurements $z_{1:t}$. **A2.** If there exists a strictly increasing function f such that $\mathcal{S}' = f(\mathcal{S})$, then the two surprise measures are equally informative about the measurable variable Z. In this case, $\mathcal{S}$ and $\mathcal{S}'$ are 'indistinguishable' (cf. Definition 2). **B.** Schematic of the theoretical relation between different measures of surprise. A line connecting two measures indicates that the two measures are indistinguishable, i.e., one is a strictly increasing function of the other, under the condition corresponding to the color and the type of the line. The conditions are shown on the bottom right of the panel: a solid black line means the two measures are always indistinguishable; a dashed black line corresponds to the condition $p_c = 0$; a solid red line corresponds to the prior marginal probability $P(.|x_{t+1}; \pi^{(0)})$ being flat; a dashed red line corresponds to the prior belief $\pi^{(0)}$ being flat; a solid blue line corresponds to the limit of $p_c \to 1$; and a dashed blue line means that the relation holds only for some special cases (e.g., for Gaussian tasks or when the observation is 1-dimensional). Table 2 summarizes which of these conditions are satisfied in several experimental paradigms used to study measures of surprise. Two lines indicate that one of the conditions is sufficient for the two measures to be indistinguishable. The text beside each line shows where in the text the existence of the mapping is proven, e.g., R1, C2, and P3 stand for Remark 1, Corollary 2, and Proposition 3, respectively. The purple box includes surprise measures that are computed in the parameter (Θ_t) space, whereas the surprise measures outside of the purple box are computed in the space of observations (Y_t). See Section 3 for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

according to the current belief of the agent. Rather, the agent should measure how much more expected the new observation is under the prior belief than under the current belief. The Bayes Factor surprise was introduced by Liakoni et al. (2021) to quantify this concept of surprise, motivated by the idea that surprise modulates the speed of learning in the brain (Frémaux & Gerstner, 2016; Iigaya, 2016).

Here, we apply their definition to our generative model. Similar to Xu et al. (2021), we define the Bayes Factor surprise of observing y_{t+1} given the cue x_{t+1} as the ratio of the marginal probability of observing y_{t+1} given x_{t+1} and $C_{t+1} = 1$ (i.e., assuming a change) to the marginal probability of observing y_{t+1} given x_{t+1} and $C_{t+1} = 0$ (i.e. assuming no change):

$$\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := \frac{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}, C_{t+1} = 1)}{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}, C_{t+1} = 0)} = \frac{P(y_{t+1}|x_{t+1}; \pi^{(0)})}{P(y_{t+1}|x_{t+1}; \pi^{(t)})}. \tag{5}$$

The name arises because $\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; \pi^{(t)})$ is the Bayes Factor (Bayarri & Berger, 1997; Kass & Raftery, 1995) used in statistics to test whether a change has occurred at time t. For a given $P(y_{t+1}|x_{t+1}; \pi^{(0)})$, the Bayes Factor surprise is a decreasing function of $P(y_{t+1}|x_{t+1}; \pi^{(t)})$: Hence, more probable events are perceived as less surprising. However, the key feature of $\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; \pi^{(t)})$ is that it measures not only how unexpected (unlikely) the observation y_{t+1} is according to the current belief $\pi^{(t)}$ but also how expected it would be if the agent had reset its belief to the prior belief. More precisely, for a given $P(y_{t+1}|x_{t+1}; \pi^{(t)})$, the Bayes Factor surprise is an increasing function of $P(y_{t+1}|x_{t+1}; \pi^{(0)})$.

Such a comparison is necessary to evaluate whether a reset of the belief (or an increase in the update rate of the belief) can be beneficial in order to have a more accurate estimate of the environment's parameters (cf. Soltani and Izquierdo (2019)). This intuition is formulated in a precise way by Liakoni et al. (2021) in their Proposition 1, where they show that, for the generative model of Fig. 2B, the exact Bayesian inference for the update of $\pi^{(t)}$ to $\pi^{(t+1)}$ upon observing y_{t+1} leads to a learning rule modulated by the Bayes Factor surprise. Proposition 1 states that this result is also true for our more general generative model (Fig. 2A).

Proposition 1 (*Extension of Proposition 1 of Liakoni et al. (2021)*)**.** *For the generative model of Definition 1, the Bayes Factor surprise can be used to write the updated (according to exact Bayesian inference) belief $\pi^{(t+1)}$, after observing y_{t+1} with the cue x_{t+1}, as*

Table 2
Indistinguishability conditions of Fig. 4 for several experimental paradigms. Publications specified by ⋄ use a generative model similar to ours to describe their experiment from the point of view of participants, even if the actual experimental condition has a slightly different structure compared to their generative model. Publications specified by ∗ include either (i) features that are not part of our generative model or (ii) additional experiments not covered by our model. See the original publications for details and Fig. 1 for a description of four of the tasks. A value $p_c > 0$ in the last column indicates a volatile environment; however, we note that participants may by default assume that the environment is volatile even in situations where the actual experimental conditions are stationary (Meyniel et al., 2016).

	Task	$\pi^{(0)}$	$P(.\|x;\pi^{(0)})$	p_c
Nassar et al. (2012, 2010)⋄	Volatile Gaussian	= flat	= flat	>0
Glaze et al. (2015)⋄,∗	Volatile 2D Gaussian	= flat	$\neq$ flat	>0
O'Reilly et al. (2013) Visalli et al. (2021)	Volatile Gaussian with outliers	= flat	= flat	>0
Squires et al. (1976) Mars et al. (2008)⋄ Maheu et al. (2019)⋄, etc.	Oddball	= flat	= flat	$=0$
Heilbron and Meyniel (2019)⋄ Meyniel (2020)⋄	Volatile oddball	= flat	= flat	>0
Ostwald et al. (2012)⋄ Lieder et al. (2013)	Roving oddball	= flat	= flat	$=0$
Gijsen et al. (2021)⋄	Volatile roving oddball	= flat	= flat	>0
Kolossa et al. (2015)⋄	Urn-ball	$\neq$ flat	$\neq$ flat	$=0$
Behrens et al. (2007)⋄ Horvath et al. (2021)⋄	Reversal bandit	= flat	= flat	>0
Rouhani and Niv (2021)∗ Findling et al. (2021)⋄	Volatile contextual bandit	= flat	= flat	>0
Gläscher et al. (2010)	Multi-step decision-making	= flat	= flat	$=0$
Liakoni et al. (2022)⋄	Multi-step decision-making with outliers	$\neq$ flat	= flat	$=0$
Xu et al. (2021)⋄	Volatile multi-step decision-making	$\neq$ flat	= flat	>0

$$\pi^{(t+1)}(\theta) = (1-\gamma_{t+1})\pi^{(t+1)}_{\text{integration}}(\theta) + \gamma_{t+1}\pi^{(t+1)}_{\text{reset}}(\theta), \tag{6}$$

where γ_{t+1} is an adaptation rate modulated by the Bayes Factor surprise

$$\gamma_{t+1} := \frac{m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})}{1+m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})} \qquad m := \frac{p_c}{1-p_c}, \tag{7}$$

and

$$\pi^{(t+1)}_{\text{integration}}(\theta) := \frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi^{(t)}(\theta)}{P(y_{t+1}|x_{t+1};\pi^{(t)})}, \qquad \pi^{(t+1)}_{\text{reset}}(\theta) := \frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi^{(0)}(\theta)}{P(y_{t+1}|x_{t+1};\pi^{(0)})}. \tag{8}$$

Therefore, the Bayes Factor surprise $\mathcal{S}_{\text{BF}}$ controls the trade-off between the integration of the new observation into the old belief (via $\pi^{(t+1)}_{\text{integration}}$) and resetting the old belief to the prior belief (via $\pi^{(t+1)}_{\text{reset}}$).

4.2. Shannon surprise

No matter if there has been an abrupt change ($C_{t+1}=1$) or not ($C_{t+1}=0$), an unlikely event may be perceived as surprising. Therefore, another way to measure the surprise of an observation is to quantify how unlikely the observation is in the eye of the agent. Shannon surprise, also known as surprisal (Barto et al., 2013), is a way to formalize this concept of surprise. It comes from the field of information theory (Shannon, 1948) and statistical physics (Tribus, 1961) and is widely used in neuroscience (Gijsen et al., 2021; Kolossa et al., 2015; Konovalov & Krajbich, 2018; Kopp & Lange, 2013; Maheu et al., 2019; Mars et al., 2008; Meyniel et al., 2016; Modirshanechi et al., 2019; Mousavi et al., 2022; Visalli et al., 2021).

Formally, for the generative model of Definition 1, one can define the Shannon surprise of observing y_{t+1} given the cue x_{t+1} as

$$\begin{aligned}\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &:= -\log\mathbb{P}^{(t)}\big(y_{t+1}|x_{t+1}\big)\\ &= -\log\Big(p_c P(y_{t+1}|x_{t+1};\pi^{(0)}) + \\ &\qquad (1-p_c)P(y_{t+1}|x_{t+1};\pi^{(t)})\Big),\end{aligned} \tag{9}$$

where the 2nd equality is a result of the marginalization

$$\mathbb{P}^{(t)}\big(y_{t+1}|x_{t+1}\big) = \sum_c \mathbb{P}^{(t)}\big(y_{t+1}, C_{t+1}=c|x_{t+1}\big). \tag{10}$$

The Shannon surprise $\mathcal{S}_{\text{Sh1}}$ measures how unexpected or unlikely y_{t+1} is considering the possibility that there might have been an abrupt change in the environment. As a result, for a fixed $P(y_{t+1}|x_{t+1};\pi^{(t)})$, the Shannon surprise is a decreasing function of $P(y_{t+1}|x_{t+1};\pi^{(0)})$ (cf. Eq. (9)): It is *less* surprising to observe an event that is more probable under the prior belief because this event is also in total more probable if we consider the possibility of an abrupt change at time $t+1$. In contrast, the Bayes Factor surprise is an increasing function of $P(y_{t+1}|x_{t+1};\pi^{(0)})$ (cf. Eq. (5)): It is *more* surprising to observe an event that is more probable under the prior belief because such events indicate higher chances that an abrupt change has occurred. This essential difference between the Shannon and the Bayes Factor surprise has been exploited by Liakoni et al. (2021) to propose experiments where these two measures of surprise make different predictions.

Experimental evidence (Nassar et al., 2012, 2010) indicates that in volatile environments like the one in Fig. 2B, human participants do not actively consider the possibility that there *may be* an abrupt change while predicting the *next* observation y_{t+1} — even though they update their belief after observing y_{t+1} by considering the possibility that there *might have been* a change before the *current* observation at time $t+1$. To arrive at a Shannon surprise measure consistent with this observation, we suggest a

second definition:

$$\begin{aligned}\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &:= -\log \mathbb{P}^{(t)}\big(y_{t+1}|x_{t+1}, C_{t+1}=0\big)\\ &= -\log P(y_{t+1}|x_{t+1};\pi^{(t)}).\end{aligned} \tag{11}$$

In other words, $\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})$ neglects the potential presence of change-points, and, therefore, it is independent of both p_c and $P(y_{t+1}|x_{t+1};\pi^{(0)})$. For a non-volatile environment that does not allow for abrupt changes ($p_c = 0$), the two definitions of Shannon surprise are identical: $\mathcal{S}_{\mathrm{Sh1}} = \mathcal{S}_{\mathrm{Sh2}}$ (Fig. 4B).

