Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Perspective

Surprise and novelty in the brain

Alireza Modirshanechi[1,2], Sophia Becker[1,2], Johanni Brea[1,2] and Wulfram Gerstner[1,2]

Abstract

Notions of surprise and novelty have been used in various experimental and theoretical studies across multiple brain areas and species. However, 'surprise' and 'novelty' refer to different quantities in different studies, which raises concerns about whether these studies indeed relate to the same functionalities and mechanisms in the brain. Here, we address these concerns through a systematic investigation of how different aspects of surprise and novelty relate to different brain functions and physiological signals. We review recent classifications of definitions proposed for surprise and novelty along with links to experimental observations. We show that computational modeling and quantifiable definitions enable novel interpretations of previous findings and form a foundation for future theoretical and experimental studies.

Addresses

[1] Brain-Mind Institute, School of Life Sciences, EPFL, Lausanne, Switzerland

[2] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

Corresponding authors: Gerstner, Wulfram (wulfram.gerstner@epfl.ch); Modirshanechi, Alireza (alireza.modirshanechi@epfl.ch)

(Modirshanechi A.), (Becker S.), (Gerstner W.)

Current Opinion in Neurobiology 2023, **82:**102758

This review comes from a themed issue on **Computational Neuroscience 2023**

Edited by **Jeanette Hellgren Kotaleski** and **Tatjana Tchumatchenko**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.conb.2023.102758

Introduction

An unexpected video interruption strengthens human memory of the video's content [1], mismatches between visual flow and locomotion facilitate synaptic changes in the mouse visual cortex [2], monkeys show faster saccades to unseen objects than to familiar ones [3], and mice have a higher breathing frequency when sniffing new odors than those already known [4].

What these four statements have in common is that they all concern situations where words like 'surprise' and 'novelty' seem applicable: The first two statements assess neural responses to violation of expectations, potentially caused by a feeling of surprise, whereas the second two statements assess behavioral responses to unfamiliar stimuli, potentially triggered by novelty of the stimuli. It hence feels tempting to rephrase the first two statements to 'surprise strengthens memory and modulates learning' and the second two to 'novelty attracts attention and drives curiosity'. However, the rephrased statements imply notably more than the original statements: They suggest common mechanisms for different experimental phenomena across different species. Such generalisations are important for moving towards a unified understanding of the brain, but they can be misleading if not justified.

Intuitive usage of 'surprise' and 'novelty' is common practice in neuroscience [5], psychology [6], and machine learning [7]. However, it has remained a mystery how humans' self-reported degree of 'surprise' when entering a new and unexpected room [8] relates to the brain activity of monkeys seeing 'surprising' fractals [9]. This is particularly worrisome as the words 'surprise' and 'novelty', sometimes used interchangeably, refer to different measurable variables in different studies [10,11]. Moreover, neural and behavioral signatures of several novelty- or surprise-related variables have been found simultaneously in single experiments [12–16].

If there are indeed common principles of how 'surprise' and 'novelty' contribute to different brain functions across brain areas and species, then we need systematic studies that enable neuroscientists to distinguish between different 'aspects' of surprise and novelty. In this paper, we argue that computational modeling and quantifiable definitions are necessary first steps for such systematic studies.

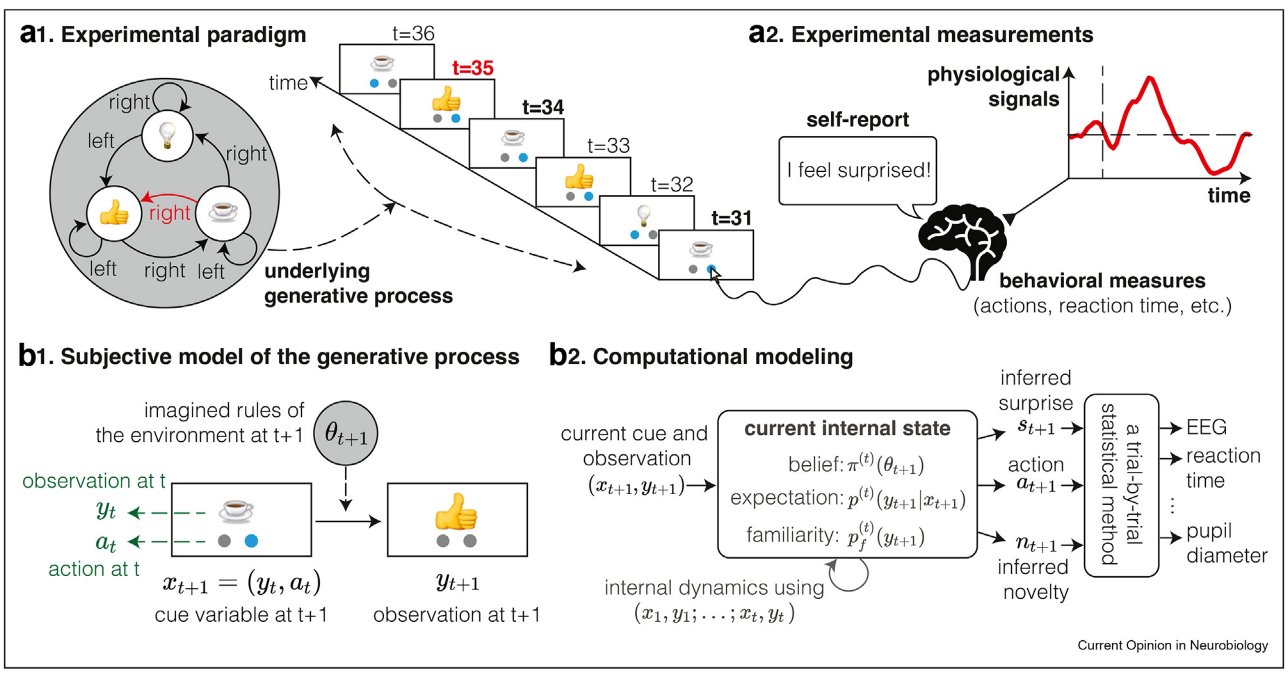
A unifying computational framework

In experimental paradigms for studying surprise and novelty, experimental subjects (human participants or animals) are presented with unlikely or infrequent observations [17,18], observations violating repeating patterns [19–22], or, in general, any observation that can *intuitively* be called 'novel' or 'surprising' (e.g., Figure 1a1). The goal of these experiments is to study how novel or surprising observations influence physiological brain signals [13,23] and action choices [16,24] (Figure 1a2).

In an example of human multi-step decision-making [16], participants see an image on a computer screen and are instructed to select an action by clicking on one of the disks below the image (Figure 1a1). The next image to appear on the screen depends on the current image and the selected action and is determined by some underlying rules that are unknown to the participants. After several trials, participants associate a particular action with a particular outcome, e.g., clicking on the right action below the coffee cup yields the light bulb as the next image ($t = 31$ and $t = 32$ in Figure 1a1). Participants will feel surprised if they see a different image than the expected one (e.g., the thumb at $t = 35$ in Figure 1a1). The experimental design is based on the idea that measurable changes in, e.g., EEG, pupil dilation, or reaction time after seeing the unexpected image can be attributed to surprise.

Computational models and quantifiable definitions allow us to go beyond mere ideas. A computational model consists of two parts: (i) an abstract description of the experimental paradigm (from the perspective of experimental subjects; Figure 1b1) and (ii) a formal description of subjects' perception and behavior (Figure 1b2). We can describe most existing experiments on surprise and novelty by using only three variables at time $t+1$ (Figure 1b1): The observation y_{t+1}, a potential cue x_{t+1}, and a set of hidden parameters θ_{t+1} (Box 1) [11].

