EPFL

Thèse n° 9716

Seeking the new, learning from the unexpected:
Computational models of surprise and novelty in the brain

Présentée le 26 avril 2024

Faculté informatique et communications
Laboratoire de calcul neuromimétique (IC/SV)
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Alireza MODIRSHANECHI

Acceptée sur proposition du jury

Dr O. Lévêque, président du jury
Prof. W. Gerstner, directeur de thèse
Prof. S. J. Gershman, rapporteur
Prof. J. Gold, rapporteur
Prof. M. Mathis, rapporteuse

École
polytechnique
fédérale
de Lausanne

2024

Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics.
— Galileo

Acknowledgements

I have lost count of times I felt so lucky to have Wulfram as my supervisor; I learned enormously from him, and I am deeply grateful for his invaluable trust, support, and criticism, for the enjoyable conversations, and for the fantastic Ph.D. experience. I would like to thank my jury members not only for judging my thesis but also for their groundbreaking works that have inspired me throughout my Ph.D. I would like to particularly thank the president of my jury, Olivier, for also being an amazing teacher and for our magical conversations about probability theory and Markov chains. I am thankful to Prof. Michael Herzog and Prof. Sebastian Haesler (for the fruitful collaborations), Prof. Emmanuel Abbe (for training me to communicate with biologists), Prof. Mats Stensrud (for introducing me to the beautiful world of causal inference), and Elke Hewitt (for her help and support in the last few years).

I feel so lucky that I had the chance to know Johanni and work with him; I am immensely grateful to him for teaching me to look for the right questions, for our phenomenal collaborations, for our genuine friendship, and for being a role model in light of 'das Gute, Wahre und Schöne'. My life and Ph.D. would not have been the same[1] without the precious friendship with Vassia; I cannot put into words my gratitude for all our memorable conversations, with laughter and tears (i.e., the 'highlights'), for her endless psychological support, and, of course, for our priceless collaboration. I am also profoundly thankful to my dear friends Valentin (especially for inspiring debates in Lausanne, Venice, and London), Berfin (for Sunday brunches, reading clubs, joint complaining sessions, and being my Ph.D. twin), and Sophia (for the pe*r*fect friendship and the wonde*r*ful collaboration).

I am genuinely grateful to all my incredible friends and colleagues in LCN (ordered by the time of their appearance in my life): Samuel (for pushing me out of my comfort zone and, of course, for the apartment!), Bernd (for the lovely dinners and the constant encouragement to bike and ski), Marco (for the joy and laughter that he brings with himself), Chiara, Martin (for our enjoyable arguments, random stories, and everyday laughs), Florian (for the sailing experience), Olivia, Noé, Christos and Flavio (for 'High Three' and for the cheerful chats, in front of a board or over drinks), Georgios, Guillaume (for never-ending scientific debates), Shuqi (for the lovely friendship and her wise words after long pauses), Kai, Louis, Ariane, Lucas (especially for revising my thesis' French abstract), and Kasper.

[1] Clarification: They would have been worse!

I am also genuinely grateful to all my incredible friends *outside* LCN (ordered by the time of their appearance in my life): Vahid (for coffees in Lausanne and gin & tonics in Tenerife), Dey (for being 'Awesome'), Romane, Marie (especially for making me like cats), Stella (for the 'Yogang', tea times, and the safe space of our friendship), Heather (for always being cheerful), and Elena. An additional special thanks goes to Laura for making my life happier and easier.

I am deeply thankful to my Iranian friends who proved that true friendships are immune to physical distance (ordered by some arbitrary categories): Siavash (for his endless curiosity and our lifelong friendship), Ali (a.k.a. Abbas) and Bahman (for the late-night drives), Mohammad (for the cheerful calls and his refreshing honesty), Kian (for the Sunday video calls), Safa (for the chill times and parties in Zurich), Elias, 'Former Room 318' as a whole, Parnian, Navid, and Setayesh.

Finally, I cannot thank my parents enough, for their unconditional love, for being incredibly selfless, caring, and supportive, for the quality times when we are under the same roof, and for our joyful and honest calls when we are far away; thank you! I would also like to thank my dearest sister, Niloufar (for still loving me despite the experience of growing up with me; I love you too!), and many members of my huge extended family, in particular (ordered by approximate age): My grandparents, Amoo Joon Hassan Agha, Amme Simin (for being the cool aunt), Maryam and Amirali (for being like role models), Daee Javad (for being like an older brother), Shadi, Ali (for being like a twin brother), Negar, Naghme, and Erwan.

Lausanne, March 11, 2024 Alireza

Abstract

Human babies have a natural desire to interact with new toys and objects, through which they learn how the world around them works, e.g., that glass shatters when dropped, but a rubber ball does not. When their predictions are proven incorrect, such as when a glass does not shatter after a fall, they feel surprised. This, in turn, impacts their subsequent decisions and makes them reconsider their beliefs, e.g., they may continue dropping the glass until they realize it does not shatter because it falls on a carpet. Similarly, human adults and other species react differently to new and surprising events compared to familiar and expected ones, possibly due to the vital importance of these events in a continuously changing world with sparse resources. The influence of novelty and surprise on the brain and behavior has been a prominent topic in neuroscience and psychology. However, quantifying surprise and novelty and their contribution to various brain functions remain unresolved and disputed.

In this thesis, I take a mathematical approach to study (i) definitions of surprise and novelty as well as (ii) their computational roles in the brain. I first present an exhaustive analysis of 18 mathematical definitions of surprise, investigating their similarities, differences, and conditions that make them indistinguishable. I classify these definitions into different categories and propose a unified framework for systematic comparison of different approaches to quantifying surprise. Within this framework, I propose a formalism that distinguishes novelty from surprise. I use this mathematical distinction to construct a Reinforcement Learning (RL) model of human behavior that describes surprise as the modulator of the learning speed ('learning from the unexpected') and novelty as the drive of goal-directed exploration ('seeking the new'). I test this model against behavioral and electroencephalogram (EEG) data of human participants and show that both surprise and novelty are crucial determinants of human behavior in volatile environments with sparse rewards. Then, I ask whether these results generalize to stochastic environments where novelty-driven exploration has proven suboptimal. To answer this question, I compare models of exploration driven by novelty and different surprise definitions in stochastic environments. Testing these models against the behavioral data of human participants shows that human exploration closely aligns with novelty-driven models, even when they are not optimal. This establishes novelty as a dominant drive of human goal-directed exploration.

This thesis offers a comprehensive comparison of various computational models and definitions of surprise and novelty, from both mathematical and experimental points of view. Our theoretical findings allow fresh insights into previous research and lay a foundation for future theoretical and experimental studies. Moreover, our computational modeling of experimental data expands our understanding of the computational roles of surprise and novelty in learning and exploration.

Keywords: Surprise, Novelty, Information Gain, Adaptive Learning, Exploration, Computational Models, Reinforcement Learning, Human Behavior, EEG

Résumé

Les bébés humains manifestent naturellement un intérêt pour l'interaction avec de nouveaux jouets et objets, contribuant à leur apprentissage du monde. Par exemple, ils apprennent que le verre se brise lors d'une chute, contrairement à une balle en caoutchouc. En cas d'incohérence, comme un verre qui ne se brise pas, les bébés manifestent de la surprise, répétant l'expérience jusqu'à comprendre qu'un tapis le prévient de se casser. De même, les humains adultes et d'autres espèces réagissent différemment aux événements nouveaux et surprenants par rapport aux événements familiers et attendus, peut-être en raison de leur importance dans un monde en constante évolution et où les ressources sont rares. Bien que l'effet de la nouveauté et de la surprise sur le cerveau et le comportement ait été largement exploré en neurosciences et en psychologie au fil des décennies, la mesure précise de ces phénomènes et leur impact sur différentes fonctions cérébrales demeurent des questions non résolues et contestées.

Dans cette thèse, j'adopte une approche mathématique pour étudier (i) les définitions de la surprise et de la nouveauté ainsi que (ii) leurs rôles computationnels dans le cerveau. Je présente tout d'abord une analyse mathématique complète de 18 définitions distinctes de la surprise, en étudiant leurs similitudes, leurs différences et les conditions qui les rendent indiscernables. Je catégorise ces définitions et crée un cadre unifié facilitant la comparaison systématique des diverses méthodes de quantification de la surprise. Dans ce cadre, je propose un formalisme qui distingue la nouveauté de la surprise. J'utilise cette distinction mathématique pour construire un modèle d'apprentissage par renforcement du comportement humain, décrivant la surprise comme le modulateur de la vitesse d'apprentissage ('apprendre de l'inattendu') et la nouveauté comme le moteur de l'exploration orientée vers un but ('chercher la nouveauté'). Je teste ce modèle à l'aide de données comportementales et d'électroencéphalogrammes (EEG) de participants humains et démontre que la surprise et la nouveauté sont des déterminants cruciaux du comportement humain dans des environnements volatils et où les récompenses sont rares. Ensuite, j'explore l'applicabilité de ces résultats aux environnements stochastiques où l'exploration guidée par la nouveauté s'est avérée sous-optimale. Pour cela, j'analyse des modèles d'exploration guidés par la nouveauté et différentes définitions de la surprise. La comparaison de ces modèles aux données comportementales des participants humains montre que l'exploration humaine s'aligne étroitement sur les modèles axés sur la nouveauté, même lorsqu'ils ne sont pas optimaux. Cela établit que la nouveauté est un moteur dominant de l'exploration humaine orientée vers un but.

Résumé

Cette thèse propose une comparaison complète de divers modèles informatiques et définitions de la surprise et de la nouveauté, d'un point de vue à la fois mathématique et expérimental. Nos résultats théoriques apportent un nouvel éclairage sur les recherches précédentes et posent les fondements de futures études théoriques et expérimentales. Notre modélisation informatique des données expérimentales élargit notre compréhension des rôles de la surprise et de la nouveauté dans l'apprentissage et l'exploration.

Mots-clés : Surprise, Nouveauté, Apprentissage adaptatif, Exploration, Modèles informatiques, Apprentissage par renforcement, Comportement humain, EEG

Contents

1 Introduction

When a movie's villain turns out to be the warm-hearted hero, when a presidential candidate, who was considered by our friends to surely lose, suddenly wins the election, or when we see the image of our old friend's face suddenly distorted through the new fisheye lens of our camera, we feel something special. Later we may want to watch other movies by the same director, we may rethink how limited our social bubble is, or we may keenly watch our friends through our new camera lens.

We probably all agree that our special feelings at such moments are due to how unexpected (surprising) or new (novel) our experiences at those moments are. Although we use the same set of adjectives to describe these moments, one may wonder if there is any 'objective' commonality between these experiences beyond introspection. Specifically, how similar are the measurable reactions of our brain to the movie's plot twist and to the election news? Through what mechanisms do these events influence our future behavior? And is it possible to predict our measurable reactions to future situations when we experience something 'similar'?

Following common practices in sensory and motor neuroscience (Fyhn et al., 2004; Henry et al., 1974; Hubel and Wiesel, 1968; O'Keefe and Dostrovsky, 1971), neuroscientists approached these questions by proposing 'experimental' definitions of the 'new' and 'unexpected'; e.g., since a rarely encountered stimulus can *intuitively* be called 'new' or 'unexpected', one may consider the occurrence probability of a stimulus as an indicator of its surprise and novelty (e.g., Duncan-Johnson and Donchin (1977); Figure 1.1). Hence, analogously to earlier research on the neuroscience of vision that studied the links between the physical features of visual stimuli and their corresponding neural signals (Figure 1.1A), neuroscientists attempted to understand surprise and novelty in the brain by studying the links between the occurrence probability of stimuli and their corresponding neural signals (Hershenhoren et al., 2014; Näätänen et al., 2007; Näätänen and Picton, 1987; Tiitinen et al., 1994; Tueting et al., 1970; Ulanovsky et al., 2003) (Figure 1.1B).

The premise of such studies is that experimental subjects infer the statistical properties

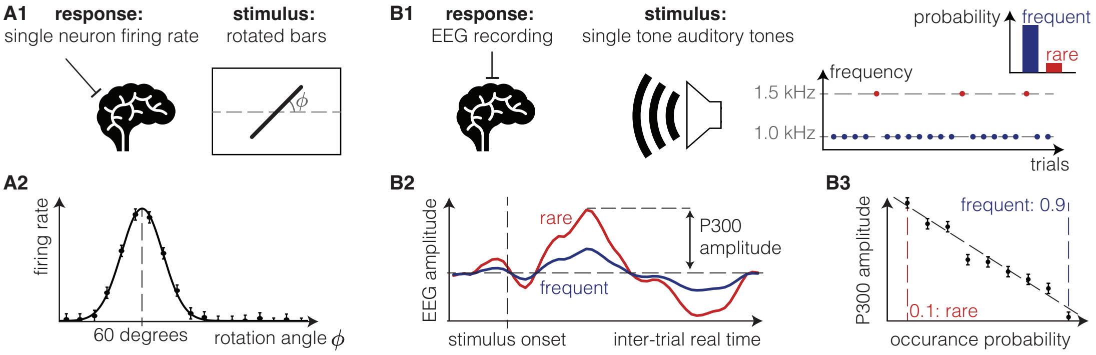


Figure 1.1: **Studies on surprise and novelty started with the same approach as the early studies of vision in neuroscience. A.** Schematic of a simplified version of the experimental paradigm used by Henry et al. (1974) to study the tuning curve of simple neurons in the cat's primary visual cortex. **A1.** Cats were presented with images of a single bar. As a response to the stimulus, the firing rate of single neurons in the primary visual cortex was recorded as a function of the rotation angle ϕ of the bar. **A2.** Toy illustration of a single neuron's tuning curve where the firing rate shows a bell-shaped response to ϕ. Such results were interpreted as evidence for some neurons working as edge-detectors with a particular orientation (in this case 60 degrees). **B.** Schematic of a simplified version of the experimental paradigm used by Duncan-Johnson and Donchin (1977) to study the surprise and novelty responses in auditory oddball tasks. **B1.** Human participants listened to a sequence of pure auditory tones with frequencies randomly sampled from 1.0 kHz (with 0.9 probability; frequent stimulus) and 1.5 kHz (with 0.1 probability; rare stimulus). As a response to the stimulus, the electroencephalogram (EEG) signals were recorded from the subjects' scalp. **B2.** Toy illustration of the average time-locked EEG signals (a.k.a. Event-Related Potential or ERP) for the frequent and rare stimuli. The signal amplitude at around 300 ms (P300 amplitude) after the stimulus was considered as the 'surprise' signal – as it took higher values for the rare than for the frequent stimuli. **B3.** The dependence of the P300 amplitude on the occurrence probability of stimuli can be seen as a tuning curve (similar to A2) for the P300 component. Such results were interpreted as evidence for some ERP components being signatures of surprise and novelty.

of their observations and accordingly assign a *subjective* surprise or novelty value to each stimulus. Because such subjective values cannot be measured by a third person, a statistical property of stimuli such as their occurrence probabilities may be considered as an approximation of these subjective values (averaged over many repetitions). Numerous alternative approaches have been proposed to define the 'new' and 'unexpected' in terms of experimental variables and operations (see section 1.1). Although this perspective has been successful in characterizing many of the neural and behavioral consequences of surprising and novel experiences (see section 1.2), it has remained unclear whether there is any common quantity that is measured by experimental variables that have served as intuitive approximations of subjective surprise and novelty values. This challenges the legitimacy of using the same terminology to describe and link different experimental phenomena believed to be related to 'surprise' or 'novelty'.

Computational models of learning and memory address this challenge by defining surprise and novelty in mathematical frameworks that can be adapted to specific experiments.

These models enable inferring the subjective surprise and novelty values of each observation, propose a common ground for different experimental approximations of these notions, and have significantly expanded our understanding of how surprise and novelty contribute to different brain functions (see section 1.3). However, there is currently no consensus on either the best modeling approach or the mathematical definitions of surprise and novelty in a given model (Baldi, 2002; Barto et al., 2013; Palm, 2012; Schmidhuber, 2010). The difficulty of a consensus is largely due to the field's focus on simplistic situations where many models and definitions are indistinguishable. Nevertheless, even subtle differences between different models and definitions result in diverging predictions in complex experimental paradigms (section 1.3). This highlights the need for more complex and theory-driven experimental paradigms that enable the comparison of different computational models and definitions of surprise and novelty.

This thesis aims (i) to provide a mathematical framework along with a systematic comparison of different definitions of surprise and novelty, to facilitate designing theory-driven experiments, and (ii) to study the computational roles of surprise and novelty in two examples of such experiments. In the rest of this introduction, I give an overview of experimental paradigms for studying surprise and novelty (section 1.1) and review related neural and behavioral studies with and without explicit use of computational models (section 1.2 and section 1.3). I finish the introduction with an overview of the thesis and how it contributes to addressing some of the open questions of the field.

1.1 How to experimentally define the new and unexpected

The general recipe for designing experiments on surprise and novelty involves two parts: (i) phases of learning and familiarization that induce expectations in experimental subjects (i.e., humans or animals) and (ii) observations that violate those expectations. In general, the observations that violate the induced expectations can be presented throughout the experiments and be even followed by further phases of learning and familiarization. In this section, I review and classify experimental paradigms into six different classes based on the type of expectations that the experimentalists *aim* to induce in the subjects throughout the phases of learning and familiarization. It is often unknown before the experiment whether experimental subjects will actually build the desired expectations; as we will discuss in section 1.2, this may remain unknown even after conducting the experiment. Toy examples of each category are illustrated in Figure 1.2.

1.1.1 Category 1. Stable observation distribution

The first category includes experimental paradigms where subjects are presented with a sequence of observations that are randomly and *independently* sampled from a *stable* distribution, where 'stable' implies that the observation statistics remain fixed throughout the experiment. The most famous example of this category in both human (Näätänen et al., 2007) and animal studies (Hershenhoren et al., 2014; Paller et al., 1988) is the binary oddball task where the observation sequence is composed of a rare and a frequent stimulus (see the auditory example in Figure 1.1B and the visual example in Figure 1.2A). In general, the observations do not need to be categorical and can, for example, be auditory tones with frequencies sampled from a Gaussian distribution (Garrido et al., 2013, 2016).

The rationale behind such experiments is that experimental subjects infer the stable distribution of observations and form expectations on which stimulus is more likely to be observed next. The unlikely stimuli are hence considered as 'unexpected'. Because the unlikely stimuli are encountered less often than the others, they are also less 'familiar' in general. It is hence impossible to dissociate surprise from novelty in these experiments without further modifications. This is reflected in the interchangeable use of related terms in oddball studies (Näätänen et al., 2007; Näätänen and Picton, 1987; Tiitinen et al., 1994; Tueting et al., 1970). One option to introduce surprising events that are not 'novel' is to omit stimuli with a fixed probability, i.e., with a fixed probability no stimulus is presented to the subjects (Yabe et al., 1997). If the omission probability is small and observations are presented with a regular timing, then the omission trials are 'unexpected' but cannot be called 'new' because no stimulus is presented to the subjects (see Heilbron and Chait (2018) for further discussions).

1.1.2 Category 2. Volatile observation distribution

The second category includes experimental paradigms where subjects are presented with a sequence of observations that are randomly and *independently* sampled from a distribution that abruptly *changes* at random points in time, i.e., the observation statistics are 'volatile' (Figure 1.2B). In general, the observation distribution can also *gradually* change between every two abrupt changes (Gershman et al., 2014). Similar to Category 1, observations can be either categorical (Foucault and Meyniel, 2023) or continuous (Glaze et al., 2018, 2015; Nassar et al., 2012, 2010). The rationale behind these experiments is similar to that of the experiments in Category 1, with the particular advantage that, if experimental subjects can rapidly update their inference about the observation distribution after each change point, then the momentarily unlikely stimuli are not necessarily the same as those encountered less frequently overall. Therefore, volatility in the observation distributions allows for dissociating between 'new' stimuli from 'unexpected' ones – and hence novelty from surprise.

1.1.3 Category 3. Never-seen-before stimuli

The third category includes experimental paradigms where subjects are first familiarized with a set of stimuli, often over days in animal experiments (i.e., training phase). They are then presented with stimuli that are chosen from a new and previously unobserved set, potentially combined with the familiar set (i.e., testing phase). For example, the training phase in an animal experiment may take several consecutive days of presenting random sequences of familiar stimuli (Figure 1.2C1). The testing phase can then be similar to an oddball task where the rare and frequent observations are chosen from the familiar set except for the occasional presentation of a new stimulus that was never presented during the training phase (Morrens et al. (2020); Figure 1.2C1).

During the testing phase, other experiments present pairs of familiar and new stimuli side-by-side (Ghazizadeh et al., 2020, 2016; Ogasawara et al., 2022; Sheldon, 1969) or allow free exploration of new objects in a familiar environment (Ahmadlou et al. (2021); Akiti et al. (2022); see Figure 1.2C2). The rationale is that differences between reactions to familiar and new stimuli must be related to their different novelty values. However, since completely new stimuli are always also unexpected, the influence of novelty and surprise on these reactions cannot be directly studied in these experiments. One indirect approach in experiments like the one in Figure 1.2C1 is to compare the reactions to the new stimuli with the reactions to the *rare* familiar stimuli that, after an extensive training phase, are unexpected but not new (Morrens et al., 2020).

1.1.4 Category 4. Between-observation associations

The fourth category includes experimental paradigms where between-observation associations can be used by subjects to predict the next observations. The most basic example is the Posner task (Posner, 1980) where experimental subjects are presented with a cue followed by a stimulus; the cue predicts, with some probabilities, which stimulus is presented next (Figure 1.2D1). A straightforward extension is a sequence of categorical observations where the distribution of the upcoming observation depends on the current observation – given by 'transition probabilities'. The transition probabilities can be either stable (Meyer and Olson, 2011) or volatile (Heilbron and Meyniel, 2019). Such experimental paradigms have been used in both human (Maheu et al., 2019; Meyniel, 2020; Todorovic and de Lange, 2012) and animal studies (Homann et al., 2022; Meyer and Olson, 2011; Zhang et al., 2022).

The category of between-observation associations also includes complex sequential patterns of categorical observations (see Barascud et al. (2016); Bekinschtein et al. (2009); Yaron et al. (2012) for some examples and Heilbron and Chait (2018) for a review) and associations of visual stimuli with location, e.g., a corridor's wallpaper (Fiser et al. (2016); see Figure 1.2D2). The rationale behind such experiments is that, with 'sufficient'

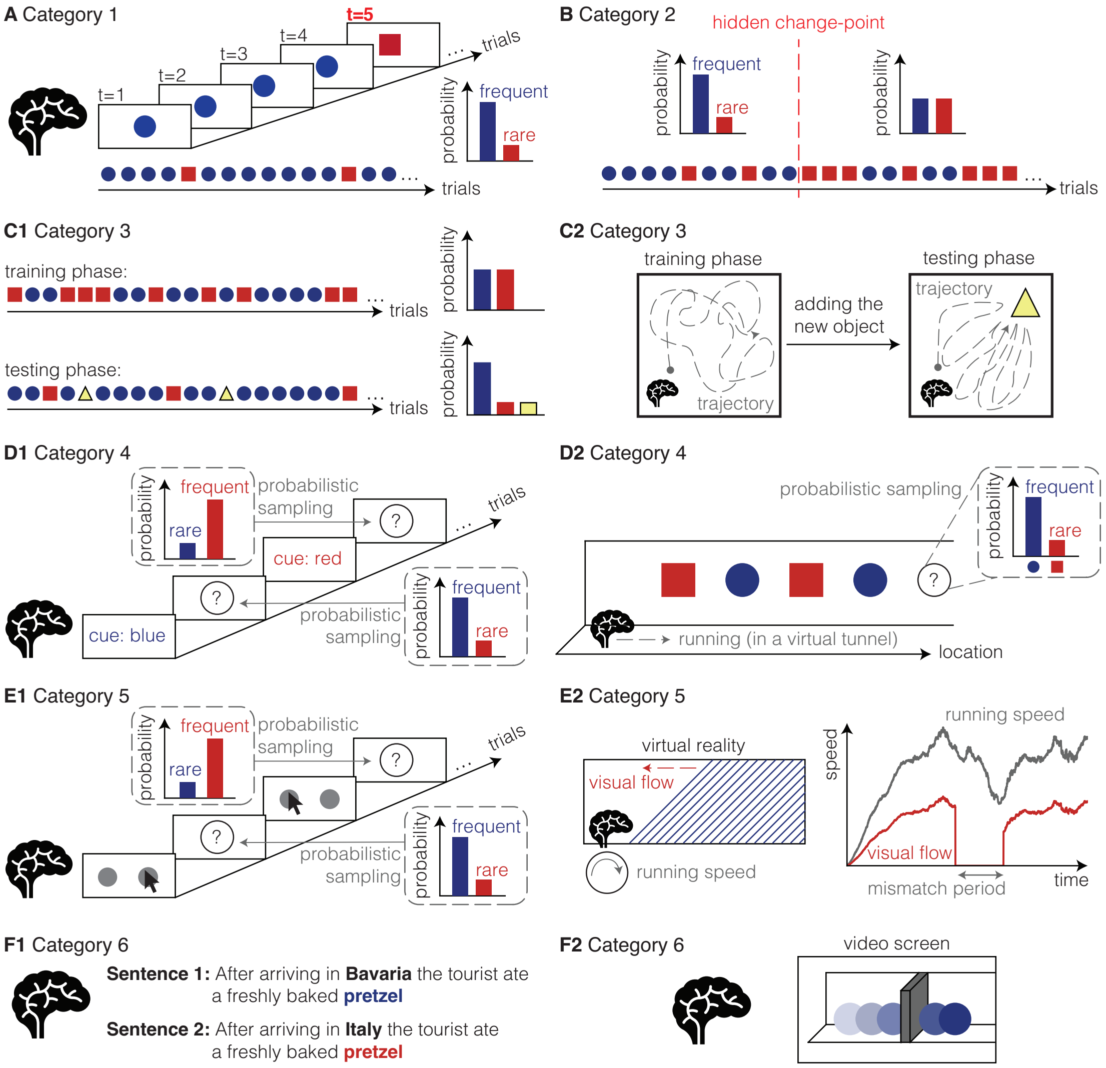


Figure 1.2: (Caption next page.)

experience, the experimental subjects infer these associations and accordingly use the current observations to predict the next. If the next observation is unlikely according to the current observations and the inferred associations, then it can be considered as ‘unexpected’ and ‘surprising’. An unlikely observation can be either an outlier or the result of a change in associations (Meyniel, 2020). It can also be due to a purely memory-based mismatch such as a sudden interruption of a video that experimental subjects previously watched during a familiarization phase (Sinclair and Barense, 2018; Sinclair et al., 2021). In principle, the correlation between surprise and novelty can be avoided by careful experimental design. Experiments in this category can combine never-seen-before stimuli (as in Category 3) with designs involving between-observation associations (e.g., see Zhang et al. (2022)).

Figure 1.2: **Toy illustration of some example experiments in each category reviewed in section 1.1. A.** A visual oddball task (similar to Figure 1.2B) as an example in Category 1: Stable observation distribution. **B.** A volatile oddball task as an example in Category 2: Volatile observation distribution. Subjects are presented with a sequence of visual stimuli as in A except that the observation distribution abruptly changes at random and unknown (to subjects) points in time (i.e., hidden change points). **C.** Category 3: Never-seen-before stimuli. **C1.** A simplified illustration of Morrens et al. (2020). Subjects are presented with a sequence of visual stimuli as in A. During the training phase, the observations are randomly sampled from two stimuli with equal probability. After days of training, the testing phase starts where observations are randomly sampled from the two known stimuli plus a never-seen-before stimulus (yellow triangle). **C2.** A simplified illustration of Ahmadlou et al. (2021); Akiti et al. (2022). During the training phase, mice freely explore an empty box to habituate to the environment. After a few days, a new object (yellow triangle) is added to the box. During the testing phase, the mice explore the box and the new object. **D.** Category 4: Between-observation associations. **D1.** An example Posner task (Posner, 1980). At each trial, subjects receive a cue predicting the next observation (blue or red). The next observation is sampled from a distribution determined by the cue, e.g., the cue's predictions are 90% correct. The subjects do not a priori know the association. **D2.** A simplified illustration of Fiser et al. (2016). Head-fixed mice run in a virtual tunnel. At each location in the tunnel, a particular visual stimulus appears on the wall (a blue circle or a red square). The first four visual stimuli are fixed over all trials – a trial ends when the mice arrive at the end of the tunnel. The fifth stimulus is randomly sampled and is more likely to be the blue circle. The mice do not a priori know the association. **E.** Category 5: Action-observation associations. **E1.** A reward-free two-armed bandit task. At each trial, subjects choose between two different actions (the mouse icon). The next observation is sampled from a distribution determined by the chosen action, e.g., the right action leads to a higher probability of the blue circle. The subjects do not a priori know the association. **E2.** A simplified illustration of Jordan and Keller (2023). Given a virtual reality setting, head-fixed mice run on a freely moving ball (left). On the screen, they see a tunnel with walls patterned with diagonal lines. The diagonal patterns move at a speed matching the running speed of the mice, as if they are running in the tunnel (right). At certain points in time, the visual bars are frozen, leading to a mismatch between the visual flow and running speed. **F.** Category 6: Core knowledge and common sense. **F1.** Example from Hodapp and Rabovsky (2021). Subjects are presented with two sentences that are only slightly different ('Bavaria' versus 'Italy'). The slight difference between the two sentences makes a common word ('pretzel') expected in one sentence (sentence 1) and unexpected in the other (sentence 2). **F2.** A simplified illustration of Stahl and Feigenson (2015). Subjects watch videos where intuitive physics is violated, e.g., a solid blue disk passes through a wall.

1.1.5 Category 5. Action-observation associations

The fifth category includes experimental paradigms where experimental subjects can predict the next observations by learning the consequences of their actions (i.e., associations between actions and observations). For example, in multi-armed bandit tasks, experimental subjects choose among some available actions and receive a reward as an outcome, sampled from a distribution that depends on the chosen action. These action-dependent reward distributions (i.e., action-observation associations) can be either stable or volatile but must be unknown to the subjects initially (Behrens et al., 2007; Findling et al., 2021; Gershman, 2019a; Horvath et al., 2021; Kao et al., 2020b; Li et al., 2019); see Figure 1.2E1 for a modified example.

The category of action-observation associations also includes action-dependent transitions between different observations[1] (Daw et al., 2011; Gläscher et al., 2010), associations between running speed and visual flow in a virtual reality task (Jordan and Keller (2023); O'Toole et al. (2023); see Figure 1.2E2), and associations between intended force and a joystick movement (Mathis et al., 2017). The rationale is the same as the one for experiments in Category 4 except that the experimental subjects have the choice to influence upcoming observations – and their expectations depend on their actions. Similarly to the experiments in Category 4, the correlation between surprise and novelty can be avoided in these experiments.

1.1.6 Category 6. Core knowledge and common sense

The sixth category includes experiments where expectations are assumed to be based on core knowledge and common sense (e.g., based on natural language or intuitive physics). Hence, these experiments bypass the familiarization and learning phases of the other categories and directly design observations that violate core knowledge and common sense. In an example from the language literature, experimental subjects read sentences in which some words are considered, by experimentalists, either 'expected' or 'unexpected' (Hodapp and Rabovsky (2021); see Figure 1.2F1). In another example from developmental psychology, babies watched videos where intuitive physics was violated, e.g., solid objects passed through each other (Stahl and Feigenson (2015); see Figure 1.2F2). Many other examples can be found in the psychology literature (see Schützwohl and Reisenzein (2012); Stone et al. (2023) for two examples and Reisenzein et al. (2019) for a review). These experiments are mainly focused on surprise and are the closest among the above-mentioned experiments to our real-life experiences of the 'unexpected'.

1.1.7 Summary

I proposed a classification of typical experimental paradigms on surprise and novelty into six categories based on the type of expectations that experimentalists either induce in experimental subjects (Categories 1-5) or assume because of core knowledge or common sense (Category 6). I reviewed some specific examples (Figure 1.2) and discussed the

[1] We use similar paradigms in chapter 3 and chapter 5.

complexity of associations in each category as well as their potential to dissociate surprise from novelty. I use the proposed classification in the next sections to review neural and behavioral signatures of surprise and novelty.

1.2 How far we can go without computational models

The focus of this section is on studies without any 'explicit' computational modeling. Specifically, I review studies that are based on statistical links between variables that are either (i) set by the experimentalists (e.g., the occurrence probability of stimuli) or (ii) 'directly' measurable in standard experiments (e.g., EEG amplitude and neural firing rate).

It is debated what it means for a variable to be 'directly' measurable, what is considered a standard experimental operation, and whether it is possible to make experimental statements without implicit assumptions about the underlying computations (Chang, 2021; Gershman, 2021). For example, consider statements based on estimated parameters of Generalized Linear Models (Walz et al., 2013) or based on the inferred 'learning rate' reflected in changes in the subject's self-reports (Nassar et al., 2010). In both cases, the variables can be defined independently of any computational models, but the logic of the definitions is essentially based on computational assumptions. Hence, the boundary between computational and non-computational studies is blurry.

To determine whether some statements are made based on explicit use of computational models, my conservative criterion in this section is whether the statements are based on variables that are extracted from *a model that could simulate an experimental subject throughout (an abstract version of) the experiment (as in Figure 1.3).* In the following subsections, we go through all six categories defined in section 1.1 and present examples of experimental results without any 'explicit' computational modeling.

1.2.1 Categories 1-2. Stable and volatile observation distributions

Oddball tasks have been extensively used to study the signatures of surprise and novelty among the Event-Related Potential (ERP) components (Duncan-Johnson and Donchin, 1977; Näätänen et al., 2007; Näätänen and Picton, 1987; Tiitinen et al., 1994; Tueting et al., 1970). ERP components are peaks and troughs of the stimulus-specific average of the EEG or MEG (magnetoencephalography) signal, time-locked to the onset of the stimulus presentation (e.g., the P300 peak visualized in Figure 1.1B2; see Luck (2014)). Both the amplitude and the latency of many of these components (e.g., N100 and P300) are correlated with how *infrequently* a stimulus is presented (Duncan-Johnson and Donchin, 1977; Näätänen et al., 2007), how *different* the rare and frequent stimuli are (Näätänen and Picton, 1987; Tiitinen et al., 1994), and in what *order* the stimuli are presented (SanMiguel et al., 2021; Squires et al., 1976). Similar correlations have been found for pupil dilation (Friedman et al., 1973; Qiyuan et al., 1985) and even single-neuron firing rates in animal studies (Hershenhoren et al., 2014; Ulanovsky et al., 2003).

The ERP components of EEG and MEG signals are also studied in tasks with stable observation distributions other than oddball tasks (Garrido et al., 2013, 2016). In a task with observations composed of auditory tones with frequencies sampled from a Gaussian distribution, unlikely frequencies resulted in a higher N100 amplitude (Garrido et al., 2016). Interestingly, the authors showed that this response increases with task-independent cognitive load (Garrido et al., 2016). The unequivocal message of the studies in Category 1 is that the brain treats the new and unexpected differently (e.g., with stronger responses) than the familiar and expected.

On the other hand, experiments with volatile observation distributions (Category 2) have been mainly used to study adaptive learning in humans. Subjects are often asked to report their prediction of the probability of the next categorical observation (e.g., as in Figure 1.2B1; Foucault and Meyniel (2023)) or the value of the next continuous observation (Nassar et al., 2012, 2010). The goal is to quantify how fast subjects update their predictions upon the presentation of unexpected observations following a change point. While most results in these experiments are based on computational models of learning in volatile environments (see section 1.3), the increased learning rate immediately after each change point is interpreted as evidence for surprise-modulation of the learning speed (Foucault and Meyniel, 2023; Nassar et al., 2012, 2010).

1.2.2 Category 3. Never-seen-before stimuli

Experiments with never-seen-before stimuli have been successful in characterizing novelty-driven behavior and neural circuitry involved in novelty processing. For example, mice have a higher breathing frequency when sniffing new odors than those already known (e.g., as in Figure 1.2C1; Morrens et al. (2020)) and show a higher desire for approaching and exploring novel objects than familiar ones (e.g., as in Figure 1.2C2; Ahmadlou et al. (2021); Sheldon (1969)). Similarly, monkeys show faster saccades to new fractals than to familiar ones (Ghazizadeh et al., 2020, 2016; Ogasawara et al., 2022). Hence, it is believed that animals show a general tendency to explore the new – see also Akiti et al. (2022) for different results and perspectives.

Physiological studies with these paradigms have been used to identify the differences and similarities of the neural circuits of novelty-seeking and reward-seeking. For example, Morrens et al. (2020) showed that dopamine neurons that respond to unexpected rewards also respond to new odors but not to rare familiar stimuli (similar to Figure 1.2C1). On the other hand, studies on novelty-seeking in monkeys revealed overlapping but parallel pathways for novelty- and reward-seeking (Ghazizadeh et al., 2020, 2016; Ogasawara et al., 2022). Ogasawara et al. (2022) identified the Zona Incerta (ZI) as a brain region whose inactivation impairs novelty-seeking but does not influence reward-seeking (see also Wang et al. (2022)). Parallel works on the neural circuitry of novelty detection show that the neural response in the monkey's Inferior Temporal (IT) cortex differs between the 1st and the 2nd exposure to a particular image, even if the low-level visual features of the image (e.g., contrast) are modified (Mehrpour et al., 2021). Collectively, these results imply that a reward-independent mechanism in the brain is responsible for detecting and exploring the new.

1.2.3 Categories 4-5. Between- and action-observation associations

Experiments with between-observation associations (Category 4) have found ERP responses to unexpected *transitions* similar but not identical to the ERP responses to infrequent stimuli in oddball tasks (Todorovic and de Lange, 2012). Similarly, single neurons in mice respond less strongly to rare stimuli in a periodic sequence of observations (i.e., when the rare stimulus is *predictable* due to between-observation associations; Category 4) than to rare stimuli in a random sequence (i.e., when the rare stimulus is *unpredictable* due to the independence of observations; Category 1) (Yaron et al., 2012). Moreover, introducing never-seen-before stimuli in experiments with between-observation associations has shown that neural responses to new stimuli overlap with but are not identical to neural responses to unexpected familiar stimuli (Garrett et al., 2023, 2020; Homann et al., 2022; Zhang et al., 2022). These results show that the brain treats surprise and novelty through related but not identical mechanisms.

Similar to studies with humans and mice, studies with monkeys have shown that, after extensive training in tasks where each observation predicts the next one with high accuracy, neurons in the monkey's IT cortex respond more strongly to unexpected transitions than to expected ones (Meyer and Olson, 2011; Meyer et al., 2014; Ramachandran et al., 2017). However, such transition-dependent surprise signals were not observed in experiments without extensive training (Solomon et al., 2021). In a similar experimental paradigm with humans, Solomon et al. (2021) found the MEG signatures of unexpected transitions only when participants actively looked for unexpected patterns. The importance of attention has also been highlighted in the correlation between pupil dilation and surprise across a variety of tasks (Alamia et al., 2019; Zhao et al., 2019). These results challenge the main premise of Category 4 that experimental subjects infer and use the between-observation associations to predict the next observations. Importantly, if an experimental measurement (e.g., population activity in the IT cortex) does not differ between expected and unexpected observations, then it is not clear whether the conclusion must be 'the measurement does not correlate with surprise' or 'the experimental subjects failed to learn the associations'.

Experiments with action-observation associations (Category 5) allow us to identify whether the experimental subjects learn the experiment's associations. For example, if experimental subjects perform the task as instructed (Daw et al., 2011; Gläscher et al., 2010) or show a rapid adaptation in their behavior after an abrupt change in the associations (Hamid et al., 2016; Mathis et al., 2017), then one can argue that they infer some information about the associations. Experiments with action-observation associations have been used to study error-based and surprise-based learning in humans and animals, both with and without computational models; see section 1.3 for the former and Jordan (2023) for a review of the latter. In a recent seminal example, Jordan and Keller (2023) studied neural activities in the mice Locus Coeruleus (LC) during mismatches between visual flow and running speed (as in Figure 1.2) and showed that the modulation of the LC activity by unexpected mismatches facilitates synaptic changes in the visual cortex. Such studies propose potential mechanisms for how unexpected and surprising events influence learning in the brain, conceptually in line with studies on surprise-modulation of memory (Sinclair and Barense, 2018; Sinclair et al., 2021).

1.2.4 Category 6. Core knowledge and common sense

Experiments based on core knowledge and common sense have been employed to study neural and behavioral signatures of surprise and novelty in realistic situations. For example, a significantly higher amplitude of the N400 ERP component was found for unexpected words in a sentence than for expected words (Hodapp and Rabovsky (2021); Stone et al. (2023); Figure 1.2F1). Lindborg et al. (2023) used different word categories as stimuli in a categorical task from Category 3 (roving-oddball paradigm) and showed an increase in the amplitude of the N400 component even for words that are unexpected only according to a between-observation association. In the example from developmental psychology in Figure 1.2F2, Stahl and Feigenson (2015) showed that 11-month infants had a desire to explore the objects that violated intuitive physics compared to a new distractor object. The infants had also a higher score in learning about hidden features of the objects that violated intuitive physics compared to a new distractor object. These results suggest that there are commonalities in behavioral and neural responses to the new and unexpected in real life and the experimental findings in controlled experimental settings.

1.2.5 Challenges and open questions

How far can we go without computational models? The answer seems to be: Quite far. Neuroscientists have identified numerous physiological signatures of surprise and novelty using different techniques (e.g., neuroimaging, pupilometry, electrophysiology), have shown that surprise and novelty modulate learning and memory in a variety of tasks, and have found strong evidence for a general desire to explore new and unexpected objects across species.

However, it is inevitable to notice that both (i) the definition of the new and unexpected and (ii) the experimental variables linked to these definitions differ across studies. A natural question is 'How do these results relate to each other?'. For example, is there a common basis for the N100 and P300 ERP components in oddball tasks (section 1.1A), the N400 component in the language tasks (section 1.1F1), and the LC activity in the mice virtual reality task (section 1.1E2)? Or is there any link between the desire of monkeys to look at new fractals, the desire of mice to approach new objects (section 1.1C2), and the desire of human babies to explore objects defying intuitive physics (section 1.1F2)?

The reviewed studies imply such links based on an implicit and unspoken belief that there must be a common, but hidden value of surprise and novelty that influences different experimental measurements across different experimental paradigms and species. However, without explicit and formal assumptions about how experimental subjects form expectations and make decisions based on these expectations, it is not clear how one can infer such hidden variables from experimental measurements. This is particularly challenging because (i) many of these experimental variables also respond to other task variables (e.g., cognitive load and extrinsic reward), (ii) not all unexpected or new observations are equally unexpected or new, and (iii) how expectations are formed can potentially be different across sensory modalities and levels of abstractions.

1.3 Computational models of surprise and novelty

In this section, I review computational models and mathematical definitions of surprise and novelty and discuss how these models have enabled neuroscientists (i) to define surprise and novelty on a trial-by-trial basis, (ii) to formally articulate and test hypotheses on how surprise and novelty contribute to learning, memory, and decision-making, and (iii) to clarify links between different experimental responses to surprise and novelty reviewed in section 1.2.

Computational studies of surprise and novelty involve two parts (Figure 1.3A): (i) an abstract and formal description of the experimental paradigm, conceptually similar to illustrations in Figure 1.2, and (ii) a computational model that can 'participate' in such an experiment by interacting with an experiment simulator. For example, a simple computational model of an oddball task (Figure 1.3A) receives a binary input (representing one of the two stimuli) at each trial and updates its estimate of the input distribution after each observation – without taking any actions. The computational model can potentially contain different components for memory (to model familiarity with different stimuli), associative learning (to model how expectations are formed), and decision-making (to model how actions are selected); see Figure 1.3A. Surprise and novelty values are outputs of the model and are mathematically defined as a function of

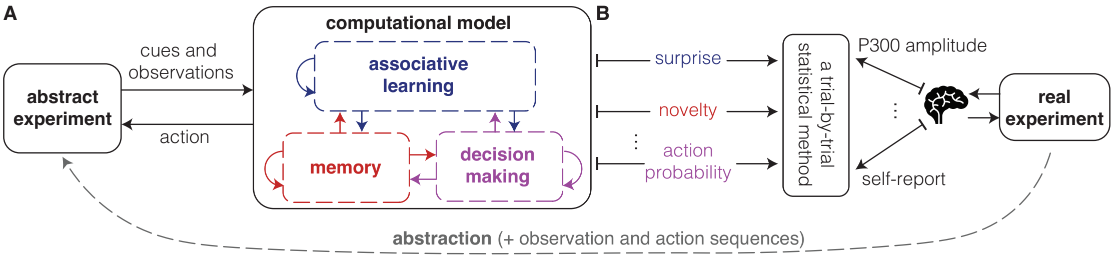


Figure 1.3: **Schematic of computational modeling of surprise and novelty in the brain.** **A.** Computational modeling of surprise and novelty consists of an abstract description of the experiment to generate cues and observations and a computational model of experimental subjects participating in the experiment. The computational model can potentially contain different components for associative learning, memory, and decision-making. In principle, given a choice of parameters, a computational model can run autonomously and 'participate' in the experiment. **B.** Internal variables of the computational model, using the same sequences of observations and actions as in the real experiment, can be used to assign a surprise and novelty value to each observation, infer the probability of taking a particular action, and extract other relevant variables that can describe, for example, learning in the model. Trial-by-trial statistical methods (middle) can be used to relate these variables to different experimental measurements collected during the real experiment (e.g., P300 amplitude and self-reported degree of surprise; right).

the input to the computational model (i.e., observations) and its internal variables (e.g., expected probability of the observation).

There are two main approaches to using computational models in understanding surprise and novelty in the brain: (i) the correlational approach and (ii) the model-testing approach. While the two are often combined, they serve different purposes. In the correlational approach, the computational model is used to infer the surprise and novelty value of each stimulus; these values are then used to explain the variability in the experimental measurements collected from real experiments (Figure 1.3B), e.g., to quantify the influence of surprise on the P300 ERP component. In the model-testing approach, the hypothesis about how surprise and novelty influence different brain functions (e.g., adaptive learning) is formalized in the design of the model, e.g., by considering different candidates for how surprise influences the associative learning module in Figure 1.3A. Different hypotheses can be tested by quantifying which model best explains experimental data (e.g., predicts the subjects' actions with the highest accuracy). Hence, computational modeling in the first approach is only used to infer the hidden value of surprise and novelty, whereas, in the second approach, it is also used to make inferences about brain functions.

In the next four subsections, I review the correlational and model-testing results on the influence of surprise and novelty on physiological measurements, learning, memory, and exploration. To minimize the overlap with the content of the next chapters (particularly chapter 2 and chapter 4), I stick to a conceptual level and do not use a precise mathematical formulation.

1.3.1 Physiological signatures

Typical computational models of oddball tasks (Figure 1.1B and Figure 1.2A) and many other sequential tasks in Categories 1, 2, and 4, use variants of Bayesian learners as the associative learning module in Figure 1.3 to estimate the probability of observing different stimuli on a trial-by-trial basis: They often define the surprise of each observation as a decreasing function of its estimated probability (i.e., less likely observations are considered more surprising), called 'Shannon' surprise or 'surprisal' (Barto et al., 2013; Faraji et al., 2018). Numerous studies with a correlational approach have shown that the Shannon surprise explains a significant trial-by-trial variability of EEG (Gijsen et al., 2021; Kolossa et al., 2015; Kopp and Lange, 2013; Mars et al., 2008; Meyniel et al., 2016; Modirshanechi et al., 2019; Mousavi et al., 2022) and MEG signals in oddball tasks (Maheu et al., 2019; Meyniel, 2020; Mousavi et al., 2022).

In a seminal work, Meyniel et al. (2016) used a combination of correlational and model-testing approaches and showed that the Shannon surprise can be seen as the hidden variable modulating the P300 EEG amplitude, reaction time, and human self-reports of surprise in variants of oddball tasks. Surprisingly, model-testing results revealed that, although the observations in oddball tasks are sampled independently of each other (as in Category 1 in section 1.1), human participants automatically assume that there are between-observation associations and learn the transition probabilities between stimuli (as in Category 4 in section 1.1). Maheu et al. (2019) confirmed this result for the late MEG components such as P300 but showed the opposite for early components such as N100. This implies that the ERP components that had been previously linked to a single notion of being 'unexpected' (as reviewed in section 1.2) are related to being 'unexpected' *due to different expectations*. Whether these components are signatures of surprise or novelty is debated (Barto et al., 2013).

An alternative definition, known as 'Bayesian' surprise, considers the amount of update in the model's predictions as a measure of surprise (Baldi, 2002; Schmidhuber, 2010). The signatures of Bayesian surprise have been compared to those of Shannon surprise (Gijsen et al., 2021; Kolossa et al., 2015; Mars et al., 2008; Mousavi et al., 2022; Ostwald et al., 2012; Visalli et al., 2021). In two separate tasks in Categories 1 and 4 (section 1.1), the P300 amplitude at the frontal and frontocentral electrodes has been reported to be explained better by Bayesian surprise than Shannon surprise, and the P300 amplitude at the parietal electrodes has been reported to be explained better by Shannon surprise than Bayesian surprise (Kolossa et al., 2015; Visalli et al., 2021), linked to the P3a and P3b sub-components of the P300 component (Polich, 2007). In addition to Bayesian and Shannon surprise, Kolossa et al. (2015) reported a significant positive correlation between another definition of surprise (Postdictive surprise; see chapter 2) and the EEG amplitudes in a later time window (i.e., the EEG slow wave component). Gijsen et al. (2021) reported similar results for another task from Category 3, another set of surprise definitions (Shannon, Bayesian, and Confidence-Corrected surprise; see chapter 2), and another set of EEG components.

Further correlates of different mathematical definitions of surprise have been found in a variety of experimental paradigms and measurements (Antony et al., 2021; Cogliati Dezza et al., 2022; Daw et al., 2011; Gläscher et al., 2010; Kao et al., 2020a,b; Kolossa et al., 2015; Konovalov and Krajbich, 2018; Kopp and Lange, 2013; Li et al., 2019; Lindborg et al., 2023; Loued-Khenissi and Preuschoff, 2020; Mars et al., 2008; McGuire et al., 2014; Meyniel, 2020; Modirshanechi et al., 2019; Mousavi et al., 2022; Nassar et al., 2012; Nour et al., 2018; Ostwald et al., 2012; O'Reilly et al., 2013; Preuschoff et al., 2011). The main takeaway is that computational models (i) have helped to link many separate experimental measurements to concrete and precise notions of surprise (and novelty), but they (ii) have also falsified the belief that there is a single notion of 'new' or 'unexpected' in the brain.

1.3.2 Surprise and novelty in learning

To characterize the role of surprise and novelty in learning, we can use a model-testing approach to identify the best candidate for the associative learning module (Figure 1.3A). Different candidate models include (approximate) Bayesian models, mechanistic and heuristic models (Barry and Gerstner, 2022; Grossman et al., 2022; Iigaya, 2016; Wilmes et al., 2023), and data-driven models (e.g., large language models (Heilbron et al., 2022; Kumar et al., 2023) or models of intuitive physics (Piloto et al., 2022)). How surprise and novelty contribute to learning can be directly incorporated into the design of mechanistic and heuristic models (e.g., Barry and Gerstner (2022); Grossman et al. (2022); Iigaya (2016); Pearce and Hall (1980); Rouhani and Niv (2021)), but, for optimal and data-driven models, it is often needed to be inferred from analyzing the model's internal dynamics.

Bayesian and approximate Bayesian models have been frequently used to analyze behavioral data of human participants in experimental paradigms with volatile observation distribution (Category 2 in section 1.1; Nassar et al. (2012, 2010)), volatile between-observation associations (Category 3 in section 1.1; Heilbron and Meyniel (2019); Meyniel (2020)), and volatile action-observation associations (Category 4 in section 1.1; Behrens et al. (2007); Kao et al. (2020b), and chapter 3). The unified message of these studies is that human behavior in volatile environments is explained better by models that have an adaptive rate of learning, modulated by a definition of surprise (e.g., Shannon surprise or the difference between subjects' predictions and true observations).

These results have been further supported by correlational studies that found neural correlates of different internal variables of Bayesian models in humans (Kao et al., 2020b; McGuire et al., 2014) and monkeys (Li et al., 2019). Studies in mice have even provided examples of these models guiding experiment design to unravel the functional roles of different brain regions in adaptive learning (Mathis et al., 2017). Alternative heuristic models with explicit surprise-modulation of learning have been tested against experimental data (Findling et al., 2021; Grossman et al., 2022; Rouhani and Niv, 2021; Rouhani et al., 2018). All these models provide concrete quantitative descriptions of learning in humans and animals and argue that 'surprise' increases the rate of learning, but they disagree on the definition of surprise. The role of novelty in learning (e.g., as experimentally discussed by Morrens et al. (2020)) has remained relatively unexplored from the perspective of computational modeling.

1.3.3 Surprise and novelty in memory

Bayesian, heuristic, and data-driven models have also been used to study the link between surprise and memory. For example, according to Bayesian models of volatile observation distribution (Category 2; Gershman et al. (2014)) and volatile between-observation associations (Category 4; Gershman et al. (2017)), whether a surprising observation makes humans and animals (i) modify their current estimate of associations, (ii) switch their estimate to another previously learned association, or (iii) allocate a new memory trace for learning a new association depends on the amplitude of their error in predicting the next observations. Mechanistic approximations have shown that such a surprise-modulated allocation and modification of memories can be implemented in a network of spiking neurons under biological constraints (Barry and Gerstner, 2022). The indicator signal of whether a new memory trace should be allocated to learn new associations can be seen as a novelty signal in these models.

Heuristic models in experiments with volatile action-observation associations (Category 5 in section 1.1) have shown a positive correlation between a definition of surprise and the memorability of task-independent stimuli (Rouhani and Niv, 2021; Rouhani et al., 2018). Similar models have shown that surprise values predict how human participants chunk their sensory observations into segments that are memorized (Rouhani et al., 2018, 2020). Data-driven models have reproduced these results in experiments based on core knowledge and common sense (Category 6 in section 1.1; Antony et al. (2021); Kumar et al. (2023)) but with different definitions of surprise.

In summary, computational models have proposed concrete mechanisms for how surprise and novelty contribute to the consolidation, modification, and segmentation of memories. However, the results of different studies appear as a portfolio of separate findings as they use different modeling assumptions and different mathematical definitions for surprise and novelty.

1.3.4 Surprise and novelty in exploration

Computational studies of surprise and novelty in exploration propose models of how these signals, derived from the memory and associative learning modules, guide action selection in the decision-making module (Figure 1.3). Recent models of exploration consider intrinsically motivated Reinforcement Learning (RL; Aubret et al. (2019); Ladosz et al. (2022); Singh et al. (2010b)) algorithms as models of decision-making modules (Gottlieb and Oudeyer, 2018; Murayama, 2022; Oudeyer, 2018). These models describe decision-making by action policies that are learned, via interaction with the environment, to collect the maximum amount of 'reward', where 'reward' is assumed to be a combination of extrinsically valuable goods (e.g., food or money) and intrinsically generated satisfactory signals (e.g., the excitement of experiencing something novel).