Proposition 2 shows that the Bayes Factor surprise $\mathcal{S}_{\mathrm{BF}}$ is related to $\mathcal{S}_{\mathrm{Sh1}}$ and $\mathcal{S}_{\mathrm{Sh2}}$:

Proposition 2 (*Relation between the Shannon Surprise and the Bayes Factor Surprise*). *For the generative model of Definition 1, the Bayes Factor surprise $\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})$ can be written as*

$$\begin{aligned}\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \frac{(1-p_c)e^{\Delta\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})}}{1-p_c e^{\Delta\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})}}\\ &= e^{\Delta\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})},\end{aligned} \tag{12}$$

where

$$\begin{aligned}\Delta\mathcal{S}_{\mathrm{Shi}}(y_{t+1}|x_{t+1};\pi^{(t)}) := \mathcal{S}_{\mathrm{Shi}}(y_{t+1}|x_{t+1};\pi^{(t)}) -\\ \mathcal{S}_{\mathrm{Shi}}(y_{t+1}|x_{t+1};\pi^{(0)})\end{aligned} \tag{13}$$

for $i \in \{1,2\}$.

Proposition 2 states that the Bayes Factor $\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})$ has a behavior similar to the *difference* in Shannon surprise (i.e., $\Delta\mathcal{S}_{\mathrm{Sh1}}$ or $\Delta\mathcal{S}_{\mathrm{Sh2}}$) as opposed to Shannon surprise itself (i.e., $\mathcal{S}_{\mathrm{Sh1}}$ or $\mathcal{S}_{\mathrm{Sh2}}$). The difference in Shannon surprise (i.e., $\Delta\mathcal{S}_{\mathrm{Sh1}}$ or $\Delta\mathcal{S}_{\mathrm{Sh2}}$) compares the Shannon surprise under the current belief with that under the prior belief. Two direct consequences of this proposition are summarized in Corollaries 1 and 2.

Corollary 1 states that the modulation of learning as presented in Proposition 1 can also be written in the form of the difference in Shannon surprise (i.e., $\Delta\mathcal{S}_{\mathrm{Sh1}}$ or $\Delta\mathcal{S}_{\mathrm{Sh2}}$).

Corollary 1. *The adaptation rate γ_{t+1} in Proposition 1 can be written as*

$$\begin{aligned}\gamma_{t+1} &= p_c \exp\Big(\Delta\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\Big)\\ \gamma_{t+1} &= \mathrm{Sigmoid}\Big(\tilde{m}\Delta\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})\Big),\end{aligned} \tag{14}$$

with $\tilde{m} := \log\frac{p_c}{1-p_c} = \log m$ (cf. Proposition 1) and $\mathrm{Sigmoid}(u) := \frac{1}{1+e^{-u}}$

Corollary 2 indicates that, under a flat prior, the Bayes Factor surprise and the two definitions of the Shannon surprise are indistinguishable from each other (Fig. 4B):

Corollary 2 (*Flat Prior Prediction*). *For the generative model of Definition 1, if the probability of observing y_{t+1} with the cue x_{t+1} is flat under the prior belief $\pi^{(0)}$ (i.e., if $P(.|x_{t+1};\pi^{(0)})$ is uniform), then there are strictly increasing mappings between $\mathcal{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})$, $\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})$, and $\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})$.*

A consequence of Corollary 2 is that experiments with flat marginal priors of the agent cannot be used to distinguish $\mathcal{S}_{\mathrm{BF}}$ from $\mathcal{S}_{\mathrm{Sh1}}$ or $\mathcal{S}_{\mathrm{Sh2}}$ (Fig. 4).

4.3. State prediction error

The State Prediction Error (SPE) was introduced by Gläscher et al. (2010) in the context of model-based reinforcement learning in Markov Decision Processes (MDPs — cf. Fig. 2E) (Sutton & Barto, 2018). Similar to the Shannon surprise, the SPE considers less probable events as the more surprising ones.

Whenever observations $y_{1:t}$ come from a discrete distribution so that we have $P_{Y|X}(y_{t+1}|x_{t+1};\theta) \in [0,1]$ for all θ, x_{t+1}, and y_{t+1}, we can generalize the definition of Gläscher et al. (2010) to the setting of our generative model. Analogously to our two definitions of Shannon surprise (cf. Eqs. (9) and (11)), we give also two definitions for SPE:

$$\begin{aligned}\mathcal{S}_{\mathrm{SPE1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &:= 1-\mathbb{P}^{(t)}\big(y_{t+1}|x_{t+1}\big)\\ &= 1-\Big(p_c P(y_{t+1}|x_{t+1};\pi^{(0)})+\\ &\qquad (1-p_c)P(y_{t+1}|x_{t+1};\pi^{(t)})\Big),\end{aligned} \tag{15}$$

and

$$\begin{aligned}\mathcal{S}_{\mathrm{SPE2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &:= 1-\mathbb{P}^{(t)}\big(y_{t+1}|x_{t+1}, C_{t+1}=0\big)\\ &= 1-P(y_{t+1}|x_{t+1};\pi^{(t)}).\end{aligned} \tag{16}$$

In non-volatile environments ($p_c = 0$), the two definitions of SPE are identical (Fig. 4B). In particular, in an MDP without abrupt changes ($p_c = 0$; Fig. 2E), both definitions are equal to $1 - \mathbb{P}^{(t)}(s_t, a_t \to s_{t+1})$, where $\mathbb{P}^{(t)}(s_t, a_t \to s_{t+1})$ is an agent's estimate (at time t) of the probability of the transition to state s_{t+1} after taking action a_t in state s_t; cf. Gläscher et al. (2010).

Proposition 3 states that both definitions ($\mathcal{S}_{\mathrm{SPE1}}$ and $\mathcal{S}_{\mathrm{SPE2}}$) can always be written as strictly increasing functions of Shannon surprise (Fig. 4B):

Proposition 3 (*Relation between the Shannon Surprise and the SPE*). *For the generative model of Definition 1, for $i \in \{1,2\}$, the state prediction error $\mathcal{S}_{\mathrm{SPEi}}(y_{t+1}|x_{t+1};\pi^{(t)})$, can be written as*

$$\mathcal{S}_{\mathrm{SPEi}}(y_{t+1}|x_{t+1};\pi^{(t)}) = 1-\exp\Big(-\mathcal{S}_{\mathrm{Shi}}(y_{t+1}|x_{t+1};\pi^{(t)})\Big). \tag{17}$$

Therefore, the SPE and the Shannon surprise are indistinguishable (Fig. 4).

5. Observation-mismatch surprise measures

5.1. Absolute and squared errors

Assume an agent predicts $\hat{y}_{t+1}$ for the next observation y_{t+1}. Then, a measure of surprise can be defined as the prediction error or the mismatch between the prediction $\hat{y}_{t+1}$ and the actual observation y_{t+1} (Nassar et al., 2012, 2010; Prat-Carrabin et al., 2021) (Fig. 3). For the sake of completeness, we discuss four possible definitions for observation-mismatch surprise measures.

Before turning to an 'observation-mismatch', we first need to define an agent's prediction for the next observation. Analogously to our two definitions for the Shannon surprise (cf. Eqs. (9) and (11)), we define two different predictions for the next observation y_{t+1} given the cue x_{t+1}[5]:

$$\begin{aligned}E_1[Y_{t+1}] := p_c\mathbb{E}_{P(.|x_{t+1};\pi^{(0)})}[Y_{t+1}]+\\ (1-p_c)\mathbb{E}_{P(.|x_{t+1};\pi^{(t)})}[Y_{t+1}]\end{aligned} \tag{18}$$

and

$$E_2[Y_{t+1}] := \mathbb{E}_{P(.|x_{t+1};\pi^{(t)})}[Y_{t+1}]. \tag{19}$$

Although $E_1[Y_{t+1}]$ is a more reasonable prediction for y_{t+1} given the fact that there is always a possibility of an abrupt change according to our generative model of the environment (Definition 1), Nassar et al. (2010) have shown that, in a Gaussian task (cf. Fig. 1A), $E_2[Y_{t+1}]$ explains human participants' predictions better than $E_1[Y_{t+1}]$.

[5] The evaluation of the full distribution $P(.|x_{t+1};\pi^{(t)})$ may not always be necessary for the computation of E_1 and E_2 (Aguilera et al., 2022; Liakoni et al., 2021; Nassar et al., 2010).

We note that the observation y_{t+1} is, in general, multi-dimensional. As two natural ways of measuring mismatch, we define the squared and the absolute error surprise, for $i \in \{1, 2\}$, as

$$\begin{aligned}\mathcal{S}_{\text{Ab},i}(y_{t+1}|x_{t+1}; \pi^{(t)}) &:= \|y_{t+1} - E_i[Y_{t+1}]\|_1 \\ \mathcal{S}_{\text{Sq},i}(y_{t+1}|x_{t+1}; \pi^{(t)}) &:= \left(\|y_{t+1} - E_i[Y_{t+1}]\|_2\right)^2,\end{aligned} \tag{20}$$

where $\|.\|_1$ and $\|.\|_2$ stand for the ℓ_1- and ℓ_2-norms (cf. Table 1), respectively, and E_1 and E_2 are defined in Eq. (18) and Eq. (19), respectively. Similar definitions have been used in neuroscience (Nassar et al., 2010; Prat-Carrabin et al., 2021) and machine learning (Burda et al., 2019; Pathak et al., 2017). In Propositions 4–6, we show for three special cases that the absolute and the squared error surprise can be written as strictly increasing functions of either each other or the SPE and the Shannon surprise (Fig. 4B).

Proposition 4 (*Relation between the Absolute and Squared Errors and the SPE for Categorical Distributions*)**.** *For the generative model of Definition 1, if Y_{t+1} is represented as one-hot coded vectors, i.e., vectors with one element equal to 1 and the others equal to 0, then we have, for $i \in \{1, 2\}$,*

$$\mathcal{S}_{\text{Abi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) = 2\mathcal{S}_{\text{SPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)}), \tag{21}$$

and

$$\begin{aligned}\mathcal{S}_{\text{Sqi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) =& 2\mathcal{S}_{\text{SPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) + \\ &\text{Conf.}\left[P(.|x_{t+1}; \pi^{(t)})\right],\end{aligned} \tag{22}$$

where $\text{Conf.}\left[P(.|x_{t+1}; \pi^{(t)})\right]$ *can be seen as a measure of confidence in the prediction (see Appendix).*

Proposition 5 (*Relation between the Squared Error Surprise and the Shannon Surprise for Gaussian Distributions — from Pathak et al. (2017)*)**.** *For the generative model of Definition 1, if the marginal distribution of $Y_{t+1} \in \mathbb{R}^N$ given the cue x_{t+1} and the belief $\pi^{(t)}$ is a Gaussian distribution with a covariance matrix equal to $\sigma I_{N\times N}$, where $I_{N\times N}$ is the $N \times N$ identity matrix, then $\mathcal{S}_{\text{Sq2}}(y_{t+1}|x_{t+1}; \pi^{(t)})$ is a strictly increasing function of $\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1}; \pi^{(t)})$.*

Proposition 6 (*Observation-Mismatch Surprise Measures for 1-D Observations*)**.** *For the generative model of Definition 1, if $Y_t \in \mathbb{R}$, then we have $\mathcal{S}_{\text{Sqi}} = \mathcal{S}^2_{\text{Abi}}$ for $i \in \{1, 2\}$ implying that the two observation-mismatch surprise measures are indistinguishable.*

We note that, according to Proposition 3, the SPE is a strictly increasing function of the Shannon surprise. Hence, for categorical distributions with one-hot coding, the SPE, the Shannon surprise, and the absolute error surprise are indistinguishable, and for Gaussian distributions with scaled identity covariance, the SPE, the Shannon surprise, and the squared error surprise are indistinguishable (Fig. 4).