Figure 1



Computational modeling of experimental paradigms studying surprise and novelty. a. The goal of experiments on surprise and novelty is to study the influence of 'novel' or 'surprising' observations (a1) on various behavioral and physiological measurements (a2). The example in a1 shows a simplified version of the task of [16]: In each trial, human participants see an image on a computer screen and select one of the two available actions (i.e., disks below the image; selected actions are shown in blue). The next image depends on the current image and the selected action and is determined by the underlying rules of the experiment that are unknown to participants (i.e., the graph on the left side; the black arrows correspond to available actions and the red one to a potentially surprising transition after an unannounced change of rules). Assuming all transitions have been experienced in the first 30 trials, observing the 'light bulb' at $t = 32$ is expected, whereas observing the 'thumb' at $t = 35$ is unexpected and potentially surprising (after taking action 'right' when seeing the 'coffee cup'). See Figure 2a and [11] for other examples. **b.** A computational model of an experiment consists of an abstract description of the experimental paradigm (b1) and a formal description of the subjects' behavior (b2). **b1.** The great majority of experiments can be described using three variables for the trial at time $t + 1$: The observation y_{t+1}, the cue x_{t+1}, and the parameter set θ_{t+1} [11]. For the example in a1, y_{t+1} is the image at time $t + 1$, $x_{t+1} = (y_t, a_t)$ is the pair of the last image y_t and action a_t, and θ_{t+1} models the transitions according to the rules imagined by the subject. **b2.** A subject is modeled by an algorithm that receives a cue x_{t+1} and an observation y_{t+1} as inputs and gives an inferred surprise value s_{t+1}, an inferred novelty value n_{t+1}, and, when required, an action a_{t+1} as outputs. The algorithm has an internal state that is iteratively updated according to some internal dynamics by using the past cues and observations $(x_1, y_1; \ldots; x_t, y_t)$. In general, the internal state includes a belief $\pi^{(t)}(\theta_{t+1})$ about the parameter set θ_{t+1}, a predictive model $p^{(t)}(y_{t+1}|x_{t+1})$ to summarise the subject's expectations (e.g., Equation (1)), and a familiarity measure $p_f^{(t)}(y_{t+1})$ to quantify the familiarity of observations (e.g., Equation (2)); see Box 1. Novelty and surprise values of each observation are evaluated according to the internal state of the algorithm as in Equation (3) and Equation (4), respectively. These values are used for trial-by-trial prediction of experimental measurements (e.g., using linear regression). See [11,16] for precise definitions and [13,25] for some examples.

The cue x_{t+1} summarises all information in time step $t+1$ that subjects may consider for predicting y_{t+1}, e.g., the pair (y_t, a_t) of observation y_t and action a_t (Figure 1b1). We always include the action a_t in the cue variable x_{t+1}; this allows us to use the same mathematical formulation for experiments with or without the possibility of selecting actions. The set of parameters θ_{t+1} summarises the hidden rules (for example action-dependent transitions in Figure 1b1) that subjects, potentially unconsciously, imagine to explain the observation y_{t+1} given x_{t+1}. The imagined rules are estimates of the 'real' rules of the experiment.

Defining novelty and surprise for the observation y_{t+1} needs a formal model of how experimental subjects perceive y_{t+1}, which is described by the second part of a computational model. All modeling studies on surprise and novelty assume that subjects use their past experiences $(x_1, y_1; \ldots; x_t, y_t)$ and some internal update dynamics to make a prediction of the next observation $\widehat{y}_{t+1}$ (Box 1) and, if required, select an action a_t accordingly (Figure 1b2) [26–29]. Depending on the model assumptions, the internal dynamics can have different levels of abstractions [30], ranging from algorithmic implementations of Bayesian inference [31–34] to detailed models of biological neural networks [35–38]. In the most general setting, the model describes (i) the belief $\pi^{(t)}(\theta_{t+1})$ of the subject about the unknown set of parameters θ_{t+1} and (ii) a predictive distribution of the next observation $p^{(t)}(y_{t+1}|x_{t+1})$ based on that belief (Box 1). The belief $\pi^{(t)}(\theta_{t+1})$ indicates the probability of θ_{t+1} to be the 'real' rule of the experiment at time $t+1$ according to the subjects' past experience up to time t. The predictive distribution $p^{(t)}(y_{t+1}|x_{t+1})$ summarises subjects' expectations of what they might observe next (Box 1). For example, in a simple case where x_{t+1} and y_{t+1} take discrete values, we can define the predictive distribution as [29,39]

$$p^{(t)}\left(y_{t+1}|x_{t+1}\right) = \frac{C^{(t)}\left(y_{t+1}|x_{t+1}\right) + \text{constant}}{C^{(t)}\left(x_{t+1}\right) + \text{constant}}, \qquad (1)$$

where $C^{(t)}(x_{t+1})$ is the count of how many times a subject has received cue x_{t+1} until time t, $C^{(t)}(y_{t+1}|x_{t+1})$ is the count of those trials that were followed by observation y_{t+1}, and constants are added to avoid having zero probabilities.

Novelty is not surprise

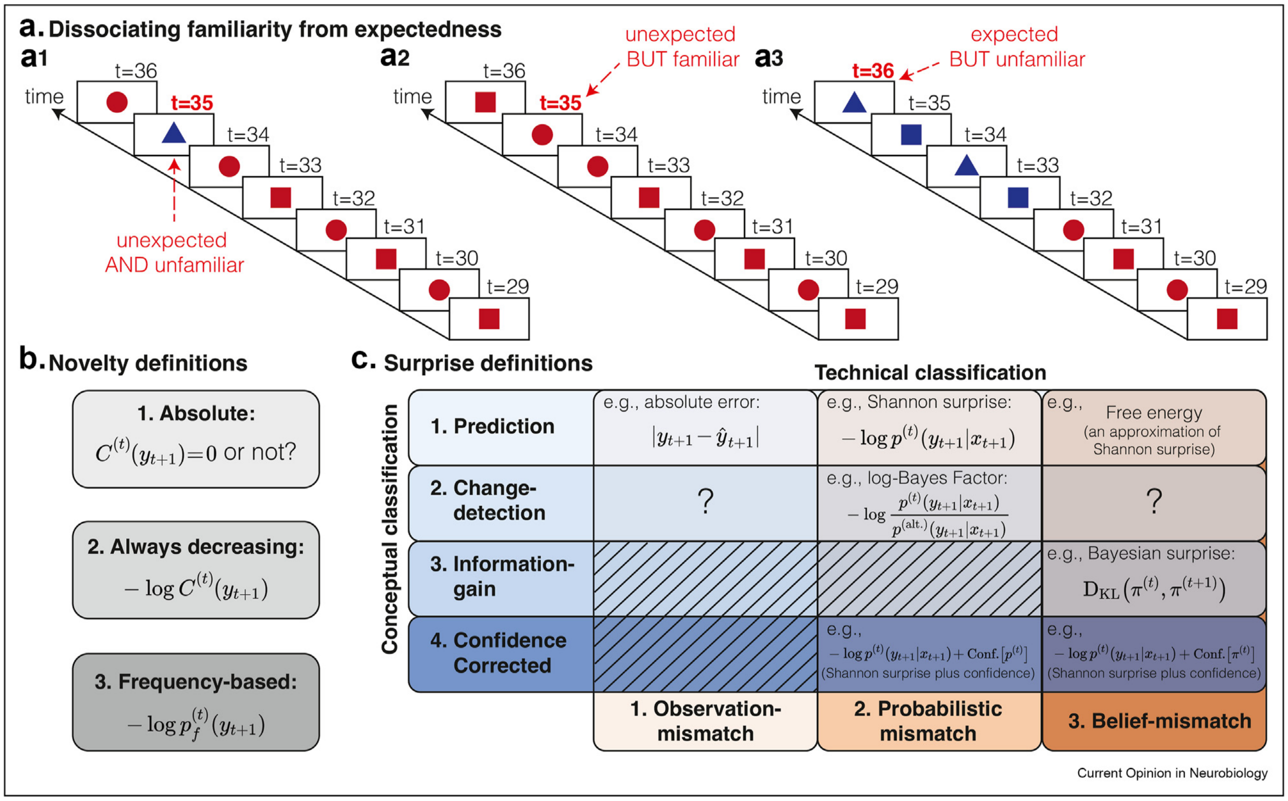
Homann et al. (2022) [22] identify a population of neurons in the mouse primary visual cortex that shows strong responses to novel stimuli but not to familiar stimuli even if the latter violate highly predictable observation patterns (Figure 2a1 versus Figure 2a2; Box 1). In the computational framework described above, this means that the physiological variables studied by [22] do not depend on the unexpectedness of y_{t+1} given the cue x_{t+1} (i.e., preceding stimuli in this case) but only on the unfamiliarity of y_{t+1} independently of any inferred regularities in the sequence of observations (Box 1).