For several experiments with action-observation associations (Category 5 in section 1.1), the exploratory actions of human participants are accurately predicted by reward signals that include mathematical definitions of novelty (Cogliati Dezza et al., 2022; Poli et al., 2022) or surprise (Horvath et al., 2021; Kobayashi et al., 2019; Nelson, 2005; Poli et al., 2022; Ten et al., 2021). For example, in a task where the participants' actions did not influence their total monetary reward, participants showed a clear preference for actions that decreased their uncertainty of total monetary value, i.e., actions whose outcomes were least predictable and, hence, most 'surprising' (Kobayashi et al., 2019). Other studies have shown that both surprise and novelty predict exploratory action choices of human participants (Cockburn et al., 2022; Poli et al., 2022), but these studies do not agree on the exact definition of surprise and novelty; this, importantly, may be due to fundamental differences in the experimental design and modeling approach.

Bayesian surprise (i.e., the amount of update in the estimates of associations after a new observation) and signals correlated with Bayesian surprise predict exploratory actions of human participants in multi-armed bandit tasks (Gershman, 2019a; Horvath et al., 2021; Schulz and Gershman, 2019) as well as their gaze orientation (Itti and Baldi, 2006, 2009). On the other hand, correlational studies on gaze orientation in human infants (Kidd et al., 2012), human young adults (Cubit et al., 2021), and monkeys (Wu et al., 2022) proposed Shannon surprise as the main predictor of gaze orientation, with an inverted-U relationship.

While intrinsically motivated RL algorithms have proven to be powerful models of human exploration, there is no consensus on the choice of intrinsic reward signal in different experiments. It is unclear why the best predictor of subjects' exploratory actions appears to be different in different experimental paradigms (see Dubey and Griffiths (2019) for further discussion and potential answers).

1.3.5 Challenges and open questions

Computational models have enabled neuroscientists to go beyond the traditional approach (as in Figure 1.1) by exploiting precise mathematical definitions to describe surprise and novelty. Computational models have provided quantitative descriptions of how surprise and novelty influence learning, memory, and exploration, and importantly, have shown that several definitions of surprise and novelty are potentially involved in the neural and behavioral responses to the new and unexpected.

Although the diversity of modeling approaches and definitions has helped to explain many experimental phenomena, the diversity also challenges generalization across different studies. In particular, it is not clear (i) how different mathematical definitions of surprise and novelty relate to each other and, given an experimental paradigm and specific task, (ii) which definitions are best suited to describe experimental measurements.

1.4 Thesis contribution

The main purpose of this thesis is to take a step towards addressing two of the main challenges of computational models of surprise and novelty reviewed above, i.e., (i) how different mathematical definitions of surprise and novelty relate to each other and (ii) which definitions are best suited to describe experimental measurements in a given task.

In **chapter 2** (based on the publication Modirshanechi et al. (2022) together with J Brea and W Gerstner), I present a detailed mathematical investigation of 18 different definitions of surprise. I study their similarities and differences, identify conditions under which they are indistinguishable, and propose a taxonomy of these definitions based on two classification schemes. These results propose a refined terminology and a solid ground for relating previous studies of surprise to each other. The unifying framework and common mathematical language also enable the design of new experimental paradigms.

In **chapter 3** (based on the publication Xu et al. (2021) together with HA Xu[2], MP Lehmann, W Gerstner[3], and MH Herzog[3]), I propose a mathematical formalism to distinguish novelty from surprise and design an augmented RL algorithm that formalizes the hypothesis that surprise modulates the rate of learning while novelty drives exploratory actions. I test this model against experimental data of human participants (collected by HA Xu) and compare it to 12 alternative algorithms. I show that the model's internal variables (including the definition of surprise and novelty) can explain significant variabilities of the EEG signals in frontal electrodes (collected by HA Xu). These results propose concrete and distinct roles for surprise and novelty in sequential decision-making tasks with sparse rewards and volatile action-observation associations (Category 5 in section 1.1). The surprise-modulated learning model used in this chapter is based on the earlier publication Liakoni et al. (2021) (with V Liakoni[4], W Gerstner, and J Brea) which is not included in the main text of the thesis for the sake of space (abstract in **Appendix D**).

[2] HA Xu and I are joint first authors.

[3] W Gerstner and MH Herzog are joint senior authors.

[4] V Liakoni and I are joint first authors.

Chapter 4 (based on the publication Modirshanechi et al. (2023a) together with S Becker, J Brea, and W Gerstner) plays the role of an intermediate discussion where I review different studies on physiological signatures of surprise and novelty in light of our taxonomy in chapter 2 and our formal separation of surprise and novelty in chapter 3. In a complementary review paper Modirshanechi et al. (2023b) (together with K Kondrakiewicz, W Gerstner, and S Haesler), I review the different roles of surprise and novelty in studies of curiosity-driven exploration. This publication is not included in the main text of the thesis for the sake of space (abstract in **Appendix E**).

In **chapter 5** (based on the preprint Modirshanechi et al. (2023c) together with WH Lin, HA Xu, MH Herzog, and W Gerstner), I build upon the formal categorizations developed throughout chapters 2-4 and compare intrinsically motivated RL algorithms based on novelty, surprise, and information gain as models of human exploration. Given a specifically designed experimental paradigm that dissociates different exploration strategies, I study the behavioral data of human participants (collected by WH Lin and HA Xu) and show that their exploratory actions can be explained best by novelty-driven exploration, even though novelty-seeking is suboptimal in our experimental paradigm. These results extend the results of chapter 3 and provide further evidence that novelty is the dominant drive of human exploration in environments with sparse rewards. Finally, I present a brief discussion of the main results of the thesis along with potential future directions in **chapter 6**.

In addition to the main publications in the thesis, **Appendix F** contains the abstracts of my other publications as a contributing author through my Ph.D. (Bellec et al., 2021; Brea et al., 2023; Esmaeili et al., 2021; Liakoni et al., 2022; Oryshchuk et al., 2024). My contribution to these publications mainly includes analyzing data as well as helping with statistical methodology, computational modeling, and interpretation of results (detailed author contribution is provided for each paper).

2 A taxonomy of surprise definitions

This chapter was published in the Journal of Mathematical Psychology (Modirshanechi et al., 2022)[1].

Authors: Alireza **Modirshanechi**, Johanni Brea, and Wulfram Gerstner

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.

Funding: This research was supported by the Swiss National Science Foundation (no. 200020 184615).

Author contribution: All authors contributed to the conceptualization of the study. AM did the formal analyses and visualization and wrote the original draft. All authors revised the text.

[1] For consistency across the thesis chapters, the mathematical notation has been slightly adjusted.

2.1 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 and 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 and 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 and Izquierdo, 2019; Yu and 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 and 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 and 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 and Baldi, 2006; Rouhani and Niv, 2021; Xu et al., 2021) 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 2.2. Then, in section 2.3 to section 2.6, 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 2.7, we build upon our theoretical analyses and propose a taxonomy of surprise measures by classifying them into four conceptually different categories.

2.2 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. subsection 2.2.2). 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 Figure 2.1 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.

2.2.1 General definition

At each discrete time $t \in \{0, 1, 2, ...\}$, the agent's model of the environment is characterized by a tuple of 4 random variables $(X_t, Y_t, \Theta_t, C_t)$ (Figure 2.2A). 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

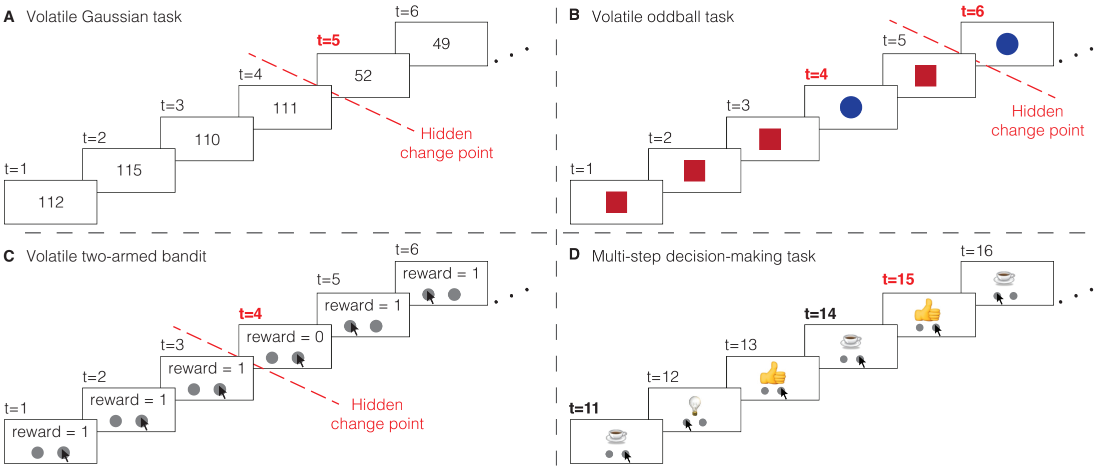


Figure 2.1: **Four typical experimental paradigms to study functional roles and physiological signatures of surprise in the brain. A.** Volatile Gaussian task (Nassar et al., 2012, 2010): Participants see a sequence of numbers randomly sampled from a Gaussian distribution whose mean is piece-wise constant 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 and 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 grey 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 and 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.

(Kolossa et al., 2015; Mars et al., 2008) (e.g., Figure 2.1), 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.,

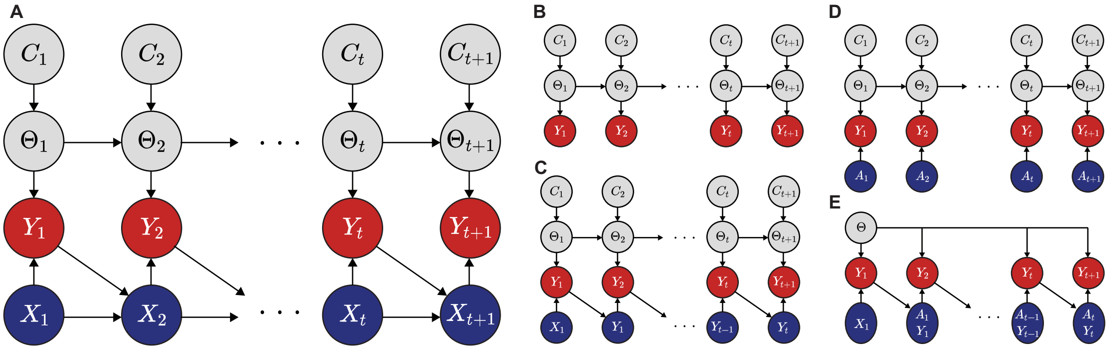


Figure 2.2: **Subjective model of the environment. A.** The Bayesian network (Barber, 2012) corresponding to the most general case of our generative model in Equation 2.1 and Equation 2.2. The arrows show conditional dependence, the grey nodes show the hidden variables ($C_{1:t+1}$ and $\Theta_{1:t+1}$), 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 Figure 2.1A (Adams and MacKay, 2007; Fearnhead and 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 Figure 2.1B (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 Figure 2.1C (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 Figure 2.1D (Daw et al., 2011; Gläscher et al., 2010; Huys et al., 2015; Lehmann et al., 2019; Schultz et al., 1997; Sutton and Barto, 2018), where the cue variable $X_t = (A_{t-1}, Y_{t-1})$ consists of previous action and observation. See subsection 2.2.2 for details.

2007; Horvath et al., 2021) (e.g., Figure 2.1C-D), and a conditioned stimulus in Pavlovian conditioning tasks (Gershman et al., 2017).

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 $\Theta_{1:t} = (\Theta_1, ..., \Theta_t)$ 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 and MacKay, 2007; Behrens et al., 2007; Fearnhead and Liu, 2007; Findling et al., 2021; Glaze et al., 2015; Heilbron and Meyniel, 2019; Liakoni et al., 2021; Meyniel et al., 2016; Nassar et al., 2012, 2010; Wilson et al., 2013; Xu et al., 2021; Yu and Cohen, 2009; Yu and Dayan, 2005), the agent assumes that the environment undergoes abrupt changes at random points in time (e.g., Figure 2.1A-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 $b^{(0)}$ 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. Figure 2.2A).

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*

$$\begin{aligned}\mathbb{P}(y_{1:t}, x_{1:t}, \theta_{1:t}, c_{1:t}) :=\ & \mathbb{P}(c_1)\mathbb{P}(\theta_1)\mathbb{P}(x_1)\mathbb{P}(y_1|x_1,\theta_1)\times \\ & \prod_{\tau=2}^{t} \mathbb{P}(c_\tau)\mathbb{P}(\theta_\tau|\theta_{\tau-1},c_\tau)\mathbb{P}(x_\tau|x_{\tau-1},y_{\tau-1})\mathbb{P}(y_\tau|x_\tau,\theta_\tau),\end{aligned} \tag{2.1}$$

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

$$\begin{aligned}\mathbb{P}(c_\tau) :=& \text{Bernoulli}(c_\tau; p_c) \\ \mathbb{P}(\theta_\tau|\theta_{\tau-1}, c_\tau) :=& b^{(0)}(\theta_\tau)\delta_{\{1\}}(c_\tau) + \delta_{\{\theta_{\tau-1}\}}(\theta_\tau)\delta_{\{0\}}(c_\tau) \\ \mathbb{P}(y_\tau|x_\tau,\theta_\tau) :=& P_{Y|X}(y_\tau|x_\tau;\theta_\tau),\end{aligned} \tag{2.2}$$

where δ is the Dirac measure (cf. Table 2.1), and $P_{Y|X}$ is a time-invariant conditional distribution of observations given cues[2]*. We do not make any assumption about $\mathbb{P}(x_1)$ and $\mathbb{P}(x_\tau|x_{\tau-1}, y_{\tau-1})$.*

See Table 2.1 for a summary of the notation.

2.2.2 Special cases and links to related works

Many of the commonly used experimental paradigms (e.g., see Figure 2.1) 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 and MacKay, 2007; Liakoni et al., 2021; Nassar et al., 2012, 2010) is obtained if we remove the cue variables

[2] The last line of Equation 2.2 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 $\tau' \in \{0, 1, 2, ...\}$.

Table 2.1: **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., $\mathbb{P}^{(t)}_W(w) := \mathbb{P}^{(t)}(W = w)$
$b^{(0)}$	Prior distribution over the environment parameter; equivalently, the distribution of Θ_t given $C_t = 1$
$b^{(t)}$	The belief about parameter Θ_t at time t, i.e., $b^{(t)}(\theta) := \mathbb{P}^{(t)}(\Theta_t = \theta)$
$P(y\|x;b^{(t)})$	The marginal probability of observation y given cue x and belief $b^{(t)}$; see Equation 2.4
$P(.\|x;b^{(t)})$	The full marginal distribution over the space of observations given cue x and belief $b^{(t)}$
$\|\|w\|\|_1$	ℓ_1-norm of the vector $w = (w_1, ..., w_N) \in \mathbb{R}^N$ defined as $\|\|w\|\|_1 := \sum_{n=1}^N \|w_n\|$
$\|\|w\|\|_2$	ℓ_2-norm of the vector $w = (w_1, ..., w_N) \in \mathbb{R}^N$ 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.

$X_{1:t}$ (Figure 2.2B). For example, in the Gaussian experiment of Nassar et al. (2010) (Figure 2.1A), Y_t is a sample from a Gaussian distribution with a mean equal to Θ_t and a known variance, and $b^{(0)}$ is a very broad uniform distribution.

The minimal model of human inference about binary sequences of Meyniel et al. (2016) (Figure 2.2C) 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 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 (Figure 2.1B).

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 $A_{1:t}$ (Figure 2.2D). For example, in the experiment of Behrens et al. (2007) (Figure 2.1C), $X_t = A_t$ 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[3].

Classic Markov Decision Processes (MDPs) (Sutton and 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}) (Figure 2.2E). In this setting, we have $\mathbb{P}(X_\tau = (a_{\tau-1}, y)|x_{\tau-1}, y_{\tau-1}) \coloneqq \delta_{\{y_{\tau-1}\}}(y)\mathbb{P}(a_{\tau-1}|y_{\tau-1})$, where $\mathbb{P}(a_{\tau-1}|y_{\tau-1})$ is called the action selection policy in Reinforcement Learning theory (Sutton and Barto, 2018) and is independent of the dynamics of the environment[4]. 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) (Figure 2.1D).

2.2.3 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 $\mathbb{P}^{(t)}(w) \coloneqq \mathbb{P}(w|y_{1:t}, x_{1:t})$. Following this notation, we define an agent's belief about the parameter Θ_t at time t as

$$b^{(t)}(\theta) \coloneqq \mathbb{P}^{(t)}(\Theta_t = \theta), \tag{2.3}$$

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. subsection 2.3.1), 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 $b^{(t+1)}(\theta) = \mathbb{P}^{(t+1)}(\Theta_{t+1} = \theta)$ can be found by normalizing the product of the *prior* belief

[3] We note that the action probability $\mathbb{P}\big(a_\tau\big)$ in bandit tasks often depends on the whole history of the agent, i.e., $a_{1:\tau-1}$ and $y_{1:\tau-1}$ (Sutton and 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 $\mathbb{P}\big(X_\tau = (a'_{1:\tau}, y'_{1:\tau-1})|x_{\tau-1}, y_{\tau-1}\big) \coloneqq \delta_{\{a_{1:\tau-1}\}}(a'_{1:\tau-1})\delta_{\{y_{1:\tau-1}\}}(y'_{1:\tau-1})\mathbb{P}\big(a'_\tau|a_{1:\tau-1}, y_{1:\tau-1}\big)$ where $\mathbb{P}\big(a'_\tau|$ $a_{1:\tau-1}, y_{1:\tau-1}\big)$ is the non-stationary action selection policy – cf. Sutton and Barto (2018).

[4] 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 $y_{1:\tau-1}$, i.e., through estimation of action values (Sutton and Barto, 2018). In these situations, we can define x_τ as the concatenation of $a_{1:\tau}$ and $y_{1:\tau-1}$.

$\mathbb{P}^{(t)}(\Theta_{t+1} = \theta)$ about Θ_{t+1} and the *likelihood* $P_{Y|X}(y_{t+1}|x_{t+1};\theta)$. In subsection 2.4.1, 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 $b^{(t)}$:

$$\begin{aligned} P(y|x;b^{(t)}) &:= \mathbb{E}_{b^{(t)}}\Big[P_{Y|X}(y|x;\Theta)\Big] \\ &= \int P_{Y|X}(y|x;\theta)b^{(t)}(\theta)d\theta, \end{aligned} \tag{2.4}$$

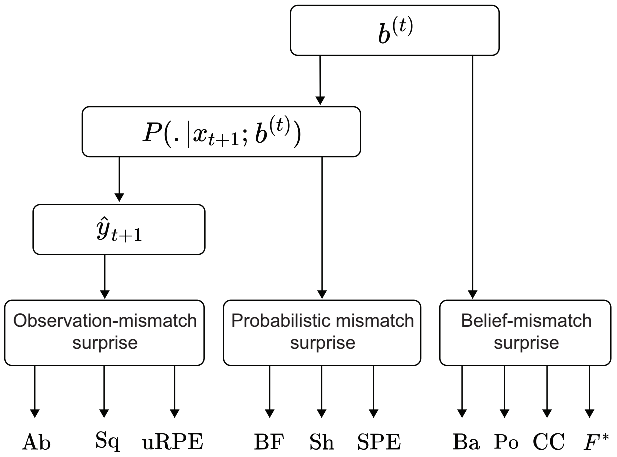
where the integration is replaced by summation whenever θ is discrete.

2.3 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 A.

2.3.1 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} (Figure 2.2A). Hence, the influence of $y_{1:t}$ and $x_{1:t}$ on the surprise of observing y_{t+1} is exclusively through the belief $b^{(t)}$, which indicates the importance of $b^{(t)}$ in surprise computation. More precisely, a surprise measure is a function $\mathsf{S} : \mathcal{Y} \times \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ that takes an observation $y_{t+1} \in \mathcal{Y}$, a cue $x_{t+1} \in \mathcal{X}$, and a belief $b^{(t)} \in \mathcal{P}$ as arguments and gives the value $\mathsf{S}(y_{t+1}|x_{t+1};b^{(t)}) \in \mathbb{R}$ as the corresponding surprise value. However, the specific form of how $b^{(t)}$ influences surprise computation changes between one measure and another. Based on how they depend on $b^{(t)}$, we divide existing surprise measures into three categories: (i) probabilistic mismatch, (ii) observation-mismatch, and (iii) belief-mismatch surprise measures (Figure 2.3). *Probabilistic mismatch* surprise measures depend on the belief $b^{(t)}$ only through the marginal probability $P(y_{t+1}|x_{t+1};b^{(t)})$; an example is the Shannon surprise (Barto et al., 2013; Tribus, 1961). In other words, probabilistic mismatch surprise depends only on the integral $P(y_{t+1}|x_{t+1};b^{(t)}) = \int P_{Y|X}(y_{t+1}|x_{t+1};\theta)b^{(t)}(\theta)d\theta$ (Equation 2.4) and is independent of other characteristics of the belief $b^{(t)}$. *Observation-mismatch* surprise measures depend on $b^{(t)}$ only through some estimate $\hat{y}_{t+1}$ of the next observation according to the marginal distribution $P(.|x_{t+1};b^{(t)})$ (cf. Table 2.1); an example is the absolute difference between y_{t+1} and $\hat{y}_{t+1}$ (Nassar et al., 2010; Prat-Carrabin et al., 2021).



Abbreviations:

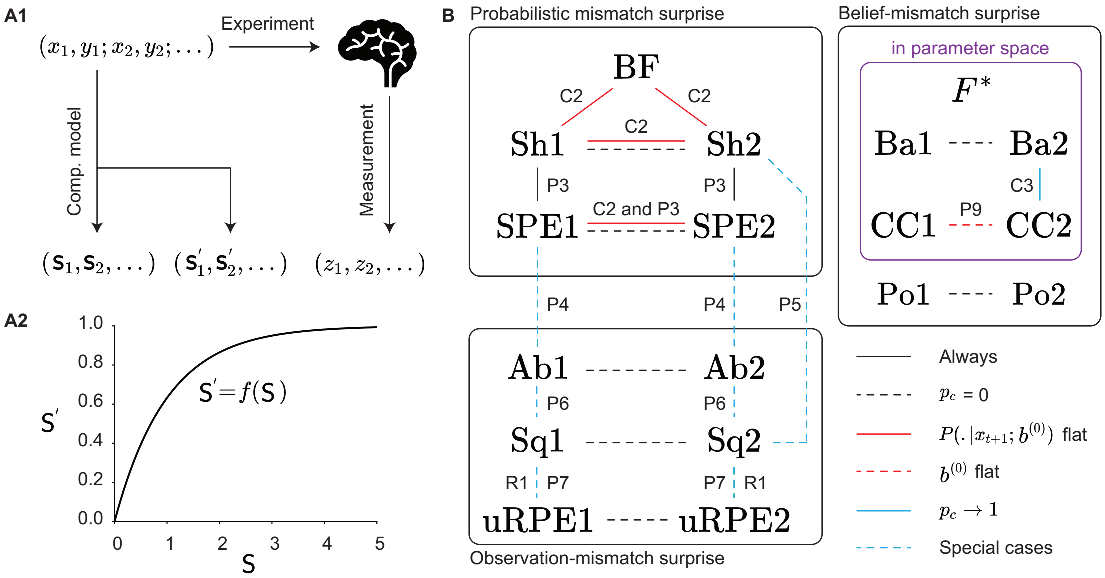
Ab	Absolute error
Sq	Squared error
uRPE	Unsigned reward prediction error
BF	Bayes Factor surprise
Sh	Shannon surprise
SPE	State prediction error
Ba	Bayesian surprise
Po	Postdictive surprise
CC	Confidence Corrected surprise
F^*	Minimized free energy

Figure 2.3: **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 $b^{(t)}$. However, the specific form of the dependence changes between one measure and another. 'Observation-mismatch' surprise measures use the marginal distribution $P(.|x_{t+1};b^{(t)})$ (cf. Table 2.1) 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 2.1). 'Probabilistic mismatch' surprise measures use the marginal probability $P(y_{t+1}|x_{t+1};b^{(t)})$ directly, without extracting a specific estimate. 'Belief-mismatch' surprise measures use the belief $b^{(t)}$ directly, without extracting the marginal probability $P(y_{t+1}|x_{t+1};b^{(t)})$. See section 2.3 for details.

In other words, observation-mismatch surprise depends only on some statistics (e.g., average or mode) of $P(.|x_{t+1};b^{(t)})$ that is used as the estimate $\hat{y}_{t+1}$ and is independent of the other characteristics of $b^{(t)}$ and $P(.|x_{t+1};b^{(t)})$. To compute the *belief-mismatch* surprise measures, however, we need to have the whole distribution $b^{(t)}$; an example is the Bayesian surprise (Baldi, 2002; Schmidhuber, 2010). In other words, neither the marginal distribution $P(.|x_{t+1};b^{(t)})$ nor the estimate $\hat{y}_{t+1}$ can solely determine the value of a belief-mismatch surprise measure.

2.3.2 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 S and 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 Figure 2.4A1. However, if there exists a strictly increasing mapping between S and S' (e.g., as in Figure 2.4A2), then the two surprise measures have the same explanatory power with respect to Z – because any function of S can be written in terms of S' and vice-versa. For example, assume that $\mathsf{S} = f(\mathsf{S}')$ for a strictly increasing function f. If an estimator of the variable Z is found using the measure S as $\hat{Z} = g(\mathsf{S})$, then we can rewrite the same estimator in terms of S' as $\hat{Z} = \tilde{g}(\mathsf{S}') = g(f(\mathsf{S}'))$.



Abbreviations:

BF : Bayes Factor surprise
Sh : Shannon surprise
SPE : State prediction error

Ab : Absolute error
Sq : Squared error
uRPE : Unsigned reward prediction error

Ba : Bayesian surprise
Po : Postdictive surprise
CC : Confidence Corrected surprise
F^* : Minimized free energy

Figure 2.4: **Indistinguishable surprise measures. A.** A typical question in human and animal experiments is whether a surprise measure S explains the variations of a behavioral or physiological variable Z better than an alternative surprise measure 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 $\mathsf{s}_{1:t}$ or $\mathsf{s}'_{1:t}$ is predicted by computational modeling. Then statistical tools are used to study whether the sequence $\mathsf{s}_{1:t}$ or $\mathsf{s}'_{1:t}$ is more informative about the sequence of measurements $z_{1:t}$. **A2.** If there exists a strictly increasing function f such that $\mathsf{S}' = f(\mathsf{S})$, then the two surprise measures are equally informative about the measurable variable Z. In this case, S and 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 $P(.|x_{t+1}; b^{(0)})$ being flat; a dashed red line corresponds to the prior belief $b^{(0)}$ 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 2.2 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 2.3 for details.

Because $g(\mathsf{S})$ and $\tilde{g}(\mathsf{S}')$ have the same explanatory power given any function g and any measure of performance, the two surprise measures S and S' are equally informative about the variable Z in this regard[5]. We formalize this idea in Definition 2.

Definition 2. *(Indistinguishability) For the generative model of Definition 1, we say* S *and* S' *are indistinguishable if there exists a strictly increasing function* $f : \mathbb{R} \rightarrow \mathbb{R}$ *such that* $\mathsf{S} = f(\mathsf{S}')$ *for all choices of belief* $b^{(t)}$*, 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 (Figure 2.4B and Table 2.2).

2.4 Probabilistic mismatch surprise measures

2.4.1 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 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 and 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):

$$\begin{aligned} \mathsf{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; b^{(t)}) &:= \frac{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}, C_{t+1} = 1)}{\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}, C_{t+1} = 0)} \\ &= \frac{P(y_{t+1}|x_{t+1}; b^{(0)})}{P(y_{t+1}|x_{t+1}; b^{(t)})}. \end{aligned} \tag{2.5}$$

The name arises because $\mathsf{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; b^{(t)})$ is the Bayes Factor (Bayarri and Berger, 1997; Kass and Raftery, 1995) used in statistics to test whether a change has occurred at time t. For a given $P(y_{t+1}|x_{t+1}; b^{(0)})$, the Bayes Factor surprise is a decreasing function of $P(y_{t+1}|x_{t+1}; b^{(t)})$: Hence, more probable events are perceived as less surprising. However,

[5] 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 and Foreman, 2014). For example, the Spearman correlation (a measure of monotonic relationship between two random variables) between S' and Z is the same as the Spearman correlation between $\mathsf{S} = f(\mathsf{S}')$ and Z, but this is not the case for Pearson correlation (a measure of linear relationship between two random variables) if f is nonlinear.

Table 2.2: **Indistinguishability conditions of Figure 2.4 for several experimental paradigms.** Publications specified by $\diamond$ 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 Figure 2.1 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	$b^{(0)}$	$P(.\|x; b^{(0)})$	p_c
Nassar et al. (2012, 2010)$^{\diamond}$	Volatile Gaussian	$=$ flat	$=$ flat	> 0
Glaze et al. (2015)$^{\diamond,*}$	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)$^{\diamond}$ Maheu et al. (2019)$^{\diamond}$, etc.	Oddball	$=$ flat	$=$ flat	$= 0$
Heilbron and Meyniel (2019)$^{\diamond}$ Meyniel (2020)$^{\diamond}$	Volatile oddball	$=$ flat	$=$ flat	> 0
Ostwald et al. (2012)$^{\diamond}$ Lieder et al. (2013)	Roving oddball	$=$ flat	$=$ flat	$= 0$
Gijsen et al. (2021)$^{\diamond}$	Volatile roving oddball	$=$ flat	$=$ flat	> 0
Kolossa et al. (2015)$^{\diamond}$	Urn-ball	$\neq$ flat	$\neq$ flat	$= 0$
Behrens et al. (2007)$^{\diamond}$ Horvath et al. (2021)$^{\diamond}$	Reversal bandit	$=$ flat	$=$ flat	> 0
Rouhani and Niv (2021)* Findling et al. (2021)$^{\diamond}$	Volatile contextual bandit	$=$ flat	$=$ flat	> 0
Gläscher et al. (2010)	Multi-step decision-making	$=$ flat	$=$ flat	$= 0$
Liakoni et al. (2022)$^{\diamond}$	Multi-step decision-making with outliers	$\neq$ flat	$=$ flat	$= 0$
Xu et al. (2021)$^{\diamond}$	Volatile multi-step decision-making	$\neq$ flat	$=$ flat	> 0

the key feature of $\mathsf{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; b^{(t)})$ is that it measures not only how unexpected (unlikely) the observation y_{t+1} is according to the current belief $b^{(t)}$ but also how expected it would be if the agent had reset its belief to the prior belief. More precisely, for a given $P(y_{t+1}|x_{t+1}; b^{(t)})$, the Bayes Factor surprise is an increasing function of $P(y_{t+1}|x_{t+1}; b^{(0)})$.

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 Figure 2.2B, the exact Bayesian inference for the update of $b^{(t)}$ to $b^{(t+1)}$ upon observing y_{t+1} leads to a learning rule modulated by the Bayes Factor surprise. Proposition 1 below states that this result is also true for our more general generative model (Figure 2.2A).

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 $b^{(t+1)}$, after observing y_{t+1} with the cue x_{t+1}, as*

$$b^{(t+1)}(\theta) = (1-\gamma_{t+1}) b^{(t+1)}_{\text{integration}}(\theta) + \gamma_{t+1} b^{(t+1)}_{\text{reset}}(\theta), \tag{2.6}$$

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

$$\gamma_{t+1} := \frac{m\mathsf{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; b^{(t)})}{1 + m\mathsf{S}_{\mathrm{BF}}(y_{t+1}|x_{t+1}; b^{(t)})} \quad \textit{with} \quad m := \frac{p_c}{1-p_c}, \tag{2.7}$$

and

$$\begin{aligned} b^{(t+1)}_{\text{integration}}(\theta) &:= \frac{P_{Y|X}(y_{t+1}|x_{t+1}; \theta) b^{(t)}(\theta)}{P(y_{t+1}|x_{t+1}; b^{(t)})}, \\ b^{(t+1)}_{\text{reset}}(\theta) &:= \frac{P_{Y|X}(y_{t+1}|x_{t+1}; \theta) b^{(0)}(\theta)}{P(y_{t+1}|x_{t+1}; b^{(0)})}. \end{aligned} \tag{2.8}$$

Therefore, the Bayes Factor surprise S_{BF} controls the trade-off between the integration of the new observation into the old belief (via $b^{(t+1)}_{\text{integration}}$) and resetting the old belief to the prior belief (via $b^{(t+1)}_{\text{reset}}$)[6].

2.4.2 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.

[6] **Thesis footnote:** The adaptation rate γ_{t+1} is also known the *posterior* Change-Point-Probability and has been interpreted as a surprise measure (Kao et al., 2020b). Given a fixed p_c, the Bayes Factor surprise S_{BF} is indistinguishable from γ (Definition 2), but they are distinguishable across experiments with different p_c, e.g., same value of S_{BF} corresponds to different values of γ for different p_c.

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 and Krajbich, 2018; Kopp and 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

$$\begin{aligned} \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)}) &:= -\log \mathbb{P}^{(t)}(y_{t+1}|x_{t+1}) \\ &= -\log\Big(p_c P(y_{t+1}|x_{t+1};b^{(0)}) + (1-p_c)P(y_{t+1}|x_{t+1};b^{(t)})\Big), \end{aligned} \tag{2.9}$$

where the 2nd equality is a result of the marginalization

$$\mathbb{P}^{(t)}(y_{t+1}|x_{t+1}) = \sum_c \mathbb{P}^{(t)}(y_{t+1}, C_{t+1}=c|x_{t+1}). \tag{2.10}$$

The Shannon surprise S_{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 $P(y_{t+1}|x_{t+1};b^{(t)})$, the Shannon surprise is a decreasing function of $P(y_{t+1}|x_{t+1};b^{(0)})$ (cf. Equation 2.9): 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 $P(y_{t+1}|x_{t+1};b^{(0)})$ (cf. Equation 2.5): 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 Figure 2.2B, 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:

$$\begin{aligned} \mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};b^{(t)}) &:= -\log \mathbb{P}^{(t)}(y_{t+1}|x_{t+1}, C_{t+1}=0) \\ &= -\log P(y_{t+1}|x_{t+1};b^{(t)}). \end{aligned} \tag{2.11}$$

In other words, $\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};b^{(t)})$ neglects the potential presence of change-points, and, therefore, it is independent of both p_c and $P(y_{t+1}|x_{t+1};b^{(0)})$. For a non-volatile environment that does not allow for abrupt changes ($p_c = 0$), the two definitions of

Shannon surprise are identical: $\mathsf{S}_{\text{Sh1}} = \mathsf{S}_{\text{Sh2}}$ (Figure 2.4B).

Proposition 2 shows that the Bayes Factor surprise S_{BF} is related to S_{Sh1} and S_{Sh2}:

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

$$\begin{aligned} \mathsf{S}_{\text{BF}}(y_{t+1}|x_{t+1}; b^{(t)}) &= \frac{(1-p_c)e^{\Delta\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})}}{1-p_c e^{\Delta\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})}} \\ &= e^{\Delta\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};b^{(t)})}, \end{aligned} \tag{2.12}$$

where

$$\Delta\mathsf{S}_{\text{Sh}i}(y_{t+1}|x_{t+1}; b^{(t)}) \coloneqq \mathsf{S}_{\text{Sh}i}(y_{t+1}|x_{t+1}; b^{(t)}) - \mathsf{S}_{\text{Sh}i}(y_{t+1}|x_{t+1}; b^{(0)}) \tag{2.13}$$

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

Proposition 2 states that the Bayes Factor $\mathsf{S}_{\text{BF}}(y_{t+1}|x_{t+1}; b^{(t)})$ has a behavior similar to the *difference* in Shannon surprise (i.e., $\Delta\mathsf{S}_{\text{Sh1}}$ or $\Delta\mathsf{S}_{\text{Sh2}}$) as opposed to Shannon surprise itself (i.e., S_{Sh1} or S_{Sh2}). The difference in Shannon surprise (i.e., $\Delta\mathsf{S}_{\text{Sh1}}$ or $\Delta\mathsf{S}_{\text{Sh2}}$) 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., $\Delta\mathsf{S}_{\text{Sh1}}$ or $\Delta\mathsf{S}_{\text{Sh2}}$).

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

$$\begin{aligned} \gamma_{t+1} &= p_c \exp\Big(\Delta\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; b^{(t)})\Big) \\ \gamma_{t+1} &= \text{Sigmoid}\Big(\tilde{m}\Delta\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1}; b^{(t)})\Big), \end{aligned} \tag{2.14}$$

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

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 (Figure 2.4B):

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* $b^{(0)}$ *(i.e., if* $P(y_{t+1}|x_{t+1}; b^{(0)})$ *is uniform), then there are strictly increasing mappings between* $\mathsf{S}_{\text{BF}}(y_{t+1}|x_{t+1}; b^{(t)})$*,* $\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; b^{(t)})$*, and* $\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1}; b^{(t)})$*.*

A consequence of Corollary 2 is that experiments with flat marginal priors of the agent cannot be used to distinguish S_{BF} from S_{Sh1} or S_{Sh2} (Figure 2.4).

2.4.3 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. Figure 2.2E) (Sutton and 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. Equation 2.9 and Equation 2.11), we give also two definitions for SPE:

$$\begin{aligned}\mathsf{S}_{\text{SPE1}}(y_{t+1}|x_{t+1};b^{(t)}) :=&1-\mathbb{P}^{(t)}(y_{t+1}|x_{t+1})\\ =&1-\Big(p_c P(y_{t+1}|x_{t+1};b^{(0)})+(1-p_c)P(y_{t+1}|x_{t+1};b^{(t)})\Big),\end{aligned} \tag{2.15}$$

and

$$\begin{aligned}\mathsf{S}_{\text{SPE2}}(y_{t+1}|x_{t+1};b^{(t)}) :=&1-\mathbb{P}^{(t)}(y_{t+1}|x_{t+1},C_{t+1}=0)\\ =&1-P(y_{t+1}|x_{t+1};b^{(t)}).\end{aligned} \tag{2.16}$$

In non-volatile environments ($p_c = 0$), the two definitions of SPE are identical (Figure 2.4B). In particular, in an MDP without abrupt changes ($p_c = 0$; Figure 2.2E), 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 (S_{SPE1} and S_{SPE2}) can always be written as strictly increasing functions of Shannon surprise (Figure 2.4B):

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 $\mathsf{S}_{\text{SPE}i}(y_{t+1}|x_{t+1};b^{(t)})$, can be written as*

$$\mathsf{S}_{\text{SPE}i}(y_{t+1}|x_{t+1};b^{(t)}) =1-\exp\Big(-\mathsf{S}_{\text{Sh}i}(y_{t+1}|x_{t+1};b^{(t)})\Big). \tag{2.17}$$

Therefore, the SPE and the Shannon surprise are indistinguishable (Figure 2.4).

2.5 Observation-mismatch surprise measures

2.5.1 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) (Figure 2.3). 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. Equation 2.9 and Equation 2.11), we define two different predictions for the next observation y_{t+1} given the cue x_{t+1}[7]:

$$E_1[Y_{t+1}] :=p_c\mathbb{E}_{P(.|x_{t+1};b^{(0)})}[Y_{t+1}] + (1-p_c)\mathbb{E}_{P(.|x_{t+1};b^{(t)})}[Y_{t+1}] \tag{2.18}$$

and

$$E_2[Y_{t+1}] := \mathbb{E}_{P(.|x_{t+1};b^{(t)})}[Y_{t+1}]. \tag{2.19}$$

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, $E_2[Y_{t+1}]$ explains human participants' predictions better than $E_1[Y_{t+1}]$.

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

$$\begin{aligned} \mathsf{S}_{\mathrm{Ab},i}(y_{t+1}|x_{t+1};b^{(t)}) &:= ||y_{t+1} - E_i[Y_{t+1}]||_1 \\ \mathsf{S}_{\mathrm{Sq},i}(y_{t+1}|x_{t+1};b^{(t)}) &:= \Big(||y_{t+1} - E_i[Y_{t+1}]||_2\Big)^2, \end{aligned} \tag{2.20}$$

where $||.||_1$ and $||.||_2$ stand for the ℓ_1- and ℓ_2-norms (cf. Table 2.1), respectively, and E_1 and E_2 are defined in Equation 2.18 and Equation 2.19, 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 (Figure 2.4B).

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 $i \in \{1,2\}$,*

$$\mathsf{S}_{\mathrm{Ab}i}(y_{t+1}|x_{t+1};b^{(t)}) = 2\mathsf{S}_{\mathrm{SPE}i}(y_{t+1}|x_{t+1};b^{(t)}), \tag{2.21}$$

[7] The evaluation of the full distribution $P(.|x_{t+1};b^{(t)})$ 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).

and

$$\mathsf{S}_{\mathrm{Sq}i}(y_{t+1}|x_{t+1};b^{(t)}) = 2\mathsf{S}_{\mathrm{SPE}i}(y_{t+1}|x_{t+1};b^{(t)}) + \mathrm{Conf.}\Big[P(.|x_{t+1};b^{(t)})\Big], \tag{2.22}$$

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

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* $b^{(t)}$ *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* $\mathsf{S}_{\mathrm{Sq}2}(y_{t+1}|x_{t+1};b^{(t)})$ *is a strictly increasing function of* $\mathsf{S}_{\mathrm{Sh}2}(y_{t+1}|x_{t+1};b^{(t)})$*.*

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* $\mathsf{S}_{\mathrm{Sq}i} = \mathsf{S}^2_{\mathrm{Ab}i}$ *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 (Figure 2.4).

2.5.2 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 and Hall, 1980; Roesch et al., 2012; Rouhani and 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 and Barto, 2018)

$$\mathrm{uRPE}(s_t, a_t \to r_{t+1}) \coloneqq |r_{t+1} - Q^{(t)}(s_t, a_t)|, \tag{2.23}$$

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 $X_{t+1} \coloneqq (S_t, A_t)$ and $Y_{t+1} \coloneqq R_{t+1}$. Then, the unsigned reward prediction error uRPE$(s_t, a_t \rightarrow r_{t+1})$ is syntactically equal to S_{Ab} (cf. Equation 2.20; note that $E_1 = E_2$ since $p_c = 0$) and indistinguishable from S_{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 (Equation 2.23) (Hayden et al., 2011; Pearce and Hall, 1980; Roesch et al., 2012; Talmi et al., 2013) is a special case of the absolute and the squared error surprise (Equation 2.20).*

However, one can go beyond contextual bandit tasks and define uRPE for a general Markov Decision Process (MDP) (Sutton and 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

$$\tilde{R}_{t+1} \coloneqq R_{t+1} + \lambda V(S_{t+1}), \tag{2.24}$$

where $\lambda \in [0, 1)$ is the discount factor in infinite-horizon reinforcement learning (Sutton and 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. Equation 2.20), we give two definitions of uRPE:

$$\mathsf{S}_{\text{uRPE}i}(y_{t+1}|x_{t+1};b^{(t)}) \coloneqq |r_{t+1} + \lambda V(s_{t+1}) - Q_i^{(t)}(s_t, a_t)|, \tag{2.25}$$

where $i \in \{1, 2\}$ and $Q_i^{(t)}(s_t, a_t) \coloneqq E_i[\tilde{R}_{t+1}]$ (cf. Equation 2.18, Equation 2.19, and Equation 2.24). Equation 2.25 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 $\mathsf{S}_{\text{uRPE}i}(y_{t+1}|x_{t+1}; b^{(t)})$ can be written as*

$$\mathsf{S}_{\text{uRPE}i}(y_{t+1}|x_{t+1}; b^{(t)}) = \mathsf{S}_{\text{Ab}i}(y_{t+1}|x_{t+1}; b^{(t)}) - \mathsf{S}_{\text{Ab}i}(s_{t+1}|x_{t+1}; b^{(t)}) \tag{2.26}$$

and

$$\Big(\mathsf{S}_{\text{uRPE}i}(y_{t+1}|x_{t+1}; b^{(t)})\Big)^2 = \mathsf{S}_{\text{Sq}i}(y_{t+1}|x_{t+1}; b^{(t)}) - \mathsf{S}_{\text{Sq}i}(s_{t+1}|x_{t+1}; b^{(t)}). \tag{2.27}$$

where $\mathsf{S}_{\text{Ab}i}(s_{t+1}|x_{t+1}; b^{(t)}) \coloneqq ||s_{t+1} - E_i[S_{t+1}]||_1$ *and* $\mathsf{S}_{\text{Sq}i}(s_{t+1}|x_{t+1}; b^{(t)}) \coloneqq ||s_{t+1} -$

$E_i[S_{t+1}]||_2^2$ *(Equation 2.20).*

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 $\mathsf{S}_{\mathrm{Sq}i}$ or $\mathsf{S}_{\mathrm{Ab}i}$, similar to the experiment of Rouhani and Niv (2021)), then $\mathsf{S}_{\mathrm{uRPE}i}$ is indistinguishable from $\mathsf{S}_{\mathrm{Ab}i}$ and $\mathsf{S}_{\mathrm{Sq}i}$ (Figure 2.4).

2.6 Belief-mismatch surprise measures

2.6.1 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 and 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 $\Theta_1 = \Theta_2 = ... = \Theta_t = \Theta$. In this case, the Bayesian surprise of observing y_{t+1} with cue x_{t+1} is defined as $\mathrm{D}_{\mathrm{KL}}[\mathbb{P}_{\Theta}^{(t)}||\mathbb{P}_{\Theta}^{(t+1)}]$ (Baldi, 2002; Baldi and 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 2.1). Hence, in non-volatile environments, Bayesian surprise measures the pseudo-distance D_{KL} between two distributions, i.e., the belief $b^{(t)} = \mathbb{P}_{\Theta}^{(t)}$ before and the belief $b^{(t+1)} = \mathbb{P}_{\Theta}^{(t+1)}$ 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 $\mathbb{P}_{\Theta_{t+1}}^{(t+1)} = b^{(t+1)}$; however, it is unclear whether $\mathbb{P}_{\Theta}^{(t)}$ should be taken as the momentary belief $\mathbb{P}_{\Theta_t}^{(t)} = b^{(t)}$ 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:

$$b^{(t)} = \mathbb{P}_{\Theta_t}^{(t)} \neq \mathbb{P}_{\Theta_{t+1}}^{(t)} = p_c b^{(0)} + (1 - p_c) b^{(t)}. \tag{2.28}$$

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

$$\mathsf{S}_{\mathrm{Ba1}}(y_{t+1}|x_{t+1}; b^{(t)}) := \mathrm{D}_{\mathrm{KL}}\Big[p_c b^{(0)} + (1 - p_c) b^{(t)}||b^{(t+1)}\Big], \tag{2.29}$$

and

$$\mathsf{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1}; b^{(t)}) := \mathrm{D}_{\mathrm{KL}}\Big[b^{(t)}||b^{(t+1)}\Big]. \tag{2.30}$$

The first definition is more consistent with the original definition of the Bayesian surprise

(Baldi, 2002; Baldi and 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 Equation 2.29 and Equation 2.30, 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 $b^{(t+1)}$. For the case of $p_c = 0$, the two definitions are identical (Figure 2.4B).

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*

$$\begin{aligned}\mathsf{S}_{\text{Ba1}}(y_{t+1}|x_{t+1};b^{(t)}) =& p_c\mathbb{E}_{b^{(0)}}\Big[\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big]+ \\ & (1-p_c)\mathbb{E}_{b^{(t)}}\Big[\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big]- \\ & \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)}),\end{aligned} \tag{2.31}$$

and

$$\begin{aligned}\mathsf{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};b^{(t)}) =& \mathbb{E}_{b^{(t)}}\Big[\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big]- \\ & \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})+ \\ & \text{D}_{\text{KL}}\Big[b^{(t)}||p_c b^{(0)} + (1-p_c)b^{(t)}\Big],\end{aligned} \tag{2.32}$$

where $\delta_{\{\theta\}}$ is the Dirac measure at θ (cf. Table 2.1).

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.,*

$$\mathsf{S}_{\text{Ba}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathbb{E}_{b^{(t)}}\Big[\mathsf{S}_{\text{Sh}}(y_{t+1}|x_{t+1};\delta_{\{\Theta\}})\Big] - \mathsf{S}_{\text{Sh}}(y_{t+1}|x_{t+1};b^{(t)}), \tag{2.33}$$

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

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 $b^{(t)}$ for computing the expectation (Figure 2.3).

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).

2.6.2 Postdictive surprise

We saw that the Bayesian surprise measures how much the new belief $b^{(t+1)}$ 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 $P(y|x_{t+1}; b^{(t+1)})$ (cf. Equation 2.4). 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 (Equation 2.29 and Equation 2.30), there are two definitions for the postdictive surprise in volatile environments:

$$\begin{aligned} \mathsf{S}_{\text{Po1}}(y_{t+1}|x_{t+1}; b^{(t)}) &:= \\ & D_{\text{KL}}\Big[p_c P(.|x_{t+1}; b^{(0)}) + (1-p_c) P(.|x_{t+1}; b^{(t)}) || P(.|x_{t+1}; b^{(t+1)})\Big], \end{aligned} \tag{2.34}$$

and

$$\mathsf{S}_{\text{Po2}}(y_{t+1}|x_{t+1}; b^{(t)}) := D_{\text{KL}}\Big[P(.|x_{t+1}; b^{(t)}) || P(.|x_{t+1}; b^{(t+1)})\Big], \tag{2.35}$$

where the dot refers to a dummy variable y that is integrated out when evaluating D_{KL} (cf. Table 2.1). Note that for $p_c = 0$, the two definitions are identical (Figure 2.4B).

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

2.6.3 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 b_{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

$$b_{\text{flat}}(\theta|y_{t+1}, x_{t+1}) := \frac{P_{Y|X}(y_{t+1}|x_{t+1}; \theta) b_{\text{flat}}(\theta)}{P(y_{t+1}|x_{t+1}; b_{\text{flat}})} = \frac{P_{Y|X}(y_{t+1}|x_{t+1}; \theta)}{\int P_{Y|X}(y_{t+1}|x_{t+1}; \theta) d\theta}. \tag{2.36}$$

If the prior $b^{(0)}$ is equal to b_{flat} (i.e., if the prior is uniform), then $b_{\text{flat}}(\theta|y_{t+1}, x_{t+1})$ is the same as $b_{\text{reset}}^{(t+1)}(\theta)$ defined in Proposition 1. Note that the prior b_{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 $b_{\text{flat}}(.|y_{t+1}, x_{t+1})$ is a proper distribution (Efron and Hastie, 2016). Using this terminology, the original definition for the Confidence Corrected surprise is (Faraji et al., 2018)

$$\mathsf{S}_{\text{CC1}}(y_{t+1}|x_{t+1}; b^{(t)}) \coloneqq \text{D}_{\text{KL}}\Big[b^{(t)}||b_{\text{flat}}(.|y_{t+1}, x_{t+1})\Big]. \tag{2.37}$$

To interpret S_{CC1}, Faraji et al. (2018) defined the commitment (or confidence) $C[b]$ corresponding to an arbitrary belief b as its negative entropy (Cover, 1999), i.e.,

$$C[b] \coloneqq \mathbb{E}_b\Big[\log b(\Theta)\Big]. \tag{2.38}$$

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

$$\begin{aligned}\mathsf{S}_{\text{CC1}}(y_{t+1}|x_{t+1}; b^{(t)}) =& \mathsf{S}_{\text{Sh}}(y_{t+1}|x_{t+1}; b^{(t)}) + \\ & \mathsf{S}_{\text{Ba}}(y_{t+1}|x_{t+1}; b^{(t)}) + \\ & C\big[b^{(t)}\big] - A(y_{t+1}, x_{t+1}),\end{aligned} \tag{2.39}$$

where $A(y_{t+1}, x_{t+1}) \coloneqq \mathsf{S}_{\text{Sh}}(y_{t+1}|x_{t+1}; b_{\text{flat}}) + C[b_{\text{flat}}]$ is independent of the current belief $b^{(t)}$. Note that because $p_c = 0$, we have $\mathsf{S}_{\text{Sh1}} = \mathsf{S}_{\text{Sh2}}$ and $\mathsf{S}_{\text{Ba1}} = \mathsf{S}_{\text{Ba2}}$. Therefore, in a non-volatile environment (i.e., $p_c = 0$), S_{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 Equation 2.39 must be replaced by Proposition 9 below.

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

$$\mathsf{S}_{\text{CC2}}(y_{t+1}|x_{t+1}; b^{(t)}) \coloneqq \text{D}_{\text{KL}}\Big[b^{(t)}||b_{\text{reset}}^{(t+1)}\Big], \tag{2.40}$$

where $b_{\text{reset}}^{(t+1)}(\theta)$ is defined in Proposition 1. Whenever $b^{(0)} = b_{\text{flat}}$, the two definitions are identical (Figure 2.2B). 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*

$$\begin{aligned}\mathsf{S}_{\mathrm{CC1}}(y_{t+1}|x_{t+1};b^{(t)}) = &\ \mathsf{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};b^{(t)}) - \mathsf{S}_{\mathrm{Sh2}}(y_{t+1}|x_{t+1};b_{\mathrm{flat}})\\ &+ \mathsf{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1};b^{(t)})\\ &- \mathrm{D}_{\mathrm{KL}}\Big[b^{(t)}||p_c b^{(0)} + (1-p_c)b^{(t)}\Big]\\ &+ C\big[b^{(t)}\big] - C\big[b_{\mathrm{flat}}\big],\end{aligned} \tag{2.41}$$

and our 2nd definition can be written as

$$\begin{aligned}\mathsf{S}_{\mathrm{CC2}}(y_{t+1}|x_{t+1};b^{(t)}) = &\ \Delta\mathsf{S}_{\mathrm{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})\\ &+ \mathsf{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1};b^{(t)})\\ &- \mathrm{D}_{\mathrm{KL}}\Big[b^{(t)}||p_c b^{(0)} + (1-p_c)b^{(t)}\Big]\\ &+ \mathrm{D}_{\mathrm{KL}}\Big[b^{(t)}||b^{(0)}\Big].\end{aligned} \tag{2.42}$$

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 Equation 2.41 and Equation 2.42). 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 Equation 2.41 and Equation 2.42). 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 Equation 2.41 and Equation 2.42). 'Relative confidence' quantifies how different the current belief is with respect to a reference belief; note that $C\big[b^{(t)}\big] - C\big[b_{\mathrm{flat}}\big] = \mathrm{D}_{\mathrm{KL}}\big[b^{(t)}||b_{\mathrm{flat}}\big]$.

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 $\mathsf{S}_{\mathrm{CC2}}$ is identical to $\mathsf{S}_{\mathrm{Ba2}}$ 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 $\mathsf{S}_{\mathrm{CC2}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathsf{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1};b^{(t)})$.*

2.6.4 Minimized free energy

Although an agent can perform computations over the joint probability distribution in Equation 2.1 and Equation 2.2, finding the belief $b^{(t+1)}(\theta)$ (i.e., the posterior distribution in Equation 2.3) 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 and 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 $b^{(t+1)}(\theta)$ 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, 2019b).