5.2. Unsigned reward prediction error

A particular form of observation-mismatch surprise in the context of reward-driven decision making is the Unsigned Reward Prediction Error (uRPE, i.e., the absolute value of Reward Prediction Error) (Hayden et al., 2011; Pearce & Hall, 1980; Roesch et al., 2012; Rouhani & Niv, 2021; Talmi et al., 2013). In this section, we first discuss the definition of the uRPE as it often appears in experimental studies and then analyze a generalized definition of the uRPE in general sequential decision-making tasks.

Many of the experimental paradigms (e.g., Hayden et al., 2011; Roesch et al., 2012; Talmi et al., 2013) for the study of uRPE can be modeled by a non-volatile (i.e., $p_c = 0$) contextual bandit task where, given a context s_t (e.g., conditioned stimulus), the agent takes an action a_t and receives a real-valued reward r_{t+1}. The uRPE corresponding to the tuple (s_t, a_t, r_{t+1}) is (Sutton & Barto, 2018)

$$\text{uRPE}(s_t, a_t \to r_{t+1}) := |r_{t+1} - Q^{(t)}(s_t, a_t)|, \tag{23}$$

where $Q^{(t)}(s_t, a_t)$ is the latest estimate of the expectation of R_{t+1} given s_t and a_t. The generative model of Definition 1 is reduced to a model of contextual bandit tasks if we put $X_{t+1} := (S_t, A_t)$ and $Y_{t+1} := R_{t+1}$. Then, the unsigned reward prediction error $\text{uRPE}(s_t, a_t \to r_{t+1})$ is syntactically equal to $\mathcal{S}_{\text{Ab}}$ (cf. Eq. (20); note that $E_1 = E_2$ since $p_c = 0$) and indistinguishable from $\mathcal{S}_{\text{Sq}}$ (Proposition 6):

Remark 1 (*Relation between the common definition of uRPE and the other two Observation-Mismatch Surprise measures*)**.** The uRPE signal that was previously investigated in many experimental studies (Eq. (23)) (Hayden et al., 2011; Pearce & Hall, 1980; Roesch et al., 2012; Talmi et al., 2013) is a special case of the absolute and the squared error surprise (Eq. (20)).

However, one can go beyond contextual bandit tasks and define uRPE for a general Markov Decision Process (MDP) (Sutton & Barto, 2018). To reduce our generative model of Definition 1 to a (potentially volatile, i.e., $p_c \geq 0$) MDP, we put the cue variable X_{t+1} equal to the state–action pair (S_t, A_t) and the observation Y_{t+1} equal to the pair of the next state S_{t+1} and the next *extended reward* $\tilde{R}_{t+1}$ that we define as

$$\tilde{R}_{t+1} := R_{t+1} + \lambda V(S_{t+1}), \tag{24}$$

where $\lambda \in [0, 1)$ is the discount factor in infinite-horizon reinforcement learning (Sutton & Barto, 2018), and $V(S_{t+1})$ is the *perceived value* of state S_{t+1}. Here, we do not discuss the exact definition of V and how it is computed; we only assume that each state s has a value $V(s)$ that is informative about the expected amount of total reward that one can collect starting from state s — see Sutton and Barto (2018) for details. Analogously to our two definitions for the absolute and the squared error surprise (cf. Eq. (20)), we give two definitions of uRPE:

$$\mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := |r_{t+1} + \lambda V(s_{t+1}) - Q_i^{(t)}(s_t, a_t)|, \tag{25}$$

where $i \in \{1, 2\}$ and $Q_i^{(t)}(s_t, a_t) := E_i[\tilde{R}_{t+1}]$ (cf. Eq. (18), Eq. (19), and Eq. (24)). Eq. (25) implies that the uRPE surprise is like the absolute error surprise if an agent focuses exclusively on the extended reward $\tilde{r}_{t+1}$ and ignores the state s_{t+1}. We make this intuition formal in Proposition 7.

Proposition 7 (*Relation between the uRPE, the Absolute Error, and Squared Error Surprise Measures*)**.** *For the generative model of Definition 1, for $i \in \{1, 2\}$, the unsigned reward prediction error $\mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)})$ can be written as*

$$\begin{aligned}\mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) =& \mathcal{S}_{\text{Abi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) - \\ &\mathcal{S}_{\text{Abi}}(s_{t+1}|x_{t+1}; \pi^{(t)})\end{aligned} \tag{26}$$

and

$$\begin{aligned}\left(\mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1}; \pi^{(t)})\right)^2 =& \mathcal{S}_{\text{Sqi}}(y_{t+1}|x_{t+1}; \pi^{(t)}) - \\ &\mathcal{S}_{\text{Sqi}}(s_{t+1}|x_{t+1}; \pi^{(t)}).\end{aligned} \tag{27}$$

where $\mathcal{S}_{\text{Abi}}(s_{t+1}|x_{t+1}; \pi^{(t)}) := \|s_{t+1} - E_i[S_{t+1}]\|_1$ and $\mathcal{S}_{\text{Sqi}}(s_{t+1}|x_{t+1}; \pi^{(t)}) := \|s_{t+1} - E_i[S_{t+1}]\|_2^2$ (Eq. (20)).

Therefore, if observation y_{t+1} does not include state s_{t+1} (e.g., in contextual bandit tasks, similar to Hayden et al. (2011), Roesch et al. (2012), Talmi et al. (2013)) or if all possible values of state s_{t+1} are equally surprising (i.e., have constant $\mathcal{S}_{\text{Sqi}}$ or $\mathcal{S}_{\text{Abi}}$, similar to the experiment of Rouhani and Niv (2021)), then $\mathcal{S}_{\text{uRPEi}}$ is indistinguishable from $\mathcal{S}_{\text{Abi}}$ and $\mathcal{S}_{\text{Sqi}}$ (Fig. 4).

6. Belief-mismatch surprise measures

6.1. Bayesian surprise

Another way to think about surprise is to define surprising events as those that change an agent's belief about the world. Bayesian surprise (Baldi, 2002; Baldi & Itti, 2010; Schmidhuber, 2010) is a way to formalize this concept of surprise. Whereas the Bayes Factor surprise measures how likely it is that the environment has changed given the new observation, the Bayesian surprise measures how much the agent's belief changes given the new observation.

Bayesian surprise (Baldi, 2002) has been originally introduced in non-volatile environments, i.e., where there is no change ($p_c = 0$) and as a result $\Theta_1 = \Theta_2 = \cdots = \Theta_t = \Theta$. In this case, the Bayesian surprise of observing y_{t+1} with cue x_{t+1} is defined as $D_{\mathrm{KL}}[\mathbb{P}_{\Theta}^{(t)} || \mathbb{P}_{\Theta}^{(t+1)}]$ (Baldi, 2002; Baldi & Itti, 2010; Schmidhuber, 2010), where D_{KL} stands for the Kullback–Leibler (KL) divergence (Cover, 1999), and $\mathbb{P}_{\Theta}^{(t)}$ is an alternative notation for the distribution of Θ conditioned on $x_{1:t}$ and $y_{1:t}$ (cf. Table 1). Hence, in non-volatile environments, Bayesian surprise measures the pseudo-distance D_{KL} between two distributions, i.e., the belief $\pi^{(t)} = \mathbb{P}_{\Theta}^{(t)}$ before and the belief $\pi^{(t+1)} = \mathbb{P}_{\Theta}^{(t+1)}$ after observing y_{t+1}. To generalize this definition to volatile environments, we have to choose two equivalent distributions that we want to compare. The natural choice for $\mathbb{P}_{\Theta}^{(t+1)}$ is $\mathbb{P}_{\Theta_{t+1}}^{(t+1)} = \pi^{(t+1)}$; however, it is unclear whether $\mathbb{P}_{\Theta}^{(t)}$ should be taken as the momentary belief $\mathbb{P}_{\Theta_t}^{(t)} = \pi^{(t)}$ or its one-step forward-propagation $\mathbb{P}_{\Theta_{t+1}}^{(t)}$ *before* the next observation y_{t+1} is integrated. If $p_c \neq 0$, the two choices are different:

$$\pi^{(t)} = \mathbb{P}_{\Theta_t}^{(t)} \neq \mathbb{P}_{\Theta_{t+1}}^{(t)} = p_c \pi^{(0)} + (1 - p_c)\pi^{(t)}. \tag{28}$$

Therefore, for the case of volatile environments, we give two definitions for the Bayesian surprise:

$$\mathcal{S}_{\mathrm{Ba1}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := D_{\mathrm{KL}}\Big[p_c \pi^{(0)} + (1 - p_c)\pi^{(t)} || \pi^{(t+1)}\Big], \tag{29}$$

and

$$\mathcal{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := D_{\mathrm{KL}}\Big[\pi^{(t)} || \pi^{(t+1)}\Big]. \tag{30}$$

The first definition is more consistent with the original definition of the Bayesian surprise (Baldi, 2002; Baldi & Itti, 2010; Schmidhuber, 2010) applied to our generative model because the belief before the observation should include the knowledge that the environment is volatile. However, the second definition looks more intuitive from the neuroscience perspective (Gijsen et al., 2021; Mousavi et al., 2022). Note that, in Eqs. (29) and (30), the observation y_{t+1} does not appear explicitly on the right hand side; the observation has, however, influenced the update of the belief to its new distribution $\pi^{(t+1)}$. For the case of $p_c = 0$, the two definitions are identical (Fig. 4B).

In Proposition 8 and Remark 2, we show that the Bayesian surprise is correlated with the difference between the Shannon surprise and its expectation (over all possible values of Θ_{t+1}).

Proposition 8 (*Relation between the Bayesian Surprise and the Shannon Surprise*)**.** *In the generative model of Definition 1, the Bayesian surprise can be written as*

$$\begin{aligned}\mathcal{S}_{\mathrm{Ba1}}(y_{t+1}|x_{t+1}; \pi^{(t)}) =& p_c \mathbb{E}_{\pi^{(0)}}\Big[\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1}; \delta_{\{\Theta\}})\Big] + \\ & (1 - p_c)\mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1}; \delta_{\{\Theta\}})\Big] - \\ & \mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \pi^{(t)}),\end{aligned} \tag{31}$$

and

$$\begin{aligned}\mathcal{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1}; \pi^{(t)}) =& \mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1}; \delta_{\{\Theta\}})\Big] - \\ & \mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \pi^{(t)}) + \\ & D_{\mathrm{KL}}\Big[\pi^{(t)} || p_c \pi^{(0)} + (1 - p_c)\pi^{(t)}\Big],\end{aligned} \tag{32}$$

where $\delta_{\{\theta\}}$ is the Dirac measure at θ (cf. Table 1).