These experimental results support the earlier proposition of Xu et al. (2021) [16] to separate notions of

Box 1. Glossary. Explanation of technical terms used to describe experiments, experimental subjects, and observations.

- When describing an **experiment**
  - **Cue** refers to information that subjects use to predict the next observation. The previously selected action (Figure 1a1) or the previous observation (Figure 2a) can be used as cues.
  - **Hidden parameters** describe the rules that generate experimental observations. A rule may imply that observation B always comes after observation A (Figure 2a). The rules are called hidden because they are not known by the subject but need to be inferred from observations. The rule in the mind of a subject may not be the same as the 'real' rule of the experiment.
  - A **Volatile experiment** is an experiment where the 'real' rule changes at unknown moments in time, e.g., [19,24].
- When describing an **experimental subject**
  - The **Belief** summarises the subject's guess about the hidden rules, based on past observations. Belief forms a probability distribution over all possible rules of the experiment.
  - **Expectations** summarise a subject's guess about possible next observations, based on the current *cue* and the current *belief.* Expectations form a probability distribution over all possible next observations.
  - A **Prediction** condenses a subject's *expectations* into a single guess for the next observation.
  - **Confidence** quantifies the certainty of a subject about either (i) the hidden rule or (ii) the next observation.
  - **Familiarity** quantifies how often a specific observation has occurred or how similar it is to other frequent observations. Familiarity does not depend on *cues.*
- When describing an **observation**
  - **Predictable** observations can in principle (i.e., if experimental rules are known) be predicted with high probability from *cues.* For example, observations in repeating or regular patterns are predictable (Figure 2a).
  - **Unexpected** observations are either unlikely given the subject's *expectations* or predicted inaccurately given the subject's *prediction.* Given sufficient experience, *predictable* observations are on average less unexpected than *unpredictable* ones. Whether an observation is unexpected depends on the *cue.*
  - **Unfamiliar** observations are those that have been encountered rarely by subjects and are not similar to other frequent observations (i.e., low *familiarity*). An *expected* observation can be unfamiliar (Figure 2a3), while a familiar observation can be *unexpected* (Figure 2a2). Whether an observation is unfamiliar does not depend on the *cue.*

Figure 2



A taxonomy of surprise and novelty definitions. Novelty quantifies the unfamiliarity of an observation (Equation (3)), whereas surprise quantifies its unexpectedness conditioned on the cue variable x_{t+1} (Equation (4)) [16]. **a.** Average familiarity and expectedness can be manipulated in an experimental paradigm where each observation $y_t = x_{t+1}$ is the predictor of the next observation y_{t+1} (e.g., [9,14,22]; Box 1). A blue triangle in the middle of a repeating sequence of red squares and circles is unexpected and unfamiliar (high surprise, high novelty; a1), whereas a misplaced red circle is unexpected although familiar (high surprise, low novelty; a2). A blue triangle observed for the second time after a switch in the observation pattern from repeating red square-red circle to repeating blue square-blue triangle is expected but not familiar (low surprise, high/medium novelty; a3). **b.** Most definitions of novelty can be classified into three groups: 1. 'Absolute novelty' considers novel observations as those never observed before ($C^{(t)}$ (y_{t+1}): the count of y_{t+1} until time t). 2. 'Always decreasing novelty' is a decreasing function of the count $C^{(t)}$ (y_{t+1}). 3. 'Frequency-based novelty' is a decreasing function of the observation-frequency $p_f^{(t)}(y_{t+1})$ (e.g., Equation (2)). **c.** A technical classification of surprise definitions (columns) [11]: 1. Observation-mismatch surprise needs only a prediction $\hat{y}_{t+1}$ of the next observation. 2. Probabilistic mismatch surprise needs the full predictive distribution $p^{(t)}$ $(y_{t+1}|x_{t+1})$. 3. Belief-mismatch surprise needs the subject's full belief distribution $\pi^{(t)}$ (θ_{t+1}); $\mathrm{D_{KL}}$ denotes Kullback-Leibler divergence. An additional conceptual classification of surprise definitions (rows) [11]: 1. Prediction surprise defines surprising events as those that violate predictions. 2. Change-detection surprise quantifies possibility of changes in θ_{t+1} and defines surprising events as those predicted inaccurately *in comparison* with an alternative predictive model $p^{(\mathrm{alt.})}$ $(y_{t+1}|x_{t+1})$ [33]. 3. Information-gain surprise defines surprising events as those that change a subject's belief. 4. Confidence-corrected surprise adds an explicit measure of confidence into a definition of surprise, e.g., Shannon surprise plus a measure of confidence; note that the categorisation as probabilistic or belief-mismatch also depends on the definition of confidence. While the two classifications are complementary, they are not fully independent: One needs $\pi^{(t)}$ to evaluate an information-gain surprise, and it is not possible to define confidence without access to $\pi^{(t)}$ or $p^{(t)}$ (hatched boxes). Question marks: Categories without any example in the literature. See [11] for a detailed mathematical treatment of different definitions, their placement in the categories, and their relationships.

surprise and novelty based on their relation to unexpectedness and familiarity: Surprising stimuli violate expectations; hence, *surprise is a measure of the unexpectedness* of y_{t+1} according to the predictive model $p^{(t)}$ $(y_{t+1}|x_{t+1})$. Novel stimuli, however, violate familiarity; hence, *novelty is a measure of the unfamiliarity* of y_{t+1} according to the familiarity $p_f^{(t)}(y_{t+1})$ (Box 1 and Figure 2). The familiarity $p_f^{(t)}(y_{t+1})$ quantifies how frequent y_{t+1} (e.g., a specific image) has been up to time t independently of the cue x_{t+1} and potential regularities in observations (see [40] for similar ideas in machine learning). For example, in cases where x_{t+1} and y_{t+1} take discrete values (same assumption as in Equation (1)), one can define familiarity as the observation frequency

$$p_f^{(t)}(y_{t+1}) = \frac{C^{(t)}(y_{t+1}) + \text{constant}}{t + \text{constant}}, \tag{2}$$

where $C^{(t)}$ (y_{t+1}) is the count of how many times a subject has observed y_{t+1} until time t, and constants are added to avoid having zero frequencies. Novelty of observation y_{t+1} defined as $n_{t+1} = -\log\ p_f^{(t)}\ (y_{t+1})$ ('frequency-based novelty'; Figure 2b) explains some significant trial-by-trial variabilities of human EEG signals [16]. More generally, novelty of y_{t+1} can be defined as

$$n_{t+1} = N^{(t)}(y_{t+1}), \tag{3}$$

where $N^{(t)}$ is a general function that (i) takes y_{t+1} as its argument, (ii) is *independent* of the cue x_{t+1}, and (iii) depends on the subject's current internal state at time t (Figure 1b2).