Formally, in variational inference, the belief $b^{(t+1)}(\theta)$ is approximated by $\hat{b}^{(t+1)}(\theta) := q(\theta;\phi^{(t+1)})$, 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:

$$\mathbb{P}_{\Theta_{t+1}}(\theta, y_{t+1}|x_{t+1}; b) := P_{Y|X}(y_{t+1}|x_{t+1};\theta)\Big(p_c b^{(0)}(\theta) + (1-p_c)b(\theta)\Big), \tag{2.43}$$

where b 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 $\mathbb{P}_{\Theta_{t+1}}(\theta_{t+1}, y_{t+1}|x_{t+1}; b^{(t)})$ and the updated belief $b^{(t+1)}(\theta)$ as $\mathbb{P}_{\Theta_{t+1}}(\theta|y_{t+1}, x_{t+1}; b^{(t)})$. The variational loss or free energy can then be defined as (Liakoni et al., 2021; Markovic et al., 2021; Sajid et al., 2021)

$$F^{(t+1)}(\phi) := \mathbb{E}_{q(.;\phi)}\Big[\log q(\Theta;\phi) - \log \mathbb{P}_{\Theta_{t+1}}(\Theta, y_{t+1}|x_{t+1}; \hat{b}^{(t)})\Big]. \tag{2.44}$$

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

$$\begin{aligned} F^{(t+1)}(\phi) =& \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; \hat{b}^{(t)}) + \mathrm{D}_{\text{KL}}\Big[q(.;\phi)||\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{b}^{(t)})\Big] \\ \geq& \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; \hat{b}^{(t)}), \end{aligned} \tag{2.45}$$

where the right side of the inequality is independent of ϕ, and $\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{b}^{(t)})$ is the exact Bayesian update of the belief (according to the generative model in Definition 1) given the latest approximation of the belief $\hat{b}^{(t)}$ (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 Equation 2.45, can be seen as an approximation of $\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; \hat{b}^{(t)})$. The parametric family of $q(.;\phi)$ and its relation to the exact belief $b^{(t+1)}$ determine how well F^* approximates $\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; \hat{b}^{(t)})$ (Figure 2.4B). More precisely, the minimized free energy measures both how unlikely the new observation is (i.e., how large $\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1}; \hat{b}^{(t)})$ is) and how imprecise the best parametric approximation of the belief $\hat{b}^{(t+1)}$ is (i.e., how large $\mathrm{D}_{\text{KL}}[\hat{b}^{(t+1)}||\mathbb{P}_{\Theta_{t+1}}(.|y_{t+1}, x_{t+1}; \hat{b}^{(t)})]$ is). Therefore, the minimized free energy is in the category of belief-mismatch surprise measures (Figure 2.3).

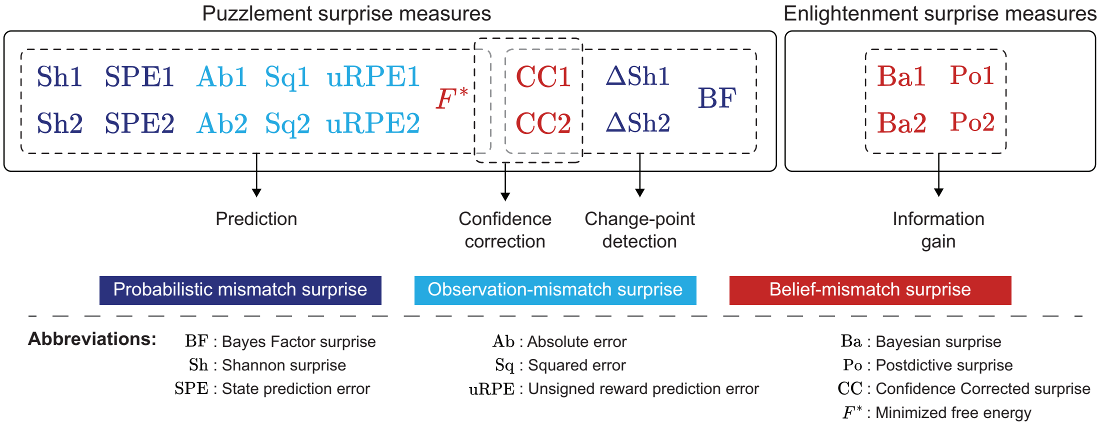

Puzzlement surprise measures
Enlightenment surprise measures
Sh1 SPE1 Ab1 Sq1 uRPE1
Sh2 SPE2 Ab2 Sq2 uRPE2
F^*
CC1
CC2
ΔSh1
ΔSh2
BF
Ba1 Po1
Ba2 Po2
Prediction
Confidence correction
Change-point detection
Information gain
Probabilistic mismatch surprise
Observation-mismatch surprise
Belief-mismatch surprise
Abbreviations:
BF : Bayes Factor surprise
Sh : Shannon surprise
SPE : State prediction error
Ab : Absolute error
Sq : Squared error
uRPE : Unsigned reward prediction error
Ba : Bayesian surprise
Po : Postdictive surprise
CC : Confidence Corrected surprise
F^* : Minimized free energy

Figure 2.5: **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 2.3, 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 Figure 2.3.

2.7 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 (Figure 2.5).

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 and Beeman, 2009). The Bayesian and the Postdictive surprise can be categorized as enlightenment surprise since both quantify information gain (Figure 2.5). Based on our theoretical analyses, however, we suggest to further divide measures of puzzlement surprise into 3 sub-categories (Figure 2.5):

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 (Figure 2.5). 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, 2021b). 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; Figure 2.5). According to these measures, the agent's puzzle is to detect environmental changes[8].

iii. 'Confidence correction 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 S_{CC1} and S_{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[9].

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 and Krajbich, 2018; Kopp and Lange, 2013; Loued-Khenissi and 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.,

[8] **Thesis footnote:** Accordingly, the change-point detection surprise definitions present a quantitative explanation for the locus coeruleus activity, which has been interpreted as a global model failure signal that modulates learning throughout the cortex (Jordan, 2023).

[9] **Thesis footnote:** After the publication of these results, we encountered further instances of confidence-corrected surprise in the psychology literature (Macedo and Cardoso, 2019; Macedo et al., 2004; Reisenzein et al., 2019; Teigen and Keren, 2003). In these studies, confidence is (implicitly) delineated as $\max_y P(y|x_{t+1}; b^{(t)})$, implying that an agent is confident if it assigns a high probability to a specific observation. It should be noted that $\max_y P(y|x_{t+1}; b^{(t)})$ bounds the definition of confidence as the negative entropy of $P(.|x_{t+1}; b^{(t)})$; this latter definition is similar to the confidence definition in Equation 2.38 but in the observation space.

2018; Ostwald et al., 2012; O'Reilly et al., 2013; Visalli et al., 2021) surprise measures (Figure 2.1). 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[10].

2.8 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 $b^{(t)}$, we divided them into three groups of probabilistic mismatch, observation-mismatch, and belief-mismatch surprise measures (Figure 2.3). 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 – Figure 2.4 and Table 2.2). We further proposed a taxonomy of surprise definitions by a conceptual classification into four main categories (Figure 2.5): (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 and Griffiths, 2020; Gottlieb and Oudeyer, 2018), memory formation (Rouhani and 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. Figure 2.4B, 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

[10] **Thesis footnote:** See chapter 4 for further discussions and links to physiological measurements.

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, S_{BF} and S_{Sh1} are distinguishable for any prior marginal distributions other than uniform distribution (Figure 2.4B), 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.

3 Novelty is not Surprise: Human exploratory and adaptive behavior in sequential decision-making

This chapter was published in PLoS Computational Biology (Xu et al., 2021)[1].

Authors: He A. Xu*, Alireza **Modirshanechi***, Marco P. Lehmann, Wulfram Gerstner**, and Michael H. Herzog**

*: HA Xu and A Modirshanechi are joint first authors.
**: W Gerstner and MH Herzog are joint senior authors.

Abstract: Classic reinforcement learning (RL) theories cannot explain human behavior in the absence of external reward or when the environment changes. Here, we employ a deep sequential decision-making paradigm with sparse reward and abrupt environmental changes. To explain the behavior of human participants in these environments, we show that RL theories need to include surprise and novelty, each with a distinct role. While novelty drives exploration before the first encounter of a reward, surprise increases the rate of learning of a world-model as well as of model-free action-values. Even though the world-model is available for model-based RL, we find that human decisions are dominated by model-free action choices. The world-model is only marginally used for planning, but it is important to detect surprising events. Our theory predicts human action choices with high probability and allows us to dissociate surprise, novelty, and reward in EEG signals.

Funding: This research was supported by the Swiss National Science Foundation No. CRSII2 147636 (Sinergia, MHH and WG) and No. 200020 184615 (WG) and by the European Union Horizon 2020 Framework Program under grant agreement No. 785907 (Human Brain Project, SGA2, MHH and WG).

Author contribution: MHH, HAX, and WG designed the experiment. HAX collected the data. AM designed the model and performed the data analysis. HAX and MPL developed preliminary models and data analysis. AM, HAX, MHH, and WG wrote the paper.

[1] For consistency across the thesis chapters, the mathematical notation has been slightly adjusted.

3.1 Introduction

Humans seek not only explicit rewards such as money or praise (Daw et al., 2011, 2005; Gläscher et al., 2010; Lehmann et al., 2019; O'Doherty et al., 2003; Pessiglione et al., 2006; Schultz et al., 1997; Wunderlich et al., 2012) but also novelty (Gershman and Niv, 2015; Jaegle et al., 2019), an intrinsic reward-like signal which is linked to curiosity (Dubey and Griffiths, 2019; Gershman and Niv, 2015; Gottlieb et al., 2013; Jaegle et al., 2019; Niv and Langdon, 2016; Schmidhuber, 2010; Singh et al., 2010a). In the theory of reinforcement learning, novelty is considered as a drive for exploration (Bellemare et al., 2016; Chentanez et al., 2005; Martin et al., 2017; Schmidhuber, 2010), and novelty-driven exploratory actions have been interpreted as steps towards building a model of the world ('world-model') which is then used for action planning (Sutton and Barto, 2018). A world-model represents implicit knowledge that links actions to observations, such as 'if I open the door to my kitchen, I will see my fridge'.

However, since the world is much more complex than any model of it, there will occasionally be a mismatch between the expectations arising from the model and the actual observation, e.g., when you return from work and the location of the fridge is suddenly empty because your room-mate has sent it off for repair. Such mismatches generate the feeling of surprise, known to manifest in pupil dilation (Nassar et al., 2012) and EEG signals (Maheu et al., 2019; Modirshanechi et al., 2019; Ostwald et al., 2012). Whereas the reward prediction error (RPE) is a mismatch between the expected reward and the actual reward, surprise is a mismatch between an expected observation and an actual observation. Behavioral experiments (Behrens et al., 2007; Heilbron and Meyniel, 2019; Nassar et al., 2012, 2010; Soltani and Izquierdo, 2019) and theories (Faraji et al., 2018; Findling et al., 2021; Liakoni et al., 2021; Soltani and Izquierdo, 2019) suggest that surprise helps humans to adapt their behavior quickly to changes in the environment, potentially by modulating synaptic plasticity (Gerstner et al., 2018; Yagishita et al., 2014; Yu and Dayan, 2005).

Surprise is fundamentally different from novelty; if you already know that your fridge would be fetched for repairing, the new arrangement of the kitchen without the fridge is novel but not surprising. However, although there is some agreement that novelty and surprise are two separate notions, it has been debated how they can be formally distinguished (Barto et al., 2013; Dubey and Griffiths, 2020; Hurley et al., 2011; Palm, 2012), whether they manifest themselves differently in EEG signals (Gijsen et al., 2021; Maheu et al., 2019; Mars et al., 2008; Modirshanechi et al., 2019), and how they influence learning and decision-making (Behrens et al., 2007; Dubey and Griffiths, 2019, 2020; Gershman and Niv, 2015; Gottlieb et al., 2013; Heilbron and Meyniel, 2019; Jaegle et al., 2019; Juechems and Summerfield, 2019; Moens and Zénon, 2019; Nassar et al., 2012, 2010; Schmidhuber, 2010).

In this study, we address three questions: First, how do surprise and novelty influence human reinforcement learning? Second, what is their relative contribution to exploratory and adaptive behavior? And third, can surprise be distinguished from novelty in human behavioral choices and event related potentials (ERP) of the electroencephalogram (EEG)? We show, via a specifically designed deep sequential decision task and a novel hybrid reinforcement learning model, that we can dissociate contributions of surprise from those of novelty and reward in human behavior and ERP.

Our key findings can be summarized in three points: (i) We find that novelty-seeking explains participants' exploratory behavior better than alternative exploration strategies such as seeking surprise or uncertainty (Achiam and Sastry, 2017; Burda et al., 2019); (ii) we observe that participants use their world-model only rarely for action planning and mainly to extract moments of surprise; and importantly, (iii) we show that surprise calculated by the world-model does not only modulate the learning of the world-model (Behrens et al., 2007; Heilbron and Meyniel, 2019; Liakoni et al., 2021; Nassar et al., 2010) but also the learning of model-free action-values. In particular, we show that such a modulation is necessary to explain participants' adaptive behavior.

3.2 Results

3.2.1 Experimental paradigm and human behavior

In order to distinguish between novelty, surprise, and reward, and to study their effects on exploratory and adaptive behavior, we designed an environment (cf. Tartaglia et al. (2017)) consisting of 10 states with 4 possible actions per state plus one goal state (Figures 3.1A and 3.1B). In the human experiments, states were represented as images on a computer screen and actions as four grey disks below the image. Before the experiment, 12 participants were shown all images of the states and were informed that their task was to find the shortest path to the goal image. Throughout the experiment, at each state, participants chose an action (by clicking on one of the grey disks) which brought them to the next image, where they then chose the next action, and so on (Figure 3.1A). Such an episode ended when the goal image was found.

Unknown to the participants, the non-goal states could be classified into the progressing states (1 to 7 in Figure 3.1B) and the trap states (8 to 10 in Figure 3.1B). At each progressing state, one action ('good' action) either brought participants to another progressing state closer to the goal or led them directly to the goal, two actions ('bad' actions) brought them to one of the trap states, and one action ('neutral' action) made them stay at the current state. At each trap state, three actions brought participants to either the same or another trap state, and one action brought them to state 1, at the beginning of the path of progressing states. The assignment of action buttons to specific transitions was random and not the same for different states, e.g., in state 1, the neutral action is action 2, whereas in state 3, the neutral action is action 3 (Figure 3.1B). Note

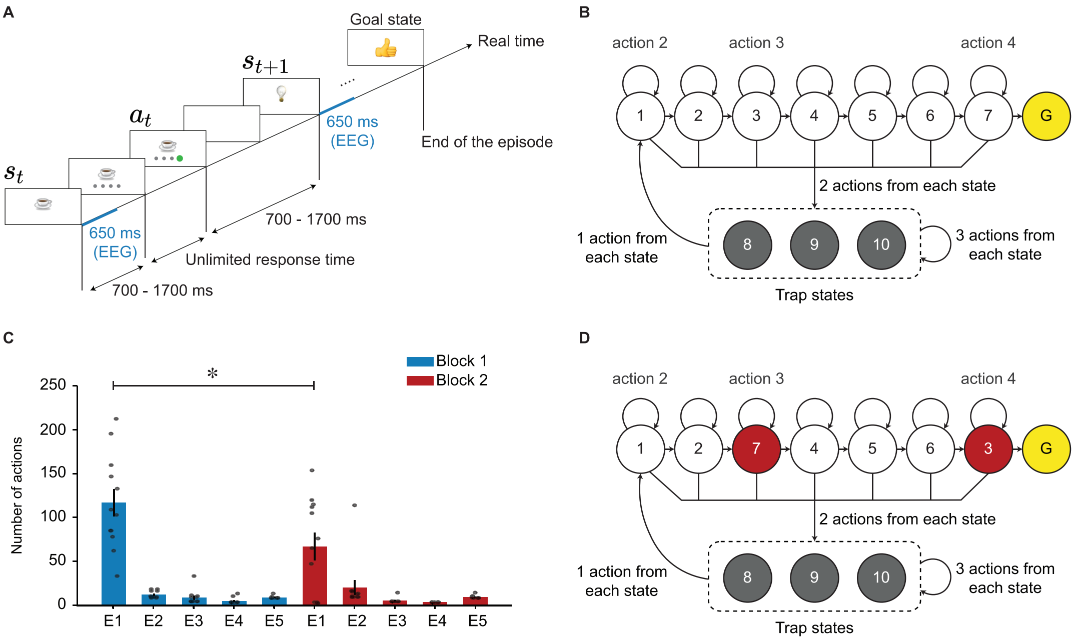


Figure 3.1: **Experimental paradigm. A.** After image onset, participants had to wait for 700-1700ms (randomly chosen) until four grey disks were presented at the bottom of the image. After clicking on one disk, a blank screen was presented for another random interval of 700 to 1700ms. The next image appeared afterwards. Different participants saw different images, but the underlying structure was identical for all participants. The goal image is a 'thumb-up' image in this example. The blue lines indicate the window of EEG analysis. **B.** Structure of the environment during block 1. There were 10 states with 4 actions each plus a goal state (G). States 1-7 are *progressing states* and states 8-10 are *trap states*. For each progressing state, one action led participants to the next progressing state, two actions led participants to one of the trap states, and one action made participants stay at the current state. The action which made participants stay at the current state is shown for states 1, 3, and 7, as an example. For each trap state, three actions led participants to one of the trap states, and one action led participants to state 1. Not all action arrows are drawn for the trap states to simplify illustration. **C.** Average number of actions of participants during block 1 (blue) and block 2 (red): The 1st episode of block 2 was significantly shorter than the 1st episode of block 1 (one-sample t-test, p-value=0.035). Error bars show the standard error of the mean, and each grey point shows the data of one participant. **D.** Environment used in block 2: The images presenting state 3 and state 7 (in red) were swapped. Other transitions remained unchanged.

that the underlying structure of the environment, the assignment of images to specific states, and the assignment of action buttons to specific transitions were unknown to the participants. We also did not tell the participants whether or not transitions were deterministic (i.e., whether the same action from a certain state always led to the same next state).

The experiment was organized in 10 episodes, i.e., it ended after the 10th time that participants found the goal state. Unknown to the participants, we divided these 10 episodes into 2 blocks of 5 episodes each; we refer to the first 5 episodes as block 1 and to the second 5 episodes as block 2. During the 1st episode of block 1, participants took between 34 and 214 actions (mean 118 and std 54) until they arrived at the goal (Figure 3.1C). They then continued for another 4 episodes, each time starting in a new initial state. The initial state for each episode was chosen randomly, but it was kept fixed across participants. After the 1st episode, participants had learnt to reach the goal in less than 20 steps (episodes 2 to 5 in Figure 3.1C). After the end of the 5th episode (the end of block 1), two states (state 3 and 7 in Figure 3.1D) were swapped, without announcing it to the participants. Participants continued for another 5 episodes with the novel layout of the environment (2nd block, Figure 3.1D).

In the 1st episode of block 1, participants explored the environment to find the goal, but they received no intermediate reward or other sign of progress while doing this. If participants followed a purely random exploration (i.e., choosing each action with 1/4 probability), it would take them on average about 10^4 actions to find the goal, starting at any non-goal state (see section B.4). This high number is an indication of the complexity and depth of our environment. Our results suggest that participants followed a non-random strategy for finding the goal (Figure 3.1C). With increasing experience, the latency of escape from the trap states was reduced (Figure 3.2A) and the good actions at progressing states were chosen with higher probability (Figure 3.2B). It is important to note that these improvements were observed in the absence of any external feedback indicating progress and before the 1st encounter of the goal state. Here, we ask whether novelty of states played a role in the way participants chose their actions and searched for the goal.

In the 1st episode of block 2, when states 3 and 7 had been swapped, participants spent a great amount of time (68 ± 16 actions on average) re-exploring the environment and searching for the goal state, but they were significantly faster in finding the goal than in the 1st episode of block 1 (Figure 3.1C). After the swap, participants continued escaping from the trap states (Figure 3.2A) and choosing the good actions at the unchanged progressing states (states 1, 2, and 4 in Figure 3.2C). Moreover, they rapidly adapted their behavior and found the new good actions at the swapped states (state 7 in Figure 3.2C). Our results indicate that participants adapted their behavior to the new situation while exploiting the knowledge they had acquired before. This observation suggests that surprise triggered by unexpected transitions helped participants to rapidly adapt their behavior. Here, we ask how surprise affects participants' adaptive behavior.

3.2.2 Defining Novelty and Surprise

In Oxford-English-Dictionary (2021a), novelty is defined as 'the quality or state of being new, original, or unusual'. Here, our focus is on the quality of being 'unusual', and by saying that a state is novel, we mean that it has not been encountered often, i.e., it is not 'usual' to encounter this state. We, therefore, assume that (i) the novelty of a state s at time t is a decreasing function of the number $C_s^{(t)}$ of encounters of state s

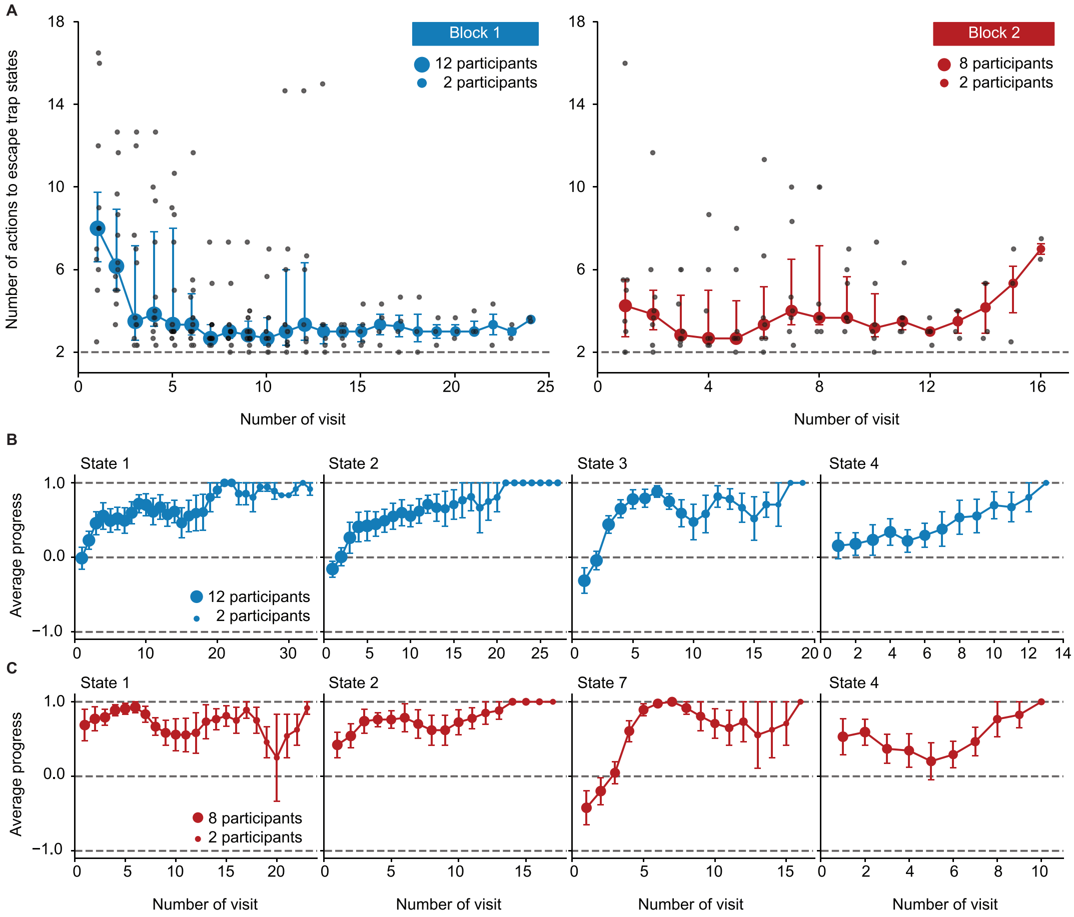


Figure 3.2: **Behavioral results for episode 1 of blocks 1 and 2.** **A.** Escape from the trap states: Median number of actions of participants between falling into a trap state and reaching state 2 in episode 1 of block 1 (left) and block 2 (right). Error bars show the 25% and 75% quantiles, and each grey point shows the data of one participant. The grey dashed lines correspond to the minimum number of actions (2) that are needed to escape the trap states. x-axis shows the number of visits of the trap states, for example, 10 means the 10th times participants fall from a progressing state into the trap states. Because of between-participant differences, not all participants visited the trap states for, e.g., 20 times. The size of circles indicates number of participants over which the average is taken. In the 1st episode of block 2 (right), four participants reached the goal state without falling into the trap states; thus, only the data for the other 8 participants is shown. A moving average of length three was applied to the data. **B.** Average progress of participants each time visiting states 1, 2, 3, and 4 in episode 1 of block 1. We assign a progress value of 1 to good actions (the ones taking participants closer to the goal), 0.5 to neutral actions (the ones making participants stay where they are), and -0.75 to bad actions (the ones taking participants to the trap states); with this assignment, average progress vanishes for random exploration. The size of circles shows the number of participants over which the average is taken, and error bars show the standard error of the mean. A moving average of length three was applied to the data. **C.** Average progress of participants each time visiting states 1, 2, 7 (swapped with 3), and 4 in episode 1 of block 2. See Figure B.3A for the average progress at the progressing states in the proximity of the goal.

until time t, e.g., a state that has been encountered 5 times is less novel than a state that has been encountered only once. Moreover, we assume that (ii) a state s that has been encountered for example $C_s^{(t)} = 5$ times within a total of $t = 5$ trials is less novel compared to a state s' that has been encountered for example $C_{s'}^{(t')} = 5$ times within a total of $t' = 500$ trials. Following assumptions (i) and (ii), we define the novelty of a state s as a decreasing function of the observation frequency

$$p_f^{(t)}(s) = \frac{C_s^{(t)} + 1}{\left(\sum_{s'} C_{s'}^{(t)}\right) + 11}. \tag{3.1}$$

$p_f^{(t)}(s)$ has two different interpretations. First, it can be seen as the empirical frequency of observing state s_t until time t. In fact, because one of the counters $C_{s'}$ increases by one at each time step, the time can be expressed as $t = \sum_{s'} C_{s'}^{(t)}$. In this interpretation, the numbers 1 in the numerator and 11 (11 is the total number of states in the environment) in the denominator correspond to the one encounter of each state before the start of the experiment. In the second interpretation, $p_f^{(t)}(s)$ can be seen as the probability of observing state s at time t, estimated in a Bayesian framework and with the assumption of independence between observations (see section B.1); measures similar to $p_f^{(t)}(s)$, sometimes called 'density models', have been used in machine learning, for example, to quantify how frequently an image has been observed (Bellemare et al., 2016). In the Bayesian interpretation, the numbers 1 in the numerator and 11 in the denominator correspond to a uniform prior that makes all states equally likely at time $t = 0$.

We define the novelty of state s at time t as

$$\mathsf{N}^{(t)}(s) = -\log p_f^{(t)}(s). \tag{3.2}$$

Consistent with the literature (Barto et al., 2013; Shannon, 1948; Tribus, 1961), we chose the logarithm in order to smooth out temporal fluctuations and compress differences in the novelty of frequent versus infrequent states (Figure 3.3A). Since our novelty measure depends on the frequencies (relative counts, $p_f^{(t)}(s)$ in Equation 3.1) rather than the raw counts ($C_s^{(t)}$), one may also interpret $\mathsf{N}^{(t)}(s)$ as a measure of 'relative novelty' or 'rareness'. See Discussion for the relation of our measure of novelty to other measures[2].

With our definition of novelty, at the beginning of the 1st episode in block 1, all states have identical novelty. Since participants often fall into one of the trap states, the novelty of the trap states decreases rapidly (Figure 3.3A). Hence, before the end of the 1st episode, the novelty is highest for states in the proximity of the goal (Figure 3.3B). This observation suggests that seeking novel states will, in our environment, effectively lead a participant closer to the goal, even *before* the participant knows where the goal is located, i.e., before encountering the goal for the first time. We conclude that novelty is

[2] **Thesis footnote:** See chapter 4 for further discussions on how the definition in Equation 3.2 links to physiological signatures and also other definitions of novelty.

potentially an important signal and will exploit this insight further below.

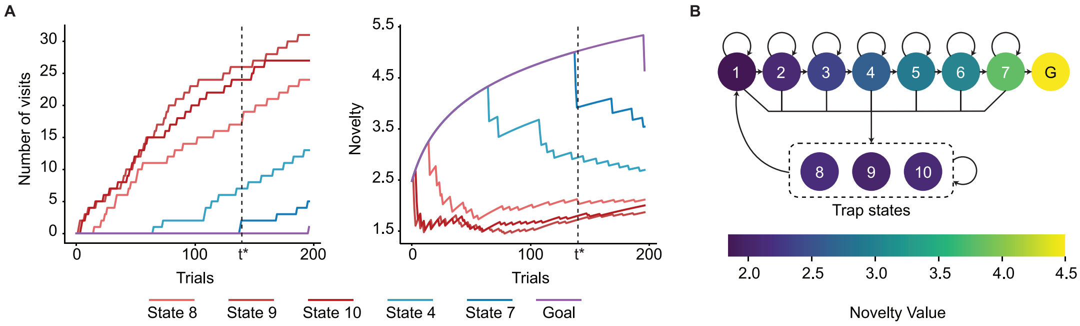


Figure 3.3: **Novelty in episode 1 of block 1. A.** The number of state visits (left panel) and novelty (right panel) as a function of time for one representative participant: The number of visits increases rapidly for the trap states and remains 0 for a long time for the states closer to the goal. Novelty of each state is defined as the negative log-probability of observing that state (see Equation 3.2 and Equation 3.1) and, hence, increases for states which are not observed as time passes. The first time participants encounter state 7 (the state before the goal state) is denoted by t^*. **B.** Average (over participants) novelty (color coded) at t^*: Novelty of each state is a decreasing function of its distance from the goal state.

Surprise is defined in the Oxford-English-Dictionary (2021b) as 'the feeling or emotion excited by something unexpected' or 'the feeling or mental state, akin to astonishment and wonder, caused by an unexpected occurrence or circumstance'. Whereas novelty is about being unusual, surprise is about being unexpected. Following this intuition, we define surprise as a measure expressing how 'unexpected' the next image (state s_{t+1}) is given the previous state s_t and the chosen action a_t. To quantify expectations, we assume that participants build an internal model of the environment ('world-model'), i.e., we hypothesize that participants estimate the probability $p^{(t)}(s_{t+1}|s_t, a_t)$ of a transition from a given state s_t to another state s_{t+1} when performing action a_t. More precisely, we assume that the world-model counts transitions from state s to s' under action a using either a leaky (Meyniel et al., 2016; Modirshanechi et al., 2019; Yu and Cohen, 2009) or a surprise-modulated (Faraji et al., 2018; Liakoni et al., 2021; Markovic et al., 2021) counting procedure, described by the pseudo-count $\tilde{C}^{(t)}_{s,a\to s'}$. The conditional probability is then

$$p^{(t)}(s_{t+1}|s_t, a_t) = \frac{\tilde{C}^{(t)}_{s_t,a_t\to s_{t+1}} + \epsilon}{\tilde{C}^{(t)}_{s_t,a_t} + 11\epsilon}, \tag{3.3}$$

where ϵ is a parameter corresponding to a prior in the Bayesian framework, 11 is the total number of states in the environment, and $\tilde{C}^{(t)}_{s_t,a_t} = \sum_{s'} \tilde{C}_{s_t,a_t\to s'}$ is the pseudo-count of taking action a_t at state s_t (see Methods and section B.1). If there is no linear or nonlinear filtering (e.g., leaky integration) applied during the counting process, pseudo-counts are equal to real counts.

Higher values of the conditional probability $p^{(t)}(s_{t+1}|s_t, a_t)$ indicate that a participant expects to experience the transition from the pair of state and action (s_t, a_t) to the next state s_{t+1} with higher probabilities and, hence perceives this transition as less surprising. Therefore, we consider the surprise of such a transition to be a decreasing function of $p^{(t)}(s_{t+1}|s_t, a_t)$. More precisely, we use a recent measure of surprise motivated by a Bayesian framework for learning in volatile environments, called the 'Bayes Factor' surprise (Liakoni et al., 2021). The Bayes Factor surprise of the transition from state s_t to state s_{t+1} after taking action a_t is

$$\mathsf{S}_{\mathrm{BF}}^{(t+1)} = \frac{\text{const.}}{p^{(t)}(s_{t+1}|s_t, a_t)}, \tag{3.4}$$

where $p^{(t)}(s_{t+1}|s_t, a_t)$ is the conditional probability of observing state s_{t+1} at time $t+1$ derived from the present world-model. Our surprise measure is an increasing function of the state prediction error (Gläscher et al., 2010) and Shannon surprise (Meyniel et al., 2016; Shannon, 1948) (see Methods) and takes high values during the 1st episode of block 2 whenever participants encounter states 3 or 7 or transit from state 3 or 7 to another state (Figures 3.4A and 3.4B). See Discussion for the relation of our measure of surprise to other measures[3].

3.2.3 The SurNoR algorithm: Distinct contributions of novelty and surprise to behavior

We hypothesize that participants use novelty to explore the environment and surprise to modulate the rate of learning. The hypothesis is formalized in the form of the Surprise-Novelty-Reward (SurNoR) algorithm and tested given the behavioral data of 12 participants.

Novelty in SurNoR plays a role analogous to that of reward. For example, in standard Temporal-Difference (TD) Learning, a reward-based Q-value $Q_{\mathrm{R}}(s, a)$ is associated with each state-action pair (s, a) (Sutton and Barto, 2018); the Q-value $Q_{\mathrm{R}}(s, a)$ estimates the mean discounted reward that can be collected under the current policy when starting from state s and action a, and the reward prediction error RPE, derived from $Q_{\mathrm{R}}(s, a)$, serves as a learning signal even for states a few steps away from the goal (Sutton and Barto, 2018). Analogously, in the SurNoR model, novelty is a reward-like signal with associated novelty-based Q-values $Q_{\mathrm{N}}(s, a)$ and an associated novelty prediction error (NPE) derived from $Q_{\mathrm{N}}(s, a)$. In the SurNoR model, the two sets of Q-values, reward-based and novelty-based, are used in a hybrid model (Daw et al., 2011; Gläscher et al., 2010) that flexibly combines model-based with model-free action selection policies (Figure 3.4C).

[3] **Thesis footnote:** Given the taxonomy of surprise definitions we presented in chapter 2, the Bayes Factor surprise in Equation 3.4 is in the category of change-point detection surprise definitions. However, given that our task in this section is a categorical task with a uniform prior marginal distribution, the Bayes Factor surprise is indistinguishable from the Shannon surprise, State Prediction Error, and the Absolute Error surprise (see Figure 2.4). See chapter 4 for further discussions on how these definitions link to physiological signatures of surprise.

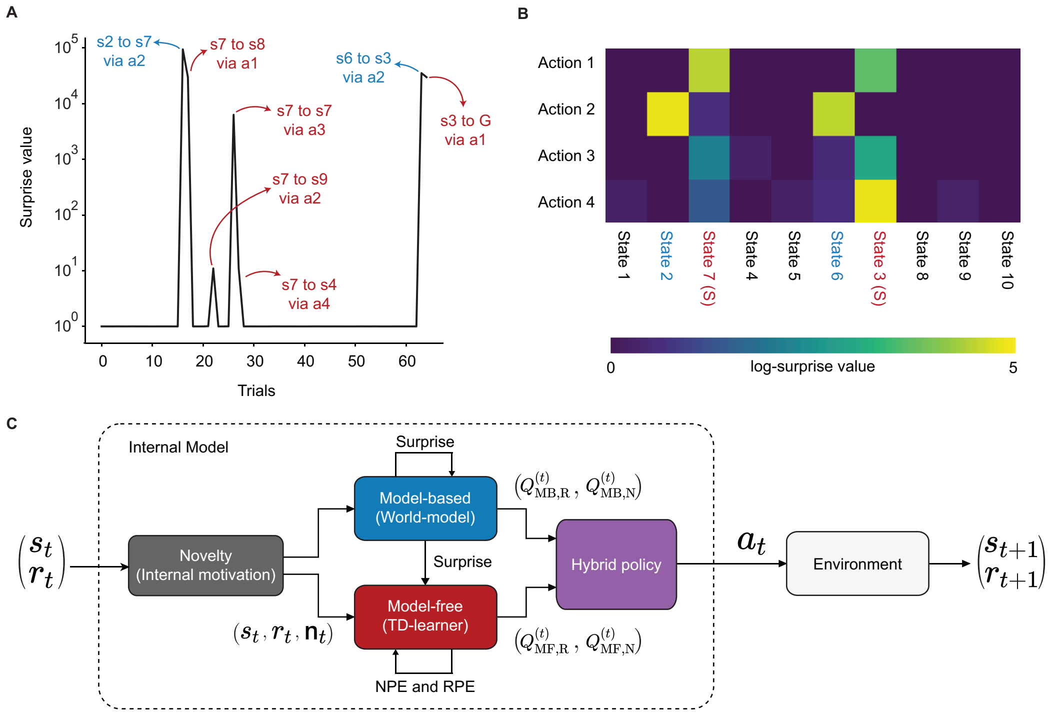


Figure 3.4: **Surprise as a modulator of the learning rate in episode 1 of block 2.** **A.** Surprise as a function of time since the start of block 2 for one representative participant: Surprise values are almost zero most of the time, because the participant has already learned the transitions in the environment during block 1. The surprising transitions are the ones to the swapped states (blue) and the ones from the swapped states (red). **B.** Maximal log-surprise values (yellow=large surprise) during the 1st episode of block 2, averaged over all participants. The swapped states are marked in red and the states before them in blue. One action from each swapped state is not surprising, i.e., the action leading participants to trap states both before and after the swap. **C.** Block diagram of the SurNoR algorithm: Information of state s_t and reward r_t at time t is combined with novelty n_t (grey block) and passed on to the world-model (blue block, implementing the model-based branch of SurNoR) and TD learner (red block, implementing the model-free branch). The surprise value computed by the world-model modulates the learning rate of both the TD-learner and the world-model. The output of each block is a pair of Q-values i.e, Q-values for estimated reward $Q_{\mathrm{MF,R}}$ and $Q_{\mathrm{MB,R}}$ as well as for estimated novelty $Q_{\mathrm{MF,N}}$ and $Q_{\mathrm{MB,N}}$. The hybrid policy (in purple) combines these values.

Surprise in SurNoR is derived from a mismatch between observations of the next state and predictions arising from the world-model embedded in the model-based branch of SurNoR. To adapt both model-based and model-free policies of the SurNoR algorithm, surprise is used in two different ways. First, high values of surprise systematically lead to a larger learning rate for the update of the world-model than smaller ones, consistent with earlier models (Liakoni et al., 2021; Soltani and Izquierdo, 2019). Second, going beyond previous models of behavior (Behrens et al., 2007; Findling et al., 2021; Heilbron and Meyniel, 2019; Nassar et al., 2012, 2010), surprise also influences the learning rate of the model-free reinforcement learning branch.

We predict that, if the behavior of participants is well described by the SurNoR algorithm, they should use an action policy that attracts them to novel states, in particular during the 1st episode of block 1. If participants do not exploit novelty, standard (potentially hybrid) reinforcement learning schemes in combination with one of several alternative exploration strategies (see next section) should be sufficient to explain the behavior. Furthermore, we predict that, if the behavior of participants is well described by the SurNoR algorithm, then surprising events during the 1st episode of block 2 should significantly change the behavior of participants; if participants do not exploit surprise, standard hybrid models combining model-based and model-free reinforcement learning (Daw et al., 2011; Gläscher et al., 2010) should be sufficient to describe the behavior.

3.2.4 Both surprise and novelty are needed to explain behavior

SurNoR has three main components: (i) action selection by Hybrid policy, (ii) exploration by novelty-seeking, and (iii) learning by surprise-modulation. To test our hypothesis, and to test whether all three components of SurNoR are necessary for explaining behavior or whether a simpler or an alternative model would have the same explanatory power, we compared SurNoR with 11 alternative algorithms plus a null algorithm based on a random choice (RC) of actions (Figure 3.5A). Three out of 11 algorithms use a hybrid policy (+Hyb), five use novelty-seeking (+N), and seven use surprise-modulation (+S).

Alternatives for (i) action selection were pure model-based (MB; 4 out of 11 algorithms) and pure model-free (MF; 4 out of 11) policies. Note that we allow for the possibility that MF algorithms are equipped with a world-model for computation of surprise but do not use this world-model for action-planning. As alternatives for (ii) exploration strategy, we used optimistic initialization (+OI; 3 out of 11) (Sutton and Barto, 2018) and uncertainty (surprise) seeking (Achiam and Sastry, 2017; Burda et al., 2019) (+U; 3 out of 11); see below for more explanations and section B.1 for details. Finally, as an alternative to surprise modulation, we used constant learning rates for learning the world-model and model-free Q-values (Maheu et al., 2019; Meyniel et al., 2016; Modirshanechi et al., 2019; Yu and Cohen, 2009) (all algorithms without +S; 4 out of 11). For the details of the alternative algorithms see section B.1.

Given the behavioral data of all 12 participants, we estimated the log-evidence of all 13 algorithms, including SurNoR (see Methods). Comparison of the algorithms' log-evidence (Figure 3.5A) shows that SurNoR explains human behavior significantly better than its alternatives. In addition, a Bayesian model selection approach with random effects (Rigoux et al., 2014; Stephan et al., 2009) indicates that the SurNoR algorithm outperforms the alternatives with a protected exceedance probability of 0.99 (Figure 3.5B and Methods).

The 1st episode of the 1st block is ideally suited to study how novelty influences behavior

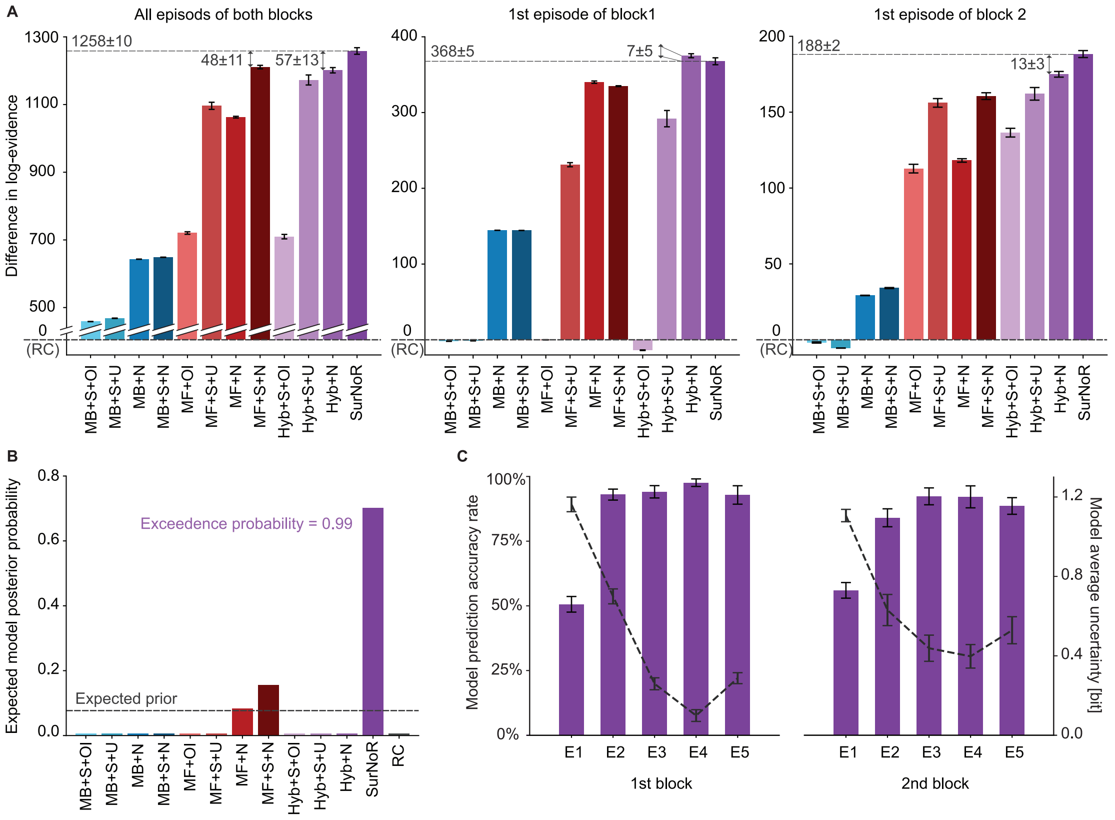


Figure 3.5: **Model comparison of model-based (MB, blue bars), model-free (MF, red bars), and hybrid algorithms (Hyb and SurNoR, purple bars).** Exploratory behavior is either induced by optimistic initialization (+OI), uncertainty-seeking (+U), unbiased random action choices (RC), or novelty-seeking (+N); e.g., a model-based algorithm with novelty seeking is denoted as MB+N. SurNoR and the model-free or hybrid algorithms annotated with '+S' use surprise to modulate the learning rate of the model-free TD learner; SurNoR and all algorithms annotated with '+S' use surprise modulation also during model building (see Methods). **A.** Difference in log-evidence (with respect to RC) for the algorithms for all episodes of both blocks (left panel), the 1st episode of block 1 (middle), and the 1st episode of block 2 (right panel). High values indicate good performance; differences greater than 3 or 10 are considered as significant or strongly significant, respectively (see Methods); a value of 0 corresponds to random action choices (RC). The random initialization of the parameter optimization procedure introduces a source of noise, and the small error bars indicate the standard error of the mean over different runs of optimization (Methods, statistical model analysis). **B.** The expected posterior model probability (Rigoux et al., 2014; Stephan et al., 2009) given the whole dataset (Methods) with random effects assumption on the models. **C.** Accuracy rate of actions predicted by SurNoR (left scale and purple bars: mean and the standard error of the mean across participant) and the average uncertainty of SurNoR (right scale and dashed grey curve: entropy of action choice probabilities).

(middle panel in Figure 3.5A). Our results show that all algorithms with novelty-seeking (+N) explain the behavior significantly better than models with random exploration strategy (RC) or optimistic initialization (MB+S+OI, MF+OI, and Hyb+S+OI), i.e., two classic approaches for exploration (Sutton and Barto, 2018). Our results also show that novelty-seeking explains behavior better than uncertainty-seeking (+U), a state-of-the-art exploration method in reinforcement learning (Achiam and Sastry, 2017; Burda et al., 2019). The models with uncertainty-seeking (MB+S+U, MF+S+U, and Hyb+S+U) use surprisal (i.e., the logarithm of our surprise measure) as an intrinsic reward as opposed to our model of novelty-seeking that uses novelty of states as an intrinsic reward.

As an alternative to novelty-seeking, participants might also solve the task simply by detecting and avoiding trap states. If so, the behavior of the participants can be explained if we replace the continuous novelty signal by a simple intrinsically generated binary signal equivalent to a negative reward. To address this issue, we tested two modified versions of the SurNoR algorithm ('Binary Novelty', see section B.1). The 1st modification detects those states that have been encountered more often than some threshold value and assigns a fixed negative reward to them. The 2nd modification considers the n most frequently encountered states as bad states and, similar to the 1st modification, assigns a fixed negative reward to them – where n is a free parameter of the algorithm. Note that in both control algorithms, the constant negative rewards are treated as an intrinsic motivation signal – similar to novelty in SurNoR-algorithm except that the signal is a binary one. We estimated the log-evidence for both control algorithms. Our results show that SurNoR outperforms the 1st control algorithm by a 244 ± 11 difference in total log-evidence and by a 235 ± 5 difference in the log-evidence of the 1st episode of block 1, and outperforms the 2nd control algorithm by a 240 ± 11 difference in total log-evidence and by a 234 ± 5 difference in the log-evidence of the 1st episode of block 1. This observation rejects the hypothesis that participants simply identify 'bad' states by some binary signal.

Surprise becomes important in the 1st episode of block 2 (right panel in Figure 3.5A). Indeed, the SurNoR model is significantly better than a hybrid model with novelty but without surprise (Hyb+N); similarly, model-free reinforcement learning with novelty and surprise (MF+S+N) is significantly better than model-free reinforcement learning with novelty alone (MF+N, right panel in Figure 3.5A). Our results show that a constant adaptation rate as implemented in standard models without surprise is not sufficient to explain the choices of participants in the episode after the swap. Rather, the rate of learning and forgetting has to be modulated by a measure of surprise.

Overall, SurNoR is better than all 12 competing algorithms by a large margin, indicating that a combination of model-based and model-free algorithms explains behavior better than each algorithm separately, consistent with the notion of parallel, model-based and model-free, policy networks in the brain (Daw et al., 2011, 2005; Gläscher et al., 2010). Going beyond these earlier studies, our results with SurNoR indicate that surprise and novelty are both necessary to explain human behavior in our task. Novelty is necessary to explain behavior during phases of exploration while surprise is necessary to explain behavior during the rapid re-adaption after a change in the environment.

3.2.5 Individual decisions are dominated by the model-free policy network

We wondered whether the SurNoR model is also able to predict the individual actions of participants. Taking the most probable action of the model in a given state as the prediction of a participant's next action in that state, SurNoR predicted the correct action in the 1st episode of the block 1 with an accuracy of 51±3% (3-fold cross validated, mean ± standard error of the mean over 12 participants, see Methods – Figure 3.5C). Note that this accuracy is achieved in the absence of any *a priori* preference of actions at initialization and is significantly higher than the accuracy rate of the naive random exploration strategy (25%, chance level).

SurNoR's predictions are also significantly better than the predictions of directed exploration through optimistic initialization (OI) or an uncertainty-seeking policy (U). Models with OI could at best predict 36±3% of the actions (for MB+S+OI), and the uncertainty-seeking strategy could at best predict 46±3% (for Hyb+S+U); one-sample t-test p-values for comparing their accuracy rates versus SurNoR's are 0.01 and 0.0025, respectively. A crucial difference between OI and novelty-based exploration is that OI prefers those actions that have been less frequently chosen in the past, while novelty-seeking prefers actions that lead to novel states, even if these are a few actions ahead and the outcome of the current action is known. Uncertainty-seeking is similar to OI because the uncertain actions are also those that have been less frequently chosen in the past.

Similarly, in the 1st episode of block 2, after the swap of states 3 and 7, the SurNoR algorithm predicts $56 \pm 3\%$ of the actions of the 12 participants (Figure 3.5C). In the remaining episodes 2-5 of the two blocks, the SurNoR algorithm predicts $89 \pm 2\%$ of the action choices (Figure 3.5C). Most of these actions move participants closer to the goal. The intrinsic uncertainty of action choices with the SurNoR model can be estimated from the entropy of the action choice probabilities across the four possible actions (Figure 3.5C). Uncertainty decreases during the first three episodes as participants become familiar with the environment, but it jumps back to higher values after the swap of states at the beginning of block 2.

To see what aspects of behavior the different components (i.e., hybrid policy, surprise, and novelty) of SurNoR capture, we fitted the parameters of the SurNoR algorithm to the behavior of the 12 participants (see Methods). Since SurNoR combines in its hybrid action selection policy a model-free with a model-based component (Figure 3.6A2), we first wanted to analyze the relative importance of each of the two components in explaining the action choices of participants; see section B.6 for a qualitative comparison of these two components. In order to evaluate the relative importance of the two components, we normalized the Q-values of both branches and determined the relative weight of each branch (see Methods) during the 1st episode and 2nd-5th episodes of each block (Figure 3.6A4). We find that the model-free branch dominates the actions. Thus the world-model is of secondary importance for action selection and is mainly used to detect surprising events.

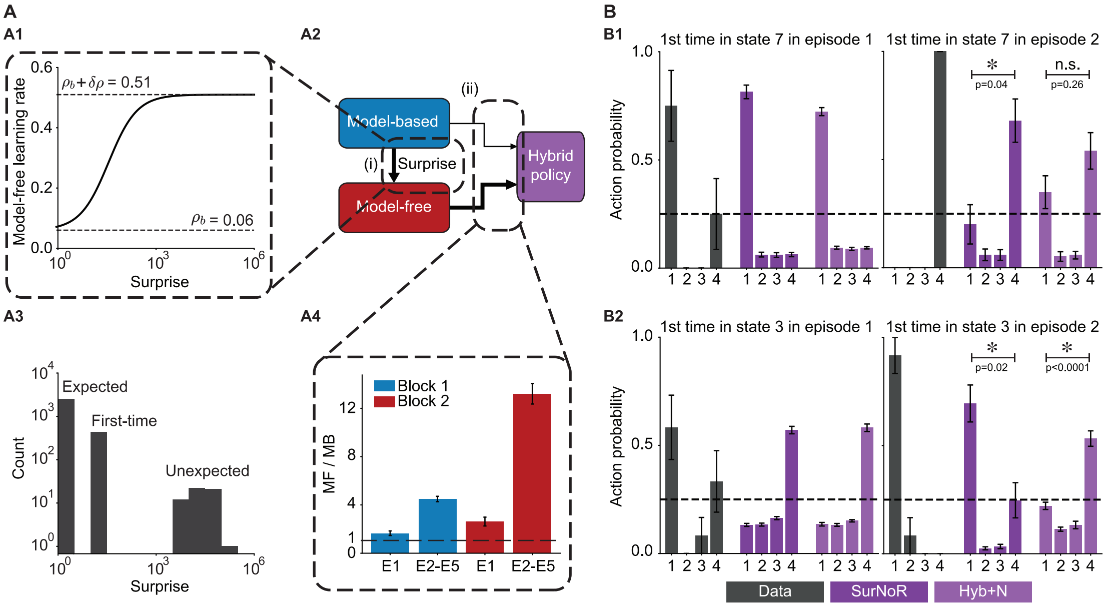


Figure 3.6: **A. Model-based surprise modulates model-free learning. A1.** The learning rate of the model-free branch as a function of the model-based surprise, after fitting parameters to the behavior of all participants (see Equation 3.9 in Methods). The model-free learning rate for highly surprising transitions is more than 8 times greater than the one for expected transitions. **A2.** Three modules from the block diagram of Figure 3.4C. There are two types of interactions between the model-based and the model-free branches of SurNoR: (i) The model-based branch modulates the learning rate of the model-free branch and (ii) the weighted (arrow thickness) outputs of the model-based and the model-free branches influence action selection (hybrid policy). **A3.** The histogram of surprise values across all trials of 12 participants. The distribution is multimodal with high surprise for the unexpected transitions in the 1st episode of block 2, medium surprise for whenever a transition is experienced for the first time, and low surprise for the expected transitions. **A4.** The relative importance of model-free (MF) compared to model-based (MB) in the weighting scheme of the hybrid policy during different episodes. Vertical axis: dominance of model-free (see Methods). Values larger than one (dashed line) indicate that the model-free branch dominates action selection. Error bars show the standard error of the mean. **B. Action choice probability indicates that surprise boosts learning during a single episode.** Action choice probabilities of 8 participants (data, grey) are compared with those of SurNoR and Hyb+N at the fist time visiting state 7 (**B1**) or state 3 (**B2**) in episodes 1 (left) and 2 (right) of block 2. Note that only 8 (out of the 12) participants encountered state 7 in the first episode of block 2 before reaching the goal. We therefore limit the data analysis to these participants. **B1.** In state 7, action 1 is the good action before the swap, and action 4 is the good action after the swap. Error bars show the standard error of the mean, and the black dashed line corresponds to random choice action probability (0.25). In episode 2, SurNoR assigns a significantly higher probability to action 4 than to action 1, while according to the Hybrid model without surprise modulation, the action probabilities of action 1 and action 4 are not significantly different. **B2.** In state 3, action 4 is the good action before the swap, and action 1 is the good action after the swap. Behavioral data and SurNoR show a more rapid re-adaptation to the good action than Hyb+N.