Remark 2. As a direct consequence of Proposition 8, when the change point probability is zero, i.e. $p_c = 0$, the Bayesian surprise is equal to the expected Shannon surprise minus the Shannon surprise, i.e.,

$$\begin{aligned}\mathcal{S}_{\mathrm{Ba}}(y_{t+1}|x_{t+1}; \pi^{(t)}) =& \mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\mathrm{Sh}}(y_{t+1}|x_{t+1}; \delta_{\{\Theta\}})\Big] - \\ & \mathcal{S}_{\mathrm{Sh}}(y_{t+1}|x_{t+1}; \pi^{(t)}),\end{aligned} \tag{33}$$

where $\mathcal{S}_{\mathrm{Ba}} = \mathcal{S}_{\mathrm{Ba1}} = \mathcal{S}_{\mathrm{Ba2}}$ and $\mathcal{S}_{\mathrm{Sh}} = \mathcal{S}_{\mathrm{Sh1}} = \mathcal{S}_{\mathrm{Sh2}}$.

There are two consequences of this observation. First, Bayesian surprise is distinguishable from Shannon surprise since it cannot be found only as a function of Shannon surprise. Second, we need access to the full belief distribution $\pi^{(t)}$ for computing the expectation (Fig. 3).

In general, surprise measures similar to the Bayesian surprise can be defined also by measuring the change in the belief via distance or pseudo-distance measures different from the KL-divergence (Baldi, 2002).

6.2. Postdictive surprise

We saw that the Bayesian surprise measures how much the new belief $\pi^{(t+1)}$ has changed after observing y_{t+1}. Kolossa et al. (2015) introduced 'postdictive surprise' with a similar idea in mind but focused on changes in the marginal distribution $P(.|x_{t+1}; \pi^{(t+1)})$ (cf. Eq. (4)). More precisely, whereas the Bayesian surprise measures the amount of update in the space of distributions over the parameters (i.e., how differently the agent thinks about the parameters), the postdictive surprise measures the amount of update in the space of distributions over the observations (i.e., how differently the agent predicts the next observations).

Analogous to our two definitions for the Bayesian surprise (Eqs. (29) and (30)), there are two definitions for the postdictive surprise in volatile environments:

$$\begin{aligned}&\mathcal{S}_{\mathrm{Po1}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := \\ &D_{\mathrm{KL}}\Big[p_c P\big(.|x_{t+1}; \pi^{(0)}\big) + (1 - p_c)P\big(.|x_{t+1}; \pi^{(t)}\big) || \\ &\quad P\big(.|x_{t+1}; \pi^{(t+1)}\big)\Big],\end{aligned} \tag{34}$$

and

$$\begin{aligned}\mathcal{S}_{\mathrm{Po2}}(y_{t+1}|x_{t+1}; \pi^{(t)}) := D_{\mathrm{KL}}\Big[&P\big(.|x_{t+1}; \pi^{(t)}\big) || \\ &P\big(.|x_{t+1}; \pi^{(t+1)}\big)\Big],\end{aligned} \tag{35}$$

where the dot refers to a dummy variable y that is integrated out when evaluating D_{KL} (cf. Table 1). Note that for $p_c = 0$, the two definitions are identical (Fig. 4B).

Although the amount of update is computed over the space of observations, $\mathcal{S}_{\mathrm{Po1}}$ and $\mathcal{S}_{\mathrm{Po2}}$ cannot be categorized as probabilistic mismatch surprise measures, since the update depends explicitly on the belief $\pi^{(t)}$. The statement is further explained in our Lemma 1 in Appendix.

6.3. Confidence corrected surprise

Since surprise arises when an expectation is violated, the violation of an agent's expectation should be more surprising when the agent is more confident about its expectation. Based on the observation that neither Shannon nor Bayesian surprise explicitly captures the concept of confidence, Faraji et al. (2018) proposed the 'Confidence Corrected Surprise' as a new measure of surprise that explicitly takes confidence into account.

To define the Confidence Corrected surprise, we first define π_{flat} as the flat (uniform) distribution over the space of parameters, i.e., over the set to which Θ_t belongs. Then, following Faraji et al. (2018), we define the normalized likelihood after observing y_{t+1} (i.e., the posterior given the flat prior) as

$$\pi_{\text{flat}}(\theta|y_{t+1}, x_{t+1}) := \frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi_{\text{flat}}(\theta)}{P(y_{t+1}|x_{t+1};\pi_{\text{flat}})} = \frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)}{\int P_{Y|X}(y_{t+1}|x_{t+1};\theta)d\theta}. \tag{36}$$

If the prior $\pi^{(0)}$ is equal to π_{flat} (i.e., if the prior is uniform), then $\pi_{\text{flat}}(\theta|y_{t+1}, x_{t+1})$ is the same as $\pi^{(t+1)}_{\text{reset}}(\theta)$ defined in Proposition 1. Note that the prior π_{flat} does not necessarily need to be a proper distribution (i.e., does not necessarily need to be normalized) as long as $\int P_{Y|X}(y_{t+1}|x_{t+1};\theta)d\theta$ is finite and the posterior $\pi_{\text{flat}}(.|y_{t+1}, x_{t+1})$ is a proper distribution (Efron & Hastie, 2016). Using this terminology, the original definition for the Confidence Corrected surprise is (Faraji et al., 2018)

$$\mathcal{S}_{\text{CC1}}(y_{t+1}|x_{t+1};\pi^{(t)}) := D_{\text{KL}}\Big[\pi^{(t)}||\pi_{\text{flat}}(.|y_{t+1}, x_{t+1})\Big]. \tag{37}$$

To interpret $\mathcal{S}_{\text{CC1}}$, Faraji et al. (2018) defined the commitment (or confidence) $C[\pi]$ corresponding to an arbitrary belief π as its negative entropy (Cover, 1999), i.e.,

$$C[\pi] := \mathbb{E}_{\pi}\Big[\log \pi(\Theta)\Big]. \tag{38}$$

Then, in a non-volatile environment (i.e., $p_c = 0$), they show that $\mathcal{S}_{\text{CC1}}$ can be written as (Faraji et al., 2018)

$$\mathcal{S}_{\text{CC1}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \mathcal{S}_{\text{Sh}}(y_{t+1}|x_{t+1};\pi^{(t)}) + \mathcal{S}_{\text{Ba}}(y_{t+1}|x_{t+1};\pi^{(t)}) + C\Big[\pi^{(t)}\Big] - A(y_{t+1}, x_{t+1}), \tag{39}$$

where $A(y_{t+1}, x_{t+1}) := \mathcal{S}_{\text{Sh}}(y_{t+1}|x_{t+1};\pi_{\text{flat}}) + C[\pi_{\text{flat}}]$ is independent of the current belief $\pi^{(t)}$. Note that because $p_c = 0$, we have $\mathcal{S}_{\text{Sh1}} = \mathcal{S}_{\text{Sh2}}$ and $\mathcal{S}_{\text{Ba1}} = \mathcal{S}_{\text{Ba2}}$. Therefore, in a non-volatile environment (i.e., $p_c = 0$), $\mathcal{S}_{\text{CC1}}$ is correlated with the sum of the Shannon and the Bayesian surprise regularized by the confidence of the agent's belief. However, such an interpretation is no longer possible in volatile environments ($p_c > 0$), and Eq. (39) must be replaced by Proposition 9.

In order to account for the information of the true prior $\pi^{(0)}$ and to avoid cases where $\pi_{\text{flat}}(.|y_{t+1}, x_{t+1})$ is not a proper distribution, we also give a 2nd definition for the Confidence Corrected surprise as

$$\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) := D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(t+1)}_{\text{reset}}\Big], \tag{40}$$

where $\pi^{(t+1)}_{\text{reset}}(\theta)$ is defined in Proposition 1. Whenever $\pi^{(0)} = \pi_{\text{flat}}$, the two definitions are identical (Fig. 2B). Proposition 9 shows how the Confidence Corrected surprise relates to the Shannon surprise, the Bayesian surprise, and the confidence in the general case.

Proposition 9 (*Relation between the Confidence Corrected Surprise, Shannon Surprise, and Bayesian Surprise*)**.** *For the generative model of Definition 1, the original definition of the Confidence Corrected surprise can be written as*

$$\begin{aligned}\mathcal{S}_{\text{CC1}}(y_{t+1}|x_{t+1};\pi^{(t)}) =\\ \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) - \mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi_{\text{flat}})\\ + \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)})\\ - D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big]\\ + C\Big[\pi^{(t)}\Big] - C\Big[\pi_{\text{flat}}\Big],\end{aligned} \tag{41}$$

and our 2nd definition can be written as

$$\begin{aligned}\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\ + \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)})\\ - D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big]\\ + D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(0)}\Big].\end{aligned} \tag{42}$$

Proposition 9 conveys three important messages. First, both definitions of the Confidence Corrected surprise depend on differences in the Shannon surprise as opposed to the Shannon surprise itself (cf. first line in Eqs. (41) and (42)). Second, both definitions depend on the difference between the Bayesian surprise (i.e., the change in the belief given the new observation) and the *a priori* expected change in the belief (because of the possibility of a change in the environment; cf. second and third lines in Eqs. (41) and (42)). Third, both definitions regularize the contributions of Shannon surprise and Bayesian surprise by the relative confidence of the current belief compared to either the flat or the prior belief (cf. the last line in Eqs. (41) and (42)). 'Relative confidence' quantifies how different the current belief is with respect to a reference belief; note that $C\Big[\pi^{(t)}\Big] - C\Big[\pi_{\text{flat}}\Big] = D_{\text{KL}}\Big[\pi^{(t)}||\pi_{\text{flat}}\Big]$.

Hence, the Confidence Corrected surprise should be distinguishable from both the Shannon and the Bayesian surprise (for $p_c < 1$). An interesting consequence of Proposition 9, however, is that $\mathcal{S}_{\text{CC2}}$ is identical to $\mathcal{S}_{\text{Ba2}}$ when the environment becomes so volatile that its parameter changes at each time step (i.e., in the limit of $p_c \to 1$):

Corollary 3. *For the generative model of Definition 1, when $p_c \to 1$, we have $\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)})$.*

6.4. Minimized free energy

Although an agent can perform computations over the joint probability distribution in Eqs. (1) and (2), finding the belief $\pi^{(t+1)}(\theta)$ (i.e., the posterior distribution in Eq. (3)) can be computationally intractable (Barber, 2012; Liakoni et al., 2021). Therefore, it has been argued that the brain uses approximate inference (instead of exact Bayesian inference) for finding the belief (Daw & Courville, 2008; Faraji et al., 2018; Findling et al., 2021; Fiser et al., 2010; Friston, 2010; Friston et al., 2017; Liakoni et al., 2021; Mathys et al., 2011). An approximation of the belief $\pi^{(t+1)}(\theta)$ can for example be found via variational inference (Blei et al., 2017; MacKay, 2003) over a family of distributions $q(\theta;\phi)$ parameterized by ϕ. Such approaches are popular in neuroscience studies of learning and inference in the brain (Friston, 2010; Friston et al., 2017; Gershman, 2019).