The central criterion proposed by Xu et al. is that definitions of surprise quantify the *unexpectedness* of y_{t+1} and must be *conditioned* on x_{t+1}, whereas definitions of novelty quantify the *unfamiliarity* of y_{t+1} and must be *independent* of x_{t+1}. Almost all existing definitions of novelty in neuroscience and psychology meet this criterion and can be written as in Equation (3) [5,10]. For example, two alternative approaches to defining novelty are to (i) consider only the first encounter of a specific observation as novel ('absolute novelty'; Figure 2b) [41,42] or (ii) define the novelty of y_{t+1} as a decreasing function of the count $C^{(t)}$ (y_{t+1}) ('always decreasing novelty'; Figure 2b) [43,44]. Note that according to novelty definitions based on observation frequency (e.g., Equation (2)), the novelty of the observation y_{t+1} increases if it has not been observed for some time.

The distinction proposed by Xu et al. enables new interpretations of earlier results: For example, the separate MEG signatures found by [13] for 'frequency-based' and 'transition-based' surprise can alternatively be interpreted as separate signatures for novelty and surprise, respectively; what has been called 'expected surprise' by [45] can be seen as novelty; and what has been called 'contextual novelty' in neuroscience [5] is a form of surprise and not novelty. These interpretations help connect otherwise separate experimental phenomena in a single coherent framework.

Finally, the perceived novelty of a stimulus does not only depend on how often the exact same stimulus has been experienced. For example, a familiar image with an altered contrast level is a novel stimulus per se, but it may be perceived as a familiar one if the subject cares only about the image identity [46]; similarly, some novel stimuli may be perceived less novel than others if they look similar to familiar stimuli. Many experimental studies support such feature-dependency in novelty responses in the brain [9,22,47]. Novelty definitions based on the simple observation frequency in Equation (2) can be generalised to account for feature-dependent novelty estimation as the familiarity measure $p_f^{(t)}$ (y_{t+1}) can be an arbitrary (non-negative and normalised) function of the stimulus. Analogously, count-based novelty definitions can account for feature-dependent novelty estimation by turning to frequency-based pseudo-counts [40,48].

A taxonomy of surprise definitions

Surprise is caused by a violation of expectations. However, even if we agree that surprise quantifies the unexpectedness of y_{t+1} conditioned on x_{t+1}, there are multiple possibilities for quantifying unexpectedness [10,12,31–33,49–51]. In general, surprise of y_{t+1} can be written as

$$s_{t+1} = S^{(t)}(y_{t+1}|x_{t+1}), \tag{4}$$

where $S^{(t)}$ is a general function that (i) takes both y_{t+1} and x_{t+1} as arguments (in contrast to Equation (3)) and (ii) depends on the subject's current internal state at time t (Figure 1b2) [11]. A recent systematic taxonomy of commonly used definitions of surprise proposes two classification schemes for these definitions [52] (Figure 2c).

The first classification is based on the minimal information, about the subject's internal state, that is needed for computing surprise with a given definition (columns in Figure 2c): 1. *Observation-mismatch* surprise is defined based on the assumption that, at each time t, an experimental subject makes a prediction $\hat{y}_{t+1}$ of the upcoming observation y_{t+1}. Observation-mismatch surprise quantifies surprise as a mismatch between y_{t+1} and $\hat{y}_{t+1}$; an example is the absolute difference $s_{t+1} = |y_{t+1} - \hat{y}_{t+1}|$, where $\hat{y}_{t+1}$ is, e.g., the mean of the predictive distribution [53]. 2. *Probabilistic mismatch* surprise depends on the full distribution $p^{(t)}$ $(y_{t+1}|x_{t+1})$ of possible outcomes and, hence, requires more information than a single prediction $\hat{y}_{t+1}$; an example is the Shannon surprise or surprisal $s_{t+1} = -\log p^{(t)}$ $(y_{t+1}|x_{t+1})$ [10]. 3. *Belief-mismatch* surprise can be evaluated only by having access to the full belief $\pi^{(t)}$ (θ_{t+1}) about the hidden parameter set θ_{t+1} and requires even more information than the full distribution $p^{(t)}$ $(y_{t+1}|x_{t+1})$; an example is the Bayesian surprise $s_{t+1} = \mathrm{D_{KL}}(\pi^{(t)}, \pi^{(t+1)})$, where $\mathrm{D_{KL}}$ denotes Kullback-Leibler divergence [31,32].

The second classification is a conceptual one (rows in Figure 2c): 1. *Prediction* surprise defines surprising events as those that violate predictions, e.g., the Shannon surprise $s_{t+1} = -\log p^{(t)}$ $(y_{t+1}|x_{t+1})$. 2. *Change-detection* surprise also defines surprising events as those that violate predictions but only *in comparison* with an alternative predictive model; an example is the difference in the Shannon surprise $s_{t+1} = \log[p^{(t)}(y_{t+1}|x_{t+1}) \,/\, p^{(\text{alt.})}(y_{t+1}|x_{t+1})]$, where $p^{(\text{alt.})}$ $(y_{t+1}|x_{t+1})$ is a prior or naive predictive model [33]. According to change-detection surprise definitions, if the observation y_{t+1} is unlikely according to both the predictive model $p^{(t)}$ and its alternative, then it is not perceived as surprising. Hence, change-detection surprise can be interpreted as a measure of relative surprise. Importantly, change-detection surprise is optimal to modulate learning in volatile environments [11,33] (Box 1), in agreement with experimental observations [19, 24, 63].

3. *Information-gain* surprise defines surprising events as those that change a subject's belief about the world, e.g., the Bayesian surprise $s_{t+1} = \mathrm{D}_{\mathrm{KL}}(\pi^{(t)}, \pi^{(t+1)})$. We note, however, that only a handful of information-gain measures [64] have been previously interpreted as measures of surprise [12,31,32]. 4. *Confidence-corrected* surprise is defined based on the argument that a given error in prediction should feel more surprising if it is made with higher confidence (Box 1); examples have been suggested both in neuroscience [50] and psychology [51].

The two classifications together propose a refined terminology necessary for a systematic study of surprise in the brain. The first classification is important to judge whether surprise computation based on different definitions can be biologically plausible. For example, evaluating observation-mismatch surprise in a recurrent network of spiking neurons might be simpler than evaluating probabilistic mismatch and belief-mismatch surprise under similar biological constraints [35,36,38]; see [65,66] for different views on the neural implementation of probabilistic inference. The first classification can thus help studies to bridge the gap between algorithmic and mechanistic neural models of 'surprise-driven' attention [67], exploration [68], and learning [28].

The second classification is important as it suggests that observations that *intuitively* feel surprising can do so because of different aspects of surprise. Importantly, experimental studies of surprise have found separate neural signatures for different definitions (Table 1). For example, Gijsen et al. (2021) [14] found independent EEG signatures of prediction, information-gain, and confidence-corrected surprise in an experimental paradigm using somatosensory roving stimuli. Similarly, Kolossa et al. (2015) [12] showed in an earlier study that even different definitions in the same surprise category (e.g., information-gain surprise) can have different neural signatures. These results suggest that the experimental phenomena previously attributed to a single broad notion of 'surprise' might relate to very different but precise definitions of surprise.

The proposed taxonomy can also provide new interpretations of existing experiments: Beyond the comparison of trial types (e.g., expected versus unexpected trials), mathematical definitions of surprise and novelty enable trial-by-trial data analysis (Table 1). For example, Zhang et al. (2022) [9] observe in monkeys that neural responses to an unexpected stimulus are different depending on whether the stimulus appears in a random unpredictable sequence or in a regular predictable sequence (Box 1). The observed difference may be an indication that surprise signals in different brain areas relate to different surprise categories rather than a single notion of surprise. Such a hypothesis can be tested by trial-by-trial data analysis combined with computational modeling.