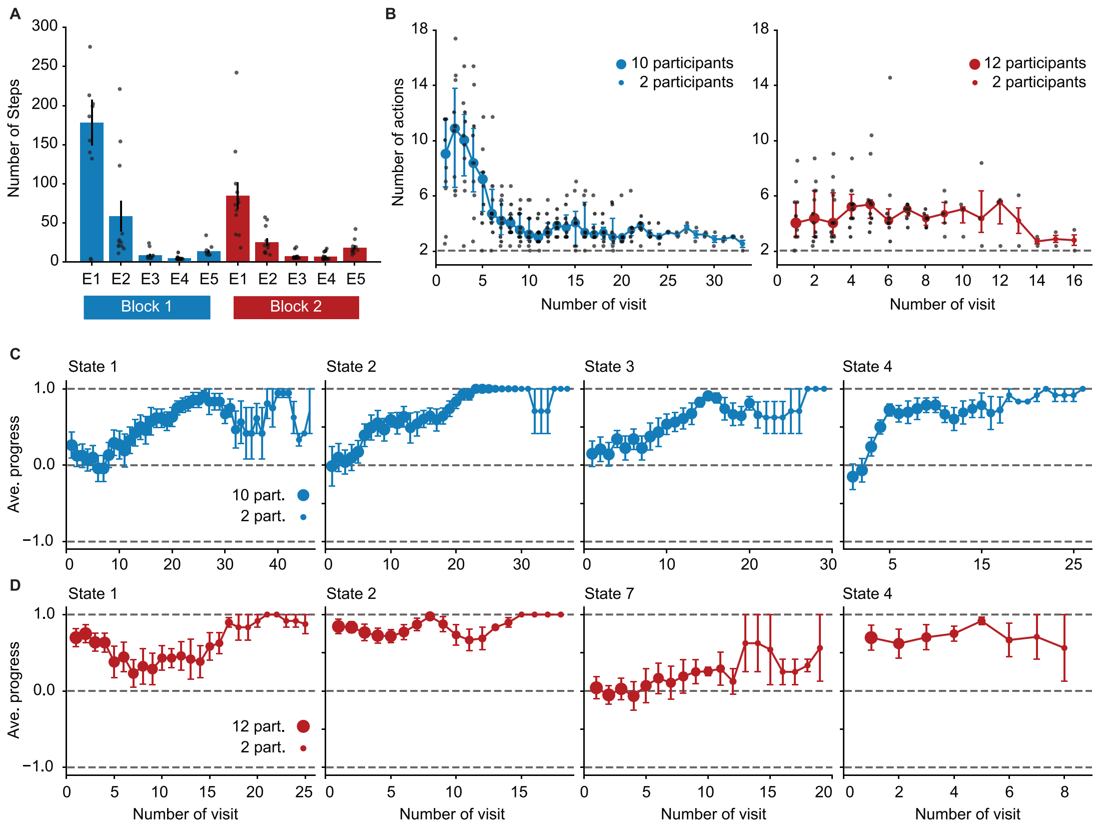


Figure 3.7: **Posterior predictive checks.** **A.** Average number of actions of all 12 simulated participants for each episode (cf. Figure 3.1C). **B.** Median number of actions of simulated participants to escape the trap states at each of their visits in episode 1 of block 1 (left) and block 2 (right) (cf. Figure 3.2A) **C.** Average progress of participants each time visiting states 1, 2, 3, and 4 in episode 1 of block 1. (cf. Figure 3.2B). **D.** Average progress of simulated participants each time visiting states 1, 2, 7 (swapped with 3), and 4 in episode 1 of block 2. (cf. Figure 3.2C). See Figure B.4 and Figure B.5 for two other sets of 12 simulated participants with different random seeds. See Figure B.3B for the average progress at the progressing states in the proximity of the goal.

Second, in order to quantify the influence of surprise on learning, we plot the learning rate (of the model-free Q-values $Q_{\mathrm{MF,R}}$ and $Q_{\mathrm{MF,N}}$) as a function of surprise (Figure 3.6A1). We find that non-surprising events lead to a small learning rate of 0.06 whereas highly surprising events induce a learning rate that is more than 8 times higher (Figures 3.6A1 and 3.6A3) indicating that surprise strongly influences the update of model-free Q-values. Moreover, we compared the action choices of participants with those of SurNoR and the model Hyb+N (i.e., SurNoR without surprise-modulation) at the swapped states in the 1st and 2nd episodes of block 2 (Figure 3.6B). Our results show that the modulation of the learning rate by surprise in SurNoR is necessary to explain the rapid adaptation of participants after the switch of states.

Finally, to see if the SurNoR model captures, in addition to other aspects, also the

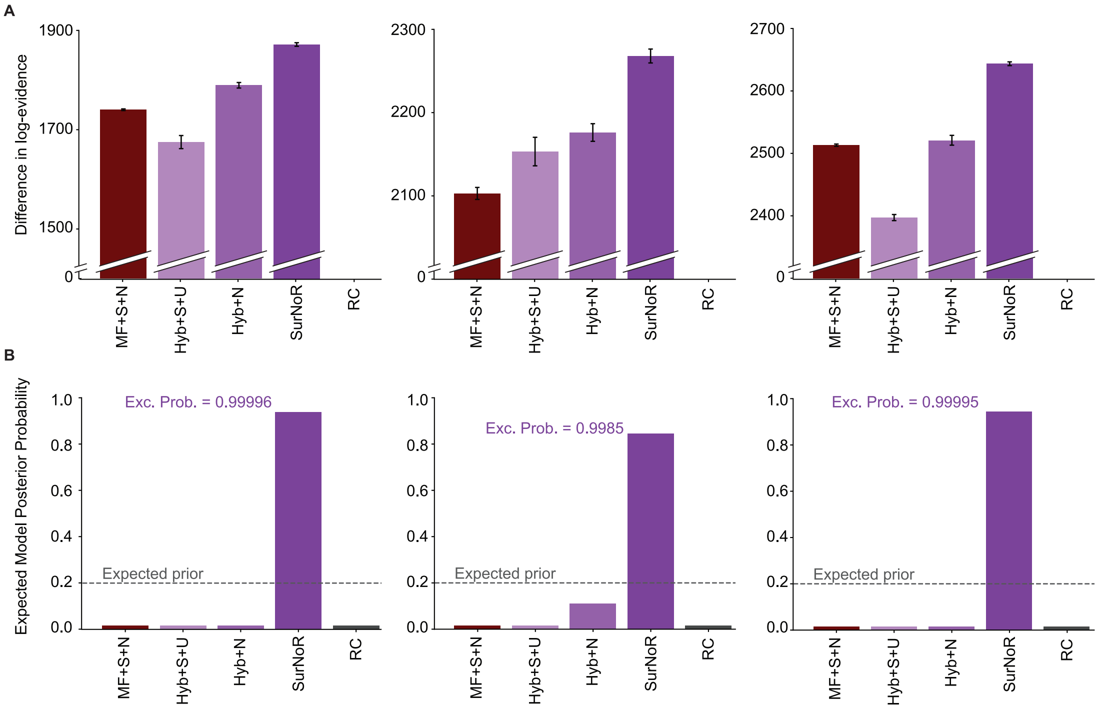


Figure 3.8: **When data is generated by SurNoR, the true model can be recovered.** We applied our model-selection method to the data of three sets of 12 simulated participants. The left column corresponds to the data shown in Figure 3.7, and the middle and the right columns correspond to the data shown in Figure B.4 and Figure B.5, respectively. We compared the SurNoR model with the strongest competitors of SurNoR: MF+S+N, Hyb+S+U, and Hyb+N (cf. Figure 3.5). **A.** Difference in log-evidence with respect to random choice (RC) and **B.** the expected posterior model probability (Rigoux et al., 2014; Stephan et al., 2009) for the algorithms for all episodes of both blocks given the data of each of the three sets (different columns) of 12 simulated participants (cf. Figures 3.5A and 3.5B).

exploratory behavior of participants (Figure 3.1C and Figure 3.2), we computed posterior predictive checks (Nassar and Frank, 2016; Wilson and Collins, 2019). To do so, we simulated SurNoR with its parameters fitted to behavior and generated data for 12 simulated participants; see Figure 3.7 for one set of 12 simulated participants and Figure B.4 and Figure B.5 for two other sets with different random seeds. Our results show that several important features of the behavior of participants are observed also in the behavior of the simulated participants: (i) They are faster in finding the goal in the first episode of block 2 than in the first episode of block 1 (Figure 3.7A), (ii) they learn to escape the trap states and to choose the good action at progressing states in the 1st episode of block 1 (Figures 3.7B and 3.7C), and (iii) after the swap, they continue choosing the same actions at unchanged states but rapidly unlearn previously learned actions at the swapped states (Figures 3.7B and 3.7D). Moreover, our model-selection approach can successfully recover the true model (SurNoR) given the data of 12 simulated participants (Figure 3.8). This observation shows that our experimental paradigm is capable of differentiating between SurNoR and its alternatives, and as few as 12 participants are sufficient for drawing conclusions based on our model-selection results (Wilson and Collins, 2019) (Figure 3.5); see section B.3, Figure B.6, and Figure B.7 for parameter recovery analysis.

In conclusion, the SurNoR algorithm is able to capture different aspects of participants' behavior and to predict individual actions with a high accuracy: it predicts $63 \pm 2\%$ of all actions and $74 \pm 3\%$ of the actions after the first time finding the goal. Our results suggests that participants (i) rely on propagation of novelty information via NPE in the first episode, (ii) base their decisions mainly on the model-free learner, and (iii) use surprise to modulate the learning rate.

3.2.6 EEG correlates with novelty and surprise

Since surprise and novelty turned out to be important and independent components of SurNoR in explaining the participants' behavior, we wondered whether they are both reflected in the ERP. We first performed a grand correlation analysis in which we pooled the more than 2500 trials of 10 participants together after normalizing their ERPs to unit energy (see Methods; two participants were excluded because of noise artifacts in the recordings). We then computed the correlations of the ERP amplitudes, for each time point after the trial onset, with the model variables 'Surprise', 'Novelty', 'Reward', 'NPE', or 'RPE' (capital initial letters indicate the 5 model variables). Note that by 'Reward' variable we mean the goal-state indicator, i.e., it is equal to one when a participant visits the goal state and zero otherwise; importantly, it should not be confused with MB or MF reward values.

We find that Surprise, Reward, and RPE show significant positive correlations with the ERP amplitudes at around 300ms after stimulus onset (Figure 3.9), in agreement with the well known correlation of the P300 amplitude with Surprise (Kolossa et al., 2015; Meyniel et al., 2016; Modirshanechi et al., 2019) and the well known correlation of the Feedback-Related Negativity (FRN) component with RPE (Holroyd and Coles, 2002; Walsh and Anderson, 2012). Moreover Novelty and NPE have, compared to Surprise, a broader positive correlation window with the ERP starting at around 200ms and ending at around 320ms after stimulus onset, and a second window with significant negative correlations from around 450ms to 550ms. Thus, Novelty and NPE have an ERP signature that is distinct from that of Surprise, Reward, or RPE (Figure 3.9).

Second, we wondered how much of the variations in the ERP amplitudes could be explained by a linear combination of our five model variables, i.e., Suprise, Novelty, NPE, RPE, and Reward. We performed a trial-by-trial multivariate linear regression (MLR), separately for each participant. To be able to more precisely identify the separate contributions of each model variable to the regression, we needed to decorrelate them from

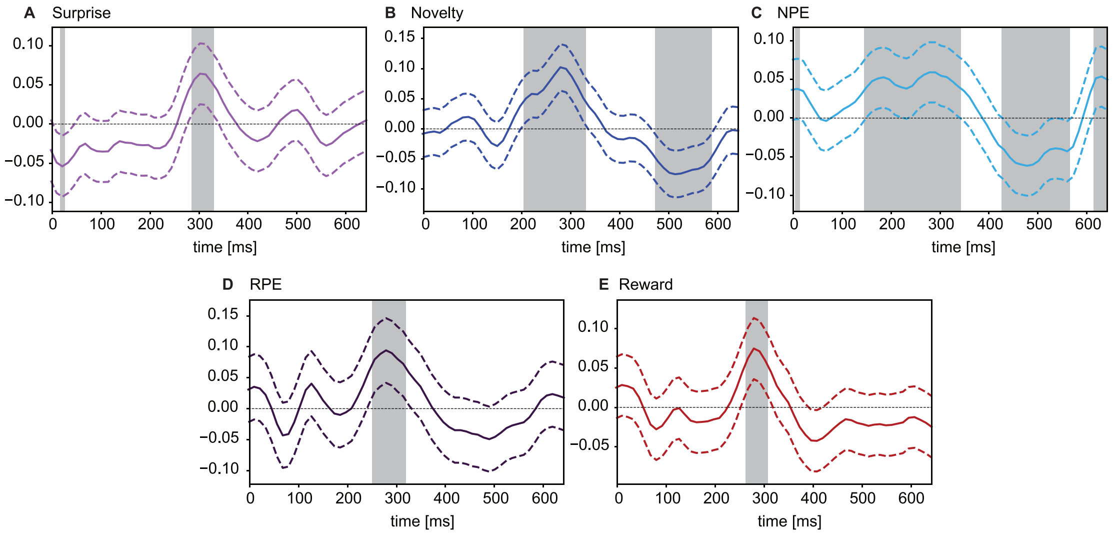


Figure 3.9: **Grand correlation analysis of normalized ERPs over all 2524 trials of 10 participant.** The dashed lines show confidence intervals. Shaded areas indicate intervals of significant correlations (FDR controlled by 0.1, one-sample t-test). Correlations of ERP with **A.** Surprise, **B.** Novelty, **C.** NPE, **D.** RPE (computed after excluding the trials from the 1st episode of the 1st block during which RPE is equal to 0) and **E.** Reward.

each other. As expected from the design of the experiment, the cross-correlations between the normalized (zero mean and unit variance) sequences of Surprise, Novelty, and NPE are negligible (see Figure B.8); however, the sequences of Reward and RPE are highly correlated with each other, mainly because Reward and RPE are both high at the goal state. Using principal components analysis over Reward and RPE, we find R_+ (the sum of RPE and Reward) and R_- (their difference) as their decorrelated combinations (see Methods). We then extracted the components of Surprise, Novelty, and NPE orthogonal to R_+ and R_- (see Methods). The resulting variables, denoted by an index $\perp$, are each orthogonal to R_+ and R_-, while staying very similar to the original signals, e.g., Surprise$_\perp$ is highly correlated with Surprise, and NPE$_\perp$ is highly correlated with NPE; see section B.2, Figure B.8, and Figure B.9 for more details.

For each participant, we considered the normalized Surprise$_\perp$, Novelty$_\perp$, NPE$_\perp$, R_+, and R_- as explanatory variables in order to predict the ERP amplitude at a given time point. We found 4 time intervals with an encoding power (adjusted R-squared, see Methods) significantly greater than zero (one-sample t-test, FDR controlled by 0.1, Figures 3.10A and 3.10B; note that the adjusted R-squared can take negative values, e.g., see baseline in Figure 3.10A). The 1st time window is around 193 ± 5ms; the P300 component can be linked to the 2nd time window which spans from 286 to 321 ± 5ms; since the 3rd time interval is long (from 392 to 487 ± 5ms), we split it into two time windows of equal size (W3a and W3b in Figure 3.10B); and the last window extends from 532 to 574 ± 5ms.

To study the contribution of surprise and novelty to encoding power, we focused on these

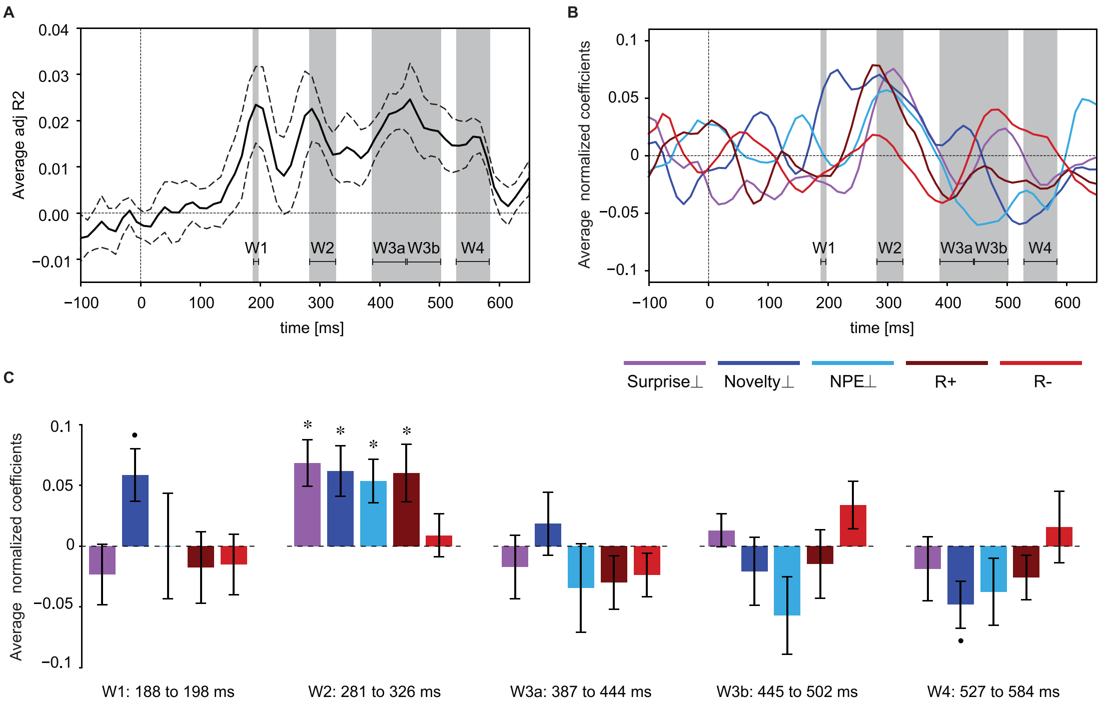


Figure 3.10: **ERP variations explained by trial-by-trial and participant-by-participant multivariate linear regression analysis.** Surprise$_\perp$ (magenta), Novelty$_\perp$ (dark blue), NEP$_\perp$ (light blue), R_+ (brown) and R_- (red) were used as explanatory variables, and the ERP amplitude at each time point was considered as the response variable. **A.** Encoding power (adjusted R-squared values) averaged over 10 participants (dashed lines show the standard error of the mean) at each time point. Shaded areas and horizontal lines indicate four time intervals (W1, ..., W4) of significant encoding power (FDR controlled by 0.1, one-sample t-test, only for the time-points after the baseline). The 3rd time interval has been split into two time windows of equal length for the analysis in C. **B.** Values of the regression coefficients (averaged over participants) for Surprise$_\perp$, Novelty$_\perp$, NEP$_\perp$, R_+, and R_- as a function of time. Errors are not shown to simplify the illustration. **C.** In each of the 5 time windows, the regression coefficients plotted in B have been averaged over time. Error bars show the standard error of the mean (across participants). Asterisks show significantly non-zero values (FDR controlled by 0.1 for each time window, one-sample t-test). The Novelty$_\perp$ coefficients in the 1st and the last time windows (dot) have p-values of 0.03 and 0.04, respectively, which are not significant after FDR correction. In the second time window, Surprise$_\perp$, Novelty$_\perp$, NEP$_\perp$, and R_+ have significantly positive coefficients.

time windows and tested the average regression coefficients of all explanatory variables in each time window in a second level analysis (Figure 3.10C). Our results show that in the the second time window (286 to 321 ± 5ms), Surprise$_\perp$, Novelty$_\perp$, NPE$_\perp$, and R_+ all have a significant positive regression coefficient in MLR (Figure 3.10C, 2nd panel, FDR controlled by 0.1). While the coefficients for Surprise$_\perp$, NPE$_\perp$, and R_+ sharply peak at around 300ms, the coefficient for Novelty$_\perp$ has a broader peak starting at around 200ms (Figure 3.10B) with a close to significant positive value during the 1st time window. This observation suggests that positive correlations of novelty with the ERP potentially extend from the 1st time window to the 2nd one, in agreement with our grand correlation analysis.

While consistent with previous studies of surprise in the ERP (Kolossa et al., 2015; Maheu et al., 2019; Meyniel et al., 2016; Modirshanechi et al., 2019), our results indicate that Surprise$_\perp$ and Novelty$_\perp$ contribute each separately to the ERP components at around 300ms. Furthermore, we find that NPE$_\perp$ is yet another independent contributor to these components. As expected from previous studies (Holroyd and Coles, 2002; Walsh and Anderson, 2012), R_+ also shows a positive correlation with the ERP amplitude at around 300ms. While the multivariate analysis based on the 5 explanatory variables shows significance in the later time windows (Figure 3.10A), individual contributions of Surprise$_\perp$ or Novelty$_\perp$ or NPE$_\perp$ alone remain below significance level even though Novelty$_\perp$ has a close to significant negative coefficient in the last window (Figure 3.10C).

To summarize, the grand correlation analysis yields time windows of significance for Novelty and NPE that start 50 to 100ms *before* those of Surprise or Reward, indicating distinct contributions. Moreover, Novelty$_\perp$ and NPE$_\perp$ explain a significant fraction of the variations of the ERP at around 300ms that is not explained by Surprise$_\perp$ and R_+ alone. Importantly, NPE has significant correlations both in the grand correlation analysis and in the regression analysis, consistent with our earlier finding that NPE is important to explain behavior.

3.3 Discussion

Combining a deep sequential decision-making task with the SurNoR model, an augmented reinforcement learning algorithm, we were able to extract the distinct contributions of surprise, novelty, and reward to human behavior. We found that the human brain (i) uses surprise to adapt their behavior to changing environments by modulating the learning rate and (ii) uses novelty as an intrinsic motivational drive to explore the world. Moreover, the model variables Suprise, Novelty, NPE, Reward, and RPE could well explain variations of the EEG amplitudes on a trial-by-trial basis.

3.3.1 Novelty is not surprise

In the SurNoR model, surprise measures how *unexpected* the next state is according to an acquired world-model *conditioned* on the current state and the chosen action; in contrast, novelty measures how *infrequent* the next state has been, *independent* of our expectations derived from a world-model. More precisely, in our formulation (Equation 3.2 and Equation 3.4), surprise and novelty have two essential differences: The first difference is that while surprise is assigned to transitions, novelty is assigned to states. If this was the only difference between novelty and surprise, one could argue that surprise and novelty are essentially the same, while one measures the infrequency of transitions and the other the infrequency of states. However, the second and more important difference is that novelty uses the exact number of encounters of each state (Equation 3.1) to measure how infrequent that state has been, while surprise uses the surprise-modulated pseudo-counts (Equation 3.3) computed by the world-model to measure how unexpected a transition is. As a consequence, if there is a sudden change in the environment, then an expected transition can rapidly become surprising, but a long time is needed for a state that has been encountered frequently to become novel again. This is consistent with ideas that novelty is more related to memory-recall and surprise is more related to predictions (Barto et al., 2013). Moreover, these ideas are also supported by recent findings showing that the brain estimates the frequency of stimuli over a much longer time-scale than the transition probabilities (Maheu et al., 2019).

Measures of surprise in neuroscience have been divided into two subgroups (Faraji et al., 2018; Gijsen et al., 2021; Hurley et al., 2011): 'puzzlement' and 'enlightenment' surprise. The conceptual definition of surprise we gave above is known as puzzlement surprise. In addition to the Bayes factor surprise (Liakoni et al., 2021) that we used here (Equation 3.4), other examples of puzzlement surprise are Shannon surprise (Shannon, 1948), minimized free energy (Friston, 2010; Friston et al., 2017), state prediction error (Gläscher et al., 2010), and the confidence corrected surprise (Faraji et al., 2018). Orthogonal to these measures, enlightenment surprise measures how much an event changes our model of the world and, as a consequence, our expectations; well-known examples are Bayesian surprise (Baldi, 2002; Itti and Baldi, 2006; Schmidhuber, 2008, 2010; Storck et al., 1995) and compression surprise (Schmidhuber, 2008, 2010). We would like to emphasize here that even a measure like Bayesian surprise (Baldi, 2002; Itti and Baldi, 2006; Schmidhuber, 2010; Storck et al., 1995) has the aforementioned differences with our definition of novelty.

Measures of novelty can also be divided into two subgroups (Barto et al., 2013): the ones that are 'memory-based' and the ones that represent 'statistical outliers'. Memory-based novelty measures focus on whether an event already exists in the memory or not (Gershman et al., 2017, 2014). Our measure of novelty (Equation 3.2) belongs to the 2nd group that considers an event as novel if it has 'a low estimated probability of occurrence' (Barto et al., 2013). Many measures of novelty that belong to this group simply consider novelty to be a decreasing function of the number of occurrences (Bellemare et al., 2016; Dubey and Griffiths, 2019; Gershman and Niv, 2015); in contrast, our measure of novelty is a decreasing function of the *frequency* of occurrence (Equation 3.1), and because of this reason one may refer to it as 'relative novelty' or a measure of 'rareness'.

3.3.2 Surprise modulates learning

As expected from previous theoretical (Faraji et al., 2018; Findling et al., 2021; Frémaux and Gerstner, 2016; Gerstner et al., 2018; Liakoni et al., 2021; Yu and Dayan, 2005) and experimental work (Behrens et al., 2007; Heilbron and Meyniel, 2019; Nassar et al., 2012, 2010), our results suggest that the human brain uses surprise to modulate the learning of its world-model. Rather unexpectedly, however, our results indicate that humans hardly use this world-model to plan behavior; instead they mainly rely on model-free TD learning with eligibility traces to choose their next actions. Importantly, although the surprise signal is triggered by a mismatch between an observation and the predictions of the world-model, the modulatory effect of surprise is not limited to readjusting the world-model but also used to modulate the learning rate of model-free TD-learning. Following the common interpretations of model-based reinforcement learning algorithms as descriptions of human planning behavior and model-free reinforcement learning algorithms as descriptions of human habitual behavior (Akam et al., 2015; Daw et al., 2011, 2005; Gläscher et al., 2010), our results suggest that (i) in the absence of surprise, humans prefer habitual behavior (potentially to reduce computational costs of decision-making (Huys et al., 2015; Kahneman, 2013)) and (ii) errors in their world-model make them reconsider their habitual behavior.

Our results extend findings that humans use hybrid policies in two-stage decision tasks (Daw et al., 2011; Gläscher et al., 2010) to the case of deep sequential decision tasks in the presence of abrupt changes. In general, the balance between model-free and model-based behaviors depends on multiple aspects and features of the task that humans are dealing with (Daw et al., 2005; Gläscher et al., 2010; Huys et al., 2015; Kool et al., 2016). For example, we observe that the model-based branch becomes more important when participants explore the environment to find the goal state, although the participants' behavior in our environment is always dominated by model-free action choices (Figure 3.6A4). Moreover, the dominance of model-free behavior in such deep tasks does not exclude that humans use model-based planning in shallow tasks that are easily comprehensible thanks to a spatial arrangement of states or explicit instructions (da Silva and Hare, 2020).

An interesting direction for future studies is to combine surprise modulation with more abstract model building algorithms, e.g., for learning the structure of neighborhood relations of an environment in the form of a graph (Whittington et al., 2020; Wu et al., 2020). Such algorithms may explain the slight difference between the participants' adaptive behavior and SurNoR's predictions (Figure 3.6B).

3.3.3 Novelty drives exploration

Our results show that exploration based on novelty-seeking can explain human behavior in our sequential decision-making task better than its alternatives: random exploration (Schulz and Gershman, 2019), optimistic initialization (Sutton and Barto, 2018), and uncertainty or surprise-seeking (Achiam and Sastry, 2017; Burda et al., 2019). In contrast to many exploration strategies that give preference to those actions for which the outcome is most uncertain, i.e., those that have been tried least (Achiam and Sastry, 2017; Burda et al., 2019; Cohen et al., 2007; Kobayashi et al., 2019; Schulz and Gershman, 2019), exploration based on novelty-seeking gives preference to actions that ultimately lead a participant to previously unvisited or less visited states, even if the participant is perfectly sure about the transition to the next state.

In general, it has been shown that exploration in humans can have more than one drive (Kobayashi et al., 2019) and that participants' desire for seeking novel events depends on their goal, inductive biases, and assumptions about their environment (Dubey and Griffiths, 2019, 2020; Gershman and Niv, 2015). In situations like our experimental paradigm where participants are sure that there exists a rewarding state but do not know how to reach it, seeking novelty and exploring the parts of the environment that have been less visited are natural ways to search for the rewarding state; however, if, for example, participants were asked to find the most accurate map of the environment, then uncertainty-seeking might be a more reasonable way to solve the task. In addition, the presence of trap states in our environment makes novelty an internally rewarding signal that helps participants to avoid 'traps'. We do not claim that novelty is always the only drive of exploration; rather, we believe that our results show that for a class of tasks similar to ours where the goal is to search for a rewarding state and novelty is a relevant signal, humans use a novelty-seeking strategy for exploration. Following the idea of information search in active sampling (Gottlieb and Oudeyer, 2018; Niv et al., 2015), we speculate that whether novelty is an informative cue (e.g., about the location of the goal) or not must be itself inferred by participants through the exploration procedure. From a different but similar perspective, we speculate that it is possible to take a normative approach and, by defining a function of curiosity (Dubey and Griffiths, 2019, 2020), show that novelty-seeking is the optimal or a close to optimal way to search for reward in a class of environments. Formulating and testing such hypotheses is an interesting direction for future studies.

The SurNoR algorithm suggests that participants treat novelty and reward as separately estimated values – as opposed to adding them into a single value estimator (Bellemare et al., 2016; Jaegle et al., 2019; Martin et al., 2017). This separation enables participants to rapidly switch from exploration to exploitation, once they have found the goal. Based on this insight, we make the following prediction: if participants find a goal state but expect a second more rewarding goal state, they will continue to explore and potentially spend a large amount of time in a novelty-rich segment of an extended version of the environment of Figure 3.1 (see section B.5).

3.3.4 Neural signatures of surprise and novelty

Our EEG analysis shows that variables of the SurNoR model can significantly explain the variations of ERP amplitudes in several time-windows: Surprise, Novelty, NPE, and Reward/RPE all significantly contribute to the encoding power in the time-window around 300ms. The positive contributions of Novelty, Surprise, and Reward/RPE in this time-window are consistent with previous studies of the P300 and the FRN component (Holroyd and Coles, 2002; Kolossa et al., 2015; Maheu et al., 2019; Mars et al., 2008; Meyniel et al., 2016; Modirshanechi et al., 2019; Walsh and Anderson, 2012). Whereas in earlier studies contributions of Novelty, Surprise, and Reward were often mixed together (Holroyd and Coles, 2002; Kolossa et al., 2015; Maheu et al., 2019; Mars et al., 2008; Meyniel et al., 2016; Modirshanechi et al., 2019; Walsh and Anderson, 2012), we have shown here separate, additive contributions of these three variables as well as a further contribution of NPE. The effect of Novelty appears in ERPs earlier (at around 200ms) than the correlations with the other variables; moreover, contributions of Novelty are distinct from those of Surprise in the time window after 400ms.

Since, in the SurNoR model, Novelty is treated analogously to an external Reward, TD-learning based on NPE along with eligibility traces rapidly diffuses information about novel states to far-away non-novel states just as TD-learning based on RPE along with eligibility traces rapidly diffuses information about rewarding states to far-away non-rewarding states. Several studies have shown that the reward-driven activity of dopaminergic neurons encodes RPE and not reward values (Kim et al., 2020a; Schultz et al., 1997; Starkweather and Uchida, 2021). Therefore, the manifestation of a separate NPE signal in neural activities may open a new door for further developments of theories and experiments on novelty-driven activity of dopaminergic neurons and other neuromodulators (Horvitz et al., 1997; Kakade and Dayan, 2002; Morrens et al., 2020; Schultz, 1998).

3.3.5 Conclusions

In conclusion, surprise and novelty are conceptually distinct concepts that also give rise to different temporal components in the ERP. Our results suggest that humans use novelty-seeking for efficient exploration and surprise for a rapid update of both their internal world-model and their model-free habitual responses.

3.4 Methods

3.4.1 Ethics Statement

The data were collected under CE 164/2014, and the protocol was approved by the Commission cantonale d'éthique de la recherche sur l'être humain. All participants were informed that they could quit the experiment at any time, and they all signed a written informed consent. All procedures complied with the Declaration of Helsinki (except for pre-registration).

3.4.2 Experimental setup

Stimuli were presented on an LCD screen that was controlled by a Windows 7 PC. Experiments were scripted in MATLAB using the Psychophysics Toolbox (Brainard and Vision, 1997).

3.4.3 Participants

14 paid participants joined the experiment. Two participants quit the experiment ($14 \pm 10\%$ of all participants), hence, we analysed data for 12 participants (5 females, aged 20-26 years, mean = 22.8, sd = 1.7). All participants were right-handed and naive to the purpose of the experiment. All participants had normal or corrected-to-normal visual acuity.

3.4.4 Stimuli and general procedure

Before starting the experiment, we showed the participants the goal image that they were required to find on a computer screen. Next, participants were presented, in random order, all the other images that they might encounter during the experiment. Thereafter, participants clicked the 'start' button to start the experiment. At each trial, participants were presented an image (state) and four grey disks below the image. Clicking on one of the disks (action) led participants to a subsequent image; for details of timing see Figure 3.1A. Participants clicked through the environment until they found the goal state which finished the episode. An episode n started at a random state $i(n)$ which was the same for all participants; in our experiment we used $i(1) = 6$, $i(2) = 9$, $i(3) = 4$, $i(4) = 5$, and $i(5) = 8$.

3.4.5 EEG recording and processing

EEG signals were recorded using an ActiveTwo Mk2 system (BioSemi B.V., The Netherlands) with 128 electrodes at a 2048Hz sampling rate. Two participants were excluded from EEG analysis because of their noisy and low quality signals caused by substantial movements during the experiment. Data were band pass filtered from 0.1Hz to 40Hz and down sampled to 256Hz. EEG data were recorded with a Common Mode Sensor (CMS) and re-referenced using the common average referencing method. We used EEGLAB (Delorme and Makeig, 2004) toolbox in MATLAB to perform the EEG preprocessing. We extracted EEG trials from 200ms before to 700ms after the state onset. Trials in which the change in voltage at any channel exceeded 35 μV per sampling point were discarded. Eye movements and electromyography (EMG) artefacts were removed by using independent component analysis (ICA). The baseline activity was removed by subtracting the mean calculated over the interval from 200ms to 0ms before the state onset. EEG data of selected prefrontal electrodes (Fz, F1, F2, AFz, FCz) were averaged for ERP analysis. We further smoothed (moving averaging with the window of length 50ms) and downsampled (to the sampling rate of 1 sample per 11.7ms) ERPs. Data were analyzed during the time window from 0 to 650ms after state onset (blue interval in Figure 3.1A). For multivariate regression analysis, a 100ms-baseline was also included for sanity check. As a result, each trial (from 100ms before to 650ms after the onset of the state) consisted of 65 time points.

3.4.6 SurNoR algorithm

We present a more detailed formulation and the psudocode of the SurNoR algorithm in section B.1. Here we outline the algorithm in brief.

We formally define the **Novelty** of a state s at time t as $\mathsf{N}^{(t)}(s) = -\log p_f^{(t)}(s)$, where $p_f^{(t)}$ is defined in Equation 3.1; see section B.1 for further discussion. When observing the image corresponding to state s_{t+1} at time $t+1$, after taking action a_t, the novelty $\mathsf{n}_{t+1} = \mathsf{N}^{(t)}(s_{t+1})$ is treated as an internal novelty-reward, completely analogous to the treatment of external rewards in reinforcement learning. This analogy between external reward and novelty is inspired by earlier experimental studies (Ghazizadeh et al., 2020; Horvitz et al., 1997; Schultz, 1998). As a result, at time $t+1$, agents receive three signals: the next state s_{t+1}, the external reward r_{t+1} (i.e., the indicator of whether s_{t+1} is the goal state or not), and the novelty n_{t+1} (indicated as the output of the grey block in Figure 3.4C).

The SurNoR algorithm has two branches, i.e., a model-based and a model-free one, which interact with each other (Figure 3.4C, blue and red blocks). The **model-based branch** computes the Bayes Factor Surprise (Liakoni et al., 2021)

$$\mathsf{S}_{\mathrm{BF}}^{(t+1)} = \frac{p_{\mathrm{reset}}(s_{t+1}|s_t, a_t)}{p^{(t)}(s_{t+1}|s_t, a_t)} \tag{3.5}$$

where $p^{(t)}(s_{t+1}|s_t, a_t)$ is the probability of observing s_{t+1} by taking action a_t in state s_t as estimated from the current world-model (cf., Equation 3.3), and $p_{\mathrm{reset}}(s_{t+1}|s_t, a_t)$ is the probability of observing s_{t+1} by taking action a_t in state s_t with the assumption that the environment has experienced an abrupt change between time t and $t+1$, so that the world-model should be reset to its prior estimate. In this work, we assume that the prior estimate $p_{\mathrm{reset}}(s_{t+1}|s_t, a_t) = 1/11$ is a uniform distribution over states and hence constant as stated in Equation 3.4; see section B.1 and Liakoni et al. (2021) for further discussion. Note that in Figure 3.4A, Figure 3.4B, and Figure 3.6 we suppressed the factor 1/11 and directly plotted $1/p^{(t)}(s_{t+1}|s_t, a_t)$ as the surprise value. As an aside we note that since the state prediction error (Gläscher et al., 2010) is defined as $SPE_{t+1} = 1 - p^{(t)}(s_{t+1}|s_t, a_t)$, the Bayes Factor Surprise can be written as $\mathsf{S}_{\mathrm{BF}}^{(t+1)} \propto 1/(1 - SPE_{t+1})$. The definition of the Bayes Factor Surprise is valid for arbitrary volatile environments (Liakoni et al., 2021). However, since in our experimental setting $p_{\text{reset}}(s_{t+1}|s_t, a_t)$ is assumed to be uniform, the Bayes Factor Surprise S_{BF} is a monotone function of Shannon Surprise and hence comparable to previous studies (Meyniel et al., 2016; Modirshanechi et al., 2019; Shannon, 1948).

The value $\mathsf{S}_{\text{BF}}^{(t+1)}$ is used in the model-based branch to update the world-model using the Variational SMiLe algorithm (Liakoni et al., 2021), an approximate Bayesian learning rule with surprise-modulated learning rate designed for volatile environments with abrupt changes. Updating the world-model is equivalent to updating the pseudo-counts $\tilde{C}^{(t)}_{s,a\to s'}$, introduced in Equation 3.3, for all possible s, a, and s'. The Variational SMiLe algorithm (Liakoni et al., 2021) yields the updates

$$\tilde{C}^{(t+1)}_{s,a\to s'} = \begin{cases} (1-\gamma_{t+1})\tilde{C}^{(t)}_{s,a\to s'} + \delta(s', s_{t+1}) & \text{if} \quad s = s_t, a = a_t \\ \tilde{C}^{(t)}_{s,a\to s'} & \text{otherwise,} \end{cases} \tag{3.6}$$

where δ is the Kronecker delta function, and γ_{t+1} is the surprise modulated adaptation factor (Liakoni et al., 2021)

$$\gamma_{t+1} = \frac{m\mathsf{S}_{\text{BF}}^{(t+1)}}{1 + m\mathsf{S}_{\text{BF}}^{(t+1)}} \in [0, 1], \tag{3.7}$$

with $m \geq 0$ a free parameter related to the volatility of the environment (Liakoni et al., 2021). Note that if the transition from s to s' caused by action a is unsurprising, then the pseudo-count of that transition is increased by one (because $\gamma = 0$ for $\mathsf{S}_{\text{BF}} = 0$). However, if this transition has a high surprise, the earlier pseudo-count is reset to zero (because $\gamma \to 1$ for $\mathsf{S}_{\text{BF}} \to \infty$) and the observed transition is counted as the first one. The updated world-model is then used to update a pair of Q-values, i.e., $Q^{(t+1)}_{\text{MB,R}}$ for Reward and $Q^{(t+1)}_{\text{MB,N}}$ for Novelty, by solving the corresponding Bellman equations with a variant of prioritized sweeping (Brea, 2017; Sutton and Barto, 2018; Van Seijen and Sutton, 2013); see section B.1 for details.

The **model-free branch** computes Reward and Novelty prediction errors, RPE_{t+1} and NPE_{t+1}. As usual, RPE is defined as $RPE_{t+1} = r_{t+1} + \lambda_R V_{\text{MF,R}}(s_{t+1}) - Q_{\text{MF,R}}(s_t, a_t)$, where λ_R is the discount factor for reward, and $V_{\text{MF,R}}(s_{t+1}) = \max_a Q_{\text{MF,R}}(s_{t+1}, a)$ is the value of the state s_{t+1}. Analogously, NPE is defined as $NPE_{t+1} = \mathsf{n}_{t+1} + \lambda_N V_{\text{MF,N}}(s_{t+1}) - Q_{\text{MF,N}}(s_t, a_t)$, where λ_N is the discount factor for novelty, and $V_{\text{MF,N}}(s_{t+1}) = \max_a Q_{\text{MF,N}}(s_{t+1}, a)$ is the novelty value of the state s_{t+1}.

A Surprise-modulated TD-learner with eligibility traces is used for updating the two separate sets of Q-values. To have the most general setting, two separate eligibility traces are used for the update of Q-values, one for reward $e^{(t)}_R$ and one for novelty $e^{(t)}_N$. The eligibility traces are initialized at zero at the beginning of each episode. The update rules for the eligibility traces after taking action a_t at state s_t is

$$\begin{aligned} e_R^{(t+1)}(s,a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_R \mu_R e_R^{(t)}(s,a) & \text{otherwise} \end{cases} \\ e_N^{(t+1)}(s,a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_N \mu_N e_N^{(t)}(s,a) & \text{otherwise}\,, \end{cases} \end{aligned} \tag{3.8}$$

where λ_R and λ_N are the discount factors defined above, and $\mu_N \in [0,1]$ and $\mu_R \in [0,1]$ are the decay factors of the eligibility traces for novelty and reward, respectively. The update rule is then $\Delta Q_{\text{MF}}^{(t+1)}(s,a) = \rho_{t+1} e^{(t+1)}(s,a) PE_{t+1}$, where $e^{(t+1)}$ is the eligibility trace (i.e., either $e_R^{(t+1)}$ or $e_N^{(t+1)}$), PE_{t+1} is the prediction error (i.e., either RPE_{t+1} or NPE_{t+1}) and

$$\rho_{t+1} = \rho_b + \gamma_{t+1} \delta\rho \tag{3.9}$$

is the surprise-modulated learning rate with parameters ρ_b for the baseline learning rate and $\delta\rho$ for the effect of Surprise.

Finally, actions are chosen by a hybrid policy (section B.1) using a softmax function of a linear combination of the values $Q_{\text{MF,R}}^{(t+1)}$, $Q_{\text{MF,N}}^{(t+1)}$, $Q_{\text{MB,R}}^{(t+1)}$, and $Q_{\text{MB,N}}^{(t+1)}$ (the purple block in Figure 3.4.A), similar, but not identical to Daw et al. (2011); Gläscher et al. (2010). The weight of $Q_{\text{MF,N}}^{(t+1)}$ and $Q_{\text{MB,N}}^{(t+1)}$ is non-zero only in the 1st episodes of blocks 1 and 2 to reduce number of parameters and make the model simpler. We tested the version with an additional free parameter for the weights of $Q_{\text{MF,N}}^{(t+1)}$ and $Q_{\text{MB,N}}^{(t+1)}$ in episodes 2-5 of blocks 1 and 2, but we did not find any significant improvement in the fit (difference in log-evidence $= 15 \pm 13$).

Overall, the SurNoR algorithm has 18 free parameters.

3.4.7 Statistical model analysis and fit to behavior

In addition to SurNoR, we considered 12 alternative algorithms with 0 to 18 free parameters and two control algorithms for SurNoR with Binary Novelty with 19 free parameters (section B.1). For each algorithm, we used 3-fold cross-validation and computed its maximum log-likelihood for each participant, similar to existing methods (Lehmann et al., 2019): (i) we divided participants into 3 folds each consisting of four participants; (ii) for participant i, we estimated the parameters of the algorithm by maximizing the likelihood function of the folds which did not include participant i; and (iii) we computed the log-likelihood for participant i using the estimated parameters. The maximization procedure was done by coordinate ascent (using grid search for each coordinate); we repeated the procedure until convergence starting from 25 different random initial points. We further repeated the whole process 4 times to have an estimation of the variability resulting from random initialization of the optimization procedure. The error bars in Figure 3.5A are calculated using these 4 samples.

Similar to studies in economics and statistics (Fong and Holmes, 2020; Rust and Schmittlein, 1985), we considered, for each participant and each algorithm, the cross-validated maximum log-likelihood (averaged over the 4 repetitions) as the log-evidence (Efron and Hastie, 2016). Similarly, the log-evidence could also be approximated by other measures like AIC or BIC (Daw, 2011), but cross-validation has been shown to have a more robust behavior (Ito and Doya, 2011). The sum (over participants) of the log-evidences for each algorithm is shown in Figure 3.5A – see Daw (2011) for a tutorial on the topic. As a convention, differences greater than 3 or 10 are considered as significant or strongly significant, respectively (Daw, 2011; Efron and Hastie, 2016). The model posterior and protected exceedance probabilities in Figure 3.5B are computed by using the participant-wise log-evidences (averaged over the 4 repetitions) and following the Bayesian model selection method of Rigoux et al. (2014); Stephan et al. (2009) (available in SPM12 toolbox for MATLAB). We used a Dirichlet distribution with parameters equal to 1 over the number of models (1/13) as the prior distribution. This choice of prior is equivalent to stating that the prior information is worth as much as the observation coming from a single participant (Efron and Hastie, 2016); it is also a default choice of prior in the VBA toolbox (Daunizeau et al., 2014).

The accuracy rate and the uncertainty in Figure 3.5C are computed by the same cross-validation procedure. We define accuracy as the ratio of the number of trials with correctly predicted actions to the total number of trials; for a given trial, whenever the action taken by the participant had the maximum probability under the policy but shared with other $n-1$ (e.g., 2) actions, we counted that trial as $1/n$ (e.g., 0.333) correctly predicted. With this procedure, the accuracy rate of the random choice algorithm is 25%. We define the uncertainty of one participant in an episode as the average of the entropy of his or her policy over all trials of that episode. Both the accuracy rate and the uncertainty were computed for each participant separately, but only the mean and the standard error across participants are reported in Figure 3.5C.

For EEG analysis, we only considered the SurNoR algorithm (i.e., the winner of statistical model selection). To have the same set of parameters for all participants, we fitted our model to the whole behavioral data set (overall 3047 actions) by maximizing total log-likelihood – similar to Daw et al. (2011). For each of 500 random initialization points, maximization was implemented as coordinate ascent until convergence (using grid search for each coordinate). Amongst the 5 local maxima with high but not significantly (< 3) different log-evidence, we kept the model which had the greatest encoding power in multivariate regression analysis of EEG. The fitted parameters are reported in section B.3.

The plots in Figure 3.6 corresponds to this set of parameters. Since the softmax operator of the hybrid policy has a free scale parameter, the effective weight of each branch of the hybrid policy in Figure 3.6A4 (i.e., model-free and model-based) is computed as the fitted weight of each component times its average difference in Q-values. For example, ω_{MF}^{eff} is equal to $\omega_{MF} \times \langle \Delta Q_{MF} \rangle$, where ω_{MF} is the weight of model-free Q-values in the hybrid policy and $\langle \Delta Q_{MF} \rangle$ is the average (over trials) of the difference between Q_{MF} of the best and the worst actions. The weight ω_{MB}^{eff} for the model-based branch is defined analogously. The dominance of the model-free branch is defined as $\omega_{MF}^{\text{eff}}/\omega_{MB}^{\text{eff}}$.

We used the same set of parameters to generate synthetic data for Figure 3.7. We simulated 200 agents with different random seeds. We considered the 62 agents ($31 \pm 3\%$ of all agents) who took more than 500 actions in any of the 10 episodes to be similar to the participants who quit the experiments ($14 \pm 10\%$ of all participants – not significantly different from $31 \pm 3\%$; p-value for two-sample t-test=0.12). Based on this criterion, we discarded 62 agents. From the remaining 138 agents, we randomly chose three subsets of 12 agents (called simulated participants in the Results section) and repeated all our behavioral analyses for the synthetic data. The results for one subset of agents is shown in Figure 3.7 and for two other subsets in Figure B.4, Figure B.5. Given the three sets of 12 simulated agents, the results for model recovery is shown in Figure 3.8, and the results for parameter recovery are reported in section B.3, Figure B.6, and Figure B.7.

3.4.8 EEG Analysis

Participant-based regression analysis

Given N trials (across all episodes of both blocks) of a given participant, the matrix X_{raw} for this participant is an N by 5 matrix whose rows correspond to trials and whose columns correspond to normalized model variables (i.e., Surprise, Novelty, NPE, RPE, and Reward). For example, if the sequence of reward prediction error values for this participant is $z_{1:N}$, then one column of the matrix X_{raw} is equal to $(z_{1:N} - \mu_z)/\sigma_z$ where μ_z is the mean and σ_z is the standard deviation of $z_{1:N}$, and one row of the matrix X_{raw} is equal to the normalized values of Surprise, Novelty, NPE, RPE, and Reward for one trial. We constructed the feature matrix X from X_{raw} by applying the following steps: (i) we put 2 columns of X to be equal to normalized Reward plus RPE and Reward minus RPE, calling them R_+ and R_-, respectively; since Reward and RPE were normalized, their sum and difference correspond to their principal components (section B.2); (ii) we orthogonalized each of the other variables to R_+ and R_-. For example, $\text{NPE}_\perp$ is NPE minus its projection on R_+ and R_-, followed by a renormalization step (see section B.2 and Figure B.8).

For each trial, time of the ERP is measured with respect to the image onset. For a given time point, we defined the target vector y as an N dimensional vector whose elements are equal to the normalized (zero mean, unit variance) amplitude of ERPs at that particular time point in different trials. Since we have 65 time points, the response matrix Y is a N by 65 matrix. We then performed multivariate linear regression (MLR) by considering $\hat{y} = X\beta$ as an estimation of y and found β by ordinary least squared error minimization. The encoding power for a single time point and for the given participant was calculated as adjusted R-squared (Miles, 2005). Note that adjusted R-squared can in principle be negative – which is the case for our regression analysis over baseline in Figure 3.10A.

Figure 3.10A shows the mean and the standard error of the mean of the encoding power over participants and for each time-point. The threshold for rejecting the null hypothesis is computed using the Benjamini and Hochberg algorithm (Efron and Hastie, 2016) for controlling false discovery rate (FDR) by 0.1. Figure 3.10B shows the average (over participants) of β values as a function of time. For Figure 3.10C, we first average the β values over time within each time window, and then evaluate their mean and their standard error of the mean (over participant). The FDR correction was done separately for each time window.

Grand correlation analysis

Similar to the approach of Makeig et al. (2004), we pooled all trials of all participants together, i.e., we concatenated X_{raw}s and Ys for different participants. However, before concatenation, to remove the difference in the between-participant variations of ERPs energy (i.e., 2nd moment), we divided ERPs of each participant by the overall squared-energy of that participant's ERPs, i.e., we replaced Y by $Y/\sqrt{\mathbb{E}[Y^2]}$. The correlations in Figure 3.9 are computed between columns of concatenated X_{raw}s and concatenated Ys. For RPE, we removed the trials corresponding to 1st episodes of the 1st blocks because RPEs are exactly equal to zero.

4 Zooming out: Surprise and novelty in the brain

This chapter was published in Current Opinion in Neurobiology (Modirshanechi et al., 2023a)[1]. In this thesis, it plays the role of intermediate summary and discussion.

Authors: Alireza **Modirshanechi**, Sophia Becker, Johanni Brea, and Wulfram Gerstner

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.

Funding: This research was supported by the Swiss National Science Foundation (no. 200020 184615).

Author contribution: All authors contributed to the conceptualization of the study. AM did the formal analyses and visualization and wrote the original draft. All authors revised the text.

[1] For consistency across the thesis chapters, the mathematical notation has been slightly adjusted.

4.1 Introduction

An unexpected video interruption strengthens human memory of the video's content (Sinclair et al., 2021), mismatches between visual flow and locomotion facilitate synaptic changes in the mouse visual cortex (Jordan and Keller, 2023), monkeys show faster saccades to unseen objects than to familiar ones (Ogasawara et al., 2022), and mice have a higher breathing frequency when sniffing new odors than those already known (Morrens et al., 2020).

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 generalizations 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 (Schomaker and Meeter, 2015), psychology (Reisenzein et al., 2019), and machine learning (Ladosz et al., 2022). However, it has remained a mystery how humans' self-reported degree of 'surprise' when entering a new and unexpected room (Schützwohl and Reisenzein, 2012) relates to the brain activity of monkeys seeing 'surprising' fractals (Zhang et al., 2022). This is particularly worrisome as the words 'surprise' and 'novelty', sometimes used interchangeably, refer to different measurable variables in different studies (Barto et al., 2013; Modirshanechi et al., 2022). Moreover, neural and behavioral signatures of several novelty- or surprise-related variables have been found simultaneously in single experiments (Gijsen et al., 2021; Kolossa et al., 2015; Maheu et al., 2019; Visalli et al., 2021; Xu et al., 2021).

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.

4.2 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 (Gläscher et al., 2010; Mars et al., 2008), observations violating repeating patterns (Fiser et al., 2016; Homann et al., 2022; Nassar et al., 2012; Ostwald et al., 2012), or, in general, any observation that can *intuitively* be called 'novel' or 'surprising' (e.g., Figure 4.1A1). The goal of these experiments is to study how novel or surprising observations influence physiological brain signals (Antony et al., 2021; Maheu et al., 2019) and action choices (Behrens et al., 2007; Xu et al., 2021) (Figure 4.1A2).