Formally, in variational inference, the belief $\pi^{(t+1)}(\theta)$ is approximated by $\hat{\pi}^{(t+1)}(\theta) := q(\theta;\phi^{(t+1)})$, where $\phi^{(t+1)}$ is the minimizer of the variational loss or free energy, i.e., $\phi^{(t+1)} := \arg\min_{\phi} F^{(t+1)}(\phi)$ (MacKay, 2003). To define $F^{(t+1)}(\phi)$, we introduce a new notation:

$$\mathbb{P}_{\Theta_{t+1}}\big(\theta, y_{t+1}|x_{t+1};\pi\big) := P_{Y|X}(y_{t+1}|x_{t+1};\theta)\Big(p_c\pi^{(0)}(\theta) + (1-p_c)\pi(\theta)\Big), \tag{43}$$

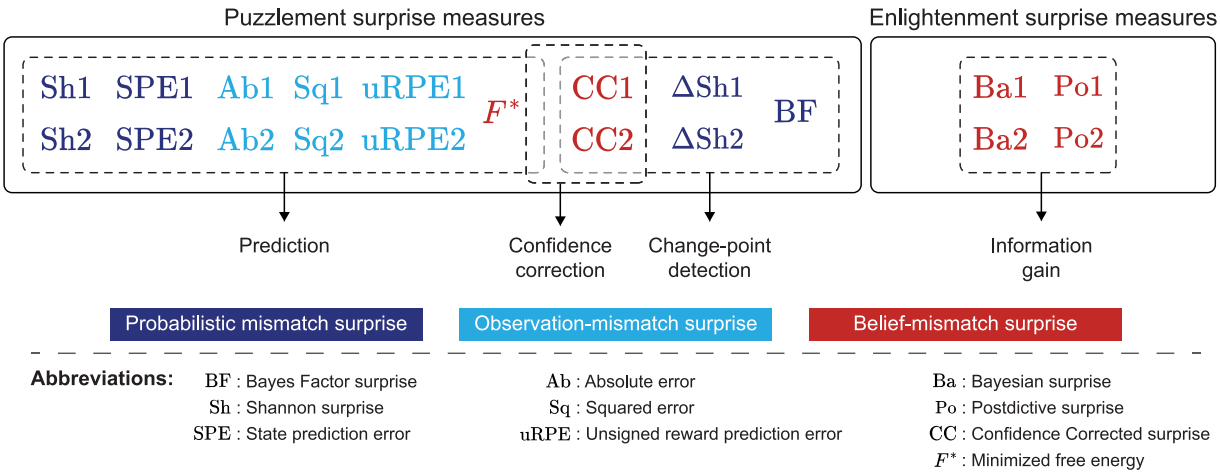

Fig. 5. Taxonomy of surprise definitions. Measures of puzzlement surprise (Faraji et al., 2018) can be further classified into 3 sub-categories of surprise measures highlighting (i) prediction, (ii) change-point detection, and (iii) confidence correction. According to surprise measures focused on prediction, the agent's puzzle is finding the most accurate prediction of the next observation. According to surprise measures focused on change-point detection, the agent's puzzle is to detect environmental changes. Surprise measures focused on confidence correction do not determine a specific puzzle (change-point detection or accurate prediction, visualized by overlapping boxes) for the agent but stress that confidence should explicitly influence puzzlement. The enlightenment surprise measures can be seen as measures of information gain. In addition to the 18 definitions of surprise discussed in Section 3, we included in the figure the difference in Shannon surprise (ΔSh1 and ΔSh2) introduced in Proposition 2. Color code shows the technical classification presented in Fig. 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where π is an arbitrary distribution over the parameter space. Using this notation, we can write the joint distribution over the observation and the parameter $\mathbb{P}^{(t)}(\theta_{t+1}, y_{t+1}|x_{t+1})$ as $\mathbb{P}_{\Theta_{t+1}}(\theta_{t+1}, y_{t+1}|x_{t+1}; \pi^{(t)})$ and the updated belief $\pi^{(t+1)}(\theta)$ as $\mathbb{P}_{\Theta_{t+1}}(\theta|y_{t+1}, x_{t+1}; \pi^{(t)})$. The variational loss or free energy can then be defined as (Liakoni et al., 2021; Markovic et al., 2021; Sajid et al., 2021)

$$F^{(t+1)}(\phi) := \mathbb{E}_{q(.;\phi)}\Big[\log q(\Theta;\phi) - \log \mathbb{P}_{\Theta_{t+1}}(\Theta, y_{t+1}|x_{t+1}; \hat{\pi}^{(t)})\Big]. \tag{44}$$

For any value of ϕ, one can show that (Blei et al., 2017; Sajid et al., 2021)

$$\begin{aligned} F^{(t+1)}(\phi) &= \mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \hat{\pi}^{(t)}) + \\ &\quad D_{\mathrm{KL}}\Big[q(.;\phi)||\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{\pi}^{(t)})\Big] \\ &\geq \mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \hat{\pi}^{(t)}), \end{aligned} \tag{45}$$

where the right side of the inequality is independent of ϕ, and $\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{\pi}^{(t)})$ is the exact Bayesian update of the belief (according to the generative model in Definition 1) given the latest approximation of the belief $\hat{\pi}^{(t)}$ (Liakoni et al., 2021; Markovic et al., 2021).

The minimized free energy $F^* := \min_\phi F^{(t+1)}(\phi)$ has been interpreted as a measure of surprise (Friston, 2010; Friston et al., 2017; Schwartenbeck et al., 2013), which, according to Eq. (45), can be seen as an approximation of $\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \hat{\pi}^{(t)})$. The parametric family of $q(.;\phi)$ and its relation to the exact belief $\pi^{(t+1)}$ determine how well F^* approximates $\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \hat{\pi}^{(t)})$ (Fig. 4B). More precisely, the minimized free energy measures both how unlikely the new observation is (i.e., how large $\mathcal{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1}; \hat{\pi}^{(t)})$ is) and how imprecise the best parametric approximation of the belief $\hat{\pi}^{(t+1)}$ is (i.e., how large $D_{\mathrm{KL}}[\hat{\pi}^{(t+1)}||\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{\pi}^{(t)})]$ is). Therefore, the minimized free energy is in the category of belief-mismatch surprise measures (Fig. 3).

7. Taxonomy of surprise definitions

In a unified framework, we discussed 10 previously proposed measures of surprise: (1) the Bayes Factor surprise; (2) the Shannon surprise; (3) the State Prediction Error; (4) the Absolute and (5) the Squared error surprise; (6) the unsigned Reward Prediction Error; (7) the Bayesian surprise; (8) the Postdictive surprise; (9) the Confidence Corrected surprise; and (10) the Minimized Free Energy. We considered different ways to define some of these measures in volatile environments and, overall, analyzed 18 different definitions of surprise. In this section, we propose a taxonomy of these 18 definitions and classify them into four main categories regarding the semantic of what they quantify (Fig. 5).

Measures of surprise in neuroscience have been previously divided into two categories (Faraji et al., 2018; Gijsen et al., 2021; Hurley et al., 2011): 'puzzlement' and 'enlightenment' surprise. Puzzlement surprise measures how puzzling a new observation is for an agent, whereas enlightenment surprise measures how much the new observation has enlightened the agent and changed its belief — a concept closely linked but not identical to the 'Aha! moment' (Dubey et al., 2021; Kounios & Beeman, 2009). The Bayesian and the Postdictive surprise can be categorized as enlightenment surprise since both quantify information gain (Fig. 5). Based on our theoretical analyses, however, we suggest to further divide measures of puzzlement surprise into three sub-categories (Fig. 5):

i. 'Prediction surprise' quantifies how unpredicted, unexpected, or unlikely the new observation is. This category includes the Shannon surprise, State Prediction Error, the Minimized Free Energy, and all observation-mismatch surprise measures (Fig. 5). According to these measures, the agent's puzzle is to find the most accurate predictions of the next observations. Surprise in natural language is defined as 'the feeling or emotion excited by something unexpected' (Oxford-English-Dictionary, 2021). If we focus on the term 'unexpected', identify it with 'unlikely under the current belief', and neglect the terms 'feeling' and 'emotion', then the quality measured by prediction surprise is closely related to the definition of surprise in natural language.

ii. 'Change-point detection surprise' quantifies relative unlikeliness of the new observation and are designed to modulate the learning rate and to identify environmental changes. This category includes the Bayes Factor surprise and the difference in Shannon surprise (cf. Corollary 1; Fig. 5). According to these measures, the agent's puzzle is to detect environmental changes.

iii. 'Confidence corrected surprise' explicitly accounts for the agent's confidence. The idea is that higher confidence (or

higher commitment to a belief) leads to more puzzlement, where the puzzle is either to detect environmental changes or to find the most accurate prediction. Faraji et al. (2018) argue, using a thought experiment, that such an explicit account for confidence is crucial to explain our perception of surprise. The only current candidates of this category are $\mathcal{S}_{\mathrm{CC1}}$ and $\mathcal{S}_{\mathrm{CC2}}$ that assume that the agent's puzzle is to detect environmental changes (cf. Proposition 9); but we anticipate that more examples in this category might be found in the future (see Modirshanechi et al. (2021) for example).

While our proposed taxonomy is solely conceptual and based on the theoretical properties of different definitions, we note that there have been a significant number of studies investigating the neural and physiological correlates of prediction (Gijsen et al., 2021; Gläscher et al., 2010; Kolossa et al., 2015; Konovalov & Krajbich, 2018; Kopp & Lange, 2013; Loued-Khenissi & Preuschoff, 2020; Maheu et al., 2019; Mars et al., 2008; Meyniel, 2020; Modirshanechi et al., 2019; Mousavi et al., 2022), change-point detection (Liakoni et al., 2022; Nassar et al., 2012; Xu et al., 2021), confidence correction (Gijsen et al., 2021), and information gain (Gijsen et al., 2021; Kolossa et al., 2015; Nour et al., 2018; O'Reilly et al., 2013; Ostwald et al., 2012; Visalli et al., 2021) surprise measures (Fig. 1). We, therefore, speculate that at least one measure from each of these categories is computed in the brain but potentially through different neural pathways and to be used for different brain functions.

8. Discussion

What does it formally mean to be surprised? And how do existing definitions of surprise relate to each other? To address these questions, we reviewed 18 definitions of surprise in a unifying mathematical framework and studied their similarities and differences. We showed that several extensions of known surprise measures to volatile environments are possible and potentially relevant; hence, further experimental evidence is needed to elucidate the relevance of precise definitions of surprise for brain research. Based on how different definitions depend on the belief $\pi^{(t)}$, we divided them into three groups of probabilistic mismatch, observation-mismatch, and belief-mismatch surprise measures (Fig. 3). We then showed how these measures relate to each other theoretically and, more importantly, under which conditions they are strictly increasing functions of each other (i.e., they become experimentally indistinguishable — Fig. 4 and Table 2). We further proposed a taxonomy of surprise definitions by a conceptual classification into four main categories (Fig. 5): (i) prediction surprise, (ii) change-point detection surprise, (iii) confidence-corrected surprise, and (iv) information gain surprise.

It is believed that surprise has important computational roles in different brain functions such as adaptive learning (Gerstner et al., 2018; Iigaya, 2016), exploration (Dubey & Griffiths, 2020; Gottlieb & Oudeyer, 2018), memory formation (Rouhani & Niv, 2021), and memory segmentation (Antony et al., 2021). Our results propose a diverse toolkit and a refined terminology to theoreticians and computational scientist to model and discuss the different functions of surprise and their biological implementation. For instance, it has been argued that the computation of observation-mismatch surprise measures is biologically more plausible than more abstract measures such as Shannon surprise (Iigaya, 2016). Our results identify conditions under which observation-mismatch surprise measures behave identically to probabilistic mismatch surprise measures that are optimal for adaptive learning (cf. Fig. 4B, Proposition 1, and Corollary 1); such insights can be exploited in future network models of adaptive behavior.