Finally, surprise can also quantify the unexpectedness of a scalar (or low dimensional) summary signal extracted from the (high dimensional) observation y_{t+1} instead of y_{t+1} itself. For example, the unsigned reward prediction error (uRPE) [69,70] measures the mismatch between the reward $r(y_{t+1})$ associated with stimulus y_{t+1} and a prediction $\widehat{r}_{t+1}$ thereof (see [11]). Similarly, an unsigned novelty prediction error (uNPE) measures the unexpectedness of the novelty value $N^{(t)}$ (y_{t+1}) of an observation y_{t+1} [16,71]. We can think of uRPE and uNPE as secondary surprise signals since they are derived from a scalar summary signal. When interpreting neural responses to 'novel' stimuli, it is hence important to consider that responses correlated with novelty may in fact be caused by errors in novelty prediction [16,71]. Moreover, subjects may assume potential associations between novelty (or similarly between surprise) and threats or rewards [43,60], which can lead to confounding effects of threats and rewards on neural responses to novelty (or surprise); hence, ideal experimental paradigms for studying neural and behavioral signatures of novelty and surprise require a dissociation of these signals from threats and rewards.

In addition, there can be multiple forms of neural responses to surprise and novelty of an abstract observation y_{t+1} depending on how it is neurally represented regarding, for example, sensory modality (e.g., auditory versus visual [59]) or the hierarchy of representations (e.g., image identity [16,46] versus primary visual features [2,22]). For example, a repeating sequence of binary observation as in Figure 2a can be presented as either a sequence of tones or a sequence of images (i.e., different modalities); Grundei et al. (2023) [59] found separate modality-specific and modality-independent EEG signatures of surprise in an experimental paradigm using somatosensory, auditory, and visual roving stimuli. Moreover, a sequence of images could consist of meaningless fractals, sketches of meaningful objects, or different visual drawing styles of always the same object, which results in the same temporal sequence of stimuli in the visual domain but at different levels of abstraction.

Towards a systematic study of surprise and novelty

Different computational roles in learning [34,72] and decision-making [73–75], broadly attributed to 'surprise' and 'novelty, may correspond to different but mathematically precise definitions of novelty and surprise and ultimately also to distinct physiological signals. This leaves us with two main questions: 1. How many fundamentally distinct physiological signals are involved in brain computations related to surprise and novelty? 2.

Table 1

Example experimental papers with more than one signal related to surprise and novelty. 'T-by-T' indicates whether trial-by-trial data analysis is performed. 'Compared signals' lists precise mathematical definitions (for trial-by-trial analysis) or the description of trial types (otherwise) that are compared. Animal studies with trial-by-trial analysis exist (e.g., [**54**,**55**]) but none with more than one definition of surprise or novelty. *Abbrevations:* CI: Calcium Imaging. Conf.: Confidence. Cort.: Cortex. DA: Dopamine. EEG: Electroencephalography. EP: Electrophysiology. Exp.: Expected. fMRI: Functional Magnetic Resonance Imaging. FP: Fiber Photometry. MEG: Magnetoencephalography. OG: Optogenetic. Seq.: Sequence. Unexp.: Unexpected.

	T-by-T	Compared signals	Subjects	Stimulus modality	Measurements
Macedo et al. (2004) [51]	Yes	1. Six definitions of prediction surprise 2. Two definitions Conf.-Corrected surprise	Humans	Questionnaire	1. Self-report
O'Reilly et al. (2013) [56]	Yes	1. Shannon surprise 2. Bayesian surprise	Humans	Visual	1. fMRI 2. Pupillometry 3. Reaction time
Kolossa et al. (2015) [12]	Yes	1. Shannon surprise 2. Bayesian surprise 3. Postdictive surprise	Humans	Visual	1. EEG
Maheu et al. (2019) [13]	Yes	1. Shannon surprise 2. Frequency-based novelty	Humans	Auditory	1. MEG
Visalli et al. (2019 & 2021 & 2023) [15,57,58]	Yes	1. Shannon surprise 2. Bayesian surprise	Humans	Visual	1. EEG [15,58] 2. fMRI [57] 3. Reaction time
Dubey and Griffiths (2019) [44]	Yes	1. Information-gain 2. Always decreasing novelty	Humans	Questionnaire	1. Action choices
Xu et al. (2021) [16]	Yes	1. Shannon/Bayes Factor surprise 2. Frequency-based novelty	Humans	Visual	1. EEG 2. Action choices
Gijsen et al. (2021) [14] and Grundei et al. (2023) [59]	Yes	1. Shannon surprise 2. Bayesian surprise 3. Conf.-Corrected surprise	Humans	1. Somatosensory [14,59] 2. Auditory [59] 3. Visual [59]	1. EEG
Modirshanechi et al. (2023) [52]	Yes	1. Shannon surprise 2. Postdictive surprise 3. Frequency-based novelty	Humans	Visual	1. Action choices
Morrens et al. (2020) [4]	No	1. New stimuli in a random seq. 2. Rare stimuli in a random seq.	Mice	Olfactory	1. FP recording of DA 2. Breathing frequency
Zhang et al. (2022) [9]	No	1. Unexp. new stimuli 2. Unexp. familiar stimuli in a random seq. 3. Unexp. familiar stimuli in a regular seq.	Monkeys	Visual	1. EP in 22 brain areas 2. Pupillometry 3. Saccade latency
Homann et al. (2022) [22]	No	1. New stimuli in a regular seq. 2. Switch of stimuli in a regular seq.	Mice	Visual	1. CI in the visual cort.
Akiti et al. (2022) [60]	No	1. Novel objects 2. Familiar objects in unexp. context	Mice	Visual	1. OG recording of DA 2. Action choices
Garrett et al. (2023) [61] (see also [62])	No	1. Exp. new stimuli 2. Unexp. new stimuli 3. Unexp. familiar stimuli 4. Omission of exp. stimuli	Mice	Visual	1. CI in the visual cort. 2. Action choices

What is the role of each physiological signal in each brain function? Addressing these questions requires interactions of theory and experiments.

Recent years have seen an increasing interest in this line of research. For example, Akiti et al. (2021) [60] show that mice exhibit different behavioral patterns when inspecting novel versus surprising objects and that striatal dopamine release modulates the inspection of novel objects differently from the inspection of surprising ones. Dubey and Griffiths (2019) [44] show that seeking novelty and information-gain (i.e., two distinct curiosity-related behavioral patterns) can be considered special cases of seeking a single 'curiosity signal' that is 'optimal' for exploration and depends on experimental conditions. Another study on exploratory behavior, on the other hand, shows that novelty-driven algorithms explain the human search for rewarding states better than algorithms driven by prediction surprise or information-gain, even when novelty-seeking is suboptimal [52]. Similar approaches can be applied to studying the influence of different aspects of surprise and novelty on learning, memory, and attention.

In conclusion, different aspects of surprise and novelty can be captured and quantified by precise definitions and well-designed experiments. The classifications in Figure 2 offer a foundation for future experimental and theoretical studies on surprise and novelty.

Author contributions

A.M. Conceptualization, Methodology, Formal analysis, Writing- Original draft, Writing-Review and Editing., S.B. Conceptualization, Writing-Review and Editing., J.B. Conceptualization, Writing-Review and Editing. W.G. Conceptualization, Writing-Review and Editing, Supervision.

Declaration of competing interest

The authors declare no competing interests.

Data availability

No data was used for the research described in the article.

Acknowledgement

A.M. thanks Vasiliki Liakoni, Martin Barry, and Valentin Schmutz for useful discussions on related topics. This research was supported by the Swiss National Science Foundation No. 200020_207426.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Sinclair Alyssa H, Manalili Grace M, Brunec Iva K, Alison Adcock R, Barense Morgan D: **Prediction errors disrupt hippocampal representations and update episodic memories**. *Proc Natl Acad Sci USA* 2021, **118**, e2117625118, https://doi.org/10.1073/pnas.2117625118.