In an example of human multi-step decision-making (Xu et al., 2021), 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 4.1A1). 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 4.1A1). Participants will feel surprised if they see a different image than the expected one (e.g., the thumb at $t = 35$ in Figure 4.1A1). 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 4.1B1) and (ii) a formal description of subjects' perception and behavior (Figure 4.1B2). We can describe most existing experiments on surprise and novelty by using only three variables at time $t + 1$ (Figure 4.1B1): The observation y_{t+1}, a potential **cue** x_{t+1}, and a set of **hidden parameters** θ_{t+1} (Table 4.1; Modirshanechi et al. (2022)). The cue x_{t+1} summarizes 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 4.1B1). 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} summarizes the hidden rules (for example action-dependent transitions in Figure 4.1B1) 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 $(x_1, y_1; ...; x_t, y_t)$ and some internal update dynamics to make a **prediction** of the next observation $\hat{y}_{t+1}$ (Table 4.1) and, if required, select an action a_t accordingly

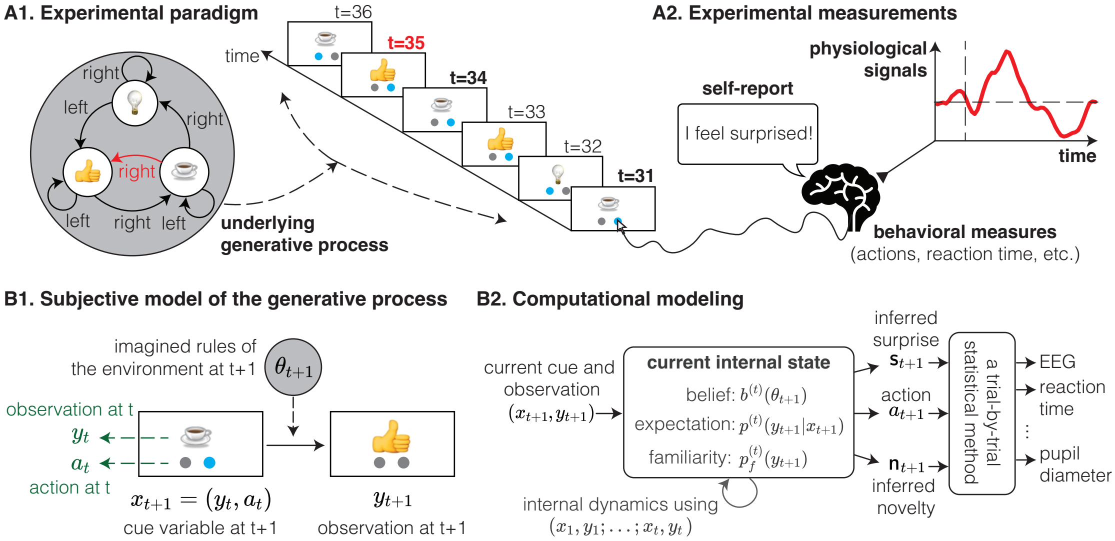


Figure 4.1: **Computational modeling of experimental paradigms studying surprise and novelty. A.** The goal of experiments on surprise and novelty is to study the influence of 'novel' or 'surprising' observations (A1) on various behavioral and physiological measurements (A2). The example in A1 shows a simplified version of the task of Xu et al. (2021): 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 4.2A and Modirshanechi et al. (2022) 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} (Modirshanechi et al., 2022). 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 $(x_1, y_1; ...; x_t, y_t)$. In general, the internal state includes a belief $b^{(t)}(\theta_{t+1})$ about the parameter set θ_{t+1}, a predictive model $p^{(t)}(y_{t+1}|x_{t+1})$ to summarize the subject's expectations (e.g., Equation 4.1), and a familiarity measure $p_f^{(t)}(y_{t+1})$ to quantify the familiarity of observations (e.g., Equation 4.2); see Table 4.1. Novelty and surprise values of each observation are evaluated according to the internal state of the algorithm as in Equation 4.3 and Equation 4.4, respectively. These values are used for trial-by-trial prediction of experimental measurements (e.g., using linear regression). See Modirshanechi et al. (2022); Xu et al. (2021) for precise definitions and Findling et al. (2021); Maheu et al. (2019) for some examples.

(Figure 4.1B2) (Friston, 2010; Meyniel et al., 2016; Soltani and Izquierdo, 2019; Yu and Dayan, 2005). Depending on the model assumptions, the internal dynamics can have different levels of abstractions (Marr, 1982), ranging from algorithmic implementations of Bayesian inference (Baldi, 2002; Liakoni et al., 2021; Piray and Daw, 2021b; Schmidhuber, 2010) to detailed models of biological neural networks (Barry and Gerstner, 2022; Berlemont and Nadal, 2022; Iigaya, 2016; Wilmes et al., 2023). In the most general setting, the model describes (i) the **belief** $b^{(t)}(\theta_{t+1})$ 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 (Table 4.1). The belief $b^{(t)}(\theta_{t+1})$ 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 $p^{(t)}(y_{t+1}|x_{t+1})$ summarizes subjects' **expectations** of what they might observe next (Table 4.1). For example, in a simple case where x_{t+1} and y_{t+1} take discrete values, we can define the predictive distribution as (Meyniel et al., 2016; Modirshanechi et al., 2019)

$$p^{(t)}(y_{t+1}|x_{t+1}) = \frac{C^{(t)}(y_{t+1}|x_{t+1}) + \text{constant}}{C^{(t)}(x_{t+1}) + \text{constant}}, \tag{4.1}$$

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.

4.3 Novelty is not surprise

Homann et al. (2022) 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 4.2A1 versus Figure 4.2A2; Table 4.1). In the computational framework described above, this means that the physiological variables studied by (Homann et al., 2022) 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 (Table 4.1).

These experimental results support the earlier proposition of Xu et al. (2021) to separate notions of 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** $p_f^{(t)}(y_{t+1})$ (Table 4.1 and Figure 4.2). 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 Bellemare et al. (2016) 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 4.1), one can define familiarity as the observation frequency

$$p_f^{(t)}(y_{t+1}) = \frac{C^{(t)}(y_{t+1}) + \text{constant}}{t + \text{constant}}, \tag{4.2}$$

Table 4.1: **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 4.1A1) or the previous observation (Figure 4.2A) 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 4.2A). 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., Behrens et al. (2007); Nassar et al. (2012).

When describing an **experimental subject**

- The **Belief** summarizes 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** summarize 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 4.2A).
- **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 4.2A3), while a familiar observation can be *unexpected* (Figure 4.2A2). Whether an observation is unfamiliar does not depend on the *cue.*

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 $\mathsf{n}_{t+1} = -\log p_f^{(t)}(y_{t+1})$ ('frequency-based novelty'; Figure 4.2B) explains some significant trial-by-trial variabilities of human EEG signals (Xu et al., 2021). More generally, novelty of y_{t+1} can be defined as

$$\mathsf{n}_{t+1} = \mathsf{N}^{(t)}\left(y_{t+1}\right), \tag{4.3}$$

where $\mathsf{N}^{(t)}$ 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 4.1B2).

The central criterion proposed by Xu et al. (2021) 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 4.3 (Barto et al., 2013; Schomaker and Meeter, 2015). 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 4.2B) (Cogliati Dezza et al., 2021; Gershman et al., 2017) 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 4.2B) (Dubey and Griffiths, 2019; Gershman and Niv, 2015). Note that according to novelty definitions based on observation frequency (e.g., Equation 4.2), 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. (2021) enables new interpretations of earlier results: For example, the separate MEG signatures found by Maheu et al. (2019) 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 Lecaignard et al. (2022) can be seen as novelty; and what has been called 'contextual novelty' in neuroscience (Schomaker and Meeter, 2015) 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 (Mehrpour et al., 2021); 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 (Homann et al., 2022; Meyer and Rust, 2018; Zhang et al., 2022). Novelty definitions based on the simple observation frequency in Equation 4.2 can be generalized to account for feature-dependent novelty estimation as the familiarity measure $p_f^{(t)}(y_{t+1})$

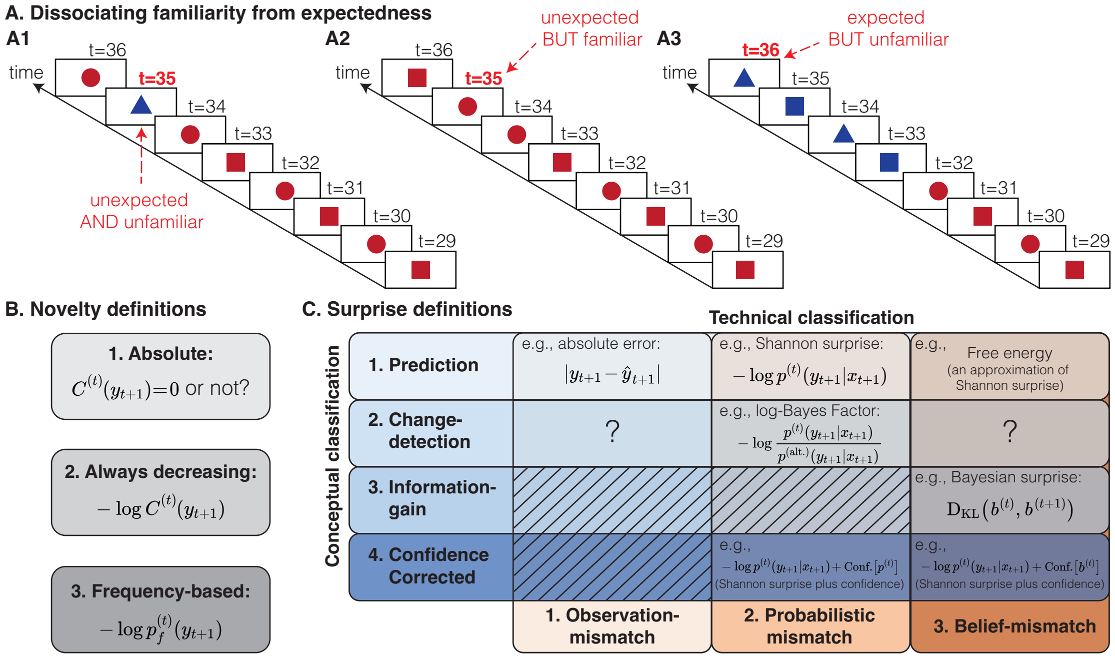

A. Dissociating familiarity from expectedness
A1
time
t=36
t=35
t=34
t=33
t=32
t=31
t=30
t=29
unexpected
AND unfamiliar
A2
unexpected
BUT familiar
A3
expected
BUT unfamiliar
B. Novelty definitions
1. Absolute:
$C^{(t)}(y_{t+1})$=0 or not?
2. Always decreasing:
$-\log C^{(t)}(y_{t+1})$
3. Frequency-based:
$-\log p_f^{(t)}(y_{t+1})$
C. Surprise definitions
Technical classification
Conceptual classification
1. Prediction
2. Change-detection
3. Information-gain
4. Confidence Corrected
e.g., absolute error:
$|y_{t+1} - \hat{y}_{t+1}|$
e.g., Shannon surprise:
$-\log p^{(t)}(y_{t+1}|x_{t+1})$
e.g., Free energy (an approximation of Shannon surprise)
?
e.g., log-Bayes Factor:
$-\log \frac{p^{(t)}(y_{t+1}|x_{t+1})}{p^{(\text{alt.})}(y_{t+1}|x_{t+1})}$
?
e.g., Bayesian surprise:
$\mathrm{D}_{\mathrm{KL}}\left(b^{(t)}, b^{(t+1)}\right)$
e.g.,
$-\log p^{(t)}(y_{t+1}|x_{t+1}) + \mathrm{Conf.}[p^{(t)}]$
(Shannon surprise plus confidence)
e.g.,
$-\log p^{(t)}(y_{t+1}|x_{t+1}) + \mathrm{Conf.}[b^{(t)}]$
(Shannon surprise plus confidence)
1. Observation-mismatch
2. Probabilistic mismatch
3. Belief-mismatch

Figure 4.2: **A taxonomy of surprise and novelty definitions.** Novelty quantifies the unfamiliarity of an observation (Equation 4.3), whereas surprise quantifies its unexpectedness conditioned on the cue variable x_{t+1} (Equation 4.4) (Xu et al., 2021). **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., Gijsen et al. (2021); Homann et al. (2022); Zhang et al. (2022); Table 4.1). 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 2nd 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 4.2). **C.** A technical classification of surprise definitions (columns) (Modirshanechi et al., 2022): 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 $b^{(t)}(\theta_{t+1})$; D_{KL} denotes Kullback-Leibler divergence. An additional conceptual classification of surprise definitions (rows) (Modirshanechi et al., 2022): 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 $p^{(\text{alt.})}(y_{t+1}|x_{t+1})$ (Liakoni et al., 2021). 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 categorization 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 $b^{(t)}$ to evaluate an information-gain surprise, and it is not possible to define confidence without access to $b^{(t)}$ or $p^{(t)}$ (hatched boxes). Question marks: Categories without any example in the literature. See Modirshanechi et al. (2022) for a detailed mathematical treatment of different definitions, their placement in the categories, and their relationships.

can be an arbitrary (non-negative and normalized) function of the stimulus. Analogously, count-based novelty definitions can account for feature-dependent novelty estimation by turning to frequency-based pseudo-counts (Bellemare et al., 2016; Jaegle et al., 2019).

4.4 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 (Baldi, 2002; Barto et al., 2013; Faraji et al., 2018; Kolossa et al., 2015; Liakoni et al., 2021; Macedo et al., 2004; Palm, 2012; Schmidhuber, 2010). In general, surprise of y_{t+1} can be written as

$$\mathsf{s}_{t+1} = \mathsf{S}^{(t)}\left(y_{t+1}|x_{t+1}\right), \tag{4.4}$$

where $\mathsf{S}^{(t)}$ is a general function that (i) takes both y_{t+1} and x_{t+1} as arguments (in contrast to Equation 4.3) and (ii) depends on the subject's current internal state at time t (Figure 4.1B2) (Modirshanechi et al., 2022). A recent systematic taxonomy of commonly used definitions of surprise proposes two classification schemes for these definitions (Modirshanechi et al., 2022) (Figure 4.2C).

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 4.2C): 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 $\mathsf{s}_{t+1} = |y_{t+1} - \hat{y}_{t+1}|$, where $\hat{y}_{t+1}$ is, e.g., the mean of the predictive distribution (Nassar et al., 2010). 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 $\mathsf{s}_{t+1} = -\log p^{(t)}(y_{t+1}|x_{t+1})$ (Barto et al., 2013). 3. *Belief-mismatch* surprise can be evaluated only by having access to the full belief $b^{(t)}(\theta_{t+1})$ 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 $\mathsf{s}_{t+1} = \mathrm{D}_{\mathrm{KL}}(b^{(t)}, b^{(t+1)})$, where D_{KL} denotes Kullback-Leibler divergence (Baldi, 2002; Schmidhuber, 2010).

The second classification is a conceptual one (rows in Figure 4.2C): 1. *Prediction* surprise defines surprising events as those that violate predictions, e.g., the Shannon surprise $\mathsf{s}_{t+1} = -\log p^{(t)}(y_{t+1}|x_{t+1})$. 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 $\mathsf{s}_{t+1} = -\log\left[p^{(t)}(y_{t+1}|x_{t+1})/p^{(\mathrm{alt.})}(y_{t+1}|x_{t+1})\right]$, where $p^{(\mathrm{alt.})}(y_{t+1}|x_{t+1})$ is a prior or naive predictive model (Liakoni et al., 2021). According to change-detection surprise definitions,

Table 4.2: **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' list 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., English et al. (2023); Rubin et al. (2016)) 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)	Yes	1. Six definitions of prediction surprise 2. Two definitions conf.-corrected surprise	Humans	Questionnaire	1. Self-report
O'Reilly et al. (2013)	Yes	1. Shannon surprise 2. Bayesian surprise	Humans	Visual	1. fMRI 2. Pupillometry 3. Reaction time
Kolossa et al. (2015)	Yes	1. Shannon surprise 2. Bayesian surprise 3. Postdictive surprise	Humans	Visual	1. EEG
Maheu et al. (2019)	Yes	1. Shannon surprise 2. Frequency-based novelty	Humans	Auditory	1. MEG
Visalli et al. (2021, 2023, 2019)	Yes	1. Shannon surprise 2. Bayesian surprise	Humans	Visual	1. EEG 2. fMRI 3. Reaction time
Dubey and Griffiths (2019)	Yes	1. Information-gain 2. Always decreasing novelty	Humans	Questionnaire	1. Action choices
Xu et al. (2021)	Yes	1. Shannon/Bayes Factor surprise 2. Frequency-based novelty	Humans	Visual	1. EEG 2. Action choices
Gijsen et al. (2021) and Grundei et al. (2023)	Yes	1. Shannon surprise 2. Bayesian surprise 3. Conf.-Corrected surprise	Humans	1. Somatosensory 2. Auditory 3. Visual	1. EEG
Modirshanechi et al. (2023c)	Yes	1. Shannon surprise 2. Postdictive surprise 3. Frequency-based novelty	Humans	Visual	1. Action choices
Morrens et al. (2020)	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)	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)	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)	No	1. Novel objects 2. Familiar objects in unexp. context	Mice	Visual	1. OG recording of DA 2. Action choices
Garrett et al. (2023) (see also Garrett et al. (2020))	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

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 (Liakoni et al., 2021; Modirshanechi et al., 2022) (Table 4.1), in agreement with experimental observations (Behrens et al., 2007; Nassar et al., 2012; Pasturel et al., 2020). 3. *Information-gain* surprise defines surprising events as those that change a subject's belief about the world, e.g., the Bayesian surprise $\mathsf{s}_{t+1} = \mathrm{D}_{\mathrm{KL}}(b^{(t)}, b^{(t+1)})$. We note, however, that only a handful of information-gain measures (Nelson, 2005) have been previously interpreted as measures of surprise (Baldi, 2002; Kolossa et al., 2015; Schmidhuber, 2010). 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** (Table 4.1); examples have been suggested both in neuroscience (Faraji et al., 2018) and psychology (Macedo et al., 2004).

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 (Barry and Gerstner, 2022; Iigaya, 2016; Wilmes et al., 2023); see Fiser et al. (2010); Knill and Pouget (2004) 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 (Itti and Baldi, 2009), exploration (Gottlieb and Oudeyer, 2018), and learning (Soltani and Izquierdo, 2019).

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 4.2). For example, Gijsen et al. (2021) 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) 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-trail data analysis (Table 4.2). For example, Zhang et al. (2022) 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 (Table 4.1). 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) (Hayden et al., 2011; Rouhani and Niv, 2021) measures the mismatch between the reward $r(y_{t+1})$ associated with stimulus y_{t+1} and a prediction $\hat{r}_{t+1}$ thereof (see Modirshanechi et al. (2022)). Similarly, an unsigned novelty prediction error (uNPE) measures the unexpectedness of the novelty value $\mathsf{N}^{(t)}(y_{t+1})$ of an observation y_{t+1} (Kakade and Dayan, 2002; Xu et al., 2021). 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 (Kakade and Dayan, 2002; Xu et al., 2021). Moreover, subjects may assume potential associations between novelty (or similarly between surprise) and threats or rewards (Akiti et al., 2022; Gershman and Niv, 2015), 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 (Grundei et al., 2023)) or the hierarchy of representations (e.g., image identity (Mehrpour et al., 2021; Xu et al., 2021) versus primary visual features (Homann et al., 2022; Jordan and Keller, 2023)). For example, a repeating sequence of binary observation as in Figure 4.2A can be presented as either a sequence of tones or a sequence of images (i.e., different modalities); Grundei et al. (2023) 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.

4.5 Towards a systematic study of surprise and novelty

Different computational roles in learning (Pearce and Hall, 1980; Piray and Daw, 2021b) and decision-making (Berlyne, 1950; Horvath et al., 2021; Schulz and Gershman, 2019), 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. 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. (2022) 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) 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 (Modirshanechi et al., 2023c). 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 4.2 offer a foundation for future experimental and theoretical studies on surprise and novelty.

5 Even if suboptimal, novelty drives human exploration

An *earlier* version of this chapter is also available as a pre-print on bioRxiv (Modirshanechi et al., 2023c).

Authors: Alireza **Modirshanechi**, Wei-Hsiang Lin, He A. Xu, Michael H. Herzog, and Wulfram Gerstner

Abstract: Human exploration has been modeled by reinforcement learning algorithms that use intrinsic rewards to guide their search for extrinsic rewards. However, different choices of intrinsic reward result in different exploration strategies, and some of these strategies are prone to sub-optimal attraction to reward-independent stochastic stimuli. Here, we ask whether humans get attracted to the same stimuli as the algorithms and, if so, which choice of intrinsic reward explains their exploration best. We design an experimental paradigm for multi-step decision-making where human participants search for rewarding states in an environment with a highly 'stochastic' but reward-free subregion. We show that (i) participants repeatedly and persistently explore the stochastic part of the environment and (ii) their action choices are best explained by algorithms driven by novelty but not by 'optimal' algorithms driven by information gain. Our results suggest that humans use suboptimal but computationally cheap strategies for exploration in complex environments.

Funding: This research was supported by the Swiss National Science Foundation No. CRSII2 147636 (Sinergia, MHH and WG), No. 200020 184615 (WG), and No. 200020 207426 (WG) and by the European Union Horizon 2020 Framework Program under grant agreement No. 785907 (Human Brain Project, SGA2, MHH and WG).

Author contribution: AM, HAX, MHH, and WG developed the study concept and designed the experiment. HAX and WL conducted the experiment and collected the data. AM designed the algorithms, did the formal analyses, and analyzed the data. AM, MHH, and WG wrote the paper.

5.1 Introduction

Humans and animals consistently explore their environments, potentially driven by objectives such as finding more valuable sources of reward (e.g., more nutritious foods or better-paid jobs) than those currently available (Cohen et al., 2007; Gottlieb et al., 2013; Kidd and Hayden, 2015; Schulz and Gershman, 2019). This exploratory behavior has been modeled by Intrinsically Motivated Reinforcement Learning (RL) algorithms (Gottlieb and Oudeyer, 2018; Jaegle et al., 2019; Modirshanechi et al., 2023b; Murayama, 2022; Oudeyer et al., 2007) that, initially inspired by research in psychology (Santucci et al., 2013; Singh et al., 2010b), were designed to solve complex machine learning tasks with sparse 'extrinsic' rewards (Bellemare et al., 2016; Haber et al., 2018; Kim et al., 2020b; Mendonca et al., 2021; Ostrovski et al., 2017; Pathak et al., 2017; Sekar et al., 2020). These algorithms use internally generated signals like 'novelty', 'surprise', or 'information gain' as 'intrinsic' rewards to guide exploratory action choices (Ladosz et al., 2022; Santucci et al., 2013; Singh et al., 2010b). Since different intrinsic rewards result in different exploration strategies (Aubret et al., 2019; Ladosz et al., 2022), a crucial challenge in computational neuroscience is to identify the intrinsic reward that best describes exploration in humans and animals (Modirshanechi et al., 2023b).

Our goal in this study is to identify the dominant drive of human *goal-directed* exploration in complex *multi-step* tasks with sparse 'extrinsic' rewards. We define exploratory actions as *goal-directed* when they are aimed at searching for extrinsic rewards. We define *multi-step* tasks as those where a single action, or even a pair of actions, does not lead to an extrinsic reward or provide final information about the resolution of a 'puzzle' (i.e., 'non-instrumental' information; Gottlieb and Oudeyer (2018)). Instead, success, in terms of reward or non-instrumental information, requires many successive actions. Exploration in such multi-step tasks had not been explored until recently (Fox et al., 2023; Xu et al., 2021) and is qualitatively different from exploration in widely studied 1-step and 2-step tasks for reward (Cockburn et al., 2022; Daw et al., 2011; Gershman, 2019a; Gläscher et al., 2010; Horvath et al., 2021; Wilson et al., 2014; Wu et al., 2018; Zajkowski et al., 2017) or non-instrumental information (Daddaoua et al., 2016; Kobayashi et al., 2019; Ogasawara et al., 2022; Poli et al., 2022; Ten et al., 2021); see Brändle et al. (2022); Gottlieb and Oudeyer (2018); Modirshanechi et al. (2023b) for recent reviews.

In particular, multi-step environments with a stochastic component enable the dissociation of different intrinsic rewards based on their behavioral signatures. Machine learning research has shown that intrinsically motivated RL agents are prone to distraction by stochasticity (the so-called 'noisy TV' problem; Aubret et al. (2019); Ladosz et al. (2022)), i.e., they are attracted to novel, surprising, or just noisy states independently of whether or not these states are rewarding (Burda et al., 2019). However, the extent of distraction varies between different algorithms (Jarrett et al., 2022; Mavor-Parker et al., 2022; Pathak et al., 2019; Savinov et al., 2019). It is well-known that artificial RL agents seeking information gain eventually lose their interest in stochasticity when exploration yields no further information, whereas RL agents seeking surprise or novelty exhibit a persistent or increasing attraction by stochasticity (Aubret et al., 2019; Ladosz et al., 2022; Schmidhuber, 2010). Here, we ask (i) whether humans get distracted in the same situations as intrinsically motivated RL agents and, if so, (ii) whether this distraction vanishes (similar to seeking information gain) or persists (similar to seeking surprise or novelty) over time.

To answer these questions, we bring ideas from machine learning (Aubret et al., 2019; Ladosz et al., 2022) to behavioral neuroscience and design a novel multi-step decision-making paradigm in a complex environment with a highly stochastic but reward-free sub-region. We test the predictions of three different intrinsically motivated RL algorithms (i.e., driven by novelty, surprise, and information gain) against the behavior of human participants and show that human behavior is both qualitatively and quantitatively consistent with that of novelty-driven RL agents: When searching for extrinsic rewards, human participants exhibit a persistent attraction to novelty signals in the stochastic sub-region. Our results provide evidence for novelty-driven RL algorithms as models of human goal-directed exploration even when novelty-seeking is suboptimal.

5.2 Results

5.2.1 Experimental paradigm

We employ a multi-step decision-making paradigm (Lehmann et al., 2019; Liakoni et al., 2022; Tartaglia et al., 2017) for navigation in an environment with 58 states plus three goal states (Figure 5.1A-B). Three actions are available in each non-goal state, and agents can move from one state to another by choosing these actions (arrows in Figure 5.1A-B). We use the term 'agents' to refer to either human participants or agents simulated by RL algorithms. In the human experiments, states are represented by images on a computer screen and actions by three disks below each image (Figure 5.1C); for simulated participants, both states and actions are abstract entities (i.e., we consider RL in a tabular setting (Sutton and Barto, 2018)). The assignment of images to states and disks to actions is random but fixed throughout the experiment. Agents are informed that there are three different goal states in the environment (G^*, G_1, or G_2 in Figure 5.1A) and that their task is to find a goal state 5 times; see Methods for how this information is incorporated in the RL algorithms. Importantly, neither human participants nor RL agents are aware of the total *number* of states or the *structure* of the environment (i.e., how states are connected to each other).

The 58 states of the environment can be classified into three groups: Progressing states (1 to 6 in Figure 5.1A), trap states (7 and 8 in Figure 5.1A), and stochastic states (S-1 to S-50 in Figure 5.1B, shown as a dashed oval in Figure 5.1A). In each progressing state, one action ('progressing' action) takes agents one step closer to the goals and another

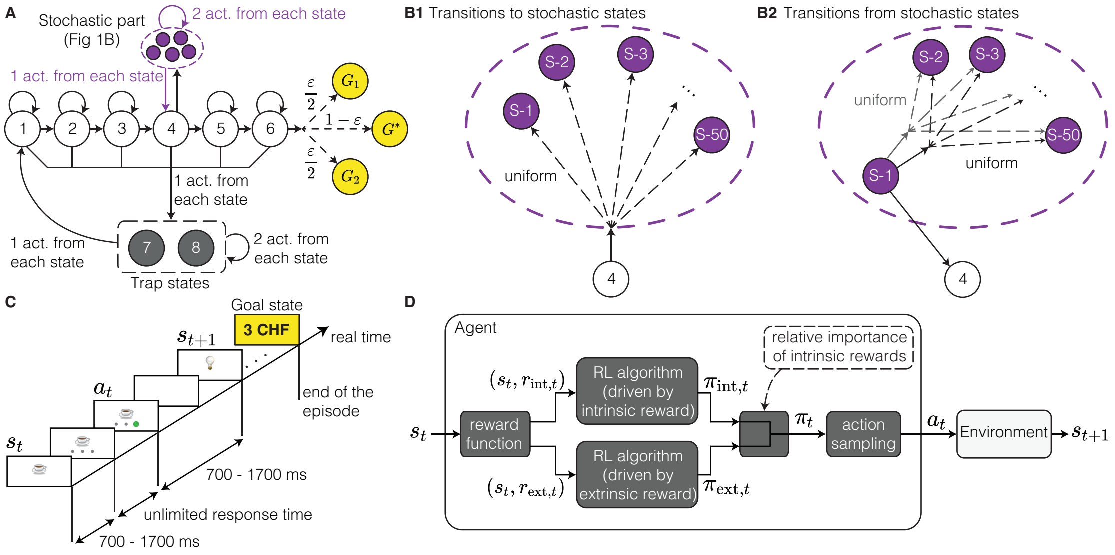


Figure 5.1: **Experimental paradigm and computational model. A.** Structure of the environment with the stochastic states merged (dashed oval; see B). Each circle represents a state and each solid arrow an action. All actions except for the ones to the stochastic part or to the goal states are deterministic. Dashed arrows indicate random transitions; values (e.g., $1-\varepsilon$) show the probabilities of each transition. We choose $\varepsilon \ll 1$ (see Methods). **B.** Structure of the stochastic part of the environment (states S-1 to S-50), i.e., the dashed oval in A. **B1.** In state 4, one action takes agents randomly (with uniform distribution) to one of the stochastic states. **B2.** In each stochastic state (e.g., state S-1 in the figure), one action takes agents back to state 4 and two actions to another randomly chosen stochastic state. **C.** Time-line of one episode in human experiments. The states are represented by images on a computer screen and actions by disks below each image. An episode ends when a goal image (i.e., '3 CHF' image in this example) is found. **D.** Block diagram of the intrinsically motivated RL algorithm. Given the state s_t at time t, the intrinsic reward $r_{\text{int},t}$ (i.e., novelty, information gain, or surprise) and the extrinsic reward $r_{\text{ext},t}$ (i.e., the monetary reward value of s_t) are evaluated by a reward function and passed to two identical (except for the reward signals) parallel RL algorithms. The two algorithms compute two policies, one for seeking intrinsic reward $\pi_{\text{int},t}$ and one for seeking extrinsic reward $\pi_{\text{ext},t}$. The two policies are then weighted according to the relative importance of the intrinsic reward and are combined to make a single hybrid policy π_t. The next action a_t is selected by sampling from π_t. See Methods for details.

action ('bad' action) takes them to one of the trap states. The third action in states 1-3 and 5-6 is a 'self-looping' action that makes agents stay in the same state. Except for the progressing action in state 6, all these actions are deterministic, meaning that they always lead to the same next state. The progressing action in state 6 is *almost* deterministic: It takes participants to the 'likely' goal state G^* with a probability of $1-\varepsilon$ and to the 'unlikely' goal states G_1 and G_2 with equal probabilities of $\frac{\varepsilon}{2} \ll 1$. In state 4, instead of a self-looping action, there is a 'stochastic' action that takes agents to a randomly chosen (with equal probability) stochastic state (Figure 5.1B1). In each stochastic state, one action takes agents back to state 4, and two stochastic actions take them to *another* randomly chosen stochastic state (Figure 5.1B2). In each trap state, all three actions are deterministic: Two actions bring agents to either the same or the other trap state and one action to state 1.

The stochastic part of the environment – which is inspired by the machine-learning literature and mimics the main features of a 'noisy TV' (Burda et al., 2019) – is a crucial difference to existing paradigms in the literature of behavioral neuroscience (Daw et al., 2011; Fox et al., 2023; Huys et al., 2015; Xu et al., 2021). Without the stochastic part, intrinsic motivation helps agents to avoid the trap states and find the goal (Xu et al., 2021), hence it helps exploration before and does not harm exploitation after finding a goal. By adding the stochastic part, we aim to identify the dominant drive of exploration and quantify its influence on the exploitation of the discovered goal versus attraction to the stochastic part.

We organize the experiment in 5 episodes: Agents are randomly initialized at state 1 or 2 and are instructed to find a goal 5 times. After finding a goal, agents are randomly re-initialized at state 1 or 2. We choose a small enough ε (Figure 5.1A) to safely assume that all agents visit only G^* while being aware that G_1 and G_2 exist (Methods).

5.2.2 Simulating intrinsically motivated agents with efficient algorithms

We simulate three intrinsically motivated RL algorithms to test whether our experimental paradigm dissociates the exploration strategies driven by different intrinsic reward signals. Agents simulated by each algorithm can navigate in an environment with an unknown number of states by seeking a combination of extrinsic and intrinsic rewards (Figure 5.1D). Intrinsic rewards are given to agents by themselves and upon visiting 'novel', 'surprising', or 'informative' states, whereas extrinsic rewards are received only when visiting the three goal states (see Methods for details). Specifically, at each time t, an agent observes state s_t and evaluates an extrinsic reward value $r_{\text{ext},t}$ (which is zero except at the goal states) and an intrinsic reward value $r_{\text{int},t}$ (e.g., novelty of state s_t). Extrinsic and intrinsic reward values are then passed to two parallel blocks of RL, each working with a single reward signal. Independently of each other, the two blocks use efficient model-based planning (Mattar and Lengyel, 2022; Van Seijen and Sutton, 2013) to propose a policy $\pi_{\text{ext},t}$ that maximizes future extrinsic rewards and $\pi_{\text{int},t}$ that maximizes future intrinsic rewards (Aubret et al., 2019; Xu et al., 2021), respectively. The two policies are combined into a hybrid policy π_t for taking the next action a_t, controlled by a set of free parameters that indicate the relative importance of intrinsic over extrinsic rewards (Methods). The degree of exploration is high if $\pi_{\text{int},t}$ dominates $\pi_{\text{ext},t}$ during action selection.

For the intrinsic reward $r_{\text{int},t}$, we choose one option from each of the three main categories of intrinsic rewards in machine learning (Aubret et al., 2019; Ladosz et al., 2022): (i) novelty (Bellemare et al., 2016; Ostrovski et al., 2017; Xu et al., 2021) quantifies how infrequent the state s_t has been until time t; (ii) information gain (Little and Sommer, 2013; Mendonca et al., 2021; Mobin et al., 2014; Sekar et al., 2020) quantifies how much the agent updates its belief about the structure of the environment upon observing the transition from the state-action pair (s_{t-1}, a_{t-1}) to state s_t; and (iii) surprise (Burda et al., 2019; Modirshanechi et al., 2022; Pathak et al., 2017) quantifies how unexpected it is to observe state s_t after taking action a_{t-1} at state s_{t-1}.

The three different intrinsic reward signals lead to three efficient intrinsically motivated algorithms and to three groups of simulated efficient agents: those (i) seeking novelty, (ii) seeking information gain, and (iii) seeking surprise. In the following section, we characterize the behavior of these simulated efficient agents (i) to test whether our experimental paradigm dissociates action choices of different intrinsically motivated RL algorithms and (ii) to gain insights about their principal differences. These simulations can also be seen as *qualitative* predictions of different algorithms for human behavior, but we note that these predictions are made by using efficient RL algorithms with perfect memory and high computational power and must *not* be taken as precise *quantitative* predictions; we present a more realistic simulation of human behavior in a later section.

5.2.3 Different intrinsically motivated algorithms exhibit principally different behavioral patterns

To avoid arbitrariness in the choice of parameters, we fine-tune the parameters of each algorithm to have on average the lowest number of actions in episode 1 (to have the most efficient exploration; Methods). As a result, different algorithms achieve similar performance during episode 1 and find the goal G^* almost equally fast (Appendix C). Hence, exploration policies driven by different intrinsic rewards cannot be qualitatively distinguished during episode 1.

Given the same set of parameters, we study how different simulated efficient agents behave in episodes 2-5 (Figure 5.2). After finding the goal G^* for the 1st time, an agent has two options: (i) return to the discovered goal state G^* (exploitation) or (ii) search for the other goal states G_1 and G_2 (exploration). In our simulations, we consider three choices for the trade-off between exploration and exploitation by changing the relative importance of $\pi_{\text{int},t}$ over $\pi_{\text{ext},t}$ (Figure 5.1D): pure exploitation (the action policy does not depend on intrinsic rewards, i.e., $\pi_t = \pi_{\text{ext},t}$), pure exploration (the action policy does not depend on extrinsic rewards, i.e., $\pi_t = \pi_{\text{int},t}$), and a mixture of both (different shades of each color in Figure 5.2). If the extrinsic reward value assigned to G_1 or G_2 is higher than the one assigned to G^*, then the policy $\pi_{\text{ext},t}$ for seeking extrinsic rewards can also contribute to exploration in episodes 2-5 (Methods). In order to characterize qualitative features essential to exploration driven by different *intrinsic* rewards, we assume a symmetry between the three goal states in the simulated efficient agents and assign the same extrinsic reward value to all goals (Methods); we drop this assumption in the next sections and quantify the additional but negligible contribution of $\pi_{\text{ext},t}$ to explaining human exploration.

For all three groups of simulated efficient agents, decreasing the relative importance of intrinsic rewards decreases both the search duration (Figure 5.2A1, C1, and E1) and the fraction of time spent in the stochastic part (Figure 5.2A2, C2, and E2). This observation implies that intrinsically motivated exploration leads to an attraction to the stochastic

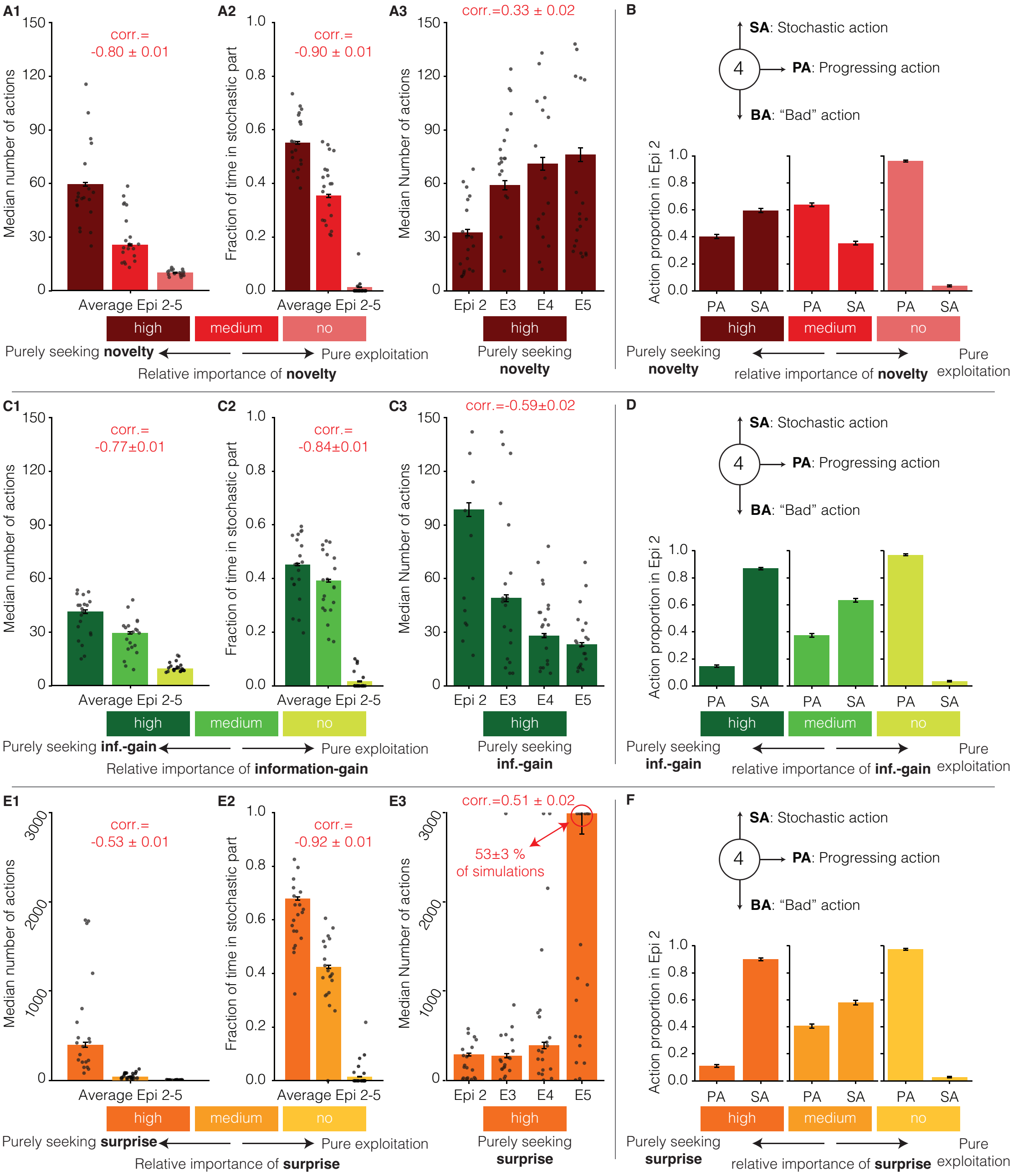


Figure 5.2: (Caption next page.)

Figure 5.2: (Previous page.) **Simulated efficient agents seeking novelty (A-B), surprise (C-D), and information gain (E-F) have principally different behavioral characteristics during episodes 2-5.** We consider three levels of importance for intrinsic rewards (Figure 5.1D): high (dark colors), medium (shaded colors), and no (light colors). For each level, we run 500 simulations of each algorithm. **A1, C1, and E1.** Median number of actions over episodes 2-5. Error bars show the standard error of the median (SEMed; evaluated by bootstrapping). Single dots show the data of 20 (randomly chosen out of 500) individual simulations to illustrate variabilities among simulations. Simulations stopped after 3000 actions even if a goal state was not reached. The Pearson correlation between the search duration and the degree of exploitation is negative (red numbers), indicating that search duration decreases if the degree of exploitation increases (Methods). **A2, C2, and E2.** Average fraction of time spent in the stochastic part of the environment during episodes 2-5. The Pearson correlation between the fraction of time spent in the stochastic part and the degree of exploitation is negative (Methods). Error bars show the standard error of the mean (SEMean) and single dots the data of 20 individual simulations. **A3, C3, and E3.** Median number of actions in episodes 2-5 for simulated efficient agents *purely* driven by intrinsic rewards (i.e., pure exploration). The Pearson correlation between the search duration and episode number is positive for seeking novelty or surprise but is negative for seeking information gain. Error bars show the SEMed and single dots the data of 20 individual simulations. **B, D, and F.** Fraction of time taking the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2. Purely seeking novelty shows a smaller difference between the preference for SA and PA in state 4 compared to purely seeking information gain or surprise. Error bars show the SEMean.

part of the environment, effectively keeping the simulated efficient agents away from the goal region beyond state 6 (Figure 5.1A). Our results thus confirm earlier findings in machine learning (Aubret et al., 2019; Burda et al., 2019) that intrinsically motivated agents get distracted by stochastic reward-independent stimuli.

While all three groups of simulated efficient agents get distracted by the stochastic part, their degree of distraction is different (different colors in Figure 5.2A3, C3, and E3). For efficient agents that purely seek information gain (i.e., pure exploration), the time spent in the stochastic part decreases over episodes (Figure 5.2C3), whereas we observe the opposite pattern for efficient agents that purely seek novelty (Figure 5.2A3) or surprise (Figure 5.2E3). In particular, efficient agents that purely seek surprise get most often (i.e., in $> 50\%$ of simulations in episode 5) stuck in the stochastic part and do not escape it within 3000 actions (Figure 5.2E3). These observations confirm the inefficiency of seeking surprise and the efficiency of seeking information gain in dealing with noise (Aubret et al., 2019).

In order to further dissociate action choices of different algorithms, we analyze the action preferences of simulated efficient agents in state 4 during episode 2 (Figure 5.2B, D, and F). For all three groups of efficient agents, increasing the relative importance of intrinsic rewards increases their preference for the stochastic action. However, for the highest importance of intrinsic rewards, the probability of choosing the progressing action is substantially lower than the probability of choosing the stochastic action for seeking surprise or information gain (15% vs. 85%; Figure 5.2D and F), whereas this difference is much smaller for seeking novelty (40% vs. 60%; Figure 5.2B).

This distinct behavior of novelty-seeking is due to the fact that novelty is defined for states, whereas surprise and information gain are defined for transitions (i.e., state-action pairs; see Methods): By the end of episode 1, the goal state has been observed only once and remains, during episode 2, relatively novel (and hence attractive for an efficient novelty-seeking agent) compared to most stochastic states, whereas there are many actions between the stochastic states that have rarely or potentially never been chosen and are, thus, attractive for an efficient agent seeking surprise or information gain.

To summarize, different intrinsically motivated algorithms exhibit principally different behavioral patterns in our experimental paradigm. We consider these behavioral patterns as *qualitative* predictions for human behavior.

5.2.4 Human participants

To characterize the key features and the dominant drive of human exploration in our experimental paradigm, we first compare the exploratory behavior of human participants with that of simulated efficient agents. For simulated efficient agents, the relative importance of the intrinsic reward for action selection (Figure 5.1D) determines the balance of exploration versus exploitation. A challenge in human experiments is that we do not have explicit control over the variable that controls the relative importance of intrinsic rewards compared to extrinsic rewards. Inspired by earlier studies (Gershman and Niv, 2015; Stojić et al., 2020; Traner et al., 2021), we conjecture that human participants who are more optimistic about finding a goal with a high value of reward are more curious to explore the environment than human participants who are less optimistic. In other words, we hypothesize that the motivation to explore and hence the relative importance of intrinsic rewards in human participants is positively correlated with their degree of 'reward optimism', where we define reward optimism as the expectancy of finding a goal of higher value than those already discovered.

Based on this hypothesis, we include a novel reward manipulation in the instructions given at the beginning of the experiment: We inform human participants that there are three different possible reward states corresponding to values of 2 Swiss Franc (CHF), 3 CHF, and 4 CHF, represented by three different images (Methods). At the beginning of the experiment, we randomly assign the three different reward values to the goal states G^*, G_1, and G_2 in Figure 5.1A, separately for each participant (without informing them), and keep the assignment fixed throughout the experiment. After this random assignment, G^* has a different value for different participants. Even though all participants receive the same instructions, participants who are randomly assigned to an environment with 4 CHF reward value for G^* do not have any monetary incentive to further explore in episodes 2-5 (4 CHF group = a low degree of reward optimism), whereas participants who are assigned to an environment with 2 CHF reward value for G^* have a monetary incentive in episodes 2-5 to further explore the environment to find a higher reward (2 CHF group = a high degree of reward optimism). We therefore expect participants in the 2 CHF group to keep searching for more valuable goals in episodes 2-5. Our goal is to characterize the exploration strategy for this search behavior. Therefore, we have three different groups of participants with three different levels of reward optimism in episodes 2-5; this information is incorporated in the RL algorithms as prior knowledge about action values and the environment structure (Methods). We note that our definition of reward optimism in the context of our experiment is in line but independent of the notion of general optimism that is quantified for individual participants in psychology (Carver et al., 2010).

Following a power analysis based on the data of simulated efficient agents (Methods), we recruited 63 human participants and collected their action choices during the 5 episodes of our experiment: 23 participants in an environment with 2 CHF reward value for G^* and two times 20 human participants in environments with 3 CHF and 4 CHF reward value for G^*, respectively. In the rest of the manuscript, we refer to each group by their reward value of G^*, e.g., the 3 CHF group is the group of human participants who were assigned to 3 CHF reward value for G^* (as in Figure 5.1C). We excluded the data of 6 human participants from further analyses since they either did not finish the experiment or had an abnormal performance (Methods).

5.2.5 Human participants exhibit a persistent attraction to stochasticity

We perform the same series of analyses on the behavior of human participants as those performed on the behavior of simulated efficient agents (Figure 5.3). In episodes 2-5, the search duration of human participants (Figure 5.3A1) and the fraction of time they spend in the stochastic part (Figure 5.3A2) are both negatively correlated with the goal value of their environment, e.g., the 2 CHF group has a longer search duration and spends more time in the stochastic part than the other two groups. Moreover, increasing the goal value increases the preference of human participants for the progressing action in state 4 during episode 2 (Figure 5.3B). These observations support our hypothesis that increasing the degree of reward optimism influences the behavior of human participants in the same way as increasing the relative importance of intrinsic rewards influences the behavior of simulated efficient agents (e.g., compare Figure 5.3A1, A2, and B with Figure 5.2A1, A2, and B, respectively).

The behavior of the 2 CHF group is particularly interesting since they are the most optimistic group of participants and have the highest motivation to search for the other goal states. The 2 CHF group exhibits a constant search duration over episodes 2-5 (zero correlation accepted by Bayesian hypothesis testing (Kass and Raftery, 1995);

Figure 5.3A3). This implies that they persistently explore the stochastic part. Moreover, during episode 2, the 2 CHF group chooses the progressing and the stochastic actions equally often (no-difference in means accepted by Bayesian hypothesis testing (Kass and Raftery, 1995); Figure 5.3B). If we assume that the high degree of reward optimism in the 2 CHF group results in a policy that is driven dominantly by intrinsic rewards (driving exploration) and only marginally by extrinsic rewards, then these observations are more similar to the qualitative predictions of seeking novelty than those of seeking information gain or surprise (compare Figure 5.3B against Figure 5.2B, D, and F).

The mismatch between the behavior of human participants (Figure 5.3) and novelty-seeking simulated efficient agents (Figure 5.2A-B) can be because (i) the action choices of the 2 CHF group are not purely exploratory, (ii) they are not as efficient as the RL algorithm used for simulating efficient agents (Figure 5.1), or (iii) novelty is not the dominant drive of human exploration. In the next section, we remove the constraints related to the first two possibilities and directly test the last one.

5.2.6 Novelty-seeking is the most probable model of human exploration

In the previous section, we observed that human participants exhibit patterns of behavior *qualitatively* similar but not identical to those of novelty-seeking simulated efficient agents. The qualitative predictions in Figure 5.2 were made based on the assumptions of (i) using efficient RL algorithms with perfect memory and high computational power, (ii) using parameters that were optimized for the best performance in episode 1, and (iii) assigning the same extrinsic reward value to different goal states. In this section, we use a more realistic model of behavior than that of efficient agents in Figure 5.2: In order to model the behavior of human participants, we use a hybrid RL model (Daw et al., 2011, 2005; Liakoni et al., 2022; Xu et al., 2021) combining model-based planning (Mattar and Lengyel, 2022) and model-free habit-formation (Gläscher et al., 2010), account for imperfect memory and suboptimal choice of parameters, and allow our algorithms to assign different extrinsic reward values to different goal states (Methods). We fit the parameters of our three intrinsically motivated algorithms to the action choices of each individual participant by maximizing the likelihood of data given parameters (Methods). Such a flexible modeling approach allows each of the three algorithms to find its closest version to the behavior of human participants, constrained on using one specific intrinsic reward signal (i.e., novelty, surprise, or information gain).

Given the fitted algorithms, we use Bayesian model-comparison (Daw, 2011; Rigoux et al., 2014) to *quantitatively* test whether human behavior is explained better by seeking novelty than seeking information gain or surprise (Methods). Our model-comparison results show that seeking novelty is the most probable model for the majority of human participants, followed by seeking information gain as the 2nd most probable model (Figure 5.4A; Protected Exceedance Probability = 0.99 and 0.01 for seeking novelty and information

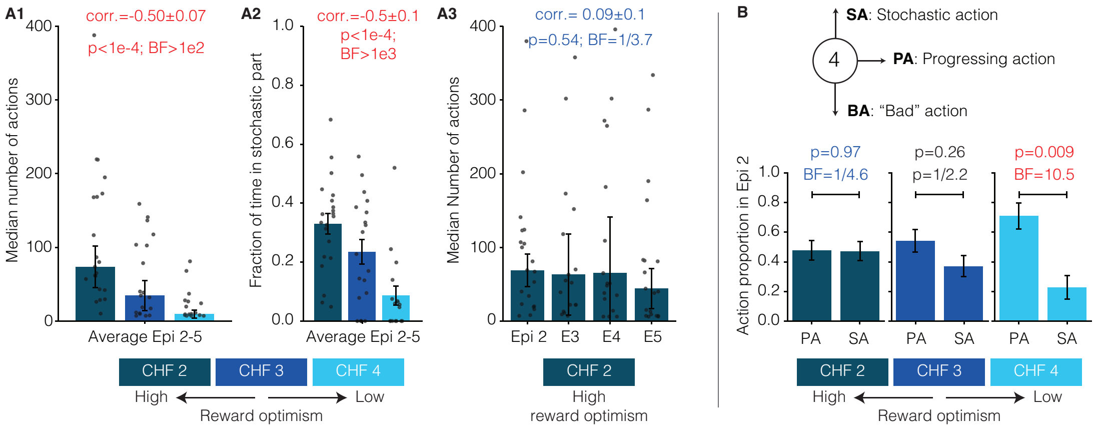


Figure 5.3: **Human participants persistently explore the stochastic part if they are highly optimistic. A.** Search duration in episodes 2-5. **A1.** Median number of actions over episodes 2-5 for the three different groups: 2 CHF (dark), 3 CHF (medium), and 4 CHF (light). Error bars show the SEMed (evaluated by bootstrapping) and single dots the data of individual participants. The Pearson correlation between the search duration and the goal value is negative (correlation test; $t = -4.2$; 95%Confidence Interval (CI) = $(-0.67, -0.27)$; Degree of Freedom (DF) = 55; Methods). **A2.** Average fraction of time spent in the stochastic part of the environment during episodes 2-5. The Pearson correlation between the fraction of time spent in the stochastic part and the goal value is negative (correlation test; $t = -4.7$; 95%CI = $(-0.70, -0.32)$; DF = 55; Methods). Error bars show the SEMean and single dots the data of individual participants. **A3.** Median number of actions in episodes 2-5 for the 2 CHF group. A Bayes Factor (BF) of 1/3.7 in favor of the null hypothesis (Kass and Raftery, 1995) suggests a zero Pearson correlation between the search duration and the episode number (one-sample t-test on individual correlations; $t = 0.63$; 95%CI = $(-0.20, 0.37)$; DF = 20). Error bars show the SEMed and single dots the data of individual participants. **C.** Fraction of time choosing the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2; see Appendix C for other progressing states. Error bars show the SEMean. The difference between PA and SA for the 4 CHF group is significant (one-sample t-test; $t = 2.99$; 95%CI = $(0.14, 0.81)$; DF = 16). A BF of 1/4.6 in favor of the null hypothesis (Kass and Raftery, 1995) suggests an equal average between PA and SA for the 2 CHF group (one-sample t-test; $t = 0.039$; 95%CI = $(-0.25, 0.26)$; DF = 20). The test for 3 CHF group is inconclusive (one-sample t-test; $t = 1.17$; 95%CI = $(-0.13, 0.47)$; DF = 18). Red p-values: Significant effects with False Discovery Rate controlled at 0.05 (Efron and Hastie, 2016) (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF$\geq$ 3). Blue BFs: Significant evidence in favor of the null hypothesis (BF$\leq$ 1/3).

gain, respectively; see Methods). This result shows that seeking novelty describes the behavior of human participants better than seeking information gain and surprise, but it does not tell us which aspects of data statistics cannot be explained by algorithms driven by information gain or surprise. To investigate this question, we use our three intrinsically motivated algorithms with their fitted parameters and simulate new participants, i.e., we perform Posterior Predictive Checks (PPC) (Nassar and Frank, 2016; Wilson and Collins, 2019). As opposed to the simulations in Figure 5.2, we do not freely choose the level of exploration in simulations for PPC. Rather, the level of exploration of each newly simulated participant is completely determined by the previously fitted parameters

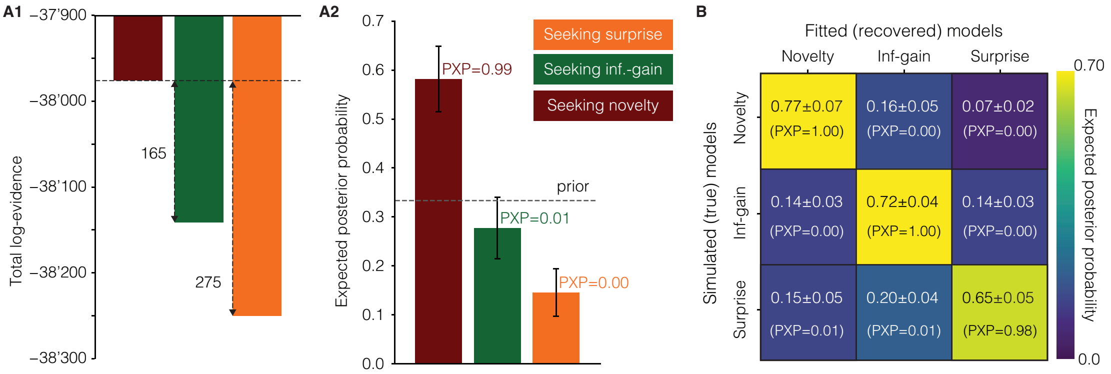


Figure 5.4: **Novelty-seeking is the most probable model of human behavior. A.** Human participants' action choices are best explained by novelty-seeking (see Methods for details). **A1.** Model log-evidence summed over all participants (i.e., assuming that different participants have the same exploration strategy but can have different parameters; see Daw (2011)) is significantly higher for seeking novelty than seeking information gain or surprise. High values indicate good performance, and differences greater than 10 are traditionally (Efron and Hastie, 2016) considered as strongly significant. **A2.** The expected posterior model probability with random effects assumption (i.e., assuming that different participants can have different exploration strategies and different parameters; see Rigoux et al. (2014)) given the data of all participants. PXP stands for Protected Exceedance Probability (Rigoux et al., 2014), i.e., the probability of one model being more probable than the others. Error bars show the standard deviation of the posterior distribution. **B.** Confusion matrix from the model recovery procedure: Each row shows the results of applying our model-fitting and -comparison procedure (as in A2) to the action choices of simulated participants by one of the three algorithms (with their parameters fitted to human data; see Methods). Color-code shows the expected posterior probability and numbers in parentheses the PXP (both averaged over 5 sets of 60 simulated participants). We could always recover the model that had generated the data (PXP ≥ 0.98), using almost the same number of simulated participants (60) as human participants (57).

from one of the 57 human participants; specifically, each simulated participant belongs to one of the three groups of human participants (e.g., the 3 CHF group), and its action choices are simulated using a set of parameters fitted to the action choices of one human participant randomly selected from the participants in that group (Methods).