Moreover, our results can be used to design novel theory-driven experiments where different measures of surprise make different predictions. Importantly, most of the previous experimental studies have focused on one measure of surprise and its role and signatures in behavioral and physiological measurements. The examples that considered more than one surprise measure (Gijsen et al., 2021; Kolossa et al., 2015; Mars et al., 2008; Mousavi et al., 2022; Ostwald et al., 2012) have mainly focused on model-selection methods to compare different models and did not look for *fundamentally* different predictions of these measures — see Visalli et al. (2021) for an exception. Even if two surprise measures are formally distinguishable, it may be that, in a given experimental set-up, the number of samples or effect size are not big enough to extract the quantitative differences between the two. For example, $\mathcal{S}_{\mathrm{BF}}$ and $\mathcal{S}_{\mathrm{Sh1}}$ are distinguishable for any prior marginal distributions other than uniform distribution (Fig. 4B), but, in practice, the distinction is hard to detect for nearly-uniform priors. Our theoretical framework enables us to go further and design experiments that enable to dissociate different surprise measures based on their *qualitatively* different predictions and to avoid experiments where different measures are either formally or practically indistinguishable (see Modirshanechi et al. (2021) for example).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

AM is grateful to Vasiliki Liakoni, Martin Barry, and Valentin Schmutz for many useful discussions in the course of the last few years, and to Andrew Barto for insightful discussions during and after EPFL Neuro Symposium 2021 on "Surprise, Curiosity and Reward: from Neuroscience to AI". This research was supported by Swiss National Science Foundation (no. 200020_184615).

Appendix. Proofs

In this appendix, we provide proofs for our Propositions and Corollaries mentioned in the main text. We also provide further results for the postdictive surprise in Lemma 1.

A.1. Proof of Proposition 1

The proof is in essence the same as the proof of Proposition 1 of Liakoni et al. (2021). We write

$$\begin{aligned}\pi^{(t+1)}(\theta) &= \mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta)\\ &= \mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta|C_{t+1}=0)\mathbb{P}^{(t+1)}(C_{t+1}=0)+\\ &\quad \mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta|C_{t+1}=1)\mathbb{P}^{(t+1)}(C_{t+1}=1).\end{aligned} \tag{A.1}$$

We use Bayes' rule and write $\mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta|C_{t+1}=0)$ (cf. the 1st term in Eq. (A.1)) as

$$\begin{aligned}&\mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta|C_{t+1}=0)\\ &=\mathbb{P}^{(t)}(\Theta_{t+1}=\theta|C_{t+1}=0,x_{t+1},y_{t+1})\\ &=\frac{\mathbb{P}^{(t)}(y_{t+1}|C_{t+1}=0,x_{t+1},\Theta_{t+1}=\theta)}{\mathbb{P}^{(t)}(y_{t+1}|C_{t+1}=0,x_{t+1})}\times\\ &\quad \mathbb{P}^{(t)}(\Theta_{t+1}=\theta|C_{t+1}=0,x_{t+1})\\ &=\frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi^{(t)}(\theta)}{P(y_{t+1}|x_{t+1};\pi^{(t)})}=\pi^{(t+1)}_{\text{integration}}(\theta),\end{aligned} \tag{A.2}$$

and similarly

$$\mathbb{P}^{(t+1)}(\Theta_{t+1}=\theta|C_{t+1}=1)=\frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta)\pi^{(0)}(\theta)}{P(y_{t+1}|x_{t+1};\pi^{(0)})} \\ =\pi^{(t+1)}_{\text{reset}}(\theta). \tag{A.3}$$

Then, for $\mathbb{P}^{(t+1)}(C_{t+1}=1)$ and $\mathbb{P}^{(t+1)}(C_{t+1}=0)=1-\mathbb{P}^{(t+1)}(C_{t+1}=1)$ we have

$$\begin{aligned}&\mathbb{P}^{(t+1)}(C_{t+1}=1)\\ &\quad=\mathbb{P}^{(t)}(C_{t+1}=1|x_{t+1},y_{t+1})\\ &\quad=\frac{p_c P(y_{t+1}|x_{t+1};\pi^{(0)})}{(1-p_c)P(y_{t+1}|x_{t+1};\pi^{(t)})+p_c P(y_{t+1}|x_{t+1};\pi^{(0)})}\\ &\quad=\frac{m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})}{1+m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})}=\gamma_{t+1}\end{aligned} \tag{A.4}$$

with $m=\frac{p_c}{1-p_c}$. Therefore, the proof is complete by substituting these terms in Eq. (A.1). ■

A.2. Proof of Proposition 2

Based on the definition of the adaptation rate γ_{t+1} (cf. Proposition 1), we have

$$\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})=\frac{1-p_c}{p_c}\frac{\gamma_{t+1}}{1-\gamma_{t+1}}. \tag{A.5}$$

For the difference in the 1st definition of the Shannon surprise (cf. Eq. (9)), we can write

$$\begin{aligned}&\Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\ &=\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})-\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(0)})\\ &=\log\Big(\frac{P(y_{t+1}|x_{t+1};\pi^{(0)})}{p_c P(y_{t+1}|x_{t+1};\pi^{(0)})+(1-p_c)P(y_{t+1}|x_{t+1};\pi^{(t)})}\Big)\\ &=\log\frac{\gamma_{t+1}}{p_c}.\end{aligned} \tag{A.6}$$

As a result, we have $\gamma_{t+1}=p_c\exp\Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})$ and hence

$$\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})=\frac{(1-p_c)\exp\Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})}{1-p_c\exp\Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})}. \tag{A.7}$$

The proof is more straightforward for the difference in the 2nd definition (cf. Eq. (11)) where we have

$$\begin{aligned}&\Delta\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})\\ &\quad=\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})-\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(0)})\\ &\quad=\log\Big(\frac{P(y_{t+1}|x_{t+1};\pi^{(0)})}{P(y_{t+1}|x_{t+1};\pi^{(t)})}\Big)=\log\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)}).\end{aligned} \tag{A.8}$$

Therefore, the proof is complete. ■

A.3. Proof of Proposition 3

Based on the definitions of the two versions of the Shannon surprise (cf. Eqs. (9) and (11)), we have

$$\begin{aligned}\mathbb{P}^{(t)}(y_{t+1}|x_{t+1})&=\exp\Big(-\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\Big),\\ P(y_{t+1}|x_{t+1};\pi^{(t)})&=\exp\Big(-\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})\Big).\end{aligned} \tag{A.9}$$

The proof is complete by using these equations and replacing the probabilities in Eqs. (15) and (16). ■

A.4. Proof of Proposition 4

For a categorical task with N categories and one-hot coded observations, we have (cf. Eqs. (18) and (19))

$$\begin{aligned}E_1[Y_{t+1}]&=\Big[p_c P(n|x_{t+1};\pi^{(0)})+(1-p_c)P(n|x_{t+1};\pi^{(t)})\Big]_{n=1}^{N}\\ E_2[Y_{t+1}]&=\Big[P(n|x_{t+1};\pi^{(t)})\Big]_{n=1}^{N}\end{aligned} \tag{A.10}$$

where $z=[z_n]_{n=1}^N$ is an N-dimensional vector with z_n the nth element. To be able to prove the proposition for $E_1[Y_{t+1}]$ and $E_2[Y_{t+1}]$ simultaneously, we define $E_i[Y_{t+1}]=[p_{i,n}]_{n=1}^N$, where $p_{1,n}=p_c P(n|x_{t+1};\pi^{(0)})+(1-p_c)P(n|x_{t+1};\pi^{(t)})$ and $p_{2,n}=P(n|x_{t+1};\pi^{(t)})$.

We show the one-hot coded vector corresponding to category $m\in\{1,\dots,N\}$ by e_m. For the absolute error surprise, we have (cf. Eq. (20))

$$\begin{aligned}\mathcal{S}_{\text{Abi}}(y_{t+1}=e_m|x_{t+1};\pi^{(t)})&=\sum_{n=1}^{N}|\delta_{m,n}-p_{i,n}|\\ &=|1-p_{i,m}|+\sum_{n=1,n\neq m}^{N}p_{i,n}=2(1-p_{i,m}),\end{aligned} \tag{A.11}$$

which is the same as $2\mathcal{S}_{\text{SPEi}}(y_{t+1}=e_m|x_{t+1};\pi^{(t)})$ (cf. Eqs. (15) and (16)).

For the squared error surprise, we have (cf. Eq. (20))

$$\begin{aligned}\mathcal{S}_{\text{Sqi}}(y_{t+1}=e_m|x_{t+1};\pi^{(t)})&=\sum_{n=1}^{N}(\delta_{m,n}-p_{i,n})^2\\ &=(1-p_{i,m})^2+\sum_{n=1,n\neq m}^{N}p_{i,n}^2\\ &=2(1-p_{i,m})+\|[p_{i,n}]_{n=1}^N\|_2^2-1,\end{aligned} \tag{A.12}$$

where we have $2(1-p_{i,m})=2\mathcal{S}_{\text{SPEi}}(y_{t+1}=e_m|x_{t+1};\pi^{(t)})$ and

$$\text{Conf.}\Big[P(.|x_{t+1};\pi^{(t)})\Big]=\|[p_{i,n}]_{n=1}^N\|_2^2-1 \tag{A.13}$$

shows the ℓ_2-norm of the estimate vector $[p_{i,n}]_{n=1}^N$ as a measure of confidence; $\|[p_{i,n}]_{n=1}^N\|_2^2$ takes its maximum value when the prediction has a probability of 1 for one category and zero for the rest and takes its minimum when it is distributed uniformly over all categories. Therefore, the proof is complete. ■

A.5. Proof of Proposition 5

Assume that $Y_{t+1}\in\mathbb{R}^N$, given the cue x_{t+1} and the belief $\pi^{(t)}$, has a Gaussian distribution with a covariance matrix $\sigma^2 I$, i.e.,

$$P(y_{t+1}|x_{t+1};\pi^{(t)})=\mathcal{N}\Big(y_{t+1};E_2[Y_{t+1}],\sigma I\Big). \tag{A.14}$$

We then have

$$\begin{aligned}\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)})&=-\log\mathcal{N}\Big(y_{t+1};E_2[Y_{t+1}],\sigma I\Big)\\ &=\frac{N}{2}\log(2\pi\sigma)+\frac{\|y_{t+1}-E_2[Y_{t+1}]\|_2^2}{2\sigma^2}\\ &=a+b\mathcal{S}_{\text{Sq},2}(y_{t+1}=e_m|x_{t+1};\pi^{(t)}),\end{aligned} \tag{A.15}$$

where $a=N\log(2\pi\sigma)/2$ and $b=1/(2\sigma^2)$. Therefore, the proof is complete. ■

A.6. Proof of Proposition 6

Using the definition of the two surprise measures in Eq. (20), we have, for $y_{t+1} \in \mathbb{R}$,

$$\begin{aligned}\mathcal{S}_{\text{Sqi}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \|y_{t+1} - E_i[Y_{t+1}]\|_2^2 \\ &= |y_{t+1} - E_i[Y_{t+1}]|^2 = \mathcal{S}_{\text{Abi}}(y_{t+1}|x_{t+1};\pi^{(t)})^2.\end{aligned} \tag{A.16}$$