2. * Jordan Rebecca, Keller Georg B: **The locus coeruleus broadcasts prediction errors across the cortex to promote sensorimotor plasticity**. *Elife* 2023, **12**:RP85111, https://doi.org/10.7554/eLife.85111.1.

An experimental study of the influence of prediction surprise on sensorimotor plasticity in the cortex through modulation of activities in the mice locus coeruleus.

3. * Ogasawara Takaya, Sogukpinar Fatih, Zhang Kaining, Feng Yang-Yang, Pai Julia, Ahmad Jezzini, Monosov Ilya E: **A primate temporal cortex–zona incerta pathway for novelty seeking**. *Nat Neurosci* 2022, **25**, https://doi.org/10.1038/s41593-021-00950-1.

An experimental study of the neural circuitry involved in novelty-seeking in monkeys.

4. Morrens Joachim, Aydin Çağatay, Aliza Janse van Rensburg, , José Esquivelzeta Rabell, Haesler Sebastian: **Cue-evoked dopamine promotes conditioned responding during learning**. *Neuron* 2020, **106**:142–153.e7, https://doi.org/10.1016/j.neuron.2020.01.012. ISSN 0896-6273.

5. Schomaker Judith, Meeter Martijn: **Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition**. *Neurosci Biobehav Rev* 2015, **55**:268–279, https://doi.org/10.1016/j.neubiorev.2015.05.002. ISSN 0149-7634.

6. Reisenzein Rainer, Horstmann Gernot, Schützwohl Achim: **The cognitive-evolutionary model of surprise: a review of the evidence**. *Topics in Cognitive Science* 2019, **11**:50–74, https://doi.org/10.1111/tops.12292.

7. Ladosz Pawel, Weng Lilian, Kim Minwoo, Oh Hyondong: **Exploration in deep reinforcement learning: a survey**. *Inf Fusion* 2022, **85**, https://doi.org/10.1016/j.inffus.2022.03.003. ISSN 1566-2535.

8. Schützwohl Achim, Reisenzein Rainer: **Facial expressions in response to a highly surprising event exceeding the field of vision: a test of Darwin's theory of surprise**. *Evol Hum Behav* 2012, **33**:657–664, https://doi.org/10.1016/j.evolhumbehav.2012.04.003. ISSN 1090-5138.

9. ** Zhang Kaining, Bromberg-Martin Ethan S, Sogukpinar Fatih, Kocher Kim, Monosov Ilya E: **Surprise and recency in novelty detection in the primate brain**. *Curr Biol* 2022, **32**:2160–2173, https://doi.org/10.1016/j.cub.2022.03.064. ISSN 0960-9822.

A systematic study of neural responses to novelty, surprise, and recency in multiple brain areas of monkeys. They find distinct but correlated responses to these different signals.

10. Barto Andrew, Mirolli Marco, Baldassarre Gianluca: **Novelty or surprise?** *Front Psychol* 2013, **4**:907, https://doi.org/10.3389/fpsyg.2013.00907. ISSN 1664-1078.

11. ** Modirshanechi Alireza, Johanni Brea, Gerstner Wulfram: **A taxonomy of surprise definitions**. *J Math Psychol* 2022, **110**, 102712, https://doi.org/10.1016/j.jmp.2022.102712. ISSN 0022-2496.

A taxonomy of 18 different definitions of surprise used in neuroscience and psychology (as in Figure 2). The authors identify conditions under which different definitions are experimentally indistinguishable.

12. Kolossa Antonio, Kopp Bruno, Tim Fingscheidt: **A computational analysis of the neural bases of Bayesian inference**. *Neuroimage* 2015, **106**:222–237, https://doi.org/10.1016/j.neuroimage.2014.11.007. ISSN 1053-8119.

13. Maheu Maxime, Dehaene Stanislas, Meyniel Florent: **Brain signatures of a multiscale process of sequence learning in humans**. *Elife* 2019, **8**, e41541, https://doi.org/10.7554/eLife.41541.

14. ** Gijsen Sam, Grundei Miro, Lange Robert T, Ostwald Dirk, Blankenburg Felix: **Neural surprise in somatosensory Bayesian learning**. *PLoS Comput Biol* 2021, **17**:1–36, https://doi.org/10.1371/journal.pcbi.1008068.

A systematic study of the EEG signatures of the Shannon, Bayesian, and Confidence-Corrected surprise in a somatosensory roving-stimulus task. The authors find separate EEG signatures for all three surprise definitions.

15. ** Visalli Antonino, Capizzi Mariagrazia, Ambrosini Ettore, Kopp Bruno, Vallesi Antonino: **Electroencephalographic**

correlates of temporal Bayesian belief updating and surprise. *Neuroimage* 2021, **231**, 117867, https://doi.org/10.1016/j.neuroimage.2021.117867. ISSN 1053-8119.

A specifically designed reaction-time task that allows dissociating experimental predictions of the Shannon surprise from those of the Bayesian surprise. The authors find separate EEG signatures for both definitions.

16. ** Xu He A, Modirshanechi Alireza, Lehmann Marco P, Gerstner Wulfram, Herzog Michael H: **Novelty is not surprise: human exploratory and adaptive behavior in sequential decision-making**. *PLoS Comput Biol* 2021, **17**, https://doi.org/10.1371/journal.pcbi.1009070.

A formal distinction between surprise and novelty (as in Figure 2) and a study of their signatures in human behavior and EEG signals. The authors show that novelty drives exploration, whereas surprise modulates the rate of learning.

17. Mars Rogier B, Debener Stefan, Gladwin Thomas E, Harrison Lee M, Haggard Patrick, Rothwell John C, Bestmann Sven: **Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise**. *J Neurosci* 2008, **28**:12539–12545, https://doi.org/10.1523/JNEUROSCI.2925-08.2008.

18. Gläscher Jan, Daw Nathaniel, Dayan Peter, O'Doherty John P: **States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning**. *Neuron* 2010, **66**:585–595, https://doi.org/10.1016/j.neuron.2010.04.016. ISSN 0896-6273.

19. Nassar Matthew R, Rumsey Katherine M, Wilson Robert C, Parikh Kinjan, Heasly Benjamin, Gold Joshua I: **Rational regulation of learning dynamics by pupil-linked arousal systems**. *Nat Neurosci* 2012, **15**:1040–1046, https://doi.org/10.1038/nn.3130.

20. Ostwald Dirk, Spitzer Bernhard, Guggenmos Matthias, Schmidt Timo T, Kiebel Stefan J, Blankenburg Felix: **Evidence for neural encoding of Bayesian surprise in human somatosensation**. *Neuroimage* 2012, **62**:177–188, https://doi.org/10.1016/j.neuroimage.2012.04.050.

21. Fiser Aris, Mahringer David, Oyibo Hassana K, Petersen Anders V, Leinweber Marcus, Keller Georg B: **Experience-dependent spatial expectations in mouse visual cortex**. *Nat Neurosci* 2016, **19**:1658–1664, https://doi.org/10.1038/nn.4385.

22. ** Homann Jan, Koay Sue A, Chen Kevin S, Tank David W, Berry II Michael J: **Novel stimuli evoke excess activity in the mouse primary visual cortex**. *Proc Natl Acad Sci USA* 2022, **119**, e2108882119, https://doi.org/10.1073/pnas.2108882119.

An experimental study of neural responses to novel stimuli in the mouse primary visual cortex. The authors show that such responses are specific to novel stimuli and are not triggered by familiar stimuli that violate observation patterns.

23. * Antony James W, Hartshorne Thomas H, Pomeroy Ken, Gureckis Todd M, Hasson Uri, McDougle Samuel D, Norman Kenneth A: **Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing**. *Neuron* 2021, **109**:377–390.e7, https://doi.org/10.1016/j.neuron.2020.10.029. ISSN 0896-6273.