Given the PPC results, we first perform model-recovery (Wilson and Collins, 2019) on the data from the simulated participants: Indeed, model recovery confirms that we can infer which algorithm has generated the action choices of simulated participants (by repeating our model-fitting and -comparison; Figure 5.4B). This implies that even the versions of different algorithms that are closest to human data can be dissociated in our experimental paradigm (average Protected Exceedance Probability ≥ 0.98 for the true model in Figure 5.4B; see Methods). Next, we perform a systematic comparison between the statistics of the action choices of human participants and those of the simulated participants (the two most discriminating statistics are reported in Figure 5.5A-B and a systematic analysis in Appendix C). Our results show that simulated participants

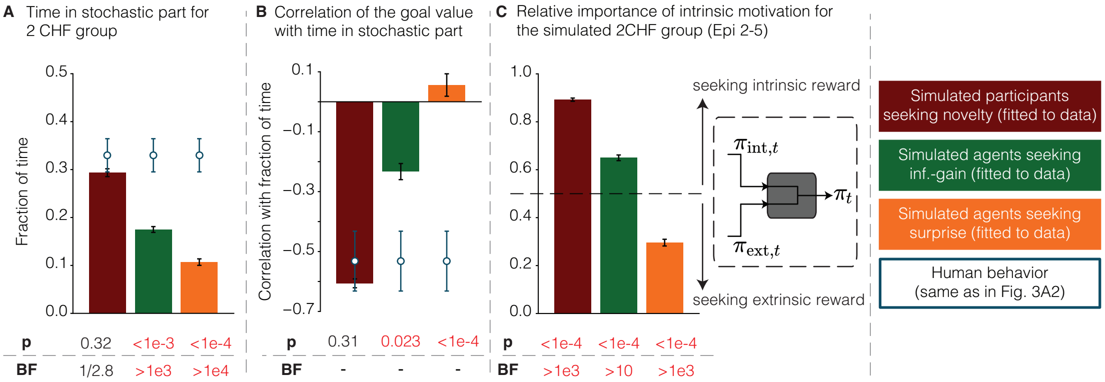


Figure 5.5: **Seeking novelty, but not surprise or information gain, can reproduce data statistics.** For each of the three intrinsic rewards, we run 1500 simulations of algorithms with parameters fitted to individual human participants; random seeds are different in each simulation. We divide the simulated participants into three groups (corresponding to the 2 CHF, 3 CHF, and 4 CHF goal values) and use the same criteria as we used for human participants to detect and remove outliers among simulated participants (Methods). **A.** Average fraction of time during episodes 2-5 spent by the 2 CHF group of human participants (blue circles, same data as in Figure 5.3A2) and the simulated participants (bars). Error bars: SEMean. P-value and BF: Comparison between the simulated and human participants (unequal variances t-test). Human participants spend a significantly greater fraction of their time in the stochastic part than simulated participants seeking information gain ($t = 4.4$; 95%CI $= (0.08, 0.23)$; DF $= 21.2$) or surprise ($t = 6.3$; 95%CI $= (0.15, 0.30)$; DF $= 21.6$). No significant difference was observed for novelty-seeking ($t = 1.0$; 95%CI $= (-0.04, 0.11)$; DF $= 22.3$). **B.** Pearson correlation between the fraction of time during episodes 2-5 spent in the stochastic part and the goal value. Human participants' data shows the same correlation value as reported in Figure 5.3A2. Error bars: Standard deviation evaluated by bootstrapping. P-values are from permutation tests (1000 sampled permutations; Bayesian testing was not applicable). **C.** The relative contribution of intrinsic rewards (i.e., the dominance of $\pi_{\text{int},t}$ over $\pi_{\text{ext},t}$; Equation 5.18 in Methods) in episodes 2-5 for the 2 CHF group of simulated participants. P-value and BF: Comparison with 0.5 (one-sample t-test). We observe a dominance of $\pi_{\text{int},t}$ for seeking novelty ($t = 58.2$; 95%CI $= (0.88, 0.91)$; DF $= 416$) and information gain ($t = 12.7$; 95%CI $= (0.63, 0.67)$; DF $= 379$) but a dominance of $\pi_{\text{ext},t}$ for seeking surprise ($t = -14.6$; 95%CI $= (0.27, 0.32)$; DF $= 327$). Red p-values: Significant effects with False Discovery Rate controlled at 0.05 (Efron and Hastie, 2016) (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3).

using novelty as intrinsic rewards reproduce all data statistics (including *the zero correlation* observed in Figure 5.3A3; see Appendix C), whereas simulated participants using information gain or surprise fail to do so. The failure of algorithms using information gain or surprise is most evident regarding the fraction of time spent in the stochastic part during episodes 2-5: 1. We observe that the 2 CHF group of simulated participants who seek information gain or surprise spends a significantly smaller fraction of their time (less than half) in the stochastic part of the environment than the 2 CHF group of human participants (Figure 5.5A). 2. Simulated participants using information gain or surprise fail to reproduce the observed negative correlation between the goal value and the fraction of time spent in the stochastic part (Figure 5.5B). We emphasize that both shortcomings are observed even though the parameters of the algorithms had been previously optimized to explain as best as possible the sequence of action choices across the whole experiment.

The failure of surprise-seeking algorithms to reproduce these statistics is due to the detrimental consequences of seeking surprise in the presence of stochasticity (e.g., as observed for the simulated efficient agents in Episode 5 of Figure 5.2E3). Hence, to stop the simulated participants from spending an enormous amount of time during episode 5 in the stochastic part of the environment, fitting surprise-seeking to action choices of human participants yields a set of parameters that causes action choices to be dominated by extrinsic reward (relative importance of surprise-seeking about 0.3 for the 2 CHF group; Figure 5.5C), which in turn cannot explain the overall high level of exploration observed in the 2 CHF group of human participants (Figure 5.5A). Similarly, the relative importance of information gain is around 0.65 when parameters of a hybrid algorithm driven by information gain are optimized to fit human behavior. A higher value of relative importance would make, during episode 2, the algorithm too attracted to the stochastic action in state 4 compared to humans (compare Figure 5.2D with Figure 5.3B). With such reduced importance of information gain, the hybrid algorithm cannot, however, explain the specific behavioral features in Figure 5.5A and B. Therefore, the attraction of human participants to the stochastic part has specific characteristics that are explained by seeking novelty but not by seeking surprise or information gain.

Taken together our results provide strong quantitative and qualitative evidence for novelty as the dominant drive of human exploration in our experiment.

5.2.7 Reward optimism correlates with the relative importance of novelty

Using novelty-seeking as the most probable model of human behavior, we can now explicitly test our hypothesis that reward optimism increases human motivation to explore by increasing the relative importance of novelty. By analyzing the parameters of our novelty-seeking algorithm fitted to the behavioral data, we observe, in agreement with our hypothesis, a significant negative correlation between the relative importance of novelty during action selection (in episodes 2-5) and the goal value participants found in episode 1 (Figure 5.6A; parameter-recovery in Figure 5.6C). Moreover, the participants in the 2 CHF group continue with an almost fully exploratory policy in episodes 2-5 indicating that they have only a small bias towards exploiting the small but known reward (Figure 5.6A).

Since our simulated participants are informed that there are three different goal states in the environment, the reward-seeking component $\pi_{\text{ext},t}$ of the action policy can also contribute to exploratory behavior, e.g., through optimistic initialization of Q-values

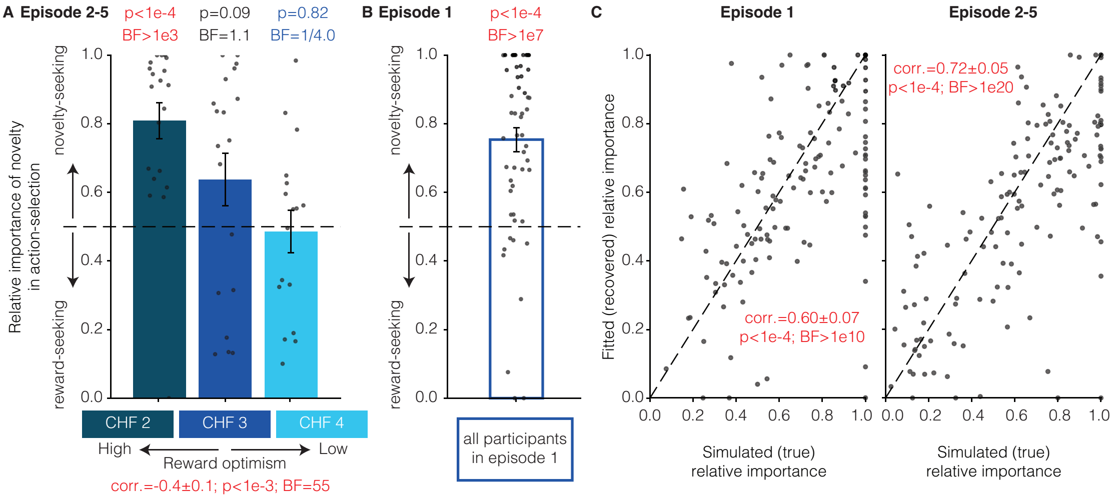


Figure 5.6: **Reward optimism increases the relative importance of novelty in action selection. A.** The relative importance of novelty-seeking in episodes 2-5 is computed for each participant after fitting the model to data (similar to Figure 5.5C but using action choices of human participants instead of simulated participants; Methods). Error bars show the SEMean and single dots the data of individual participants. We observe a significant negative correlation between the relative importance of novelty and the goal value (correlation test; $t = -3.6$; 95%CI $= (-0.63, -0.20)$; DF $= 55$). P-values and BFs on top: Comparison with 0.5 (one-sample t-test). We observe a significant dominance of $\pi_{\text{int},t}$ for the 2 CHF group ($t = 5.9$; 95%CI $= (0.70, 0.92)$; DF $= 20$). A BF of $1/4.0$ in favor of the null hypothesis (Kass and Raftery, 1995) suggests an equal contribution of $\pi_{\text{ext},t}$ and $\pi_{\text{int},t}$ for the 4 CHF group ($t = -0.23$; 95%CI $= (0.35, 0.62)$; DF $= 16$). The test for 3 CHF group is inconclusive ($t = 1.8$; 95%CI $= (0.48, 0.80)$; DF $= 18$). **B.** The relative importance of novelty-seeking in episode 1 implies a significant dominance of novelty-seeking against optimistic initialization for exploration ($t = 7.3$; 95%CI $= (0.68, 0.82)$; DF $= 56$). **C.** Parameter-recovery (Wilson and Collins, 2019) using the action choices of 150 ($= 50$ per group) simulated participants seeking novelty (Methods). The comparison between the true contribution of novelty-seeking to action selection (computed with the parameters used for simulations) and the recovered contribution (computed with the parameters fitted to the simulated action choices) shows that the relative importance of novelty-seeking is on average identifiable in our experimental paradigm: Positive correlations both for episode 1 ($t = 9.0$; 95%CI $= (0.48, 0.70)$; DF $= 148$) and episodes 2-5 ($t = 12$; 95%CI $= (0.63, 0.79)$; DF $= 148$). Red p-values: Significant effects with False Discovery Rate controlled at 0.05 (Efron and Hastie, 2016) (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3). Blue BFs: Significant evidence in favor of the null hypothesis (BF$\leq 1/3$).

(Sutton and Barto, 2018) or prior assumptions about the state-transitions (see Methods for a theoretical analysis). To study the extent of this contribution, we focus on episode 1 where this effect is most easily detectable: We observe a dominant influence of novelty-seeking on action selection (Figure 5.6B). This implies that, to explain human behavior, the knowledge of the existence of different goal states must drive exploration through a novelty-seeking policy instead of the optimistic initialization of a reward-seeking policy.

5.3 Discussion

We designed a novel experimental paradigm to study human goal-directed exploration in multi-step stochastic environments with sparse rewards. We made two main observations: (i) Human participants who are optimistic about finding higher rewards than those already discovered are persistently attracted to stochasticity; and (ii) this persistent attraction is explained better by seeking novelty than seeking information gain or surprise, even though seeking information gain is theoretically more robust in dealing with stochasticity.

How humans explore their environments has been a long-lasting question in neuroscience and psychology (Cohen et al., 2007; Schulz and Gershman, 2019; Wilson et al., 2021). Experimental studies have shown that humans use a combination of random and directed exploration (Gershman, 2019a; Wilson et al., 2014), potentially linked to different neural mechanisms (Dubois et al., 2021; Tomov et al., 2020; Wittmann et al., 2008; Zajkowski et al., 2017). Theoretical studies have proposed distinct motivational signals as potential drives of directed exploratory actions (Friston et al., 2017; Klyubin et al., 2005; Modirshanechi et al., 2023b; Murayama, 2022; Schulz and Gershman, 2019). While human exploration is in general driven by a mixture of these signals (Brändle et al., 2023; Cockburn et al., 2022; Kobayashi et al., 2019; Poli et al., 2022), a particular signal can dominate exploration in specific tasks with (Gershman and Niv, 2015; Giron et al., 2023; Horvath et al., 2021; Meder and Nelson, 2012; Wu et al., 2018) and without extrinsic rewards (Cubit et al., 2021; Dubey and Griffiths, 2019; Itti and Baldi, 2009; Kidd et al., 2012; Ten et al., 2021; Wu et al., 2022). However, most results on human exploration are limited to 1-step decision-making tasks (e.g., multi-armed bandits), and it is unclear whether they can be generalized to more complex and realistic situations (Brändle et al., 2022; Modirshanechi et al., 2023b).

To bridge a link between exploration in 1-step and multi-step tasks, we showed in an earlier study (Xu et al., 2021) that novelty dominantly drives human exploration in complex but *deterministic* environments with sparse rewards, i.e., situations where novelty-seeking is empirically shown to be an efficient and close-to-optimal exploration strategy (Bellemare et al., 2016; Ostrovski et al., 2017). Observations (i) and (ii) above provide further evidence for novelty as the dominant drive of human goal-directed exploration even in situations where seeking novelty is not optimal. Specifically, after episode 1, participants can reasonably assume that the task is solvable, i.e., if they have succeeded in finding the 2 CHF reward, then they should be able to also find the higher rewards. Hence, the fact that the participants in the 2 CHF group continue the search during episodes 2-5 is expected and economically rational, but our results show that their novelty-driven search strategy is suboptimal. Further experimental studies are needed to investigate the implications of our results for other types of human exploratory behavior. In particular, it is unclear whether goal-directed exploration, as studied here, shares some drives and mechanisms with reward-free exploration strategies in, e.g., reactive orienting and passive viewing (Kidd et al., 2012; Morrens et al., 2020), navigation (Montgomery, 1953, 1954), and non-instrumental decision-making tasks (Daddaoua et al., 2016; Kobayashi et al., 2019; Ogasawara et al., 2022).

Our results appear to be in contradiction with the long-lasting belief that humans are not prone to the 'noisy TV' problem (Gottlieb et al., 2013; Mavor-Parker et al., 2022; Schmidhuber, 2010). It is important, however, to note that the stochasticity in our environment is different from passively watching a noisy grey-flickering TV screen. Rather the environment allows participants to take actions that are in spirit similar to exploring different TV channels, where each channel contains images or videos – similar to the recent realizations of 'noisy TV' in machine learning (Burda et al., 2019). In this context, our experimental paradigm is a model experiment of recent social media where users spend hours on the 'endless scrolling option' to watch new videos (Montag et al., 2019, 2021) – despite the availability of alternative activities with 'extrinsic' rewards.

Accordingly, our results challenge the optimality of human exploration (Dubey and Griffiths, 2019; Singh et al., 2010b) and imply that algorithmic advances in machine learning may not contribute to finding better models of human exploration. However, we note that, for computing novelty, an agent only needs to track the state frequencies over time and does not need any knowledge of the environment's structure (Methods); hence computing novelty is computationally cheaper than computing information gain. This suggests that a potentially higher level of distraction by novelty in humans may be the price of spending less computational power. In other words, novelty-seeking in the presence of stochasticity may not be a globally optimal strategy for exploration but can be an optimal strategy given a set of prior assumptions and computational constraints, i.e., a 'resource rational' policy (Bhui et al., 2021; Binz and Schulz, 2022; Lieder and Griffiths, 2020).

In addition to observations (i) and (ii), we found that the relative importance of novelty- and reward-induced behaviors in human participants is correlated with the degree of reward optimism. This is in line with the known influence of environmental variables on an agent's preference for novelty (Akiti et al., 2022; Gershman and Niv, 2015; Stojić et al., 2020). In particular, theories of 'motivation crowding effect' (Frey and Jegen, 2001) and 'undermining effect' (Deci et al., 1999; Murayama et al., 2010) suggest that the absolute value of extrinsic reward might contribute, in addition to the reward optimism, to the observed negative correlation in Figure 5.6A, predicting that even if participants were confident that there is no other goal state in the environment, the 2 CHF group would spend more time in the stochastic part than the 4 CHF group – simply because 2 CHF is not an attractive reward anyway. A potential future direction is to investigate the interplay of novelty and reward in various experimental environments with various reward distributions and sources of stochasticity.

Optimism in psychology has been defined as a 'variable that reflects the extent to which people hold generalized favorable expectancies for their future' (Carver et al., 2010) and has been linked to several neural and behavioral characteristics (Carver et al., 2010; Sharot et al., 2011; Strunk et al., 2006). While the traditional approach to measure optimism is through self-tests (Scheier et al., 1994), more recently statistical inference using RL (Lefebvre et al., 2017) and Bayesian (Gesiarz et al., 2019; Stankevicius et al., 2014) models of behavior have been proposed to quantify variables correlated with traditional measurements. While there are multiple traditional ways to incorporate the notion of optimism into the RL framework (Methods), seeking intrinsic rewards has also been interpreted in the machine learning community as an 'optimistic policy' for exploration (Ghavamzadeh et al., 2015). Our results show that the preference for an intrinsic reward is indeed correlated with a notion of optimism defined in the context of our experiment as the expectancy of finding a goal of higher value in episodes 2-5 ('reward optimism' in Figure 5.6A). Moreover, the persistent exploration of the stochastic part of our environment observed in the behavior of human participants (Figure 5.3B3) is conceptually consistent with the known phenomena of optimism bias (Sharot, 2011) and optimistic belief updating in humans (Garrett and Sharot, 2017; Palminteri and Lebreton, 2022; Sharot et al., 2011).

Even though notions of 'novelty', 'surprise', and 'information gain' are frequently used in neuroscience (Baldi, 2002; Kolossa et al., 2015; Xu et al., 2021), psychology (Maguire et al., 2011; Nelson, 2005; Reisenzein et al., 2019), and machine learning (Aubret et al., 2019; Ladosz et al., 2022; Schmidhuber, 2010), there is no consensus on the precise definitions of these notions as scientific terms (Barto et al., 2013; Modirshanechi et al., 2022). Our results in this paper are based on the specific mathematical formulations that we have chosen (Methods), but we expect our conclusions to be invariant to the precise choice of definitions as long as (i) novelty quantifies infrequency of *states* (Xu et al., 2021), e.g., defined based on density models in machine learning (Bellemare et al., 2016; Ostrovski et al., 2017); (ii) surprise quantifies mismatches between observations and agents' expectations, where the expectations are made based on the previous *state-action* pair, including all measures of prediction surprise (Modirshanechi et al., 2022) and typical measures of prediction error in machine learning (Burda et al., 2019; Pathak et al., 2017); and (iii) information gain quantifies improvements in the agents' *world-model* and vanishes by accumulation of experience, e.g., including Bayesian (Baldi, 2002) and Postdictive surprise (Kolossa et al., 2015) and measures of disagreement and progress-rate in machine learning (Kim et al., 2020b; Mendonca et al., 2021; Oudeyer, 2018; Pathak et al., 2019; Sekar et al., 2020).

In conclusion, our results show (i) that human decision-making is influenced by an interplay of intrinsic with extrinsic rewards that is controlled by reward optimism and (ii) that novelty-seeking RL algorithms can successfully model this interplay in tasks where humans search for rewarding states.

5.4 Methods

5.4.1 Ethics statement

The data for human experiment were collected under CE 164/2014, and the protocol was approved by the 'Commission cantonale d'éthique de la recherche sur l'être humain'. All participants were informed that they could quit the experiment at any time, and they all signed a written informed consent. All procedures complied with the Declaration of Helsinki (except for pre-registration).

5.4.2 Experimental procedure for human participants

63 participants joined the experiment. Data of 6 participants were removed (see below) and, thus, data of 57 participants (27 female, mean age 24.1 ± 4.1 years) were included in the analyses. All participants were naïve to the purpose of the experiment and had normal or corrected-to-normal visual acuity. The experiment was scripted in MATLAB using the Psychophysics Toolbox (Brainard and Vision, 1997).

Before starting the experiment, the participants were informed that they need to find either one of the 3 goal states 5 times. They were shown the 3 goal images and informed that different images had different reward values of 2 CHF, 3 CHF, and 4 CHF. Specifically, they were given the example that 'if you find the 2 CHF goal twice, 3 CHF goal once, and 4 CHF goal twice, then you will be paid $2 \times 2 + 1 \times 3 + 2 \times 4 = 15$ CHF'; see 'Informing RL agents of different goal states and modeling optimism' for how simulated efficient agents and simulated participants were given this information. At each trial, participants were presented an image (state) and three grey disks below the image (Figure 5.1C). Clicking on a disk (action) led participants to a subsequent image which was chosen based on the underlying graph of the environment in Figure 5.1A-B (which was unknown to the participants). Participants clicked through the environment until they found one of the goal states which finished an episode (Figure 5.1C).

The assignment of images to states and disks to actions was random but kept fixed throughout the experiment and among participants. Exceptionally, we did not make the assignment for the actions in state 4 before the start of the experiment. Rather, for each participant, we assigned the disk that was chosen in the 1st encounter of state 4 to the stochastic action and the other two disks randomly to the bad and progressing actions, respectively (Figure 5.1A). With this assignment, we made sure that all human participants would visit the stochastic part at least once during episode 1. The same protocol was used for simulated efficient agents and simulated participants.

Before the start of the experiment, we randomly assigned the different goal images (corresponding to the three reward values) to different goal states G^*, G_1, and G_2 in Figure 5.1A, separately for each participant. The image and hence the reward value were then kept fixed throughout the experiment. In other words, we randomly assigned different participants to different environments with the same structure but different assignments of reward values. We, therefore, ended up with 3 groups of participants: 23 in the 2 CHF group, 20 in the 3 CHF group, and 20 in the 4 CHF group. The probability of encountering a goal state other than G^* is controlled by the parameters ε. We considered ε to be around machine precision 10^{-8}, so we have $(1-\varepsilon)^{5\times 63} \approx 1 - 10^{-5} \approx 1$, meaning that all 63 participants would be taken almost surely to the goal state G^* in all 5 episodes. We note, however, that a participant could in principle observe any of the 3 goals if they could choose the progressing action at state 6 sufficiently many times because $\lim_{t\to\infty}(1-\varepsilon)^t = 0$.

2 participants (in the 2 CHF group) did not finish the experiment, and 4 participants (1 in the 3 CHF group and 3 in the 4 CHF group) took more than 3 times group-average number of actions in episodes 2-5 to finish the experiment. We considered this as a sign of being non-attentive and removed these 6 participants from further analyses.

The sample size was determined by a power analysis performed on the data of the efficient simulations done for Figure 5.2 (see 'Efficient model-based planning for simulated participants' for the simulation details). Our goal was to have a statistical power of more than 80% (with a significance level of 0.05) for correlations in panels Figure 5.2A, C, and E as well as for the differences for the highest importance of intrinsic rewards in Figure 5.2D and F.

The correction for multiple hypotheses testing was done by controlling the False Discovery Rate at 0.05 (Efron and Hastie, 2016) over all 22 null hypotheses that are tested in Figure 5.3, Figure 5.5, and Figure 5.6 (p-value threshold: 0.034). All Bayes Factors (abbreviated BF in the figures) were evaluated using Schwartz approximation (Kass and Raftery, 1995) to avoid any assumptions on the prior distribution. We note that evaluating the Bayes Factors using priors suggested by Rouder and Morey (2012); Rouder et al. (2009) does not change our conclusions. We also note that using the Spearman correlation instead of the Pearson correlation in Figure 5.2A, C, and E, Figure 5.3A, and Figure 5.6A does not change our conclusions.

5.4.3 Full hybrid model

We first present the most general case of our algorithm as visualized in Figure 5.1D and then explain the special cases used for simulating efficient agents (Figure 5.2) and for modeling human behavior (Figure 5.4-Figure 5.6). We used ideas from non-parametric Bayesian inference (Ghahramani, 2013) to design an intrinsically motivated RL algorithm for environments where the total number of states is unknown. We present the final results here and present the derivations and pseudo-code in Appendix C.

We indicate the sequence of actions and states until time t by $s_{1:t}$ and $a_{1:t}$, respectively, and define the **set of all known states** at time t as

$$\mathcal{S}^{(t)} = \Big\{ s : \exists\, t' \in \{1, ..., t\} \text{ s.t. } s = s_{t'} \Big\} \cup \{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}, \tag{5.1}$$

where $\tilde{G}_i$s are our three different goal states – $\tilde{G}_0$ corresponds to the 2 CHF goal, $\tilde{G}_1$ to the 3 CHF goal, and $\tilde{G}_2$ to the 4 CHF goal. Note that $\{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}$ represents the images of the goal states and not their locations G^*, G_1, and G_2 and that the assignment of images to locations is unknown to the model. Hence, since $t = 0$, the simulated efficient agents and the simulated participants are aware of the existence of multiple goal states in the environment. In a more general setting, $\{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}$ should be replaced by the set of all states whose images were shown to participants prior to the start of the experiment. After a transition to state $s_{t+1} = s'$ resulting from taking action $a_t = a$ at state $s_t = s$, the reward functions R_{ext} and $R_{\text{int},t}$ evaluate the reward values $r_{\text{ext},t+1}$ and $r_{\text{int},t+1}$. We define the **extrinsic reward function** R_{ext} as

$$R_{\text{ext}}(s, a \to s') = \delta_{s',\tilde{G}_0} + r_1^* \delta_{s',\tilde{G}_1} + r_2^* \delta_{s',\tilde{G}_2}, \tag{5.2}$$

where δ is the Kronecker delta function, and we assume (without loss of generality) a subjective extrinsic reward value of 1 for $\tilde{G}_0$ (2 CHF goal) and subjective extrinsic reward values of $r_1^* \geq 1$ and $r_2^* \geq 1$ for $\tilde{G}_1$ and $\tilde{G}_2$, respectively. The prior information of human participants about the difference in the monetary reward values of different goal states can be modeled in simulated participants by varying r_1^* and r_2^* (see 'Informing RL agents of different goal states and modeling optimism'). We discuss choices of $R_{\text{int},t}$ in the next section.

As a general choice for the RL algorithm in Figure 5.1D, we consider a hybrid of model-based and model-free policy (Daw et al., 2011; Gläscher et al., 2010; Liakoni et al., 2022; Xu et al., 2021). The **model-free (MF) component** uses the sequence of states $s_{1:t}$, actions $a_{1:t}$, extrinsic rewards $r_{\text{ext},1:t}$, and intrinsic rewards $r_{\text{int},1:t}$ (in the two parallel branches in Figure 5.1D) and estimates the extrinsic and intrinsic Q-values $Q^{(t)}_{\text{MF,ext}}$ and $Q^{(t)}_{\text{MF,int}}$, respectively. Traditionally, MF algorithms do not need the total number of states (Sutton and Barto, 2018), thus the MF component of our algorithm remains similar to that of previous studies (Lehmann et al., 2019; Xu et al., 2021): At the beginning of episode 1, we initialize Q-values at $Q^{(0)}_{\text{MF,ext}}$ and $Q^{(0)}_{\text{MF,int}}$. Then, the estimates are updated recursively after each new observation. After the transition $(s_t, a_t) \to s_{t+1}$, the agent computes extrinsic and intrinsic reward prediction errors $RPE_{\text{ext},t+1}$ and $RPE_{\text{int},t+1}$, respectively:

$$\begin{aligned} RPE_{\text{ext},t+1} &= r_{\text{ext},t+1} + \lambda_{\text{ext}} V^{(t)}_{\text{MF,ext}}(s_{t+1}) - Q^{(t)}_{\text{MF,ext}}(s_t, a_t) \\ RPE_{\text{int},t+1} &= r_{\text{int},t+1} + \lambda_{\text{int}} V^{(t)}_{\text{MF,int}}(s_{t+1}) - Q^{(t)}_{\text{MF,int}}(s_t, a_t), \end{aligned} \tag{5.3}$$

where λ_{ext} and $\lambda_{\text{int}} \in [0, 1)$ are the discount factors for extrinsic and intrinsic reward seeking, respectively, and $V^{(t)}_{\text{MF,ext}}(s_{t+1}) = \max_{a'} Q^{(t)}_{\text{MF,ext}}(s_{t+1}, a')$ and $V^{(t)}_{\text{MF,int}}(s_{t+1}) =$

$\max_{a'} Q^{(t)}_{\mathrm{MF,int}}(s_{t+1}, a')$ are the extrinsic and intrinsic V-values (Sutton and Barto, 2018) of the state s_{t+1}, respectively. We use two separate eligibility traces (Lehmann et al., 2019; Sutton and Barto, 2018) for the update of Q-values, one for extrinsic reward $e_{\mathrm{ext},t}$ and one for intrinsic reward $e_{\mathrm{int},t}$, both initialized at zero at the beginning of each episode. The update rules for the eligibility traces after taking action a_t at state s_t is

$$\begin{aligned} e_{\mathrm{ext},t+1}(s,a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_{\mathrm{ext}}\mu_{\mathrm{ext}} e_{\mathrm{ext},t}(s,a) & \quad \text{otherwise} \end{cases} \\ e_{\mathrm{int},t+1}(s,a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_{\mathrm{int}}\mu_{\mathrm{int}} e_{\mathrm{int},t}(s,a) & \quad \text{otherwise}\,, \end{cases} \end{aligned} \tag{5.4}$$

where λ_{ext} and λ_{int} are the discount factors defined above, and μ_{ext} and $\mu_{\mathrm{int}} \in [0,1]$ are the decay factors of the eligibility traces for the extrinsic and intrinsic rewards, respectively. The update rule is then $\Delta Q^{(t+1)}_{\mathrm{MF}}(s,a) = \rho e_{t+1}(s,a) RPE_{t+1}$, where e_{t+1} is the eligibility trace (i.e., either $e_{\mathrm{ext},t+1}$ or $e_{\mathrm{int},t+1}$), RPE_{t+1} is the reward prediction error (i.e., either $RPE_{\mathrm{ext},t+1}$ or $RPE_{\mathrm{int},t+1}$), and $\rho \in [0,1)$ is the learning rate.

The **model-based (MB) component** builds a world-model that summarizes the structure of the environment by estimating the probability $p^{(t)}(s'|s,a)$ of the transition $(s,a) \to s'$. To do so, an agent counts the transition $(s,a) \to s'$ recursively and using a leaky integration (Liakoni et al., 2021; Yu and Cohen, 2009):

$$\tilde{C}^{(t+1)}_{s,a,s'} = \begin{cases} \kappa \tilde{C}^{(t)}_{s,a,s'} + \delta_{s',s_{t+1}} & \text{if}\, s = s_t\,,\ a = a_t \\ \tilde{C}^{(t)}_{s,a,s'} & \text{otherwise,} \end{cases} \tag{5.5}$$

where δ is the Kronecker delta function, $\tilde{C}^{(0)}_{s,a,s'} = 0$, and $\kappa \in [0,1]$ is the leak parameter and accounts for imperfect memory and model-building in humans. If $\kappa = 1$, then $\tilde{C}^{(t+1)}_{s,a,s'}$ is the exact count of transition $(s,a) \to s'$. These counts are used to estimate the transition probabilities

$$p^{(t)}(s'|s,a) = \begin{cases} \frac{\epsilon_{\mathrm{obs}} + \tilde{C}^{(t)}_{s,a,s'}}{\epsilon_{\mathrm{new}} + \epsilon_{\mathrm{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if} \quad s' \in \mathcal{S}^{(t)}\,, \\ \frac{\epsilon_{\mathrm{new}}}{\epsilon_{\mathrm{new}} + \epsilon_{\mathrm{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if} \quad s' = s_{\mathrm{new}}\,, \end{cases} \tag{5.6}$$

where $\tilde{C}^{(t)}_{s,a} = \sum_{s'} \tilde{C}^{(t)}_{s,a,s'}$ is the counts of taking action a at state s, $\epsilon_{\mathrm{obs}} \in \mathbb{R}^+$ is a free parameter for the prior probability of transition to a known state (i.e., states in $\mathcal{S}^{(t)}$), and $\epsilon_{\mathrm{new}} \in \mathbb{R}^+$ is a free parameter for the prior probability of transition to a new state (i.e., states not in $\mathcal{S}^{(t)}$) – see Appendix C for derivations. Choosing $\epsilon_{\mathrm{new}} = 0$ is equivalent to assuming there is no unknown state in the environment, for which the estimate in Equation 5.6 is reduced to the classic Bayesian estimate of transition probabilities in bounded discrete environments (Liakoni et al., 2022; Xu et al., 2021). The transition probabilities are then used in a novel variant of prioritized sweeping (Sutton and Barto, 2018; Van Seijen and Sutton, 2013) adapted to deal with an unknown number of states.

The prioritized sweeping algorithm computes a pair of Q-values, i.e., $Q^{(t)}_{\text{MB,ext}}$ for extrinsic and $Q^{(t)}_{\text{MB,int}}$ for intrinsic rewards, by solving the corresponding Bellman equations (Sutton and Barto, 2018) with $T_{PS,\text{ext}}$ and $T_{PS,\text{int}}$ iterations, respectively for $Q^{(t)}_{\text{MB,ext}}$ and $Q^{(t)}_{\text{MB,int}}$. See Supplementary Material for details.

Finally, actions are chosen by **a hybrid softmax policy** (Sutton and Barto, 2018): The probability of taking action a in state s at time t is

$$\begin{aligned}\pi_t(a|s) \propto \exp\Big[&\beta_{\text{MB,ext}} Q^{(t)}_{\text{MB,ext}}(s,a) \;+ \beta_{\text{MF,ext}} Q^{(t)}_{\text{MF,ext}}(s,a) + \\ &\beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) \;+ \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)\Big],\end{aligned} \tag{5.7}$$

where $\beta_{\text{MB,ext}} \in \mathbb{R}^+$, $\beta_{\text{MF,ext}} \in \mathbb{R}^+$, $\beta_{\text{MB,int}} \in \mathbb{R}^+$, and $\beta_{\text{MF,int}} \in \mathbb{R}^+$ are free parameters (i.e., inverse temperatures of the softmax policy (Sutton and Barto, 2018)) expressing the contribution of each Q-value to action-selection. For Figure 5.1D, we defined

$$\begin{aligned}\pi_{\text{ext},t}(a|s) &\propto \exp\Big[\frac{\beta_{\text{MB,ext}}}{\beta_{\text{MB,ext}} + \beta_{\text{MF,ext}}} Q^{(t)}_{\text{MB,ext}}(s,a) \;+ \frac{\beta_{\text{MF,ext}}}{\beta_{\text{MB,ext}} + \beta_{\text{MF,ext}}} Q^{(t)}_{\text{MF,ext}}(s,a)\Big] \\ \pi_{\text{int},t}(a|s) &\propto \exp\Big[\frac{\beta_{\text{MB,int}}}{\beta_{\text{MB,int}} + \beta_{\text{MF,int}}} Q^{(t)}_{\text{MB,int}}(s,a) \;+ \frac{\beta_{\text{MF,int}}}{\beta_{\text{MB,int}} + \beta_{\text{MF,int}}} Q^{(t)}_{\text{MF,int}}(s,a)\Big],\end{aligned} \tag{5.8}$$

and as a result $\pi_t \propto \pi_{\text{ext},t}^{\beta_{\text{MB,ext}}+\beta_{\text{MF,ext}}} \cdot \pi_{\text{int},t}^{\beta_{\text{MB,int}}+\beta_{\text{MF,int}}}$.

In general, the contribution of seeking extrinsic reward and seeking intrinsic reward as well as the MB and MF branches to action-selection depends on different factors, including time passed since the beginning of the experiment (Gläscher et al., 2010; Huys et al., 2015), cognitive load (Piray and Daw, 2021a), and whether the location of reward is known (Xu et al., 2021). Here, we make a simplistic assumption that these contributions (expressed as the 4 inverse temperatures) are constant within but potentially different between the two phases of the experiment:

- Phase 1: Before finding the goal state in episode 1, we consider $\beta_{\text{MB,ext}} = \beta^{(1)}_{\text{MB,ext}}$, $\beta_{\text{MF,ext}} = \beta^{(1)}_{\text{MF,ext}}$, $\beta_{\text{MB,int}} = \beta^{(1)}_{\text{MB,int}}$, and $\beta_{\text{MF,int}} = \beta^{(1)}_{\text{MF,int}}$ as four independent free parameters chosen independently for each agent.
- Phase 2: After finding the goal, i.e., in all episodes after episode 1, we consider $\beta_{\text{MB,ext}} = \beta^{(2)}_{\text{MB,ext}}$, $\beta_{\text{MF,ext}} = \beta^{(2)}_{\text{MF,ext}}$, $\beta_{\text{MB,int}} = \beta^{(2)}_{\text{MB,int}}$, and $\beta_{\text{MF,int}} = \beta^{(2)}_{\text{MF,int}}$ as another four independent free parameters chosen independently for each agent.

See 'Relative importance of novelty in action-selection' for how these inverse temperatures relate to the influence of intrinsic and extrinsic rewards on action-choices (Figure 5.5C and Figure 5.6).

Summary of free parameters: To summarize, the full hybrid algorithm has 22 free parameters:

$$\Phi = \{r_1^*, r_2^*, Q^{(0)}_{\text{MF,ext}}, Q^{(0)}_{\text{MF,int}}, \lambda_{\text{ext}}, \lambda_{\text{int}}, \mu_{\text{ext}}, \mu_{\text{int}}, \rho, \kappa, \epsilon_{\text{new}}, \epsilon_{\text{obs}}, T_{PS,\text{ext}}, T_{PS,\text{int}}, \\ \beta^{(1)}_{\text{MB,ext}}, \beta^{(2)}_{\text{MB,ext}}, \beta^{(1)}_{\text{MB,int}}, \beta^{(2)}_{\text{MB,int}}, \beta^{(1)}_{\text{MF,ext}}, \beta^{(2)}_{\text{MF,ext}}, \beta^{(1)}_{\text{MF,int}}, \beta^{(2)}_{\text{MF,int}}\}, \tag{5.9}$$

where r_1^* and r_2^* are subjective values of the 3 CHF goal and the 4 CHF goal, respectively (with the 2 CHF goal being the reference goal with a value of 1), $Q^{(0)}_{\text{MF,ext}}$ and $Q^{(0)}_{\text{MF,int}}$ are the initial values for MF Q-values, λ_{ext} and λ_{int} are the discount factors, μ_{ext} and μ_{int} are the decay rates of the eligibility traces, ρ is the MF learning rate, κ is the leak parameter for model-building, ϵ_{new} and ϵ_{obs} are prior parameters for model-building, $T_{PS,\text{ext}}$ and $T_{PS,\text{int}}$ are the numbers of iterations for prioritized sweeping, and $\beta^{(1)}_{\text{MB,ext}}$, $\beta^{(2)}_{\text{MB,ext}}$, $\beta^{(1)}_{\text{MB,int}}$, $\beta^{(2)}_{\text{MB,int}}$, $\beta^{(1)}_{\text{MF,ext}}$, $\beta^{(2)}_{\text{MF,ext}}$, $\beta^{(1)}_{\text{MF,int}}$, and $\beta^{(2)}_{\text{MF,int}}$ are the inverse temperatures of the softmax policy.

5.4.4 Different choices of intrinsic reward

The intrinsic reward function $R_{\text{int},t}$ maps a transition $(s, a) \to s'$ to an intrinsic reward value, i.e., $r_{\text{int},t+1} = R_{\text{int},t}(s_t, a_t \to s_{t+1})$. In this section, we present our 3 choices of $R_{\text{int},t}$.

Novelty: For an agent seeking novelty (red in Figure 5.2, Figure 5.4, and Figure 5.5), we define the intrinsic reward function as

$$R_{\text{int},t}(s, a \to s') = -\log p_f^{(t)}(s'), \tag{5.10}$$

where $p_f^{(t)}(s') = \frac{1+\tilde{C}^{(t)}_{s'}}{1+|\mathcal{S}^{(t)}|+\sum_{s''}\tilde{C}^{(t)}_{s''}}$ is the state frequency with $\tilde{C}^{(t)}_{s'}$ the pseudo-count of encounters of state s' up to time t (similar to Equation 5.5): $\tilde{C}^{(t+1)}_{s'} = \kappa\tilde{C}^{(t)}_{s'} + \delta_{s',s_{t+1}}$ with $\tilde{C}^{(0)}_{s'} = 0$. With this definition, that generalizes earlier works (Xu et al., 2021) to the case where the number of states is unknown, the least novel states are those that have been encountered most often (i.e., with highest $\tilde{C}^{(t)}_{s'}$). Moreover, novelty is at its highest value for the unobserved states as we have $\tilde{C}^{(t)}_{s'} = 0$ for any unobserved state $s' \notin \mathcal{S}^{(t)}$. Similar intrinsic rewards have been used in machine learning (Bellemare et al., 2016; Ostrovski et al., 2017).

Surprise: For an agent seeking surprise (orange in Figure 5.2, Figure 5.4, and Figure 5.5), we define the intrinsic reward function as the Shannon surprise (a.k.a. surprisal) (Modirshanechi et al., 2022)

$$R_{\text{int},t}(s, a \to s') = -\log p^{(t)}(s'|s, a), \tag{5.11}$$

where $p^{(t)}(s'|s, a)$ is defined in Equation 5.6. With this definition, the expected (over s')

intrinsic reward of taking action a at state s is equal to the entropy of the distribution $p^{(t)}(s'|s,a)$ (Cover, 1999). If $\epsilon_{\text{new}} < \epsilon_{\text{obs}}$, then the most surprising transitions are the ones to unobserved states. Similar intrinsic rewards have been used in machine learning (Burda et al., 2019; Pathak et al., 2017).

information gain: For an agent seeking information gain (green in Figure 5.2, Figure 5.4, and Figure 5.5), we define the intrinsic reward function as

$$R_{\text{int},t}(s,a \to s') = \mathrm{D}_{\text{KL}}\Big[p^{(t)}(.|s,a)||p^{(t+1)}(.|s,a)\Big], \tag{5.12}$$

where D_{KL} is the Kullback-Leibler divergence (Cover, 1999), and $p^{(t+1)}$ is the updated world-model upon observing $(s,a) \to s'$. The dots in Equation 5.12 denote the dummy variable over which we integrate to evaluate the Kullback-Leibler divergence. Note that if $s' \notin \mathcal{S}^{(t)}$, then there are some technical problems in the naïve computation of D_{KL} – since $p^{(t)}$ and $p^{(t+1)}$ have different supports. We deal with these problems using a more fundamental definition of D_{KL} using the Radon–Nikodym derivative; see Appendix C for derivations and see Mobin et al. (2014) for an alternative heuristic solution. Note that the information gain in Equation 5.12 has been also interpreted as a measure of surprise (called 'Postdictive surprise' (Kolossa et al., 2015)), but it has a behavior radically different from that of the Shannon surprise introduced above (Equation 5.11) – see Modirshanechi et al. (2022) for an elaborate treatment of the topic. Importantly, the expected (over s') information gain corresponding to a state-action pair (s,a) converges to 0 as $\tilde{C}^{(t)}_{s,a} \to \infty$ (see Appendix C for the proof). Similar intrinsic rewards have been used in machine learning (Mobin et al., 2014; Pathak et al., 2019; Schmidhuber, 2010; Sekar et al., 2020).

5.4.5 Informing RL agents of different goal states and modeling optimism

Human participants had been informed that there were different goal states in the environment with different monetary reward values. This information was aimed to motivate participants to further explore the environment after they received the first reward at the end of episode one. This information is incorporated into our hybrid algorithms through a few mechanisms, where some include explicit information about the goal states but some others only an implicit notion of optimism.

Our main focus throughout the paper has been on modeling reward optimism by balancing intrinsic rewards against extrinsic rewards (Figure 5.2, Figure 5.5, and Figure 5.6). In particular, assigning different values to $\beta_{\text{MB,ext}}$, $\beta_{\text{MF,ext}}$, $\beta_{\text{MB,int}}$, and $\beta_{\text{MF,int}}$ (c.f. Equation 5.7) during the two phases of the experiment enables us to implicitly make the relative importance of intrinsic rewards depend on the difference between the reward value of the discovered goal r_{G^*} and the known reward values r_1^* and r_2^* of the other goal states (Equation 5.2). Our results for the fitted relative importance of intrinsic reward across different groups of human participants (Figure 5.6A) support this very assumption, which implies that the influence of reward optimism on the action-choices is via regulation of the balance between two separate policies, one for seeking intrinsic and one for seeking extrinsic rewards.

However, there are two other alternative mechanisms, purely based on seeking extrinsic rewards, that can contribute to reward-optimism in our hybrid algorithms: The model-based and model-free optimistic initialization. In this section, we discuss these mechanisms and how they balance exploration versus exploitation. We note that our results in Figure 5.4 and Figure 5.6 (particularly Figure 5.6B) show that these two mechanisms alone are not enough and that a novelty-seeking module is necessary to explain the behavior of human participants; otherwise, all three intrinsically motivated algorithms would have the same probability of generating human data – because the purely reward-seeking algorithm with optimistic initialization is a special case of all three intrinsically motivated algorithms that we compared. In other words, if optimistic initialization alone were sufficient to explain human behavior, then all three algorithms would perform equally well in Figure 5.4 and the best fit would indicate a relative importance of 0 for novelty in Figure 5.6.

Model-based optimistic initialization.MB optimistic initialization is an explicit approach to model reward-optimism through designing the world-model. The MB branch of the hybrid algorithm finds the extrinsic Q-values $Q^{(t)}_{\text{MB,ext}}$ by solving the Bellman equations

$$Q^{(t)}_{\text{MB,ext}}(s,a) = \bar{R}^{(t)}_{\text{ext}}(s,a) + \lambda_{\text{ext}} \sum_{s'} p^{(t)}(s'|s,a) \max_{a'} Q^{(t)}_{\text{MB,ext}}(s',a'), \tag{5.13}$$

where $p^{(t)}(s'|s,a)$ is estimated transition probability in Equation 5.6, and

$$\begin{aligned} \bar{R}^{(t)}_{\text{ext}}(s,a) &= \sum_{s'} p^{(t)}(s'|s,a) R_{\text{ext}}(s,a \to s') \\ &= p^{(t)}(\tilde{G}_0|s,a) + r_1^* p^{(t)}(\tilde{G}_1|s,a) + r_2^* p^{(t)}(\tilde{G}_2|s,a) \end{aligned} \tag{5.14}$$

is the average immediate extrinsic reward expected to be collected by taking action a in state s (see Equation 5.2). Hence, the knowledge of the existence of three different goal states with three different rewards has an explicit influence on the MB branch of our algorithms. For example, because no transitions to any of the goal states have been experienced during episode 1, we have

$$\bar{R}^{(t)}_{\text{ext}}(s,a) = \frac{\epsilon_{\text{obs}}(1 + r_1^* + r_2^*)}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}}. \tag{5.15}$$

This equation has two important implications. First, $\bar{R}^{(t)}_{\text{ext}}(s,a)$ is an increasing function of ϵ_{obs}. This implies that the expected reward of a transition during episode 1 increases by increasing the prior probability of transition to states in $\mathcal{S}^{(t)}$. This is a direct consequence of our Bayesian approach to estimating the world-model. Second, $\bar{R}^{(t)}_{\text{ext}}(s,a)$ is a decreasing function of $\tilde{C}^{(t)}_{s,a}$. This implies that the expected reward of a state-action pair decreases by experience. Importantly, $\bar{R}^{(t)}_{\text{ext}}(s,a)$ converges to 0 as $\tilde{C}^{(t)}_{s,a} \to \infty$, which makes a link between exploration driven by the MB optimistic initialization and exploration driven by information gain.

During episodes 2-5, the exact theoretical analysis of the MB optimistic initialization is rather complex. However, using a few approximation steps for episode 2, we can find a condition for whether the MB extrinsic Q-values show a preference for exploring or leaving the stochastic part (Appendix C). The condition involves a comparison between the discounted reward value of the discovered goal state $\lambda^2_{\text{ext}} r_{G^*}$ and an optimistic estimate of a reward-to-be-found $R^{(t)}_{\text{Stoch.}}$ in the stochastic part that depends on r^*_1, r^*_2, λ_{ext}, ϵ_{obs}, $|\mathcal{S}^{(t)}|$, and the average pseudo-count $\bar{C}^{(t)}$ of state-action pairs in the stochastic part (Appendix C). We can show that if $r_{G^*} < r^*_2$, then increasing r^*_2 would eventually result in a preference for staying in the stochastic part: If the reward value of a goal state is much greater than the value of the discovered goal state, then the agent prefers to keep exploring the stochastic part. However, for any value of r^*_2 and r_{G^*}, increasing $\bar{C}^{(t)}$ would eventually result in a preference for leaving the stochastic part and going towards the already discovered goal: After a sufficiently long and unsuccessful exploration phase, agents will eventually give up exploration. This is another qualitative link between exploration based on the MB optimistic initialization and exploration driven by information gain. This qualitative link leads to the conclusion that an agent with only the MB optimistic initialization cannot explain human behavior for the same reason that an agent with intrinsic reward based on information gain cannot explain human behavior.

Model-free optimistic initialization. As opposed to the MB branch of the hybrid algorithm, the MF branch does not have any explicit knowledge about the existence of different goal states and their values. However, the initial value $Q^{(0)}_{\text{MF,ext}}$ of the MF extrinsic Q-values quantifies an expectation of the reward values in the environment prior to any interaction with the environment. During episode 1, no extrinsic reward is received by the agent, hence, for a small enough learning rate ρ and an optimistic initialization $Q^{(0)}_{\text{MF,ext}} > 0$, the extrinsic reward prediction errors are always negative (Equation 5.3). As a result, $Q^{(t)}_{\text{MF,ext}}(s,a)$ decreases as an agent keeps taking action a in state s, which motivates the agent to try new actions. This is a well-known mechanism for directed exploration in the machine learning community (Sutton and Barto, 2018). Similar to the MB optimistic initialization, the effect of the MF optimistic initialization fades out over time – which makes them both similar to exploration driven by information gain.

During episode 2-5, the exact theoretical analysis of the MF optimistic initialization is complex and dependent on an agent's exact trajectory (because of the eligibility traces). However, whether the MF extrinsic Q-values show a preference for exploring or leaving the stochastic part essentially depends on the reward value of the discovered goal state r_{G^*} and the initialization value $Q^{(0)}_{\mathrm{MF,ext}}$. For example, if an agent, starting at $s1$, takes the perfect trajectory of $s1 \to s2 \to s3 \to s4 \to s5 \to s6 \to G^*$ in episode 1, then, given a unit decay rate of the eligibility traces (i.e., $\mu_{\mathrm{ext}} = 1$), it is easy to see that, in the 1st visit of state 4 in episode 2, the agent prefers the stochastic/bad action over the progressing action if $r_{G^*} < \frac{1}{\lambda^2_{\mathrm{ext}}}(1 - \lambda_{\mathrm{ext}})(1 + \lambda_{\mathrm{ext}} + \lambda^2_{\mathrm{ext}})Q^{(0)}_{\mathrm{MF,ext}}$. This implies that, even though the MF branch is not explicitly aware of different goal states and their reward values, it is still able to model a type of reward optimism through initialization of Q-values. Nevertheless, since model fitting reveals an importance factor significantly greater than 0.5 (Figure 5.6), the effective reward optimism generated by optimistic initialization is not strong enough to explain human behavior.

5.4.6 Efficient model-based planning for simulated participants

For simulating efficient agents in Figure 5.2, we set $\varepsilon = 0$ (see Figure 5.1A) and used a pure MB version of our algorithm with 13 parameters:

$$\{r^*_1, r^*_2, \lambda_{\mathrm{ext}}, \lambda_{\mathrm{int}}, \kappa, \epsilon_{\mathrm{new}}, \epsilon_{\mathrm{obs}}, T_{PS,\mathrm{ext}}, T_{PS,\mathrm{int}}, \beta^{(1)}_{\mathrm{MB,ext}}, \beta^{(2)}_{\mathrm{MB,ext}}, \beta^{(1)}_{\mathrm{MB,int}}, \beta^{(2)}_{\mathrm{MB,int}}\}. \tag{5.16}$$

We considered perfect model-building by assuming $\kappa = 1$ and almost perfect planning by assuming $T_{PS,\mathrm{ext}} = T_{PS,\mathrm{int}} = 100$. We chose discount factors λ_{ext} and λ_{int} as well as prior parameters ϵ_{new} and ϵ_{obs} in the range of fitted parameters reported by Xu et al. (2021): $\lambda_{\mathrm{ext}} = 0.95$, $\lambda_{\mathrm{int}} = 0.70$, $\epsilon_{\mathrm{new}} = 10^{-5}$ and $\epsilon_{\mathrm{obs}} = 10^{-4}$. To relatively separate the effect of optimistic initialization (Sutton and Barto, 2018) from seeking intrinsic reward in episode 2-5, we assumed the same value of reward for all goals, i.e., $r^*_1 = r^*_2 = 1$. Finally, we considered $\beta^{(1)}_{\mathrm{MB,ext}} = 0$ to have pure intrinsic reward seeking in episode 1.