Therefore, the proof is complete. ■

A.7. Proof of Proposition 7

Using the definition of the uRPE and the absolute error surprise in Eqs. (20) and (25), we have

$$\begin{aligned}\mathcal{S}_{\text{Abi}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \|y_{t+1} - E_i[Y_{t+1}]\|_1 \\ &= |\tilde{r}_{t+1} - E_i[\tilde{R}_{t+1}]| + \|s_{t+1} - E_i[S_{t+1}]\|_1 \\ &= \mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1};\pi^{(t)}) + \mathcal{S}_{\text{Abi}}(s_{t+1}|x_{t+1};\pi^{(t)}),\end{aligned} \tag{A.17}$$

which complete the proof for the absolute error surprise. Then, we can similarly write

$$\begin{aligned}\mathcal{S}_{\text{Sqi}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \|y_{t+1} - E_i[Y_{t+1}]\|_2^2 \\ &= |\tilde{r}_{t+1} - E_i[\tilde{R}_{t+1}]|^2 + \|s_{t+1} - E_i[S_{t+1}]\|_2^2 \\ &= \mathcal{S}_{\text{uRPEi}}(y_{t+1}|x_{t+1};\pi^{(t)})^2 + \mathcal{S}_{\text{Sqi}}(s_{t+1}|x_{t+1};\pi^{(t)}).\end{aligned} \tag{A.18}$$

Therefore, the proof is complete. ■

A.8. Proof of Proposition 8

For the 1st definition of the Bayesian surprise (cf. Eq. (29)), we have

$$\begin{aligned}\mathcal{S}_{\text{Ba1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= D_{\text{KL}}\Big[\mathbb{P}^{(t)}_{\Theta_{t+1}}||\mathbb{P}^{(t+1)}_{\Theta_{t+1}}\Big] \\ &= \mathbb{E}_{\mathbb{P}^{(t)}}\Big[\log\frac{\mathbb{P}^{(t)}(\Theta_{t+1})}{\mathbb{P}^{(t+1)}(\Theta_{t+1})}\Big].\end{aligned} \tag{A.19}$$

We know

$$\mathbb{P}^{(t)}_{\Theta_{t+1}} = p_c\pi^{(0)} + (1-p_c)\pi^{(t)}, \tag{A.20}$$

and

$$\begin{aligned}\mathbb{P}^{(t+1)}(\theta_{t+1}) &= \frac{\mathbb{P}^{(t)}(\theta_{t+1})P_{Y|X}(y_{t+1}|x_{t+1};\theta_{t+1})}{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1})} \\ &\Rightarrow \\ \frac{\mathbb{P}^{(t+1)}(\theta_{t+1})}{\mathbb{P}^{(t)}(\theta_{t+1})} &= \frac{P_{Y|X}(y_{t+1}|x_{t+1};\theta_{t+1})}{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1})}.\end{aligned} \tag{A.21}$$

We, therefore, have

$$\begin{aligned}\mathcal{S}_{\text{Ba1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= -p_c\mathbb{E}_{\pi^{(0)}}\Big[\log P_{Y|X}(y_{t+1}|x_{t+1};\Theta)\Big] \\ &\quad - (1-p_c)\mathbb{E}_{\pi^{(t)}}\Big[\log P_{Y|X}(y_{t+1}|x_{t+1};\Theta)\Big] \\ &\quad + \log\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}),\end{aligned} \tag{A.22}$$

which is equivalent to (cf. Eqs. (9) and (11))

$$\begin{aligned}\mathcal{S}_{\text{Ba1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= p_c\mathbb{E}_{\pi^{(0)}}\Big[\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big] \\ &\quad + (1-p_c)\mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big] \\ &\quad - \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}).\end{aligned} \tag{A.23}$$

For the 2nd definition of the Bayesian surprise (cf. Eq. (30)), we have

$$\begin{aligned}\mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(t+1)}\Big] \\ &= \mathbb{E}_{\pi^{(t)}}\Big[\log\frac{\pi^{(t)}(\Theta)}{\pi^{(t+1)}(\Theta)}\Big].\end{aligned} \tag{A.24}$$

We use Eqs. (28) and (A.21) and write

$$\begin{aligned}\mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= -\mathbb{E}_{\pi^{(t)}}\Big[\log P_{Y|X}(y_{t+1}|x_{t+1};\Theta)\Big] \\ &\quad + \log\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}) \\ &\quad + \mathbb{E}_{\pi^{(t)}}\Big[\log\frac{\pi^{(t)}(\Theta)}{p_c\pi^{(0)}(\Theta) + (1-p_c)\pi^{(t)}(\Theta)}\Big],\end{aligned} \tag{A.25}$$

which is equivalent to (cf. Eqs. (9) and (11))

$$\begin{aligned}\mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big] \\ &\quad - \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) \\ &\quad + D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big].\end{aligned} \tag{A.26}$$

Therefore, the proof is complete. ■

A.9. Proof of Proposition 9

First, we prove the statement for the 2nd definition of the Confidence Corrected surprise (cf. Eq. (40)) for which we have

$$\begin{aligned}\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(t+1)}_{\text{reset}}\Big] \\ &= \mathbb{E}_{\pi^{(t)}}\Big[\log\frac{\pi^{(t)}(\Theta)}{\pi^{(t+1)}_{\text{reset}}(\Theta)}\Big].\end{aligned} \tag{A.27}$$

Using the definition of $\pi^{(t+1)}_{\text{reset}}$ in Proposition 1, we can write

$$\begin{aligned}\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= -\mathbb{E}_{\pi^{(t)}}\Big[\log P_{Y|X}(y_{t+1}|x_{t+1};\Theta)\Big] \\ &\quad + \log P(y_{t+1}|x_{t+1};\pi^{(0)}) \\ &\quad + \mathbb{E}_{\pi^{(t)}}\Big[\log\frac{\pi^{(t)}(\Theta)}{\pi^{(0)}(\Theta)}\Big],\end{aligned} \tag{A.28}$$

which is equivalent to (cf. Eqs. (9) and (11))

$$\begin{aligned}\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big] \\ &\quad - \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(0)}) \\ &\quad + D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(0)}\Big].\end{aligned} \tag{A.29}$$

Now, we can replace $\mathbb{E}_{\pi^{(t)}}\Big[\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big]$ by using Eq. (A.26) and have

$$\begin{aligned}\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) \\ &\quad - \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(0)}) \\ &\quad + \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)}) \\ &\quad - D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big] \\ &\quad + D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(0)}\Big],\end{aligned} \tag{A.30}$$

which is the same as Eq. (42). For the 1st definition of the Confidence Corrected surprise (cf. Eq. (37)), we can repeat all steps to have

$$\begin{aligned}\mathcal{S}_{\text{CC1}}(y_{t+1}|x_{t+1};\pi^{(t)}) &= \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) \\ &\quad - \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi_{\text{flat}}) \\ &\quad + \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)}) \\ &\quad - D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big] \\ &\quad + D_{\text{KL}}\Big[\pi^{(t)}||\pi_{\text{flat}}\Big].\end{aligned} \tag{A.31}$$

If $\pi^{(t)}$ is absolutely continuous with respect to π_{flat}, then we have $D_{\text{KL}}\Big[\pi^{(t)}||\pi_{\text{flat}}\Big] = C\Big[\pi^{(t)}\Big] - C\Big[\pi_{\text{flat}}\Big]$, which completes the proof. ■

A.10. Proof of Corollary 1

The corollary is the direct conclusion of Eqs. (A.6) and (A.8). ■

A.11. Proof of Corollary 2

Let us show the set of possible observations by $\mathcal{Y}$. We assume that $\mathcal{Y}$ is bounded, i.e., $|\mathcal{Y}| < \infty$. By assumption, we have $P(y_{t+1}|x_{t+1};\pi^{(0)}) = 1/|\mathcal{Y}|$. We therefore (using Eq. (5), Eq. (9), and Eq. (11)) have

$$\begin{aligned}&\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&\quad= \log\frac{m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})}{1+m\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})}+\log\frac{|\mathcal{Y}|}{p_c},\\&\mathcal{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \log\mathcal{S}_{\text{BF}}(y_{t+1}|x_{t+1};\pi^{(t)})+\log|\mathcal{Y}|.\end{aligned} \tag{A.32}$$

Both mappings are strictly increasing. Therefore, the proof is complete. ■

A.12. Proof of Corollary 3

In the limit of $p_c \to 1$, we have $\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(0)})$ (cf. Eq. (9)) which implies that $\Delta\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})$ (cf. Proposition 2) in Eq. (42) is equal to 0. Similarly, in the limit of $p_c \to 1$, we have $D_{\text{KL}}\Big[\pi^{(t)}||p_c\pi^{(0)} + (1-p_c)\pi^{(t)}\Big] = D_{\text{KL}}\Big[\pi^{(t)}||\pi^{(0)}\Big]$. Therefore, in the limit of $p_c \to 1$ and given Eq. (42), we have $\mathcal{S}_{\text{CC2}}(y_{t+1}|x_{t+1};\pi^{(t)}) = \mathcal{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};\pi^{(t)})$. ■

A.13. Theoretical results for the postdictive surprise

Lemma 1 (*Relation Between the Postdictive Surprise and the Shannon Surprise*)**.** *In the generative model of Definition 1, the postdictive surprise can be written as*

$$\begin{aligned}&\mathcal{S}_{\text{Po1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&= \mathbb{E}_{P\left(.|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\right)}\Big[\mathcal{S}_{\text{Sh2}}\Big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}|Y,x_{t+1}}\Big)\Big]\\&- \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\end{aligned} \tag{A.33}$$

and

$$\begin{aligned}&\mathcal{S}_{\text{Po2}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&= \mathbb{E}_{P\left(.|x_{t+1};\pi^{(t)}\right)}\Big[\mathcal{S}_{\text{Sh2}}\Big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}|Y,x_{t+1}}\Big)\Big]\\&- \mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&+ D_{\text{KL}}\Big[P\big(.|x_{t+1};\pi^{(t)}\big)||P\big(.|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)\Big],\end{aligned} \tag{A.34}$$

where $\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}} := \mathbb{P}^{(t)}_{\Theta_{t+1}}\big(.|Y_{t+1}=y,x_{t+1}\big)$ *is the belief at time* $t+1$ *if we observe* $Y_{t+1}=y$ *with the cue* x_{t+1}.

According to Lemma 1, the postdictive surprise is equal to the difference between the expected (over all values of Y_{t+1}) Shannon surprise of $Y_{t+2} = y_{t+1}$ given $X_{t+2} = x_{t+1}$ and the Shannon surprise of y_{t+1} given x_{t+1}.