A study of how information-gain surprise influences memory segmentation as well as fMRI and pupillometry signals.

24. Behrens Timothy EJ, Woolrich Mark W, Walton Mark E, Rushworth Matthew FS: **Learning the value of information in an uncertain world**. *Nat Neurosci* 2007, **10**:1214–1221, https://doi.org/10.1038/nn1954.

25. * Findling Charles, Chopin Nicolas, Koechlin Etienne: **Imprecise neural computations as a source of adaptive behaviour in volatile environments**. *Nat Human Behav* 2021, **5**:99–112, https://doi.org/10.1038/s41562-020-00971-z.

A measure of information-gain surprise is interpreted as a modulator of 'computational noise' for learning in volatile environments. They show that such heuristics explain human behavior in different experimental conditions better than optimal algorithms derived separately for each condition.

26. Yu Angela J, Dayan Peter: **Uncertainty, neuromodulation, and attention**. *Neuron* 2005, **46**:681–692, https://doi.org/10.1016/j.neuron.2005.04.026. ISSN 0896-6273.

27. Friston Karl: **The free-energy principle: a unified brain theory?** *Nat Rev Neurosci* 2010, **11**:127–138, https://doi.org/10.1038/nrn2787.

28. Soltani Alireza, Izquierdo Alicia: **Adaptive learning under expected and unexpected uncertainty**. *Nat Rev Neurosci* 2019, **20**:635–644, https://doi.org/10.1038/s41583-019-0180-y.

29. Meyniel Florent, Maheu Maxime, Dehaene Stanislas: **Human inferences about sequences: a minimal transition probability model**. *PLoS Comput Biol* 2016, **12**:1–26, https://doi.org/10.1371/journal.pcbi.1005260.

30. Marr David: *Vision: a computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc.; 1982.

31. Baldi Pierre: *A computational theory of surprise*. Boston, MA: Springer US; 2002:1–25, https://doi.org/10.1007/978-1-4757-3585-7_1. ISBN 978-1-4757-3585-7.

32. Jürgen Schmidhuber: **Formal theory of creativity, fun, and intrinsic motivation (1990–2010)**. *IEEE Transactions on Autonomous Mental Development* 2010, **2**:230–247, https://doi.org/10.1109/TAMD.2010.2056368.

33. * Liakoni Vasiliki, Modirshanechi Alireza, Gerstner Wulfram, Johanni Brea: **Learning in volatile environments with the Bayes factor surprise**. *Neural Comput* 2021, **33**:1–72, https://doi.org/10.1162/neco_a_01352.

A computational study showing that modulation of the learning rate by change-detection surprise emerges naturally from Bayesian inference in volatile environments and appears in a majority of previously proposed algorithms for adaptive learning (e.g., [19,53,76,77]).

34. Piray Payam, Daw Nathaniel D: **A model for learning based on the joint estimation of stochasticity and volatility**. *Nat Commun* 2021, **12**:6587, https://doi.org/10.1038/s41467-021-26731-9.

35. Barry Martin LLR, Gerstner Wulfram: **Fast adaptation to rule switching using neuronal surprise**. *bioRxiv* 2022, https://doi.org/10.1101/2022.09.13.507727. 2022–09.

36. Iigaya Kiyohito: **Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system**. *Elife* 2016, **5**, e18073, https://doi.org/10.7554/eLife.18073.

37. Berlemont Kevin, Jean-Pierre Nadal: **Confidence-controlled hebbian learning efficiently extracts category membership from stimuli encoded in view of a categorization task**. *Neural Comput* 2022, **34**:45–77, https://doi.org/10.1162/neco_a_01452. ISSN 0899-7667.

38. Wilmes Katharina A, Petrovici Mihai A, Sachidhanandam Shankar, Senn Walter: **Uncertainty-modulated prediction errors in cortical microcircuits**. *bioRxiv* 2023, https://doi.org/10.1101/2023.05.11.540393.

39. Modirshanechi Alireza, Kiani Mohammad Mahdi, Hamid Aghajan: **Trial-by-trial surprise-decoding model for visual and auditory binary oddball tasks**. *Neuroimage* 2019, **196**:302–317, https://doi.org/10.1016/j.neuroimage.2019.04.028.

40. Bellemare Marc, Srinivasan Sriram, Ostrovski Georg, Tom Schaul, Saxton David, Munos Remi: **Unifying count-based exploration and intrinsic motivation**. In Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R. *Advances in neural information processing systems*, vol. 29. Curran Associates, Inc.; 2016.

41. Gershman Samuel J, Monfils Marie-H, Norman Kenneth A, Niv Yael: **The computational nature of memory modification**. *Elife* 2017, **6**, e23763, https://doi.org/10.7554/eLife.23763. ISSN 2050-084X.

42. Noel Xavier, Cleeremans Axel, Yu Angela J, Irene Cogliati Dezza: **Distinct motivations to seek out information in healthy individuals and problem gamblers**. *Transl Psychiatry* 2021, **11**: 408, https://doi.org/10.1038/s41398-021-01523-3.

43. Gershman Samuel J, Niv Yael: **Novelty and inductive generalization in human reinforcement learning**. *Topics in Cognitive Science* 2015, **7**:391–415, https://doi.org/10.1111/tops.12138.

44. Dubey Rachit, Griffiths Thomas L: **Reconciling novelty and complexity through a rational analysis of curiosity**. *Psychol Rev* 2019, **127**:455–476, https://doi.org/10.1037/rev0000175.

45. Lecaignard Françoise, Bertrand Olivier, Caclin Anne, Mattout Jérémie: **Neurocomputational underpinnings of expected surprise**. *J Neurosci* 2022, **42**:474–486, https://doi.org/10.1523/JNEUROSCI.0601-21.2021. ISSN 0270-6474.

46. ** Mehrpour Vahid, Meyer Travis, Simoncelli Eero P, Rust Nicole C: **Pinpointing the neural signatures of single-exposure visual recognition memory**. *Proc Natl Acad Sci USA* 2021, **118**, e2021660118, https://doi.org/10.1073/pnas.2021660118.
An experimental study of the neural activity in the monkey inferotemporal (IT) cortex during a single-exposure memory task. The authors show whether monkeys decide a stimulus is novel or not can be linearly decoded from the population activity in IT.

47. Meyer Travis, Rust Nicole C: **Single-exposure visual memory judgments are reflected in inferotemporal cortex**. *Elife* 2018, **7**, e32259, https://doi.org/10.7554/eLife.32259. ISSN 2050-084X.

48. Jaegle Andrew, Mehrpour Vahid, Rust Nicole: **Visual novelty, curiosity, and intrinsic reward in machine learning and the brain**. *Curr Opin Neurobiol* 2019, **58**:167–174, https://doi.org/10.1016/j.conb.2019.08.004.

49. Palm Günther: *Novelty, information and surprise*. Springer Science & Business Media; 2012.

50. Faraji Mohammadjavad, Preuschoff Kerstin, Gerstner Wulfram: **Balancing new against old information: the role of puzzlement surprise in learning**. *Neural Comput* 2018, **30**:34–83, https://doi.org/10.1162/neco_a_01025.

51. Macedo Luis, Reisezein Rainer, Cardoso Amilcar: **Modeling forms of surprise in artificial agents: empirical and theoretical study of surprise functions**. *Proceedings of the Annual Meeting of the Cognitive Science Society* 2004, **26**.

52. Modirshanechi Alireza, Lin Wei-Hsiang, Xu He A, Herzog Michael H, Gerstner Wulfram: **The curse of optimism: a persistent distraction by novelty**. *bioRxiv* 2023, https://doi.org/10.1101/2022.07.05.498835.