After fixing parameter values for 10 out of 13 parameters in Equation 5.16, we fine-tuned $\beta^{(1)}_{\mathrm{MB,int}}$ to minimize the average length of episode 1 (to find the goal as fast as possible; see Appendix C). For episodes 2-5, we first set $\beta^{(2)}_{\mathrm{MB,int}} = 0$ and $\beta^{(2)}_{\mathrm{MB,ext}} = 10$ to have a non-deterministic policy for purely seeking extrinsic reward after the 1st encounter of the goal (the lightest shade of colors in Figure 5.2). Different shades of color in Figure 5.2 corresponds to different choices of $\omega \in [0, 1]$ for $\beta^{(2)}_{\mathrm{MB,int}} = \omega\beta^{(1)}_{\mathrm{MB,int}}$ and $\beta^{(2)}_{\mathrm{MB,ext}} = (1 - \omega) \cdot 10$. More precisely, we used $\omega = 0$ for the darkest color (pure extrinsic reward seeking), $\omega = 1$ for the lightest color (pure intrinsic reward seeking), and $\omega = 0.7$ for the one in between. Higher values of ω indicates higher relative importance of the intrinsic reward.

5.4.7 Model-fitting and model-comparison

To compare seeking different intrinsic rewards based on their explanatory power, we considered our full hybrid (with both MF and MB components) algorithm – except that we put $T_{PS,\text{ext}} = T_{PS,\text{int}} = 100$ to decrease number of parameters, based on the results of Xu et al. (2021) showing the negligible importance of this parameter. As a result, we had 20 free parameters for each of the three intrinsic rewards (i.e., novelty, information gain, and surprise). For each intrinsic reward $R \in \{\text{novelty}\,, \text{inf-gain}\,, \text{surprise}\}$ and for each participant $n \in \{1, ..., 57\}$, we estimated the algorithm's parameters by maximizing likelihood of data given parameters:

$$\hat{\Phi}_{n,R} = \arg\max_{\Phi} P(D_n|\Phi, R_n = R) \tag{5.17}$$

where D_n is the data of participant n, R_n is the intrinsic reward assigned to participant n, $P(D_n|\Phi, R_n = R)$ is the probability of D_n being generated by our intrinsically motivated algorithm seeking $R_n = R$ with its parameter equal to Φ (see Equation 5.9), and $\hat{\Phi}_{n,R}$ is the set of estimated parameters that maximizes that probability. For optimization, we used Subplex algorithm (Rowan, 1990) as implemented in Julia NLopt package.

Because all algorithms have the same number of parameters, we considered the maximum log-likelihood as the model log-evidence, i.e., for intrinsic reward R and participant n, we consider $\log P(D_n|R_n = R) \approx \log P(D_n|\hat{\Phi}_{n,R}, R_n = R)$ – which is equal to a shifted Schwarz approximation of the model log-evidence (also called BIC) (Daw, 2011; Kass and Raftery, 1995). Figure 5.4A1 shows the total log-evidence $\sum_n \log P(D_n|R_n = R)$. With the fixed effects assumption at the level of models (i.e., assuming that $R_1 = R_2 = ... = R_{57} = R^*$), the total log-evidence is equal to the log posterior probability $\log P(R^* = R|D_{1:57})$ of R being the intrinsic reward used by all participants (plus a constant). See Daw (2011); Wilson and Collins (2019) for tutorials.

We also considered the Bayesian model selection method of Rigoux et al. (2014) with the random effects assumption, i.e., assuming that participant n uses the intrinsic reward $R_n = R$, which is not necessarily the same as the one used by other participants, with probability P_R. We performed Markov Chain Monte Carlo sampling (using Metropolis Hasting algorithm (Efron and Hastie, 2016) with uniform prior and 40 chains of length $10'000$) for inference and estimated the joint posterior distribution

$$P(R_{1:57}, P_{\text{novelty}}, P_{\text{inf-gain}}, P_{\text{surprise}}|D_{1:57}).$$

Figure 5.4A2 shows the expected posterior probability $\mathbb{E}[P_R|D_{1:57}]$ as well as the protected exceedance probabilities $P(P_R > P_{R'} \text{ for all } R' \neq R|D_{1:57})$ computed by using the participant-wise log-evidences.

The boxplots of the fitted parameters of novelty-seeking are shown in Appendix C. The same set of parameters were used for model-recovery in Figure 5.4B, posterior predictive checks in Figure 5.5, computing the relative importance of novelty in Figure 5.6A-B (see 'Relative importance of novelty in action-selection'), and parameter recovery in Figure 5.6C.

5.4.8 Posterior predictive checks, model-recovery, and parameter-recovery

For each intrinsic reward $R \in \{\text{novelty}, \text{inf-gain}, \text{surprise}\}$ and participant group $\mathcal{G} \in \{2 \text{ CHF}, 3 \text{ CHF}, 4 \text{ CHF}\}$, we repeated the following two steps 500 times: 1. We sampled participant n from group $\mathcal{G}$ with probability $\frac{P(R_n=R|D_{1:57})}{\sum_{m\in\mathcal{G}} P(R_m=R|D_{1:57})}$. 2. We ran a 5-episode simulations in our environment using the intrinsic reward R and the parameter $\hat{\Phi}_{n,R}$, i.e., we sampled a trajectory D from $P(D|\hat{\Phi}_{n,R}, R_n = R)$ (with the G^* of the environment corresponding to the group $\mathcal{G}$). As a result, we ended up with 1500 simulated participants (with randomly sampled parameters) for each algorithm.

We considered the simulated participants who took more than 3000 actions in any of the 5 episodes to be similar to the human participants who quit the experiment and excluded them from further analyses: 238 ($\sim$ 16%) of simulated participants seeking novelty, 166 ($\sim$ 11%) of those seeking information gain, and 374 ($\sim$ 25%) of those seeking surprise. We note that, even with the marginal influence of surprise on action-selection (Figure 5.5C), one fourth of participants seeking surprise cannot escape the stochastic part in less than 3000 actions. Moreover, we excluded, separately for each algorithm, the simulated participants who took more than 3 times group-average number of actions in episodes 2-5 to finish the experiment (i.e., the same criterion that we used to detect non-attentive human participants): 45 ($\sim$ 3%) of simulated participants seeking novelty, 77 ($\sim$ 5%) of those seeking information gain, and 27 ($\sim$ 2%) of those seeking surprise. We then analyzed the remaining participants (1217 simulated participants seeking novelty, 1257 seeking information gain, and 1099 seeking surprise) as if they were real human participants. Figure 5.5 and its supplements in Appendix C show the data statistics of simulated participants in comparison to human participants.

Given the participants simulated by each of the three intrinsically motivated algorithms, we fitted all three algorithms to the action-choices of 150 simulated participants (50 from each participant group, i.e., 2 CHF, 3 CHF, and 4 CHF). Then, we applied the Bayesian model selection method of Rigoux et al. (2014) to 5 randomly chosen sub-populations of these 150 simulated participants (each with 60 participants, i.e., 20 from each participant group). Figure 5.4B shows the results of the model-comparison averaged over these 5 repetitions. Figure 5.6C shows the relative importance of novelty in action-selection (see Equation 5.18) for each of the 150 simulated participants estimated using the original parameters (which were used for simulations) and the recovered parameters (which were found by re-fitting the algorithms to the simulated data).

5.4.9 Relative importance of novelty in action-selection

The relative importance of novelty in action-selection depends not only on the inverse-temperatures $\beta_{\text{MB,ext}}$, $\beta_{\text{MF,ext}}$, $\beta_{\text{MB,int}}$, and $\beta_{\text{MF,int}}$ but also on the variability of Q-values; for example, if the extrinsic Q-values $Q^{(t)}_{\text{MB,ext}}(s,a)$ and $Q^{(t)}_{\text{MF,ext}}(s,a)$ are the same for all state-action pairs, then, independently of the values of the inverse-temperatures, the action is taken by a pure novelty-seeking policy – because the policy in Equation 5.7 can be re-written as $\pi^{(t)}(a|s) \propto \exp\left[\beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) + \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)\right]$. Thus, to measure the contribution of different components of action-selection to the final policy, we need to consider the variations in their Q-values as well.

In this section, we propose a variable $\omega_{\text{i2e}} \in [0,1]$ for quantifying the relative importance of seeking intrinsic reward in comparison to seeking extrinsic reward. We first define total intrinsic and extrinsic Q-values as $Q^{(t)}_{\text{ext}}(s,a) = \beta_{\text{MB,ext}} Q^{(t)}_{\text{MB,ext}}(s,a) + \beta_{\text{MF,ext}} Q^{(t)}_{\text{MF,ext}}(s,a)$ and $Q^{(t)}_{\text{int}}(s,a) = \beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) + \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)$, respectively. We further define the state-dependent variations in Q-values as $\Delta Q^{(t)}_{\text{ext}}(s) = \max_a Q^{(t)}_{\text{ext}}(s,a) - \min_a Q^{(t)}_{\text{ext}}(s,a)$ and $\Delta Q^{(t)}_{\text{int}}(s) = \max_a Q^{(t)}_{\text{int}}(s,a) - \min_a Q^{(t)}_{\text{int}}(s,a)$ as well as their temporal average $\Delta\bar{Q}_{\text{ext}} = \left\langle \Delta Q^{(t)}_{\text{ext}}(s_t) \right\rangle$ and $\Delta\bar{Q}_{\text{int}} = \left\langle \Delta Q^{(t)}_{\text{int}}(s_t) \right\rangle$, where $\langle . \rangle$ shows the temporal average. $\Delta\bar{Q}_{\text{ext}}$ and $\Delta\bar{Q}_{\text{int}}$ show the average difference between the most and least preferred action with respect to seeking extrinsic and intrinsic reward, respectively. Therefore, a feasible way to measure the influence of seeking intrinsic reward on action-selection is to define ω_{i2e} as

$$\omega_{\text{i2e}} = \frac{\Delta\bar{Q}_{\text{int}}}{\Delta\bar{Q}_{\text{ext}} + \Delta\bar{Q}_{\text{int}}}. \tag{5.18}$$

Figure 5.6A shows the value ω_{i2e} in episode 2-5 computed for each human participant (dots) and averaged over different groups (bars), and Figure 5.6B shows the same for episode 1. Figure 5.5C shows the value ω_{i2e} in episode 2-5 for the 2 CHF group of simulated participants. See Appendix C for a similar approach for quantifying the relative importance of the MB and MF policies in action-selection.

6 Conclusion and future directions

In this thesis, I conducted a mathematical analysis of various definitions and computational models of surprise and novelty. I showed that these theoretical results facilitate the interpretation of past research in a different light and inform experimental designs to dissociate different computational roles of surprise and novelty in the brain. I studied two particular examples of such experiments and quantified the contribution of surprise and novelty to human adaptive and exploratory behavior. Each chapter offered an individual discussion section with an outline of potential future directions. The goal of this general conclusion is to present a wider perspective on the next steps toward demystifying the role of surprise and novelty in the brain. I hope that the results of this thesis can help us, as a community, move forward in these directions.

Examining several definitions of surprise and novelty raises an immediate question: How many fundamentally unique physiological signals are involved in the brain computations related to surprise and novelty? More specifically, we can ask: How does the computation of surprise and novelty vary across different sensory modalities (e.g., visual versus auditory; Grundei et al. (2023)) and levels of abstraction (e.g., low-level visual features versus high-level image categories; Richter et al. (2023))? Given a specific sensory modality and a level of abstraction, what are the characteristics of surprise and novelty signals in different brain regions (e.g., Gijsen et al. (2021); Kolossa et al. (2015); Visalli et al. (2019))? What is the link between the local neural circuitry of surprise and novelty computation in sensory areas and the global signals transmitted through neuromodulators like dopamine, norepinephrine, or serotonin (e.g., Jordan (2023))? And, what is the relationship between these neural responses and the subjective perception of surprise in human participants measured by self-reports (Maguire et al., 2011; Reisenzein et al., 2019)?

Addressing such questions will ideally help us narrow down the set of 'brain-related' mathematical definitions of surprise and novelty, which in turn help identify their respective computational role in different cognitive functions of the brain. One goal of this thesis was to show that careful experiment design enables dissociating the predictions of different definitions of surprise and novelty for, e.g., goal-directed exploration (chapters 3 and 5). Similar methodologies can be used to study the contribution of different definitions to curiosity-driven exploration (see Appendix E) or behavioral patterns that are influenced by multiple motivational signals. Theory-driven experiments can then be used for such cases to answer questions like: What are the principal drives (e.g., novelty, information gain, nutrition, sexual drive, etc.) of decision-making (e.g., Ahmadlou et al. (2021); Kobayashi et al. (2019))? How, and at which stage of processing, do these signals interact to influence decision-making (e.g., Bromberg-Martin et al. (2024); Ogasawara et al. (2022))? How do environmental factors (e.g., task instruction, optimism, prior knowledge) influence this interaction (e.g., Meder and Nelson (2012); Traner et al. (2021))? And, importantly, how do answers to these questions change across species (e.g., Bromberg-Martin et al. (2024)) or even across individual subjects (e.g., Kelly et al. (2021))?

Along the same line of research, similar methodologies can be used to design experiments concerning the role of surprise and novelty in learning and memory. A fundamental challenge in this context involves identifying the distinct mechanisms by which surprise and novelty influence memory consolidation (Rouhani and Niv, 2021), modification (Sinclair and Barense, 2018), and segmentation (Antony et al., 2021). It remains unclear whether these mechanisms are also involved in the surprise modulation of learning speed (Gershman et al., 2017; Glaze et al., 2018; Jordan and Keller, 2023). Further questions then arise concerning the influence of attention, task instruction, cognitive load, and similar factors on each of these mechanisms for learning and memory modulation (e.g., Solomon et al. (2021); Zhao et al. (2019)). Computational modeling can particularly help this line of research by providing insights into the cognitive computations underlying these mechanisms. For example, since change detection surprise is proven to be optimal for adaptive learning, the mechanisms involved in adaptive learning are likely to entail a comparison between two contrasting predictive models (see chapter 2 and Appendix D). Such insights can support the interpretation of neural data within the scope of behavioral measurements (Niv, 2021).

Finally, similarly to every subfield of computational neuroscience, computational models of surprise and novelty are distributed across different levels of analysis, e.g., across computational, algorithmic, and mechanistic levels as termed in Marr's vocabulary (Marr, 1982) and normative, heuristics, and data-driven modeling levels as discussed in the Introduction (chapter 1). Bridging these different levels of modeling is a key step toward a unified understanding of how surprise and novelty influence the brain and behavior.

A Appendix to chapter 2

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

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

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

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

and similarly

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

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

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

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

A.2 Proof of Proposition 2

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

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

For the difference in the 1st definition of the Shannon surprise (c.f. Equation 2.9), we can write

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

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

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

The proof is more straightforward for the difference in the 2nd definition (c.f. Equation 2.11) where we have

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

Therefore, the proof is complete. ■

A.3 Proof of Proposition 3

Based on the definitions of the two versions of the Shannon surprise (c.f. Equation 2.9 and Equation 2.11), we have

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

The proof is complete by using these equations and replacing the probabilities in Equation 2.15 and Equation 2.16. ■

A.4 Proof of Proposition 4

For a categorical task with N categories and one-hot coded observations, we have (c.f. Equation 2.18 and Equation 2.19)

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

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 $p_{1,n} = p_c P(n|x_{t+1}; b^{(0)}) + (1-p_c)P(n|x_{t+1}; b^{(t)})$ and $p_{2,n} = P(n|x_{t+1}; b^{(t)})$.

We show the one-hot coded vector corresponding to category $m \in \{1, ..., N\}$ by e_m. For the absolute error surprise, we have (c.f. Equation 2.20)

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

which is the same as $2\mathsf{S}_{\mathrm{SPE}i}(y_{t+1} = e_m|x_{t+1}; b^{(t)})$ (c.f. Equation 2.15 and Equation 2.16).

For the squared error surprise, we have (c.f. Equation 2.20)

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

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

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

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 $b^{(t)}$, has a Gaussian distribution with a covariance matrix $\sigma^2 I$, i.e.,

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

We then have

$$\begin{aligned}\mathsf{S}_{\text{Sh2}}(y_{t+1}|x_{t+1};b^{(t)}) &= -\log \mathcal{N}\Big(y_{t+1}; E_2[Y_{t+1}], \sigma I\Big)\\ &= \frac{N}{2}\log\big(2\pi\sigma\big) + \frac{||y_{t+1}-E_2[Y_{t+1}]||_2^2}{2\sigma^2}\\ &= a_0 + a_1 \mathsf{S}_{\text{Sq},2}(y_{t+1}=e_m|x_{t+1};b^{(t)}),\end{aligned} \tag{A.15}$$

where $a_0 = N\log\big(2b\sigma\big)/2$ and $a_1 = 1/(2\sigma^2)$. Therefore, the proof is complete. ■

A.6 Proof of Proposition 6

Using the definition of the two surprise measures in Equation 2.20, we have, for $y_{t+1} \in \mathbb{R}$,

$$\begin{aligned}\mathsf{S}_{\text{Sq}i}(y_{t+1}|x_{t+1};b^{(t)}) &= ||y_{t+1} - E_i[Y_{t+1}]||_2^2\\ &= |y_{t+1} - E_i[Y_{t+1}]|^2 = \mathsf{S}_{\text{Ab}i}(y_{t+1}|x_{t+1};b^{(t)})^2.\end{aligned} \tag{A.16}$$

Therefore, the proof is complete. ■

A.7 Proof of Proposition 7

Using the definition of the uRPE and the absolute error surprise in Equation 2.20 and Equation 2.25, we have

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

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

$$\begin{aligned}\mathsf{S}_{\mathrm{Sq}i}(y_{t+1}|x_{t+1};b^{(t)}) &= ||y_{t+1} - E_i[Y_{t+1}]||_2^2 \\ &= |\tilde{r}_{t+1} - E_i[\tilde{R}_{t+1}]|^2 + ||s_{t+1} - E_i[S_{t+1}]||_2^2 \\ &= \mathsf{S}_{\mathrm{uRPE}i}(y_{t+1}|x_{t+1};b^{(t)})^2 + \mathsf{S}_{\mathrm{Sq}i}(s_{t+1}|x_{t+1};b^{(t)}).\end{aligned} \tag{A.18}$$

Therefore, the proof is complete. ■

A.8 Proof of Proposition 8

For the 1st definition of the Bayesian surprise (c.f. Equation 2.29), we have

$$\mathsf{S}_{\mathrm{Ba1}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathrm{D}_{\mathrm{KL}}\Big[\mathbb{P}^{(t)}_{\Theta_{t+1}}||\mathbb{P}^{(t+1)}_{\Theta_{t+1}}\Big] = \mathbb{E}_{\mathbb{P}^{(t)}}\Big[\log\frac{\mathbb{P}^{(t)}(\Theta_{t+1})}{\mathbb{P}^{(t+1)}(\Theta_{t+1})}\Big]. \tag{A.19}$$

We know

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

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

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

which is equivalent to (c.f. Equation 2.9 and Equation 2.11)

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

For the 2nd definition of the Bayesian surprise (c.f. Equation 2.30), we have

$$\mathsf{S}_{\mathrm{Ba2}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathrm{D}_{\mathrm{KL}}\Big[b^{(t)}||b^{(t+1)}\Big] = \mathbb{E}_{b^{(t)}}\Big[\log\frac{b^{(t)}(\Theta)}{b^{(t+1)}(\Theta)}\Big]. \tag{A.24}$$

We use Equation 2.28 and Equation A.21 and write

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

which is equivalent to (c.f. Equation 2.9 and Equation 2.11)

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

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 (c.f. Equation 2.40) for which we have

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

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

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

which is equivalent to (c.f. Equation 2.9 and Equation 2.11)

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

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

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

which is the same as Equation 2.42. For the 1st definition of the Confidence Corrected surprise (c.f. Equation 2.37), we can repeat all steps to have

$$\begin{aligned}\mathsf{S}_{\text{CC1}}(y_{t+1}|x_{t+1};b^{(t)}) =& \,\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})\\ &- \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b_{\text{flat}})\\ &+ \mathsf{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};b^{(t)})\\ &- \text{D}_{\text{KL}}\Big[b^{(t)}||p_c b^{(0)} + (1-p_c)b^{(t)}\Big]\\ &+ \text{D}_{\text{KL}}\Big[b^{(t)}||b_{\text{flat}}\Big].\end{aligned} \tag{A.31}$$

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

A.10 Proof of Corollary 1

The corollary is the direct conclusion of Equation A.6 and Equation 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 $P(y_{t+1}|x_{t+1};b^{(0)}) = 1/|\mathcal{Y}|$. We therefore (using Equation 2.5, Equation 2.9, and Equation 2.11) have

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

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 $\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(0)})$ (c.f. Equation 2.9) which implies that $\Delta\mathsf{S}_{\text{Sh1}}(y_{t+1}|x_{t+1};b^{(t)})$ (c.f. Proposition 2) in Equation 2.42 is equal to 0. Similarly, in the limit of $p_c \to 1$, we have $\text{D}_{\text{KL}}\Big[b^{(t)}||p_c b^{(0)} + (1-p_c)b^{(t)}\Big] = \text{D}_{\text{KL}}\Big[b^{(t)}||b^{(0)}\Big]$. Therefore, in the limit of $p_c \to 1$ and given Equation 2.42, we have $\mathsf{S}_{\text{CC2}}(y_{t+1}|x_{t+1};b^{(t)}) = \mathsf{S}_{\text{Ba2}}(y_{t+1}|x_{t+1};b^{(t)})$. ■

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*

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

and

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

where $\mathbb{P}^{(t)}_{\Theta_{t+1}|y,x_{t+1}} := \mathbb{P}^{(t)}_{\Theta_{t+1}}(.|Y_{t+1} = y, x_{t+1})$ 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 S_{Po1} for which we have (c.f. Equation 2.34)

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

where

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

and, using Bayes' rule,

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

Using the Bayes' rule and the definition of the marginal probability (c.f. Equation 2.4), we can find

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

that is equal to

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

and as a result (using Equation 2.9 and Equation 2.11)

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

which, using Equation A.35, makes the proof complete.

To prove the 2nd equality, we note that (c.f. Equation 2.35)

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

and

$$\log\frac{P(y|x_{t+1};b^{(t)})}{P(y|x_{t+1};b^{(t+1)})} = \log\frac{P(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}})}{P(y|x_{t+1};b^{(t+1)})} + \log\frac{P(y|x_{t+1};b^{(t)})}{P(y|x_{t+1};\mathbb{P}^{(t)}_{\Theta_{t+1}})}. \tag{A.42}$$

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

B Appendix to chapter 3

B.1 SurNoR and alternative algorithms

B.1.1 Surprise-Novelty-Reward (SurNoR) algorithm

The SurNoR algorithm (Alg. 1) combines surprise signals with novelty and reward so as to explore and learn the environment and exploit rewards. A simple block diagram of the algorithm is shown in Figure 3.4C of the main text. In the SurNoR algorithm, a model-based and a model-free branch interact with each other. The output of each branch is a pair of Q-values for estimated novelty and estimated reward. The model-based branch updates model-based Q-values using a world-model that is estimated online, while the model-free branch uses a surprise-modulated TD-learner for updating the model-free Q-values. Finally, actions are selected following a hybrid policy that combines model-free and model-based Q-values - see Daw et al. (2011); Gläscher et al. (2010) for similar approaches. In this section, we describe the SurNoR algorithm in detail. For the sake of clarity and coherence, we repeat here some details already explained in the main text.

Formalization of the environment. The state and the action at time t are random variables S_t and A_t which take values in the finite sets $\mathcal{S}$ and $\mathcal{A}$, respectively. In the particular case of our experiment, we have $\mathcal{S} = \{1, ..., 10, G\}$ and $\mathcal{A} = \{1, ..., 4\}$. Taking a Bayesian perspective, we consider the transition probability matrix as another random variable Θ, i.e.

$$\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a, \Theta = \theta) = \theta_{s,a}(s'). \tag{B.1}$$

where the values of $\theta_{s,a}(s')$ for combinations of s, a, and s' are unknown and needed to be estimated from experience. Since our environment is deterministic, except for the switch of two states before the start of block 2, the real transition probabilities are

$$\theta^{\text{real}}_{s,a}(s') = \delta(s', T(s, a)), \tag{B.2}$$

where $T(s, a)$ denotes the target state of the transition from state s given action a, and the Kronecker δ is defined as $\delta(x, x') = 1$ if $x = x'$ and zero otherwise; $T(s, a)$ corresponds

Algorithm 1 Pseudocode for SurNoR

```
Specify $\mathcal{S}$ and $\mathcal{A}$
Specify Episode (Epi) and Block
# Parameter specification
Specify $\{\epsilon, m, \lambda_R, \lambda_N, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}, T_{\text{PS}}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho, \omega_{\text{scale}}, \omega_0, \omega_{11}, \omega_{12}\}$.
if Block $= 1$ and Epi $= 1$, then Put $\beta = \beta_1$, $\omega = \omega_{11}$ and $\beta_N = \beta_{N1}$.
if Block $= 1$ and Epi $\neq 1$, then Put $\beta = \beta_1$, $\omega = \omega_0$ and $\beta_N = 0$.
if Block $= 2$ and Epi $= 1$, then Put $\beta = \beta_2$, $\omega = \omega_{12}$ and $\beta_N = \beta_{N2}$.
if Block $= 2$ and Epi $\neq 1$, then Put $\beta = \beta_2$, $\omega = \omega_0$ and $\beta_N = 0$.
# Initialization
Put $e_R^{(1)}(s,a) = e_N^{(1)}(s,a) = 0$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
if Epi $= 1$ and Block $= 1$ then
    Put $C_s^{(1)} = 0$, $U_R^{(1)}(s) = 0$, $U_N^{(1)}(s) = \frac{\log(|\mathcal{S}|)}{1-\lambda}$, $\forall s \in \mathcal{S}$.
    Put $Q_{\text{MB,R}}^{(1)}(s,a) = 0$, $Q_{\text{MB,N}}^{(1)}(s,a) = U_N^{(1)}(s)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Put $Q_{\text{MF,R}}^{(1)}(s,a) = 0$, $Q_{\text{MF,N}}^{(1)}(s,a) = Q_{N0}$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    Put $\alpha_{s,a}^{(1)}(s') = \epsilon$, $\forall (s,s',a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$.
else
    Initialize $C_s^{(1)}$, $U_R^{(1)}(s)$, $U_N^{(1)}(s)$, $Q_{\text{MB,R}}^{(1)}(s,a)$, $Q_{\text{MB,N}}^{(1)}(s,a)$, $Q_{\text{MF,R}}^{(1)}(s,a)$, $Q_{\text{MF,N}}^{(1)}(s,a)$ and $\alpha_{s,a}^{(1)}(s')$ with their latest values in the previous Episode.
end if
Initialize state $S_1 = s_1$ and update counts $C_s^{(1)} \leftarrow C_s^{(1)} + \delta(s, s_1)$.
$t \leftarrow 1$.
# Going through the task
while $s_t \neq s_{Goal}$ do
    # Making action
    Compute $Q_{\text{MF}}^{(t)}(s,a) = Q_{\text{MF,R}}^{(t)}(s,a) + \beta_N Q_{\text{MF,N}}^{(t)}(s,a)$.
    Compute $Q_{\text{MB}}^{(t)}(s,a) = Q_{\text{MB,R}}^{(t)}(s,a) + \beta_N Q_{\text{MB,N}}^{(t)}(s,a)$.
    Sample $a_t$ from $\pi(A_t = a | S_t = s) \propto \exp\left\{\beta\left[\omega\left(\omega_{\text{scale}} Q_{\text{MF}}^{(t)}(s,a)\right) + (1-\omega)\, Q_{\text{MB}}^{(t)}(s,a)\right]\right\}$.
    Observe $S_{t+1} = s_{t+1}$.
    # Updating internal variables
    Update counts $C_s^{(t+1)} = C_s^{(t)} + \delta(s, s_{t+1})$ and novelty $\mathsf{N}^{(t+1)}(s) = \log \frac{t+|\mathcal{S}|}{C_s^{(t+1)}+1}$.
    Update $\alpha^{(t+1)}$, $U_R^{(t+1)}$, $U_N^{(t+1)}$, $Q_{\text{MB,R}}^{(t+1)}$, and $Q_{\text{MB,N}}^{(t+1)}$ using the model-based branch in Alg. 2.
    Update $e_N^{(t+1)}$, $e_R^{(t+1)}$, $Q_{\text{MF,R}}^{(t+1)}$, and $Q_{\text{MF,N}}^{(t+1)}$ using the model-free branch in Alg. 4.
    # Going to the next step
    $t \leftarrow t + 1$.
end while
```

to the arrows in Figure 3.1B and Figure 3.1D in the main text. The target state depends on the block number in our experiment. Note that $T(s, a)$ is unknown to the participants and to the SurNoR algorithm.

Definition of novelty. While a participant moves in the environment, the count

$$C_s^{(t)} = |\{t' : 1 \leq t' \leq t \text{ and } s_{t'} = s\}|$$

indicates how often state s has been encountered up to time t. We assume that at each time t, participants are able to estimate the empirical frequency $p_f^{(t)}(s)$ of encountering state $s \in \mathcal{S}$, formally defined as

$$p_f^{(t)}(s) = \frac{C_s^{(t)} + 1}{t + |\mathcal{S}|}, \tag{B.3}$$

where $|\mathcal{S}|$ is the total number of states (i.e., 11 for our experiment). Note that the participants know the total number of states, due to the pre-experiment introduction. The empirical frequency in Equation B.3 is equal to the expected probability of observing state s given $s_{1:t}$ under the assumption of a uniform prior over the probabilities of observing different states: before the start of the experiment all $|\mathcal{S}|$ states have the same prior probability $p_f^{(0)}(s) = 1/|\mathcal{S}|$.

We define the novelty of the state s at time t as the negative logarithm of the empirical frequency

$$\mathsf{N}^{(t)}(s) = -\log(p_f^{(t)}(s)). \tag{B.4}$$

In our algorithm, novelty acts just like an internally generated reward or exploration bonus (see subsection 'Formalizing model-based Q-values'). The main difference between our definition of novelty and most of the previously proposed measures of 'exploration bonus' (Achiam and Sastry, 2017; Burda et al., 2019; Kolter and Ng, 2009; Little and Sommer, 2013; Martin et al., 2017; Mobin et al., 2014; Pathak et al., 2017) is in their dependency upon states *and* actions: while the usual exploration bonus measures are functions of state-action pairs, we define our novelty as a function of states only. Our choice is more consistent with the behavior of participants in our experiment, since a reasonable strategy of participants is to visit the states that they rarely encounter in the experiment as opposed to testing all actions in all states. From this perspective, our novelty measure is similar to the exploration bonus proposed by Bellemare et al. (2016).

In three of our alternative algorithms (see Section 'Alternative algorithms') we use a state-of-the-art exploration strategy (Achiam and Sastry, 2017; Burda et al., 2019) which defines exploration bonus (or internal reward) as a function of the pairs of states and actions. We compare these algorithms with SurNoR (see Figure 3.5 in the main text).

Model-based branch of SurNoR

The pseudocode for the model-based branch is shown in Alg. 2. In this subsection, the details are explained.

Algorithm 2 Pseudocode for the model-based branch of SurNoR

```
# Surprise and adaptation rate
Compute $\mathsf{S}^{(t+1)} = \hat{\theta}^{(1)}_{s_t,a_t}(s_{t+1})/\hat{\theta}^{(t)}_{s_t,a_t}(s_{t+1})$.
Compute $\gamma_{t+1} = m\mathsf{S}^{(t+1)}/(1 + m\mathsf{S}^{(t+1)})$.
# Updating the belief
Update $\alpha^{(t+1)}_{s_t,a_t}(s) = (1 - \gamma_{t+1})\alpha^{(t)}_{s_t,a_t}(s) + \gamma_{t+1}\epsilon + \delta(s_{t+1}, s)$, $\forall s \in \mathcal{S}$.
Update $\alpha^{(t+1)}_{s,a}(s') = \alpha^{(t)}_{s,a}(s')$, $\forall s \neq s_t, a \neq a_t$, and $s' \in \mathcal{S}$.
Update $\hat{\theta}^{(t+1)}$ as $\hat{\theta}^{(t+1)}_{s,a}(s') = \alpha^{(t+1)}_{s,a}(s')/\sum_{\tilde{s}' \in \mathcal{S}} \alpha^{(t+1)}_{s,a}(\tilde{s}')$.
# Updating the values
Update $Q^{(t+1)}_{\text{MB,N}}(s,a)$ and $U^{(t+1)}_{N}(s)$ using Alg. 3 and $\mathsf{N}^{(t+1)}(s)$ as rewards.
if Epi = 1 and Block = 1 and $s_t \neq s_{Goal}$ then
    Update $Q^{(t+1)}_{\text{MB,R}}(s,a) = U^{(t+1)}_{R}(s) = 0$.
else
    Update $Q^{(t+1)}_{\text{MB,R}}(s,a)$ and $U^{(t+1)}_{R}(s)$ using Alg. 3 and $R(s) = \delta(s, s_{Goal})$ as rewards.
end if
```

World-model. The participants knew that there were 11 states and 4 possible actions in each state. However, they were not aware of the actual transition probability matrix. In particular, they did not know whether the environment is deterministic or stochastic. Therefore, we define a participant's model of the world as an approximation $\hat{b}$ of the posterior distribution of the transition probability matrix, similar to the approach of (Faraji et al., 2018; Friston, 2010; Friston et al., 2017),

$$\hat{b}^{(t)}(\theta) \approx \mathbb{P}(\Theta = \theta | S_{1:t} = s_{1:t}, A_{1:t-1} = a_{1:t-1}). \tag{B.5}$$

In the following, we call $\hat{b}$ the belief of the participant. We assume that a participant estimates the transition probabilities by a weighted average

$$\hat{\theta}^{(t)} = \mathbb{E}_{\hat{b}^{(t)}}[\Theta] = \int \theta \hat{b}^{(t)}(\theta) d\theta, \tag{B.6}$$

where the weighting factor is given by the belief $\hat{b}^{(t)}$. For convenience, the transition probability $\hat{\theta}^{(t)}_{s,a}(s')$ is written as $p^{(t)}(s'|s,a)$ in the main text, e.g., in Equation 3.4 and Equation 3.3.

For exact Bayesian inference one needs to explicitly specify the generative model which governs the transition. Particularly, the dynamics of Θ over time should be known, e.g., whether it is fixed, continuously drifting, or experiencing abrupt changes (Behrens et al., 2007; Liakoni et al., 2021; Mathys et al., 2011; Nassar et al., 2010). However, rather than making explicit assumptions about the generative model as a starting point for exact

Bayesian inference, we work with a general (parametric, see the next part) distribution $\hat{b}^{(t)}$ which is updated by an appropriate learning algorithm after each observation, similar to approaches in machine learning (Masegosa et al., 2017; Ozkan et al., 2013).

Beliefs as Dirichlet distributions. We assume that the transition probabilities from different state-action pairs are independent of each other, i.e.

$$\hat{b}^{(t)}(\theta) = \prod_{s \in \mathcal{S}, a \in \mathcal{A}} \hat{b}^{(t)}(\theta_{s,a}), \tag{B.7}$$

where $\theta_{s,a}$ is defined as in Equation B.1. As a natural[1] choice for a probability distribution over transition probabilities, we consider the belief $\hat{b}^{(t)}(\theta_{s,a})$ to be a Dirichlet distribution with parameter $\alpha_{s,a}^{(t)}$:

$$\hat{b}^{(t)}(\theta_{s,a}) = \mathrm{Dir}(\theta_{s,a}; \alpha_{s,a}^{(t)}). \tag{B.8}$$

As a result, at each time t, the belief of participants about their environment can be summarized in the set $\alpha^{(t)} = \{\alpha_{s,a}^{(t)}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}\}$. We consider the parameter of the prior belief $\hat{b}^{(1)}$ (i.e., $\alpha^{(1)}$) to be the same for all transitions, i.e.,

$$\alpha^{(1)} = \{\alpha_{s,a}^{(1)}(s') = \epsilon, \quad \forall (s, s', a) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}\}, \tag{B.9}$$

where $\epsilon > 0$ is a free parameter. With this choice of prior, $\hat{\theta}_{s,a}^{(1)}$ (i.e., the prior estimate of the transition probabilities from the pair of state s and action a) is a uniform distribution over states. Furthermore, the free parameter ϵ expresses how deterministic the transitions are from the point of view of a participant, i.e., smaller values of ϵ indicate a more deterministic interpretation of the environment.

Using a Dirichlet distribution for the belief $\hat{b}^{(t)}$ and Equation B.6, a participant's estimation of the transition probabilities is

$$\hat{\theta}_{s,a}^{(t)}(s') = \frac{\alpha_{s,a}^{(t)}(s')}{\sum_{\tilde{s}' \in \mathcal{S}} \alpha_{s,a}^{(t)}(\tilde{s}')}. \tag{B.10}$$

Note that, the pseudo-counts $\tilde{C}_{s,a \to s'}^{(t)}$ in Equation 3.3 of the main text is equal to $\alpha_{s,a}^{(t)}(s') - \epsilon$.

Definitions of surprise. We work with the 'Bayes Factor' surprise S_{BF} (Liakoni et al., 2021). Consider the transition initiated at time t, i.e., $(S_t = s, A_t = a) \to (S_{t+1} = s')$. The Bayes Factor surprise corresponding to this transition is (Liakoni et al., 2021)

$$\mathsf{S}_{\mathrm{BF}}^{(t+1)} = \frac{\hat{\theta}_{s,a}^{(1)}(s')}{\hat{\theta}_{s,a}^{(t)}(s')}. \tag{B.11}$$

[1] If transition probabilities are stationary and have a uniform (or in general any Dirichlet) prior, exact Bayesian inference yields a Dirichlet distribution.

Due to the particular form of the prior $\hat{b}^{(1)}$ that we chose, $\hat{\theta}^{(1)}_{s,a}(s')$ is constant. As a result, the surprise $\mathsf{S}^{(t+1)}_{\text{BF}}$ at time $t+1$ is proportional to the inverse of the estimated probability $\hat{\theta}^{(t)}_{s,a}(s')$ of the transition initiated at time t. In Equation 3.5 of the main text, $\hat{\theta}^{(t)}_{s,a}(s')$ is written as $p^{(t)}(s'|s,a)$, and $\hat{\theta}^{(1)}_{s,a}(s')$ is written as $p_{\text{reset}}(s'|s,a)$.

We note that in the particular case of our behavioral paradigm, the Shannon surprise (Shannon, 1948) is just the shifted logarithm of the 'Bayes Factor' surprise, i.e., $\mathsf{S}^{(t+1)}_{\text{Sh}} = \log \mathsf{S}^{(t+1)}_{\text{BF}} + \log |\mathcal{S}|$. Furthermore, the state prediction error (SPE) (Gläscher et al., 2010) is an increasing function of the 'Bayes Factor' surprise, i.e., $\text{SPE}_{t+1} = 1 - \frac{1}{|\mathcal{S}|\,\mathsf{S}^{(t+1)}_{\text{BF}}}$. Hence, surprise-modulated learning rates in the SurNoR algorithm can alternatively be expressed in terms of $\mathsf{S}^{(t+1)}_{\text{BF}}$ or $\mathsf{S}^{(t+1)}_{\text{Sh}}$ or SPE_{t+1}.

Surprise-modulated update of the belief. Learning the world-model corresponds to updating the parameters of the Dirichlet distribution after each transition. Consider the transition $(S_t = s, A_t = a) \rightarrow (S_{t+1} = s')$ initiated at time t which generates a surprise $\mathsf{S}^{(t+1)}_{\text{BF}}$ at time $t+1$. The surprise-modulated adaptation rate is defined as (Liakoni et al., 2021)

$$\gamma(\mathsf{S}^{(t+1)}_{\text{BF}}, m) = \frac{m\mathsf{S}^{(t+1)}_{\text{BF}}}{1 + m\mathsf{S}^{(t+1)}_{\text{BF}}} \in [0,1], \tag{B.12}$$

where $m > 0$ is a positive free parameter. The parameter m controls the sharpness of the transition.

With this modulated adaptation rate, the change in a participant's belief is given by an update of the Dirichlet parameters $\alpha^{(t+1)}_{\tilde{s},\tilde{a}}(\tilde{s}')$ for all $(\tilde{s}, \tilde{s}', \tilde{a}) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A}$ (Liakoni et al., 2021)

$$\alpha^{(t+1)}_{\tilde{s},\tilde{a}}(\tilde{s}') = \begin{cases} (1-\gamma_{t+1})\alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') + \gamma_{t+1}\alpha^{(1)}(\tilde{s}') + \delta(s',\tilde{s}') & \text{if} \quad \tilde{s} = s, \tilde{a} = a \\ \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') & \text{otherwise} \end{cases}, \tag{B.13}$$

where $\gamma_{t+1} = \gamma(\mathsf{S}^{(t+1)}_{\text{BF}}, m)$. The update rule becomes the same as the one in Equation 3.6 of the main text if we replace $\alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}')$ by $\tilde{C}^{(t)}_{\tilde{s},\tilde{a}\rightarrow\tilde{s}'} + \epsilon$. The update rule expresses the new belief as a mix between two possibilities, represented by the current parameters $\alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}')$ and the prior $\alpha^{(1)}(\tilde{s}')$, weighted with $1-\gamma_{t+1}$ and γ_{t+1}, respectively. In the case of a large surprise, the value of γ_{t+1} is close to one, and as a result, the current parameters are forgotten. The update makes a step based on the currently observed transition, expressed by the Kronecker-δ in the first line. The parameters of transitions from the pairs of the states and actions different form the current one (i.e., s and a) are not changed (second line). The update rule of Equation B.13 is called Variational Surprise Minimizing Learning (VarSMiLe) rule in (Liakoni et al., 2021).

Formalizing model-based Q-values. The world-model of the participants is summarized by their beliefs $\hat{b}^{(t)}(\theta)$ about the transition matrix of the environment. The belief is used for evaluation of two sets of Q-values (Sutton and Barto, 2018), one for novelty N and the other one for the external reward R.

Novelty $\mathsf{N}^{(t)}(s)$ of state s at time t (cf. Equation B.4) is useful to guide behavior during exploration. Analogous to the common framework in reinforcement learning (Sutton and Barto, 2018) where information of a reward at state s' is propagated by the Bellman equation to states $s \neq s'$, we use a Bellman equation to propagate the novelty of state s' to other states $s \neq s'$ by using the model of the world. More specifically, for the model-based branch, we assign to each state-action pair a novelty-based value $Q^{(t)}_{\mathrm{MB,N}}(s,a)$ which is an estimation of the accumulated future discounted novelty that can be gained by taking action a in state s. The Bellman equation is

$$Q^{(t)}_{\mathrm{MB,N}}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s') \Big(\mathsf{N}^{(t)}(s') + \lambda_N \max_{a' \in \mathcal{A}} Q^{(t)}_{\mathrm{MB,N}}(s',a') \Big), \tag{B.14}$$

where $\hat{\theta}^{(t)}_{s,a}(s')$ are the estimated transition probabilities, and $\lambda_N \in [0,1]$ is a discount factor for novelty. The Bellman equation assigns a value to the action a in state s as long as a novel state is likely to be reached within the next few steps - even if the immediately neighboring states are not novel. The discount rate λ_N controls the time horizon of 'future novelty'. For $\lambda_N \to 0$, only the novelty of the immediately following state matters; for $\lambda_N \to 1$, the time horizon becomes infinitely long.

Rewards $R(s)$ of states $s \in \mathcal{S}$ guide behavior during exploitation. In the theory of reinforcement learning, reward information is summarized in values $Q^{(t)}_{\mathrm{MB,R}}(s,a)$ that are estimations of the accumulated future discounted reward that can be collected when starting at state s with action a. The Q-values are given by the Bellman equation

$$Q^{(t)}_{\mathrm{MB,R}}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s') \Big(R(s') + \lambda_R \max_{a' \in \mathcal{A}} Q^{(t)}_{\mathrm{MB,R}}(s',a') \Big), \tag{B.15}$$

where $\lambda_R \in [0,1]$ is the discount factor for reward, which is not necessarily equal to the discount factor for novelty λ_N. Note that in our environment $R(s) = 0$ at all states except at the goal. Since the scale of the reward is arbitrary, we set $R(s_{Goal}) = 1$. As a result, the reward function is $R(s) = \delta(s, s_{Goal})$.

The total model-based Q-value is a linear combination of the Q-values for novelty $Q^{(t)}_{\mathrm{MB,N}}(s,a)$ and reward $Q^{(t)}_{\mathrm{MB,N}}(s,a)$,

$$Q^{(t)}_{\mathrm{MB}}(s,a) = Q^{(t)}_{\mathrm{MB,R}}(s,a) + \beta_N Q^{(t)}_{\mathrm{MB,N}}(s,a), \tag{B.16}$$

where $\beta_N \geq 0$ is a free parameter controlling the trade-off between exploitation and exploration, i.e., between reward-seeking and novelty-seeking behavior.

In our model, β_N depends on whether participants are in the exploration phase or the exploitation phase. This dependency is simplified as follows: Since novelty is the main drive in the 1st episode of the 1st block, we keep β_N fixed at a value β_{N1} throughout this episode. However, at the end of the 1st episode of the 1st block, once participants have found the goal and do not need further exploration, we set $\beta_N = 0$ and keep it at zero for all remaining episodes of the 1st block.

Surprise increases rapidly at the first mismatch that participants face in the 1st episode of the 2nd block, when they encounter an unexpected transition. We hypothesize that the unexpected transitions make them realize that something has changed in the environment and they do not know anymore a path to the goal state and need to re-explore the environment and search for the goal in the absence of any external reward; hence, we assume that the huge surprise signal triggers renewed exploration and we therefore set $\beta_N = \beta_{N2}$ for the 1st episode of the 2nd block. With the same arguments as for the 1st block, we set β_N to zero for the remaining episodes of the 2nd block. β_{N1} and β_{N2} are free parameters of the model.

We also tested a variant of SurNoR with an additional free parameter $\beta_N = \beta_{N-2to5}$ for the weights of $Q^{(t+1)}_{\text{MF,N}}$ and $Q^{(t+1)}_{\text{MB,N}}$ in episodes 2-5 of blocks 1 and 2, but we did not find any significant improvement in the fit (difference in log-evidence $= 15 \pm 13$).

Note that, for model comparison, we use the same assumptions for all other alternative algorithms that either seek novelty or uncertainty - see Section 'Alternative algorithms'.

Updating model-based Q-value. Since solving the non-linear equations B.14 and B.15 for computing two separate sets of model-based Q-values (i.e., $Q^{(t)}_{\text{MB,N}}(s,a)$ and $Q^{(t)}_{\text{MB,R}}(s,a)$ for all $(\tilde{s},\tilde{a}) \in \mathcal{S} \times \mathcal{A}$) is computationally costly, we use a variant (Algorithm 3) of Prioritized Sweeping (Brea, 2017; Sutton and Barto, 2018; Van Seijen and Sutton, 2013).

The idea of the algorithm, for example for updating $Q^{(t)}_{\text{MB,R}}(s,a)$, is to define a set of $|\mathcal{S}|$ mirror variables $U^{(t)}_{\text{R}}(s)$, and rewrite Equation B.15 as

$$\begin{aligned} Q^{(t)}_{\text{MB,R}}(s,a) &= \sum_{s'\in\mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s')\Big(R(s') + \lambda_R U^{(t)}_{\text{R}}(s')\Big) \\ U^{(t)}_{\text{R}}(s') &= \max_{a'\in\mathcal{A}} Q^{(t)}_{\text{MB,R}}(s',a'). \end{aligned} \tag{B.17}$$

At the transition from time step $t-1$ to time step t several iterations take place. The algorithm first puts $U^{(t)}_{\text{R}}(s) = U^{(t-1)}_{\text{R}}(s)$ and updates $Q^{(t)}_{\text{MB,R}}(s,a)$ for all s,a with the current values of $U^{(t)}_{\text{R}}(s)$ using the 1st equation. The size of the update step for the value of a state s' is measured as $\Delta V(s') = \max_{a'\in\mathcal{A}} Q^{(t)}_{\text{MB,R}}(s',a') - U^{(t)}_{\text{R}}(s')$. The states s' are then ordered in a priority queue with the state of biggest $|\Delta V(s')|$ at the top. The algorithm updates the values of $U^{(t)}_{\text{R}}(s')$ of the top priority state using the 2nd equation. This results in further updates $\Delta Q^{(t)}_{\text{MB,R}}(s,a) = \hat{\theta}^{(t)}_{s,a}(s')\lambda_R \Delta V(s')$ for all s,a induced by the first equation. After these updates the priority list is resorted. Updating ends after

T_{PS} iterations where $T_{\text{PS}} \in \mathbb{N}$ is a free parameter of the algorithm.

The values of $Q^{(1)}_{\text{MB,R}}(s,a)$, $U^{(1)}_{\text{R}}(s)$, $Q^{(1)}_{\text{MB,N}}(s,a)$, and $U^{(1)}_{\text{N}}(s)$ are initialized consistent with the Bellman equations under the prior world-model (uniform distribution for all transitions) and the prior reward (zero) and novelty values ($\log|\mathcal{S}|$). For details, see Algorithms 1 and 3.

Algorithm 3 Pseudocode for the modified version of Prioritized Sweeping Algorithm for one time-step at time $t+1$

```
# Specifying whether the update is for the internal or the external reward
Put $\lambda = \lambda_R$ for reward and $\lambda = \lambda_N$ for novelty.
Put $Q^{(t)} = Q^{(t)}_{\text{MB,R}}$, $U^{(t)} = U^{(t)}_R$, and Reward $= R$ for reward, and put $Q^{(t)} = Q^{(t)}_{\text{MB,N}}$,
$U^{(t)} = U^{(t)}_N$, and Reward $= \mathsf{N}^{(t+1)}$ for novelty.
# Applying the effect of the latest observation on $Q$-values using previous $U$-values
for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
    $Q^{(t+1)}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t+1)}_{s,a}(s')\Big(\text{Reward}(s') + \lambda U^{(t)}(s')\Big)$
end for
# Making the priority queue
for $s \in \mathcal{S}$ do
    $U^{(t+1)}(s) = U^{(t)}(s)$
    $\text{Prior}(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$
end for
# Updating $U$-values for $T_{\text{PS}}$ steps
for $T_{\text{PS}}$ iterations do
    $s' = \arg\max_{s \in \mathcal{S}} \text{Prior}(s)$
    $\Delta V = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a) - U^{(t+1)}(s')$
    $U^{(t+1)}(s') = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a)$
    # Applying the effect of the update of $U$-values on $Q$-values
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        $Q^{(t+1)}(s,a) \leftarrow Q^{(t+1)}(s,a) + \lambda \hat{\theta}^{(t+1)}_{s,a}(s') \Delta V$
    end for
    # Updating the priority queue
    for $s \in \mathcal{S}$ do
        $\text{Prior}(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$
    end for
end for
```

SurNoR model-free branch

The pseudocode for the model-free branch is shown in Alg. 4. In this subsection, the details are explained.

Formalizing model-free Q-values. Analogous to the model-based branch, we define $Q^{(t)}_{\text{MF,R}}(s,a)$ and $Q^{(t)}_{\text{MF,N}}(s,a)$ as values of the state-action pairs corresponding to the

Algorithm 4 Pseudocode for the model-free branch of SurNoR

```
# Surprise-modulated learning rate
Compute $\rho_{t+1} = \rho_b + \gamma_{t+1}\delta\rho$.
# Prediction errors
Compute $RPE_{t+1} = R(s_{t+1}) + \lambda_R \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,R}}(s_{t+1}, a') - Q^{(t)}_{\text{MF,R}}(s_t, a_t)$.
Compute $NPE_{t+1} = \mathsf{N}^{(t)}(s_{t+1}) + \lambda_N \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,N}}(s_{t+1}, a') - Q^{(t)}_{\text{MF,N}}(s_t, a_t)$.
# Update of the eligibility traces
Update $e^{(t+1)}_N(s_t, a_t) = 1$, and $e^{(t+1)}_N(s, a) = \lambda_N \mu_N e^{(t)}_N(s, a)$, $\forall s \neq s_t, a \neq a_t$.
Update $e^{(t+1)}_R(s_t, a_t) = 1$, and $e^{(t+1)}_R(s, a) = \lambda_R \mu_R e^{(t)}_R(s, a)$, $\forall s \neq s_t, a \neq a_t$.
# TD-learners
Update $Q^{(t+1)}_{\text{MF,R}}(s, a) = Q^{(t)}_{\text{MF,R}}(s, a) + \rho_{t+1} e^{(t+1)}_R(s, a) RPE_{t+1}$, $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$.
Update $Q^{(t+1)}_{\text{MF,N}}(s, a) = Q^{(t)}_{\text{MF,N}}(s, a) + \rho_{t+1} e^{(t+1)}_N(s, a) NPE_{t+1}$, $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$.
```

external reward R and novelty N, respectively. In contrast to the model-based branch, the model-free Q-values are updated using TD-learning (Sutton and Barto, 2018; Watkins and Dayan, 1992), for which the model of the world is not directly used - see the paragraph 'Updating model-free Q-values'.

Analogous to the total model-based Q-values, we define the total model-free Q-values as

$$Q^{(t)}_{\text{MF}}(s, a) = Q^{(t)}_{\text{MF,R}}(s, a) + \beta_N Q^{(t)}_{\text{MF,N}}(s, a), \tag{B.18}$$

where $\beta_N \geq 0$ has the same value as the one used in Equation B.16.

Reward and novelty prediction errors. A crucial signal in model-free reinforcement learning is the reward prediction error (RPE), defined as the difference between the expected 'reward' of a state-action pair and its real 'reward' (Sutton and Barto, 2018). Since we defined two separate sets of Q-values, one for the external reward and one for novelty (which plays the role of an 'internal reward'), we also define two separate corresponding prediction errors.

Consider the transition $(S_t = s, A_t = a) \to (S_{t+1} = s')$. The RPE at time $t+1$ is defined as

$$RPE_{t+1} = R(s') + \lambda_R \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,R}}(s', a') - Q^{(t)}_{\text{MF,R}}(s, a), \tag{B.19}$$

and similarly, the novelty prediction error (NPE) at time $t+1$ is defined as

$$NPE_{t+1} = \mathsf{N}^{(t)}(s') + \lambda_N \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,N}}(s', a') - Q^{(t)}_{\text{MF,N}}(s, a), \tag{B.20}$$

where λ_R and λ_N are the same discount factors as the ones used in the model-based branch.