Proof. We first prove the equality for $\mathcal{S}_{\text{Po1}}$ for which we have (cf. Eq. (34))

$$\begin{aligned}&\mathcal{S}_{\text{Po1}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&\quad= D_{\text{KL}}\Big[P\big(.|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)||P\big(.|x_{t+1};\pi^{(t+1)}\big)\Big]\\&\quad= \mathbb{E}_{P\left(.|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\right)}\Big[\log\frac{P\big(Y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}{P\big(Y|x_{t+1};\pi^{(t+1)}\big)}\Big],\end{aligned} \tag{A.35}$$

where

$$P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big) = \int P_{Y|X}(y|x_t;\theta)\mathbb{P}^{(t)}\big(\Theta_{t+1}=\theta\big)d\theta, \tag{A.36}$$

and, using Bayes' rule,

$$\begin{aligned}P\big(y|x_{t+1};\pi^{(t+1)}\big) &= \int P_{Y|X}(y|x_{t+1};\theta)\pi^{(t+1)}(\theta)d\theta\\&= \int P_{Y|X}(y|x_{t+1};\theta)\frac{\mathbb{P}^{(t)}\big(\Theta_{t+1}=\theta\big)P_{Y|X}(y_{t+1}|x_{t+1};\theta)}{P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}d\theta.\end{aligned} \tag{A.37}$$

Using the Bayes' rule and the definition of the marginal probability (cf. Eq. (4)), we can find

$$\begin{aligned}\frac{P\big(y|x_{t+1};\pi^{(t+1)}\big)}{P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)} &= \frac{1}{P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}\\&\times\int P_{Y|X}(y_{t+1}|x_{t+1};\theta)\frac{\mathbb{P}^{(t)}\big(\Theta_{t+1}=\theta\big)P_{Y|X}(y|x_{t+1};\theta)}{P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}d\theta\end{aligned} \tag{A.38}$$

that is equal to

$$\begin{aligned}&\frac{\int P_{Y|X}(y_{t+1}|x_{t+1};\theta)\mathbb{P}^{(t)}\big(\Theta_{t+1}=\theta|Y_{t+1}=y,x_{t+1}\big)d\theta}{P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}\\&= \frac{\int P_{Y|X}(y_{t+1}|x_{t+1};\theta)\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}}(\theta)d\theta}{P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}\\&= \frac{P\Big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}}\Big)}{P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)},\end{aligned} \tag{A.39}$$

and as a result (using Eqs. (9) and (11))

$$\begin{aligned}&\log\frac{P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}{P\big(y|x_{t+1};\pi^{(t+1)}\big)}\\&= -\log P\Big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}}\Big)\\&\quad+\log P\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)\\&=\mathcal{S}_{\text{Sh2}}\big(y_{t+1}|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}}\big)\\&\quad-\mathcal{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};\pi^{(t)}),\end{aligned} \tag{A.40}$$

which, using Eq. (A.35), makes the proof complete.

To prove the 2nd equality, we note that (cf. Eq. (35))

$$\begin{aligned}&\mathcal{S}_{\text{Po2}}(y_{t+1}|x_{t+1};\pi^{(t)})\\&\quad= D_{\text{KL}}\Big[P\big(.|x_{t+1};\pi^{(t)}\big)||P\big(.|x_{t+1};\pi^{(t+1)}\big)\Big]\\&\quad= \mathbb{E}_{P\left(.|x_{t+1};\pi^{(t)}\right)}\Big[\log\frac{P\big(Y|x_{t+1};\pi^{(t)}\big)}{P\big(Y|x_{t+1};\pi^{(t+1)}\big)}\Big],\end{aligned} \tag{A.41}$$

and

$$\begin{aligned}&\log\frac{P\big(y|x_{t+1};\pi^{(t)}\big)}{P\big(y|x_{t+1};\pi^{(t+1)}\big)} =\\&\log\frac{P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}{P\big(y|x_{t+1};\pi^{(t+1)}\big)}+\log\frac{P\big(y|x_{t+1};\pi^{(t)}\big)}{P\big(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}}\big)}.\end{aligned} \tag{A.42}$$

Therefore, using Eq. (A.40) and the definition of D_{KL}, the proof is complete. ■

References

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

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50.

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2021). Behavioral physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, *109*(2), 377–390.

Baldi, P. (2002). *A computational theory of surprise* (pp. 1–25). Boston, MA: Springer US.

Baldi, P., & Itti, L. (2010). Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, *23*(5), 649–666.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, *4*(907).

Bayarri, M., & Berger, J. O. (1997). *Measures of surprise in bayesian analysis*. Duke University.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2019). Large-scale study of curiosity-driven learning. In *International conference on learning representations*.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in Neural Information Processing Systems*, *20*, 369–376.

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Dubey, R., & Griffiths, T. L. (2019). Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review*, *127*(3), 455–476.

Dubey, R., & Griffiths, T. L. (2020). Understanding exploration in humans and machines by formalizing the function of curiosity. *Current Opinion in Behavioral Sciences*, *35*, 118–124.

Dubey, R., Ho, M. K., Mehta, H., & Griffiths, T. (2021). Aha! moments correspond to meta-cognitive prediction errors. PsyArXiv.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

Faraji, M., Preuschoff, K., & Gerstner, W. (2018). Balancing new against old information: the role of puzzlement surprise in learning. *Neural Computation*, *30*(1), 34–83.

Fearnhead, P., & Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*(4), 589–605.

Findling, C., Chopin, N., & Koechlin, E. (2021). Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, *5*, 99–112.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

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

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, *29*(1), 1–49.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons Behavior, Data Analysis, and Theory*, *2*(3), 1–10.

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *ELife*, (6), Article e23763.

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in Neural Circuits*, *12*.

Gijsen, S., Grundei, M., Lange, R. T., Ostwald, D., & Blankenburg, F. (2021). Neural surprise in somatosensory bayesian learning. *PLoS Computational Biology*, *17*(2), 1–36.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Glaze, C. M., Kable, J. W., & Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *ELife*, *4*, Article e08825.

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*, 758–770.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., & Platt, M. L. (2011). Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience*, *31*(11), 4178–4187.

Heilbron, M., & Meyniel, F. (2019). Confidence resets reveal hierarchical adaptive learning in humans. *PLoS Computational Biology*, *15*(4), Article e1006972.

Horvath, L., Colcombe, S., Milham, M., Ray, S., Schwartenbeck, P., & Ostwald, D. (2021). Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task. *Computational Brain & Behavior*.

Huettel, S. A., Mack, P. B., & McCarthy, G. (2002). Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*, *5*(5), 485–490.

MIT press, Hurley, M. M., Dennett, D. C., Adams Jr, R. B., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Iigaya, K. (2016). Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *ELife*, *5*, Article e18073.

Imada, T., Hari, R., Loveless, N., McEvoy, L., & Sams, M. (1993). Determinants of the auditory mismatch response. *Electroencephalography and Clinical Neurophysiology*, *87*(3), 144–153.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems, vol. 18*. MIT Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

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

Konovalov, A., & Krajbich, I. (2018). Neurocomputational dynamics of sequence learning. *Neuron*, *98*(6), 1282–1293e4.

Kopp, B., & Lange, F. (2013). Electrophysiological indicators of surprise and entropy in dynamic task-switching environments. *Frontiers in Human Neuroscience*, *7*(300).

Kounios, J., & Beeman, M. (2009). The aha! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science*, *18*(4), 210–216.

Lehmann, M. P., Xu, H. A., Liakoni, V., Herzog, M. H., Gerstner, W., & Preuschoff, K. (2019). One-shot learning and behavioral eligibility traces in sequential decision making. *ELife*, *8*, Article e47463.

Liakoni, V., Lehmann, M. P., Modirshanechi, A., Brea, J., Lutti, A., Gerstner, W., & Preuschoff, K. (2022). Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. *NeuroImage*, *246*, Article 118780.

Liakoni, V., Modirshanechi, A., Gerstner, W., & Brea, J. (2021). Learning in volatile environments with the bayes factor surprise. *Neural Computation*, *33*(2), 1–72.

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., & Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Computer Biology*, *9*(2), Article e1002911.

Loued-Khenissi, L., & Preuschoff, K. (2020). Information theoretic characterization of uncertainty distinguishes surprise from accuracy signals in the brain. *Frontiers in Artificial Intelligence*, *3*(5).

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

Maheu, M., Dehaene, S., & Meyniel, F. (2019). Brain signatures of a multiscale process of sequence learning in humans. *ELife*, *8*, Article e41541.

Markovic, D., Stojic, H., Schwoebel, S., & Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. *Neural Networks*, *144*, 229–246.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, *28*(47), 12539–12545.

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, *5*(39).

Meyniel, F. (2020). Brain dynamics for confidence-weighted learning. *PLoS Computational Biology*, *16*, 1–27.

Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human inferences about sequences: A minimal transition probability model. *PLoS Computational Biology*, *12*, 1–26.

Modirshanechi, A., Brea, J., & Gerstner, W. (2021). Surprise: a unified theory and experimental predictions. *bioRxiv*, http://dx.doi.org/10.1101/2021.11.01.466796.

Modirshanechi, A., Kiani, M. M., & Aghajan, H. (2019). Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks. *NeuroImage*, *196*, 302–317.

Mousavi, Z., Kiani, M. M., & Aghajan, H. (2022). Spatiotemporal signatures of surprise captured by magnetoencephalography. *Frontiers in Systems Neuroscience*, *16*.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154, Special Issue: Dynamic Decision Making.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H. B., Coello, C., Wall, M. B., Dolan, R. J., & Howes, O. D. (2018). Dopaminergic basis for signaling belief updates but not surprise, and the link to paranoia. *Proceedings of the National Academy of Sciences*, *115*(43), E10167–E10176.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660–E3669.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., & Blankenburg, F. (2012). Evidence for neural encoding of bayesian surprise in human somatosensation. *NeuroImage*, *62*(1), 177–188.

Oxford-English-Dictionary (2021). Surprise. OED Online.

Palm, G. (2012). *Novelty, information and surprise*. Springer Science & Business Media.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International conference on machine learning - volume 70* (pp. 2778–2787). JMLR.org..

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552.

Prat-Carrabin, A., Wilson, R. C., Cohen, J. D., & Silveira, R. Azeredo. da. (2021). Human inference in changing environments with temporal structure. *Psychological Review*, *128*(5), 879–912.

Preuschoff, K., t Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*(115).

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! neural correlates of pearce–hall and rescorla–wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190–1200.

Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife*, *10*, e61077.

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1430–1443.

Rubin, J., Ulanovsky, N., Nelken, I., & Tishby, N. (2016). The representation of prediction error in auditory cortex. *PLoS Computational Biology*, *12*(8), Article e1005058.

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, *33*(3), 674–712.

Schmidhuber, J. (2010). Formal theory of creativity fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *2755306*, 1593–1599.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration novelty, surprise, and free energy minimization. *Frontiers in Psychology*, *4*(710).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sinclair, A. H., & Barense, M. D. (2018). Surprise and destabilize: prediction error influences episodic memory reconsolidation. *Learning & Memory*, *25*(8), 369–381.

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*(10), 635–644.

Squires, K. C., Wickens, C., Squires, N. K., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, *1934258*, 1142–1146.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction*. MIT Press.

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors not reward prediction errors. *Journal of Neuroscience*, *33*(19), 8264–8269.

Tribus, M. (1961). In D. Van Nostrand (Ed.), *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*.

Visalli, A., Capizzi, M., Ambrosini, E., Kopp, B., & Vallesi, A. (2021). Electroencephalographic correlates of temporal bayesian belief updating and surprise. *NeuroImage*, *231*, Article 117867.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to bayesian inference in change-point problems. *PLoS Computational Biology*, *9*(7), Article e1003150.

Xu, H. A., Modirshanechi, A., Lehmann, M. P., Gerstner, W., & Herzog, M. H. (2021). Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLoS Computational Biology*, *17*(6).

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior? In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems, vol. 21*. Curran Associates, Inc.

Yu, A. J., & Dayan, P. (2005). Uncertainty neuromodulation, and attention. *Neuron*, *46*(4), 681–692.