53. Nassar Matthew R, Wilson Robert C, Heasly Benjamin, Gold Joshua I: **An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment**. *J Neurosci* 2010, **30**:12366–12378, https://doi.org/10.1523/JNEUROSCI.0822-10.2010.

54. English Gwendolyn, Ghasemi Nejad Newsha, Sommerfelt Marcel, Fatih Yanik Mehmet, von der Behrens Wolfger: **Bayesian surprise shapes neural responses in somatosensory cortical circuits**. *Cell Rep* 2023, **42**, 112009, https://doi.org/10.1016/j.celrep.2023.112009. ISSN 2211-1247.

55. Rubin Jonathan, Ulanovsky Nachum, Nelken Israel, Tishby Naftali: **The representation of prediction error in auditory cortex**. *PLoS Comput Biol* 2016, **12**, e1005058, https://doi.org/10.1371/journal.pcbi.1005058.

56. O'Reilly Jill X, Schüffelgen Urs, Cuell Steven F, Behrens Timothy EJ, Mars Rogier B, Rushworth Matthew FS: **Dissociable effects of surprise and model update in parietal and anterior cingulate cortex**. *Proc Natl Acad Sci USA* 2013, **110**:E3660–E3669, https://doi.org/10.1073/pnas.1305373110.

57. Visalli Antonino, Capizzi Mariagrazia, Ambrosini Ettore, Mazzonetto Ilaria, Vallesi Antonino: **Bayesian modeling of temporal expectations in the human brain**. *Neuroimage* 2019, **202**, 116097, https://doi.org/10.1016/j.neuroimage.2019.116097. ISSN 1053-8119.

58. Visalli Antonino, Capizzi Mariagrazia, Ambrosini Ettore, Kopp Bruno, Vallesi Antonino: **P3-like signatures of temporal predictions: a computational eeg study**. *Exp Brain Res* 2023, https://doi.org/10.1007/s00221-023-06656-z.

59. ** Grundei Miro, Schröder Pia, Gijsen Sam, Blankenburg Felix: **EEG mismatch responses in a multimodal roving stimulus paradigm provide evidence for probabilistic inference across audition, somatosensation, and vision**. *Hum Brain Mapp* 2023, **44**:3644–3668, https://doi.org/10.1002/hbm.26303.
A study of the EEG signatures of surprise in roving-stimulus tasks with visual, auditory, and somatosensory stimuli. The authors found modality-specific signatures localised mainly in the sensory cortices and modality-independent signatures localised mainly in the frontal cortex, the former occurring 100–200 ms and the latter 300–350 ms after the stimulus onset. Trial-by-trial analysis of data suggests potential links to Shannon, Bayesian, and Confidence-Corrected surprise.

60. ** Akiti Korleki, Tsutsui-Kimura Iku, Xie Yudi, Alexander Mathis, Markowitz Jeffrey E, Anyoha Rockwell, Robert Datta Sandeep, Mathis Mackenzie Weygandt, Uchida Naoshige, Watabe-Uchida Mitsuko: **Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction**. *Neuron* 2022, **110**:3789–3804.e9, https://doi.org/10.1016/j.neuron.2022.08.022. ISSN 0896-6273.
An experimental study of mice exploring novel and surprising objects. The authors observe different patterns of assessment and interaction with respect to novel versus surprising objects and show that dopamine release in the tail of the striatum during the assessment phase is a predictor of whether mice avoid or interact with novel objects.

61. Garrett Marina, Groblewski Peter, Piet Alex, Ollerenshaw Doug, Najafi Farzaneh, Yavorska Iryna, Adam Amster, Bennett Corbett, Buice Michael, Caldejon Shiella, *et al.*: **Stimulus novelty uncovers coding diversity in visual cortical circuits**. *bioRxiv* 2023, https://doi.org/10.1101/2023.02.14.528085. 2023–02.

62. Garrett Marina, Manavi Sahar, Roll Kate, Ollerenshaw Douglas R, Groblewski Peter A, Ponvert Nicholas D, Kiggins Justin T, Casal Linzy, Mace Kyla, Ali Williford, Arielle Leon, Jia Xiaoxuan, Ledochowitsch Peter, Buice Michael A, Wakeman Wayne, Mihalas Stefan, Olsen Shawn R: **Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells**. *Elife* 2020, **9**, e50340, https://doi.org/10.7554/eLife.50340. ISSN 2050-084X.

63. Pasturel Chloé, Montagnini Anna, Perrinet Laurent Udo: **Humans adapt their anticipatory eye movements to the volatility of visual motion properties**. *PLoS Comput Biol* 2020, **16**, https://doi.org/10.1371/journal.pcbi.1007438.

64. Nelson Jonathan D: **Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain**. *Psychol Rev* 2005, **112**:979–999, https://doi.org/10.1037/0033-295X.112.4.979.

65. Knill David C, Pouget Alexandre: **The Bayesian brain: the role of uncertainty in neural coding and computation**. *Trends Neurosci* 2004, **27**:712–719, https://doi.org/10.1016/j.tins.2004.10.007. ISSN 0166-2236.

66. Fiser József, Berkes Pietro, Orbán Gergő, Lengyel Máté: **Statistically optimal perception and learning: from behavior to neural representations**. *Trends Cognit Sci* 2010, **14**:119–130, https://doi.org/10.1016/j.tics.2010.01.003. ISSN 1364-6613.

67. Itti Laurent, Baldi Pierre: **Bayesian surprise attracts human attention**. *Vis Res* 2009, **49**:1295–1306, https://doi.org/10.1016/j.visres.2008.09.007. ISSN 0042-6989.

68. Gottlieb Jacqueline, Oudeyer Pierre-Yves: **Towards a neuroscience of active sampling and curiosity**. *Nat Rev Neurosci* 2018, **19**:758–770, https://doi.org/10.1038/s41583-018-0078-0.

69. Hayden Benjamin Y, Heilbronner Sarah R, Pearson John M, Platt Michael L: **Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior**. *J Neurosci* 2011, **31**:4178–4187, https://doi.org/10.1523/JNEUROSCI.4652-10.2011. ISSN 0270-6474.

70. * Rouhani Nina, Niv Yael: **Signed and unsigned reward prediction errors dynamically enhance learning and memory**. *Elife* 2021, **10**, e61077, https://doi.org/10.7554/eLife.61077.
An elaborated study of how signed and unsigned reward prediction errors enhance human memory in a contextual bandit task.

71. Kakade Sham, Dayan Peter: **Dopamine: generalization and bonuses**. *Neural Network* 2002, **15**:549–559, https://doi.org/10.1016/s0893-6080(02)00048-5.

72. John M Pearce, Geoffrey Hall: **A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli**. *Psychol Rev* 1980, **87**:532–552, https://doi.org/10.1037/0033-295X.87.6.532.

73. Berlyne Daniel E: **Novelty and curiosity as determinants of exploratory behaviour**. *British Journal of Psychology. General Section* 1950, **41**:68–80, https://doi.org/10.1111/j.2044-8295.1950.tb00262.x.

74. Horvath Lilla, Stanley Colcombe, Milham Michael, Ray Shruti, Schwartenbeck Philipp, Ostwald Dirk: **Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task**. *Computational Brain & Behavior* 2021, https://doi.org/10.1007/s42113-021-00112-3.

75. Schulz Eric, Gershman Samuel J: **The algorithmic architecture of exploration in the human brain**. *Curr Opin Neurobiol* 2019, **55**:7–14, https://doi.org/10.1016/j.conb.2018.11.003.

76. Fearnhead Paul, Liu Zhen: **On-line inference for multiple changepoint problems**. *J Roy Stat Soc B* 2007, **69**:589–605, https://doi.org/10.1111/j.1467-9868.2007.00601.x.

77. Adams Ryan Prescott, MacKay David JC: *Bayesian online changepoint detection*. 2007. arXiv preprint arXiv:0710.3742.