Eligibility trace. To keep track of the previously chosen state-action pairs, and to include them in the update rule, we use eligibility traces (Gerstner et al., 2018; Lehmann et al., 2019; Sutton and Barto, 2018). To have the most general setting, we define two separate eligibility traces, one for the external reward $e_R^{(t)}(s,a)$ and one for novelty (the internal reward) $e_N^{(t)}(s,a)$ for all state-action pairs (s,a). We initialize the eligibility traces at zero and reset their values to zero at the beginning of each episode. After the transition $(S_t = s, A_t = a) \to (S_{t+1} = s')$, the eligibility traces are updated to

$$\begin{aligned} e_R^{(t+1)}(s',a') &= \begin{cases} 1 & \text{if } \quad s' = s, a' = a \\ \lambda_R \mu_R e_R^{(t)}(s',a') & \qquad \text{otherwise} \end{cases} \\ e_N^{(t+1)}(s',a') &= \begin{cases} 1 & \text{if } \quad s' = s, a' = a \\ \lambda_N \mu_N e_N^{(t)}(s',a') & \qquad \text{otherwise}\,, \end{cases} \end{aligned} \tag{B.21}$$

where λ_R and λ_N are the discount factors defined above, and $\mu_N \in [0,1]$ and $\mu_R \in [0,1]$ are free parameters expressing how fast eligibility traces decay in time.

Surprise modulation of model-free learning rate. Usual TD learning algorithms use a constant (or decreasing in time) learning rate for updating Q-values (Sutton and Barto, 2018). However, the model-free branch of SurNoR has a learning rate modulated by the model-based branch. This novel interaction between model-based and model-free modules has not been explored by previous hybrid models in neuroscience, e.g., (Daw et al., 2011; Gläscher et al., 2010).

We define the surprise modulated model-free learning rate ρ_t as

$$\rho_t = \rho_b + \gamma(\mathsf{S}_{\text{BF}}^{(t)}, m)\delta\rho, \tag{B.22}$$

where $\gamma(\mathsf{S}_{\text{BF}}^{(t)}, m)$ is the surprise modulated adaptation rate of the model-based branch defined in Equation B.12, $\rho_b \in [0,1]$ is the baseline learning rate (when there is no surprise, i.e., if $\mathsf{S}_{\text{BF}}^{(t)} = 0$), and $\delta\rho \in [0, 1-\rho_b]$ is the maximum possible variation of the learning rate due to the surprise modulation. As a result, the learning rate value ρ_t ranges between ρ_b (when $\mathsf{S}_{\text{BF}}^{(t)} = 0$) and $\rho_b + \delta\rho$ (when $\mathsf{S}_{\text{BF}}^{(t)} \to \infty$).

Updating model-free Q-value. The model-free Q-values for external reward are initialized to zero, $Q_{\text{MF,R}}^{(1)}(s,a) = 0$ for all s, a. This initialization avoids any potential bias towards optimistic initialization (OI). The reason for this choice is to have novelty as the only exploration drive during the 1st episode of the 1st block. We separately test the effect of the initialization of reward-based Q-values in three alternative algorithm which use OI of $Q_{\text{MF,R}}^{(1)}(s,a)$ as a drive for exploration (Sutton and Barto, 2018) - see Section 'Alternative algorithms'. However, to consider the most general case, we initialize the model-free Q-values for novelty at $Q_{\text{MF,N}}^{(1)}(s,a) = Q_{N0}$ with a free parameter $Q_{N0} \geq 0$.

At each time step $t+1$, the model-free Q-values are updated with a TD-learning algorithm

$$\begin{aligned} Q_{\mathrm{MF,R}}^{(t+1)}(s,a) &= Q_{\mathrm{MF,R}}^{(t)}(s,a) + \rho_{t+1} e_R^{(t+1)}(s,a) RPE_{t+1} \\ Q_{\mathrm{MF,N}}^{(t+1)}(s,a) &= Q_{\mathrm{MF,N}}^{(t)}(s,a) + \rho_{t+1} e_N^{(t+1)}(s,a) NPE_{t+1}. \end{aligned} \tag{B.23}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Hybrid policy

The policy for action selection is based on a linear combination of Q-values, similar to Daw et al. (2011); Gläscher et al. (2010). We use a softmax policy (Sutton and Barto, 2018) and consider the probability of choosing action a in state s as

$$\pi(A_t = a|S_t = s) = \frac{1}{Z(s)} \exp\Big\{\beta \Big[\omega \left(\omega_{\text{scale}} Q_{\mathrm{MF}}^{(t)}(s,a)\right) \;+\; (1-\omega)\, Q_{\mathrm{MB}}^{(t)}(s,a)\Big]\Big\}, \tag{B.24}$$

where $Z(s)$ is the normalization constant (that ensures that $\sum_a \pi(A_t = a|S_t = s) = 1$), $\omega_{\text{scale}} \geq 0$ is a free parameter to correct the potentially different scaling of the model-based and model-free values, and $\omega \in [0,1]$ is a free parameter to balance the relative contribution of the model-based and model-free branches on the policy. When $\omega = 1$, the policy is purely model-free (but includes the effect of surprise modulation on the TD-learning learning rate), and when $\omega = 0$, the policy is purely model-based. Note that ω_{MF} and ω_{MB} mentioned in the main texts are equal to $\omega \times \omega_{\text{scale}}$ and $1-\omega$, respectively. The reverse temperature $\beta \geq 0$ controls the sharpness of policy (the larger β the more deterministic is the policy).

As it was shown by Gläscher et al. (2010), ω can vary in time. Specific to our experiment, we consider ω to be piece-wise constant in time: 1. $\omega = \omega_{11}$ for the 1st episode of the 1st block, when participants are in the pure exploration phase, 2. $\omega = \omega_{12}$ for the 1st episode of the 2nd block, when the goal is lost, and 3. $\omega = \omega_0$ for the rest of the experiments (i.e., episodes 2 to 5 for both blocks), when participants are in the exploitation phase. Moreover, we allow the value of β to be different for the 1st and the 2nd block, β_1 and β_2 respectively. By doing so, we allow the model to change its confidence in action selection after observing the sudden change in the environment.

Note that, for model comparison, we use the same assumptions for all other alternative algorithms that use hybrid policy - see Section ‘Alternative algorithms’.

Summary of free parameters

SurNoR has 18 free parameters, summarized as

$$\{\epsilon, m, \lambda_R, \lambda_N, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}, T_{\mathrm{PS}}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho, \omega_{\text{scale}}, \omega_0, \omega_{11}, \omega_{12}\}. \tag{B.25}$$

ϵ is used for initialization of the belief in Equation B.9. m is used for modulation of the adaptation rate in Equation B.12. λ_R and λ_N are discount factors used in the definitions and the updates of Q-values. β_1 and β_2 are the inverse temperatures controlling the sharpness of the hybrid policy in Equation B.24. β_{N1} and β_{N2} are used for balancing novelty against external reward in equations B.16 and B.18. T_{PS} is used for Prioritized Sweeping in Algorithm 3. μ_R and μ_N are used for controlling the decay of eligibility traces in Equation B.21. Q_{N0} is used for initialization of $Q_{\text{MF,N}}$. ρ_b and $\delta\rho$ are used for the baseline learning rate of the model-free branch and its surprise modulation in Equation B.22. ω_{scale} is used for correcting the potential different scaling of the model-based and model-free values in Equation B.24, and ω_0, ω_{11}, and ω_{12} are used for balancing model-free against model-based in the hybrid policy of Equation B.24.

B.1.2 Alternative algorithms

To statistically test the effect of surprise and novelty, we implemented 12 alternative algorithms explained in this section. Their key features are summarized in Table B.1. The modified versions of SurNoR which do not seek novelty but assign negative reward to the most frequent states are explained at the end.

Model-based alternatives. Four out of 12 algorithms are purely model-based. They all use world-model and prioritized sweeping to calculate model-based Q-values. However, they have different approaches for learning the world-model and different strategies for exploration.

(i) MB+S+N: This algorithm has both features of SurNoR in using surprise modulation for model-building and novelty-seeking for exploration, but it does not use a parallel TD-learner. MB+S+N is a reduced version of SurNoR with $\mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{\text{scale}} = \omega_0 = \omega_{11} = \omega_{12} = 0$, which is equivalent to the model-based branch of SurNoR. MB+S+N has 9 free parameters $\{\epsilon, m, \lambda_R, \lambda_N, T_{PS}, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}\}$.

(ii) MB+N: This algorithm is a modified version of MB+S+N; it uses novelty-seeking for exploration, but it does not have surprise modulation for learning the world-model. MB+N uses leaky integration to update the belief parameters (analogous to Equation B.13),

$$\alpha^{(t+1)}_{\tilde{s},\tilde{a}}(\tilde{s}') = \begin{cases} \kappa_{\text{Leak}}\alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') + \delta(s', \tilde{s}') & \text{if} \quad \tilde{s} = s, \tilde{a} = a \\ \alpha^{(t)}_{\tilde{s},\tilde{a}}(\tilde{s}') & \text{otherwise} \end{cases}, \tag{B.26}$$

where $\kappa_{\text{Leak}} \in [0, 1]$ is a constant free parameter. Such a learning rule has been used previously to model human behavior (Maheu et al., 2019; Meyniel et al., 2016; Modirshanechi et al., 2019; Yu and Cohen, 2009). Overall, MB+N has 9 free parameters $\{\epsilon, \kappa_{\text{Leak}}, \lambda_R, \lambda_N, T_{PS}, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}\}$. It cannot be considered as a special case of SurNoR, but it can be implemented in the framework of the SurNoR algorithm by using Equation B.26 instead of Equation B.13 for updating the belief and by putting $m = \mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{\text{scale}} = \omega_{11} = \omega_{12} = \omega_0 = 0$.

(iii) MB+S+U: This algorithm is similar to MB+S+N; it uses surprise modulation for learning the world-model, but it seeks uncertainty instead of novelty for exploration. Following the ideas from Achiam and Sastry (2017); Burda et al. (2019), we define a set of uncertainty-based Q-values, analogous to the SurNoR's novelty-based Q-values (Equation B.14), as

$$Q^{(t)}_{\text{MB,U}}(s,a) = \sum_{s' \in \mathcal{S}} \hat{\theta}^{(t)}_{s,a}(s')\Big(- \log \hat{\theta}^{(t)}_{s,a}(s') + \lambda_U \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MB,U}}(s',a')\Big), \tag{B.27}$$

where $-\log \hat{\theta}^{(t)}_{s,a}(s')$, sometimes called surprisal (equal to Shannon surprise) is considered as the intrinsic reward of the transition $(s,a) \to s'$. The model MB+S+U is implemented by modifying SurNoR in 3 steps: 1. Replacing Equation B.14 by Equation B.27 and using $Q^{(t)}_{\text{MB,U}}(s,a)$ instead of $Q^{(t)}_{\text{MB,N}}(s,a)$ in all equations. 2. Replacing λ_N by λ_U, β_{N1} by β_{U1}, and β_{N2} by β_{U2}. 3. Putting $\mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{\text{scale}} = \omega_0 = \omega_{11} = \omega_{12} = 0$. MB+S+U has 9 free parameters $\{\epsilon, m, \lambda_R, \lambda_U, T_{PS}, \beta_1, \beta_2, \beta_{U1}, \beta_{U2}\}$.

(iv) MB+S+OI: This algorithm removes the novelty-seeking block of MB+S+N and uses optimistic initialization for exploration, i.e., it updates reward-based Q-values also in the 1st episode of block 1 even before observing the goal states. MB+S+OI can be implemented by modifying SurNoR in 2 steps: 1. Removing the 'if' condition in the lines 7-10 of Alg. 2 and keeping only line 10. 2. Putting $\lambda_N = \beta_{N1} = \beta_{N2} = \mu_R = \mu_N = Q_{N0} = \rho_b = \delta\rho = \omega_{\text{scale}} = \omega_0 = \omega_{11} = \omega_{12} = 0$. This algorithm has 6 free parameters $\{\epsilon, m, \lambda_R, T_{PS}, \beta_1, \beta_2\}$.

Model-free alternatives. Four out of 12 algorithms are model-free. All of them use a TD-learner for learning model-free Q-values. However, the ones with surprise-modulation are also equipped with a world-model, but the world model is not used for computing a set of model-based Q-values, and the policy is not hybrid.

(v) MF+S+N: This algorithm is equivalent to the model-free branch of SurNoR, but it also uses the world-model in the model-based branch for surprise-computation. MF+S+N can be seen as a reduced version of SurNoR by putting $T_{PS} = 0$ and $\omega_{\text{scale}} = \omega_{11} = \omega_{12} = \omega_0 = 1$. It has 13 free parameters $\{\epsilon, m, \lambda_R, \lambda_N, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}, \mu_R, \mu_N, Q_{N0}, \rho_b, \delta\rho\}$.

(vi) MF+N: This algorithm is a reduced version of SurNoR by putting $m = T_{PS} = \delta\rho = 0$ and $\omega_{\text{scale}} = \omega_{11} = \omega_{12} = \omega_0 = \epsilon = 1$, which is equivalent to the model-free branch of SurNoR without any surprise modulation. It can also be seen as a modified version of the famous $Q(\lambda)$ algorithm (Sutton and Barto, 2018) (with $\lambda = \mu_R$ in our notation) but with novelty as an exploration bonus (instead of optimistic initialization). The model MF+N has overall 10 free parameters $\{\lambda_R, \lambda_N, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}, \mu_R, \mu_N, Q_{N0}, \rho_b\}$.

(vii) MF+S+U: The relation between MF+S+U and MB+S+U is the same as the relation between MF+S+N and MB+S+N. All the features of MF+S+U (including the surprise modulation of the learning rate of the model-free) except for its exploration strategy are the same as the ones of MF+S+N. For exploration, MF+S+U seeks uncertainty instead of novelty. Similar to what we did for MB+S+U, we followed the ideas from Achiam and Sastry (2017); Burda et al. (2019) and defined the uncertainty-based Q-values as in Equation B.27. Then, we define the Uncertainty Prediction Error (UPE), analogous to the SurNoR's NPE (Equation B.20), as

$$UPE_{t+1} = -\log \hat{\theta}^{(t)}_{s,a}(s') + \lambda_U \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,U}}(s', a') - Q^{(t)}_{\text{MF,U}}(s, a), \tag{B.28}$$

and then we update the uncertainty-based model-free Q-values as

$$Q^{(t+1)}_{\text{MF,U}}(s, a) = Q^{(t)}_{\text{MF,U}}(s, a) + \rho_{t+1} e^{(t+1)}_U (s, a) UPE_{t+1}, \tag{B.29}$$

where $e^{(t+1)}_U(s, a)$ is the uncertainty eligibility trace with a decay factor μ_U. Then MF+S+U can be implemented by modifying SurNoR in three steps: 1. Replacing $Q^{(t)}_{\text{MF,N}}(s, a)$ by $Q^{(t)}_{\text{MF,U}}(s, a)$ in all equations. 2. Replacing λ_N by λ_U, β_{N1} by β_{U1}, β_{N2} by β_{U2}, and μ_N by μ_U. 3. Putting $T_{PS} = 0$ and $\omega_{\text{scale}} = \omega_{11} = \omega_{12} = \omega_0 = 1$. The model MF+S+U has 13 free parameters $\{\epsilon, m, \lambda_R, \lambda_U, \beta_1, \beta_2, \beta_{U1}, \beta_{U2}, \mu_R, \mu_U, Q_{U0}, \rho_b, \delta\rho\}$.

(iix) MF+OI: This algorithm is our simplest algorithm, and neither surprise nor novelty is used in it. MF+OI is equivalent to $Q(\lambda)$(Sutton and Barto, 2018), with $\lambda = \mu_R$ in our notation. It uses optimistic initialization for exploration by putting $Q^{(0)}_{MF,R} = Q_{R0}$, where Q_{R0} is a free parameter. It can be seen as a modified version of SurNoR by initializing $Q^{(0)}_{MF,R} = Q_{R0}$ and putting $m = \lambda_N = \beta_{N1} = \beta_{N2} = T_{PS} = \mu_N = Q_{N0} = \delta\rho = 0$ and $\omega_{\text{scale}} = \omega_{11} = \omega_{12} = \omega_0 = \epsilon = 1$. The model MF+OI has overall 6 free parameters $\{\lambda_R, Q_{R0}, \rho_b, \beta_1, \beta_2, \mu_R\}$.

Hybrid alternatives. Three out of 12 algorithms are hybrid, meaning they use both model-free and model-based Q-values for decision-making.

(ix) Hyb+N: This algorithm uses MB+N and MF+N in parallel and combines their Q-values (in the same fashion as in SurNoR) in a hybrid policy. It has overall 17 free parameters $\{\epsilon, \kappa_{\text{Leak}}, \lambda_R, \lambda_N, \beta_1, \beta_2, \beta_{N1}, \beta_{N2}, T_{PS}, \mu_R, \mu_N, Q_{N0}, \rho_b, \omega_{\text{scale}}, \omega_{11}, \omega_{12}, \omega_0\}$.

(x) Hyb+S+U: This algorithm uses MB+S+U and MF+S+U in parallel and combines their Q-values (in the same fashion as in SurNoR) in a hybrid policy. Hyb+S+U is as complex as SurNoR and has overall 18 free parameters $\{\epsilon, m, \lambda_R, \lambda_U, \beta_1, \beta_2, \beta_{U1}, \beta_{U2}, T_{PS}, \mu_R, \mu_U, Q_{U0}, \rho_b, \delta\rho, \omega_{\text{scale}}, \omega_{11}, \omega_{12}, \omega_0\}$.

(xi) Hyb+S+OI: This algorithm uses MF+OI (but with surprise modulation of the learning rate of the model-free branch) and MB+S+OI in parallel and combines their Q-values (in the same fashion as in SurNoR) in a hybrid policy. Hyb+S+OI has overall

14 free parameters $\{\epsilon, m, \lambda_R, \beta_1, \beta_2, T_{PS}, \mu_R, Q_{R0}, \rho_b, \delta\rho, \omega_{\text{scale}}, \omega_{11}, \omega_{12}, \omega_0\}$.

Table B.1: **Summary of the key features of all models.**

	Algorithm	World-model	Hybrid-policy	Novelty	Surprise	Param.
i	MB+S+N	✓	✗	✓	✓	9
ii	MB+N	✓	✗	✓	✗	9
iii	MB+S+U	✓	✗	✗	✓	9
iv	MB+S+OI	✓	✗	✗	✓	6
v	MF+S+N	✓	✗	✓	✓	13
vi	MF+N	✗	✗	✓	✗	10
vii	MF+S+U	✓	✗	✗	✓	13
iix	MF+OI	✗	✗	✗	✗	6
ix	Hyb+N	✓	✓	✓	✗	17
x	Hyb+S+U	✓	✓	✗	✓	18
xi	Hyb+S+OI	✓	✓	✗	✓	14
xii	RC	✗	✗	✗	✗	0
xiii	BinaryNovelty	✓	✓	(✓)	✓	19
xiv	**SurNoR**	✓	✓	✓	✓	18

Null model.

(xii) RC (Random Choice): According to this algorithm, participants choose actions with uniform distribution, i.e., each action is selected with a probability equal to $\frac{1}{|\mathcal{A}|} = 0.25$. We used this model as a reference to quantify the effect of our novelty-seeking exploration in the 1st episode of the 1st block. This algorithm does not have any free parameter.

Control modifications of SurNoR. (xiii) Binary Novelty: The two control algorithms mentioned in the main text are exactly the same as SurNoR except for a change in the intrinsic motivation signal that drives exploration. While in the SurNoR algorithm the continuous-valued novelty signal defined in Equation B.3 and Equation B.4 serves as the intrinsic reward, in the control algorithms the intrinsic reward of state s at time t is binary: in the first control algorithm it is considered to be -1 if the count $C_s^{(t)} \geq C_{\text{thr}}$ and 0 otherwise, where C_{thr} is a new free parameter, i.e., the algorithm considers the states that are encountered more than C_{thr} times as bad states and assigns a constant negative reward to them. Similarly, in the 2nd control algorithm, the intrinsic reward of state s at time t is considered to be -1 if state s is among the n most frequently encountered states and 0 otherwise, where n is a new free parameter, i.e., the algorithm considers the n most frequently encountered states as bad states. Therefore, the pseudo-code of the control algorithms is the same as the pseudo-code of SurNoR in Alg. 1 but with 2 modifications: (i) $U_N^{(1)}(s)$ is initialized at a value 0. (ii) The definition of novelty is changed to

$$\mathsf{N}^{(t)}(s) = \begin{cases} -1 & \text{if} \quad C_s^{(t)} \geq C_{\text{thr}} \\ 0 & \quad\quad \text{otherwise} \end{cases} \tag{B.30}$$

for the 1st algorithm and to

$$\mathsf{N}^{(t)}(s) = \begin{cases} -1 \quad \text{if} \quad C_s^{(t)} \in n\,\text{highest counts} \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{B.31}$$

for the 2nd algorithm. Overall, both algorithms have 19 free parameters, i.e., 18 free parameters of SurNoR plus C_{thr} for the 1st and n for the 2nd control algorithm.

B.2 EEG preprocessing and control analyses

B.2.1 PCA over Reward and RPE

We show that PCA on the two normalized variables Reward and RPE yields R_+ and R_-.

Lemma. For two random variables X_1 and X_2 with zero mean (i.e., $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$), unit variance (i.e., $\mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = 1$), and correlation $r = \mathbb{E}(X_1 X_2)$, the new variables $X_+ = (X_1 + X_2)/\sqrt{2}$ and $X_- = (X_1 - X_2)/\sqrt{2}$ are the projections on the two normalized principal components of the correlation matrix.

Proof: The 2×2 correlation matrix C has diagonal elements $c_{11} = c_{22} = 1$ because of the normalization of each variable to unit variance and off-diagonal elements $c_{21} = c_{12} = r$ according to the assumption and symmetry of correlations. The normalized eigenvectors of the correlation matrix are then $e_+ = (1,1)^T/\sqrt{2}$ and $e_- = (1,-1)^T/\sqrt{2}$ with eigenvalues $\lambda_\pm = 1 \pm r$. ■

Therefore, if Reward and RPE are normalized, then $R_+ =$ Reward+RPE and $R_- =$ Reward−RPE are their principal components. Furthermore, for positive correlations $r > 0$, the first principal component is R_+.

B.2.2 Correlation and orthogonalization for EEG analysis

Novelty is nearly decorrelated from Surprise and NPE, but Reward and RPE are highly correlated with each other and also correlated with Novelty and NPE (Figure B.8A). PCA on the variables Reward and RPE yields new decorrelated variables R_+ and R_- (Section B.2.1). After projecting Surprise, Novelty, and NPE on the space orthogonal to R_+ and R_- (Figure B.8B), all five variables are decorrelated (Figure B.8C, left part). Nevertheless, the new variables Surprise$_\perp$, Novelty$_\perp$, and NPE$_\perp$ remain very similar to the original variables Surprise, Novelty, and NPE as indicated by correlations equal to or above 0.89 (Figure B.8C, right part).

The regression analysis without PCA and without orthogonalization (Figure B.9) yields results quite similar to the one with PCA and with orthogonalization (Figure 3.10 in the main text); the main difference is that the significant positive correlation of a reward-related variable (R_+) with the EEG amplitude in the 3rd time-window disappears. The reason is the very high correlation between Reward and RPE (Figure B.8A), which leads to imprecise estimation of regression coefficients (compare Figure B.9B and Figure B.9C with Figure 3.10B and Figure 3.10C in the main text). Please note that since the regressors in Figure B.9 are a linear combination of the regressors in Figure 3.10, the adjusted R-squared is the same for both analyses (compare Figure B.9A with Figure 3.10A in the main text).

B.3 Fitted parameters and parameter-recovery

B.3.1 Fitted Parameters

The optimal parameters after fitting the SurNoR model to behavior are summarized in Table B.2 (corresponding to light green lines in Figure B.6). The reported error for each parameter is the maximum of its standard deviation approximated by Laplace approximation (MacKay, 2003) and the optimization precision. Laplace approximation was done for each dimension separately, i.e. the covariance matrix was assumed to be diagonal to avoid the problems arising from approximation of the full Hessian matrix in high-dimensional spaces. Therefore, the reported errors can be seen as lower bounds for the real errors.

B.3.2 Simulated data, model recovery, and parameter recovery

The summary of the analyses of the data of two sets of 12 simulated participants (different from the one shown in Figure 3.7 in the main text) is shown in Figure B.4 and Figure B.5. While we observe some variabilities in data generated by different random seeds, the main aspects of participants' behavior are captured by the model (Figure 3.7 in the main text, Figure B.4, and Figure B.5). The true model was successfully recovered in all 3 different cases (Figure 3.8 in the main text).

To check whether parameters of SurNoR are recoverable in our experimental paradigm, we fitted SurNoR to the three sets of 12 simulated participants (corresponding to the data shown in Figure 3.7 in the main text, Figure B.4, and Figure B.5). The recovered parameters are shown in the log-likelihood landscape in Figure B.6. The true parameters were successfully recovered with reasonable errors given the measured curvature of the log-likelihood function.

B.3.3 Robustness of model-variables in EEG analysis

In the EEG analysis in Figure 3.10 of the main text, we used model variables (Surprise, Novelty, NPE, Reward, and RPE) that were calculated from the SurNoR algorithm

Table B.2: **SurNoR parameters fitted to the behavioral data of all participants.** This set of parameters was used for EEG analysis and illustrations in Figure 3.6 in the main text. The values correspond to the light green lines in Figure B.6.

Param.	Value	Error
ϵ	2e−4	2e−4
m	0.31	0.02
λ_R	0.97	0.01
λ_N	0.70	0.05
β_1	4.7	0.2
β_2	1.50	0.06
β_{N1}	0.145	0.006
β_{N2}	0.220	0.015
T_{PS}	10	1
μ_R	0.94	0.01
μ_N	0.80	0.05
Q_{N0}	0.50	0.35
ρ_b	0.06	0.02
$\delta\rho$	0.45	0.05
ω_{scale}	5.8	0.2
ω_0	0.75	0.05
ω_{11}	0.20	0.05
ω_{12}	0.15	0.05

with one specific parameter choice. We wanted to check the robustness of these model variables with respect to changes in the parameters of SurNoR. Novelty and Reward are by definition independent of the SurNoR parameters. To check the robustness of the other variables, we used the 3 sets of recovered parameters (Figure B.6) and recalculated, for each of our 12 participants, the time course of Surprise, NPE, and RPE for each of the three recovered parameter sets. The model variables extracted given the recovered parameters were extremely highly correlated (> 0.97) with the model-variables extracted given the fitted parameters (Figure B.7).

B.4 The analysis of random exploration

For any stationary policy (e.g., random choice), the sequence of states $\{S_1, S_2, ...\}$ forms a stationary Markov chain. Let us define the random variable $\mathcal{T}$ as the time of the 1st encounter of the goal state, i.e., $\mathcal{T}$ is the length of the 1st episode. We connect the expected number τ_i of actions to find the goal starting from state $S_1 = i$ (with $i \in \{1, ..., 10\}$) to the expected number of actions τ_j in the possible *next* states $S_2 = j$,

$$\tau_i = \mathbb{E}[\mathcal{T}|S_1 = i] = 1 + \sum_{j=1}^{10} p_{ij}\tau_j, \tag{B.32}$$

where p_{ij} is the probability of transitioning from state i to state j (dependent on the stationary policy), and we have already exploited that the goal state does not contribute in the sum because τ_G is by definition zero.

For a random policy (0.25 probability for each of the four actions) and the layout of the environment in Figure 3.1 in the main text, we find $\tau_{\text{trap}} := \tau_8 = \tau_9 = \tau_{10} = \tau_1 + 4$, because it takes on average 4 actions to leave the trap states. Similarly, from state 7, you have a probability of 1/4 to reach the goal in one step, but you can also remain in state 7 or go to one of the trap states. Evaluating all possibilities we arrive at

$$\begin{aligned}
\tau_{\text{trap}} &= \tau_1 + 4, \\
\tau_{i+1} &= 3\tau_i - 2\tau_{\text{trap}} - 4 \text{ for } i \in \{1, ..., 6\}, \\
\tau_7 &= \frac{4}{3} + \frac{2}{3}\tau_{\text{trap}}.
\end{aligned} \tag{B.33}$$

By solving this set of linear equations, we find

$$\begin{aligned}
&\tau_1 = 13116 \ , \ \tau_2 = 13104 \ , \ \tau_3 = 13068 \ , \ \tau_4 = 12960 \\
&\tau_5 = 12636 \ , \ \tau_6 = 11664 \ , \ \tau_7 = 8748 \\
&\tau_8 = \tau_9 = \tau_{10} = \tau_{\text{trap}} = 13120.
\end{aligned} \tag{B.34}$$

The results of calculation show that, starting from state 6 (which is the starting state of the first episode in our experiments), it takes on average more than 10000 actions to find the goal with a random policy.

B.5 Precise statement of the prediction in ‘Discussion’

Consider an extended version of our environment in Figure B.1 which includes a new (and not necessarily finite) set of states (i.e., the purple states in Figure B.1) that can be accessed from state 4 in the middle of the direct path to the goal. Assume that a participant has found the goal state G at the end of the first episode. In episodes 2 to 5 two different situations may arise. (i) If participants believe that the yellow goal in Figure B.1 is the only (or the most) rewarding state in the environment, then they should ideally stop exploration as soon as they have found the goal and go straight to the goal in subsequent episodes. (ii) If participants wonder whether there may exist another state with a higher value of reward than state G, then they will spend a large amount of time in novelty-rich states like the purple states in Figure B.1. Our prediction, based on the SurNoR model presented in the main text, is that both situations can be observed in the behavioral data and that the difference depends on the prior knowledge given to the participant about the environment before the start of the experiment.

The environment of Figure B.1 also provides a critical test for alternative algorithms of SurNoR. Importantly, in the SurNoR model, information on novelty and external reward

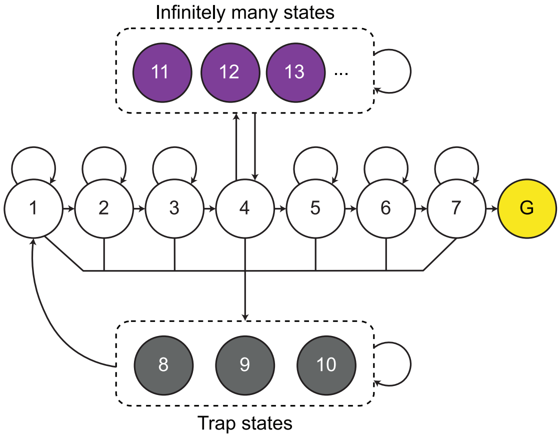


Figure B.1: **An example of the extended version of our environment mentioned in the 'Discussion' section of the main text.** The existence of a set novelty-rich states may distract participants from exploiting the reward at the goal in the episodes after the 1st episode.

are summarized in two separate set of Q-values. Consider an alternative model where the novelty is treated as an internal reward and is added to the external reward in a *single* set of Q-values. This is equivalent to adding Q-values of novelty and reward with a *fixed* factor β_N (see section B.1).

Since the novelty of the purple states is constantly increasing, no matter the values of β and ω in section B.1, any fixed and non-zero value of β_N (see section B.1) will eventually drive the agent back towards novelty-seeking and hence exploration of the purple states. This statement holds for non-deterministic model-free, model-based, or hybrid models.

A straightforward way to avoid being distracted from exploiting the external reward is to stop seeking novelty after finding the goal for the first time. This is done in the SurNoR algorithm by reducing β_N, i.e., by reducing the relative importance of novelty.

B.6 Qualitative differences between model-based (MB) and model-free (MF) branches of SurNoR

The difference in log-evidence of SurNoR and MF+N+S (Figure 3.5A in the main text) shows that both branches, MF and MB, are need for explaining behavior, although the participants' decisions are dominated by the MF action choices (Figure 3.6A4 in the main text). We wanted to find out at which point in the experiment the policy of MB is different from MF.

To do so, we focus on the 1st episode of block 1 and analyze the MB and MF branches of SurNoR with its parameters fitted to behavior (section B.3). We analyze below two different situations where MB and MF have different preferences and show that in both cases a hybrid policy explains the behavior better than either MF or MB separately. Note that our analyses are based on the specific set of fitted parameters.

After the 1st failure

Consider an agent and assume that during the first visit of state 1 it has chosen one of the bad actions. We wondered which action the agent chooses the next time it visits state 1: Does it repeat its last action or choose another action? We, therefore, analyzed the behavior of participants, SurNoR, the MB branch of SurNoR, and the MF branch of SurNoR in this situation for state 1, 2, and 3 (Figure B.2A). In all three states, while the MB branch of SurNoR favors changing the action (and exploring the ones not chosen before), the MF branch of SurNoR favors repeating the same action as chosen the first time. SurNoR combines the two and prefers changing the action, which is consistent with the behavior of participants.

The reason that the MB branch prefers a new action is that it is aware of the existence of many other states that are more novel than any of the trap states, and hence, it assigns a larger novelty value $Q_{\text{MB,N}}$ to unexplored actions than the previously chosen one. On the other hand, because the initial value Q_{N0} of MF Q-values is very small (section B.3 and Figure B.6), the first encounter of a trap state increases the MF novelty value $Q_{\text{MF,N}}$ of the chosen state-action pair. Taken together, the MB branch literally plans to find more novel states while the MF branch follows what have been internally rewarding before. Hence, the MB branch is important to correctly guide action choices after the first failure.

After the nth success

Consider an agent and assume that it has chosen n times the good action in state 1. We wondered which action the agent chooses the next time it visits state 1: Does it further repeat the good action or choose another action? We, therefore, analyzed the behavior of participants, SurNoR, the MB branch of SurNoR, and the MF branch of SurNoR in this situation for n equal to 2, 6, and 9 (Figure B.2B). As n increases, the MF branch of SurNoR increases its preference for staying with the good action, while the MB branch of SurNoR still favors exploring the environment (although with less confidence compared to the case discussed above, Figure B.2A). SurNoR combines the two, and, as n increases, gets closer to the MF preferences and to the participants' behavior.

In summary, in the 1st situation, participants are more MB, while in the 2nd situation, they are more MF. The take-home message is that the hybrid policy flexibly combines the two and explains the participants' behavior in both situations.

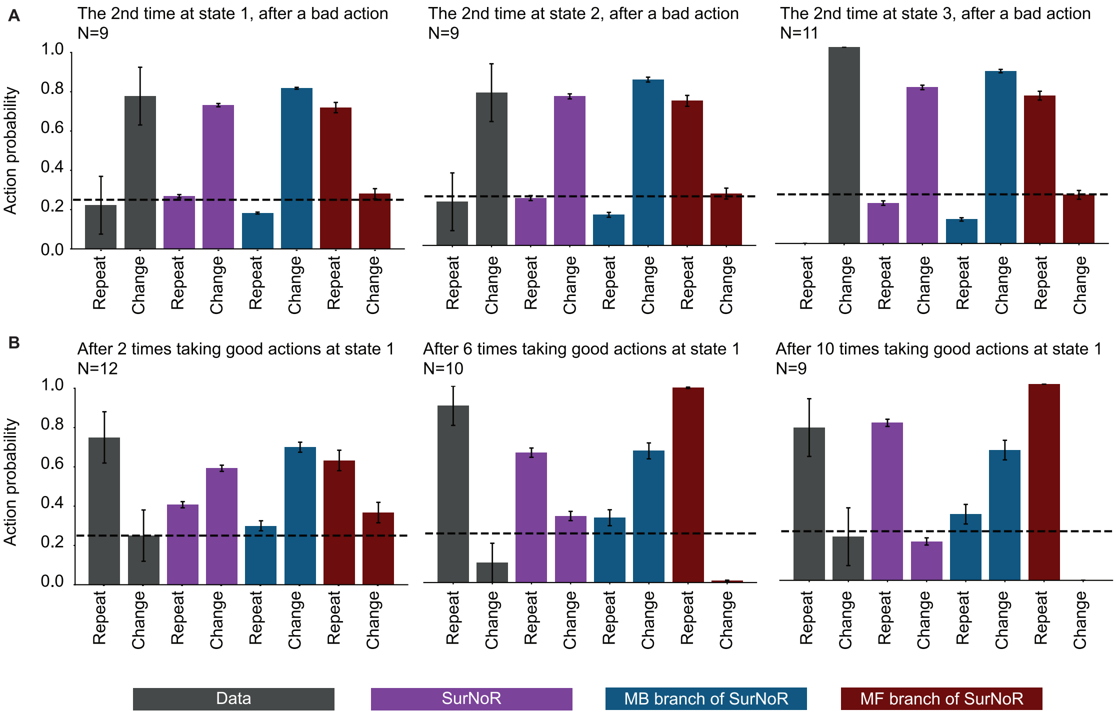


Figure B.2: **The MB and MF branches of SurNoR have different preferences in different situations.** Participants behave more MB in some situations and more MF in the others, and a hybrid policy can capture their behavior in both situations. **A.** Average probability of repeating the previous action or changing to another one in states 1, 2, and 3, with the condition that the 1st time the agent took the bad action. Error bars show the standard error of the mean. **B.** Average probability of repeating the good action or changing to another one after 2, 6, and 9 times taking the good action in state 1. Error bars show the standard error of the mean.

B.7 Supplementary Figures

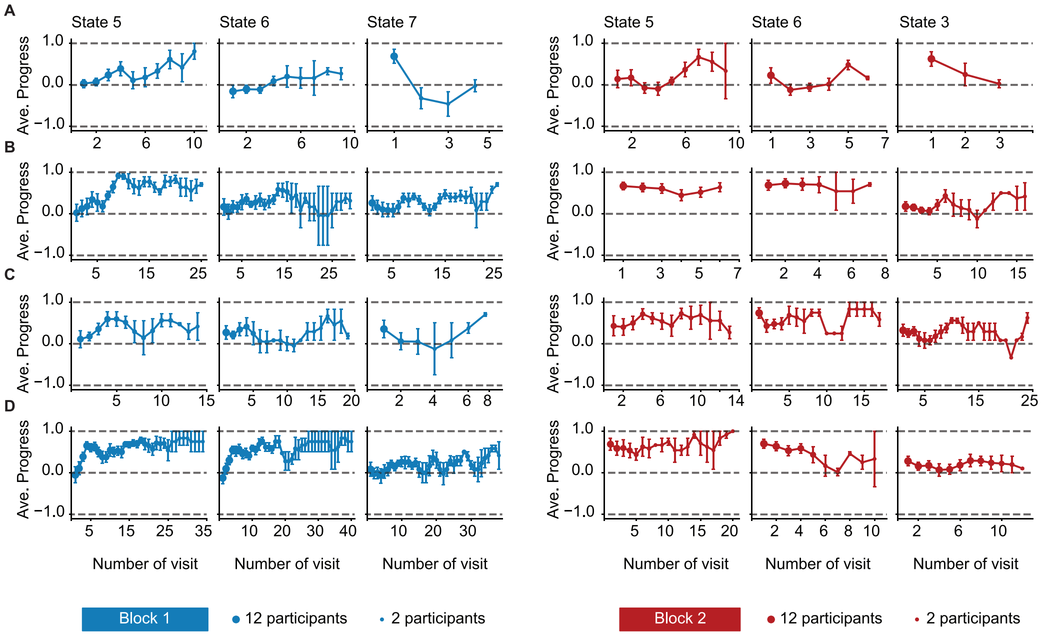


Figure B.3: **Average progress at states in proximity of the goal (complement to Figure 3.2 and Figure 3.7 of the main text).** Average progress of participants (**A**) and simulated participants (**B-D**) each time visiting states 5, 6, and 7 in episode 1 of block 1 (blue) or states 5, 6, and 3 in episode 1 of block 2 (red). Panel **A** corresponds to the experimental data shown in Figure 3.2, panel B to the simulated data shown in Figure 3.7, panel **C** to the simulated data shown in Figure B.4, and panel **D** to the simulated data shown in Figure B.5. While states 1 and 2 are visited by most participants at least 15 times (c.f. Figure 3.2, main text), only very few participants visit the states close to the goal more than 5 times, and as a result, total learning progress between the start and the end of the episode is smaller and data is noisy. Similar observations can be made for the simulated participants (cf. Figure 3.2, main text; size of circles indicates number of participants). Since the number of visits of states close to the goal is small, the noise-induced differences between different simulations runs are large (compare the runs in **B**, **C**, and **D**). In particular, at the state before the goal state (state 7 in block 1 and state 3 in block 2), the first time that (simulated) participants choose the good action, they reach the goal state and finish the episode. Therefore, the average progress for the state before the goal is always calculated across those (simulated) participants who did not choose the good action when they visited that state previously.

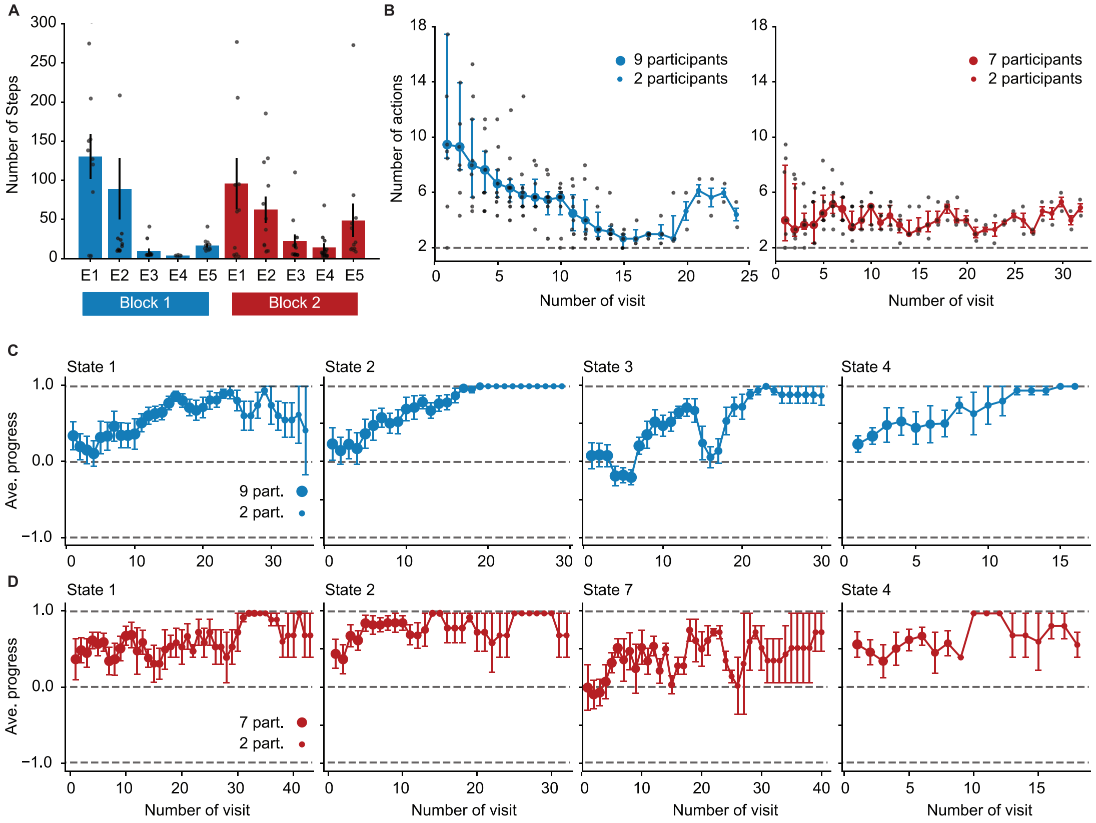


Figure B.4: **Posterior predictive checks;** the structure of figure is the same as Figure 3.7. and their only difference is in the random seed used for generating data. **A.** Average number of actions of 12 simulated participants at each episode (c.f. Figure 3.1C). **B.** Median number of actions of simulated participants to escape the trap states at each of their visits in episode 1 of block 1 (left) and block 2 (right) (c.f. Figure 3.2A) **C.** Average progress of participants each time visiting states 1, 2, 3, and 4 in episode 1 of block 1. (c.f. Figure 3.2B). **D.** Average progress of simulated participants each time visiting states 1, 2, 7 (swapped with 3), and 4 in episode 1 of block 2. (c.f. Figure 3.2C). See Figure B.3C for the average progress at the progressing states in the proximity of the goal.

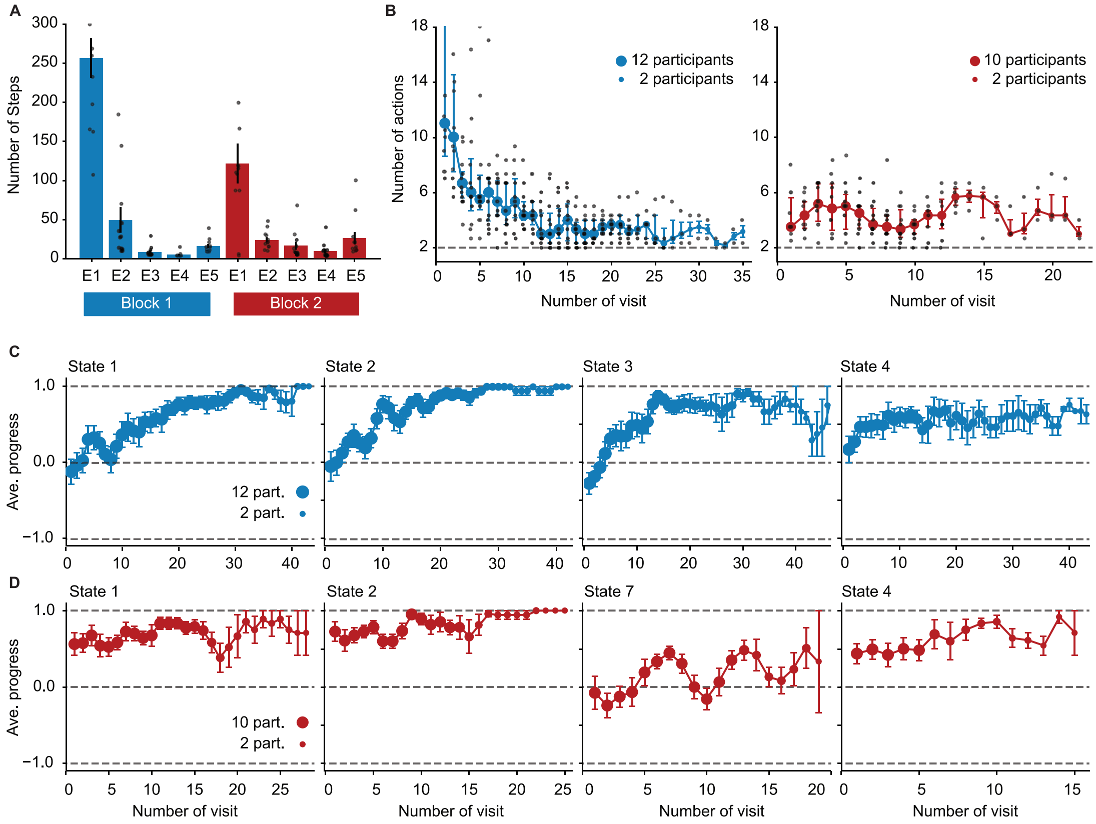


Figure B.5: **Posterior predictive checks;** the structure of figure is the same as Figure 3.7, and their only difference is in the random seed used for generating data. **A.** Average number of actions of 12 simulated participants at each episode (c.f. Figure 3.1C). **B.** Median number of actions of simulated participants to escape the trap states at each of their visits in episode 1 of block 1 (left) and block 2 (right) (c.f. Figure 3.2A) **C.** Average progress of participants each time visiting states 1, 2, 3, and 4 in episode 1 of block 1. (c.f. Figure 3.2B). **D.** Average progress of simulated participants each time visiting states 1, 2, 7 (swapped with 3), and 4 in episode 1 of block 2. (c.f. Figure 3.2C). See Figure B.3D for the average progress at the progressing states in the proximity of the goal.

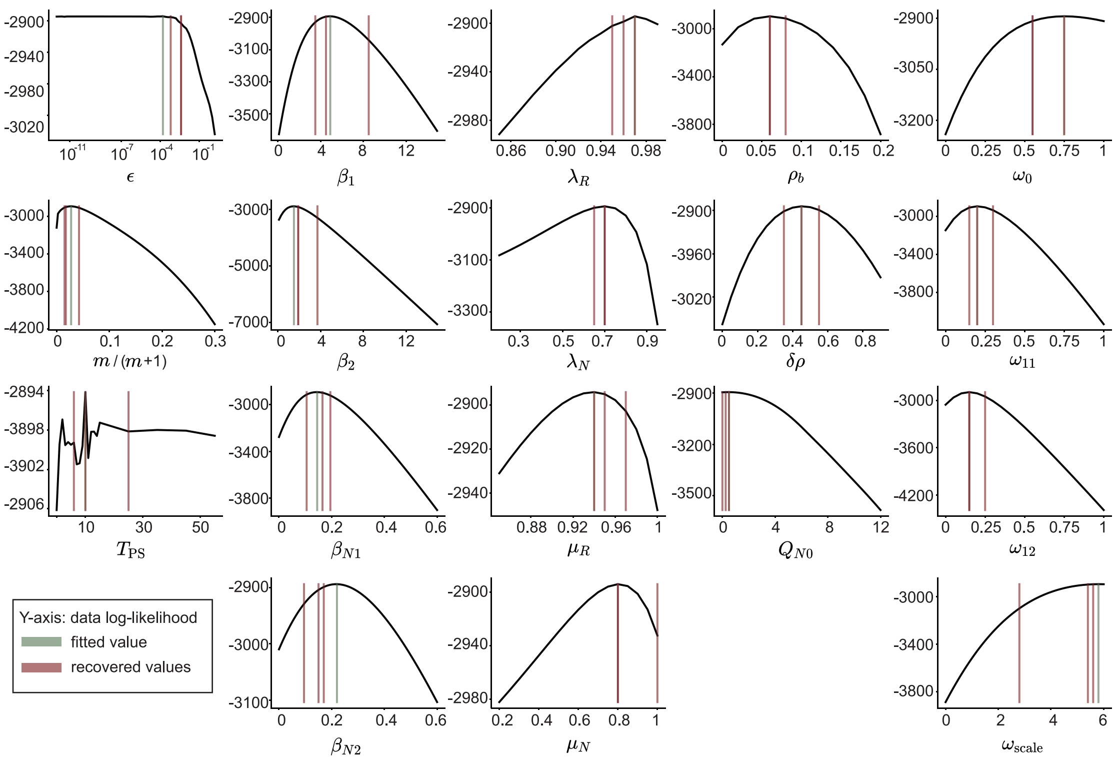


Figure B.6: **Parameter recovery results and log-likelihood landscape.** The solid black curve in each panel shows the log-likelihood of the behavioral data of 12 participants as a function of one of the free parameters of SurNoR, while the other parameters are fixed at their fitted values (section B.3). The fitted value for each parameter corresponds to the peak of the log-likelihood function and is specified by the light green lines. The recovered parameters for 3 different sets of 12 simulated participants (corresponding to the data shown in Figure 3.7 in the main text, Figure B.4, and Figure B.5) are shown by the light red lines. Note that the procedure of fitting parameters to the generated data were exactly the same as the procedure of fitting parameters to the real data, i.e., we searched in the 18-dimensional space of parameters. The 1st column corresponds to parameters mainly related to model-building and model-based planning (ϵ, m, and T_{PS}); the 2nd column corresponds to the softmax policy temperatures and the parameters controlling the exploration and exploitation trade-off (β_1, β_2, β_{N1}, and β_{N2}); the 3rd column corresponds to the discount factors and the decay rates of eligibility traces (λ_R, λ_N, μ_R, and μ_N); the 4th column corresponds to parameters mainly related to model-free learning (ρ_b, $\delta\rho$, and Q_{N0}); and the last column corresponds to the parameters controlling the trade-off between model-based and model-free policies (ω_0, ω_{11}, ω_{12}, and ω_{scale}). See section B.3 for details.

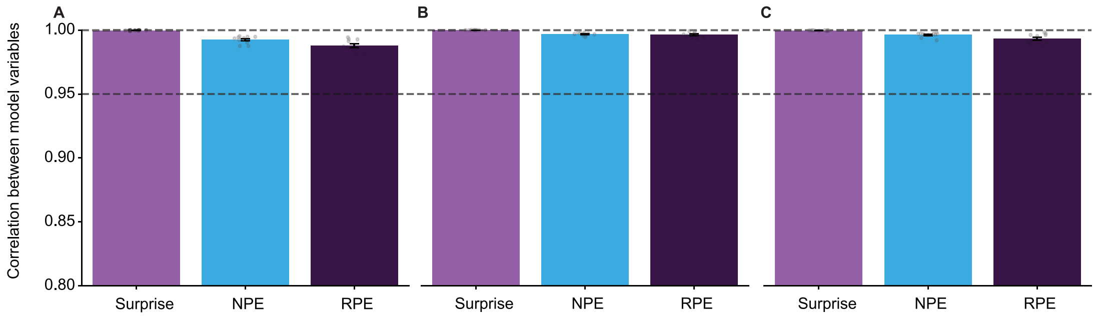


Figure B.7: **Robustness of model-variables.** Average correlation between the model-variables extracted given the fitted parameters (the ones used for EEG analyses) and model-variables extracted given the recovered parameters corresponding to **A.** Figure 3.7 in the main text, **B.** Figure B.4, and **C.** Figure B.5. Error bars show the standard error of the mean, and each grey point shows data of one participant. See section B.3 for details.

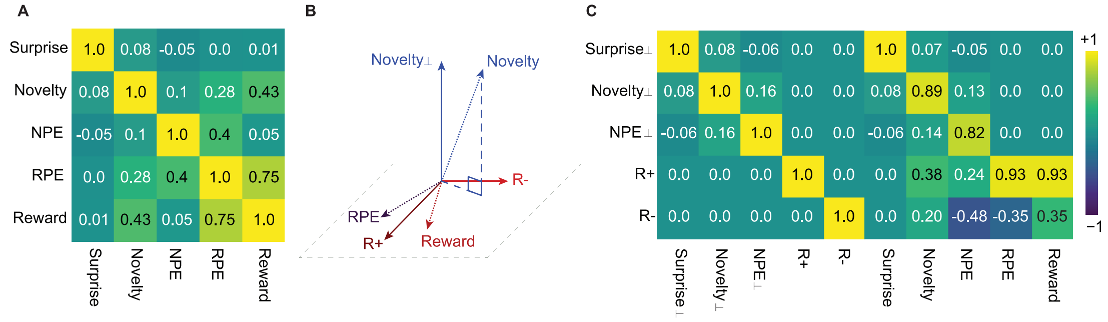


Figure B.8: **Correlations (averaged over participants) between relevant variables. A.** The cross-correlations between Surprise, Novelty, NPE, RPE, and Reward during the behavioral task. **B.** Novelty$_\perp$ is the projection of Novelty onto the subspace orthogonal to the plane spanned by Reward and RPE. The variables R_+ and R_- are the (normalized) sum and difference of RPE and Reward, respectively. An analogous orthogonalization is applied to Surprise and NPE. **C.** The cross-correlation matrix of the orthogonalized variables and the original ones. Surprise$_\perp$, Novelty$_\perp$, and NPE$_\perp$ are highly correlated with their raw values but have zero correlation with reward and RPE. See section B.2 for details.

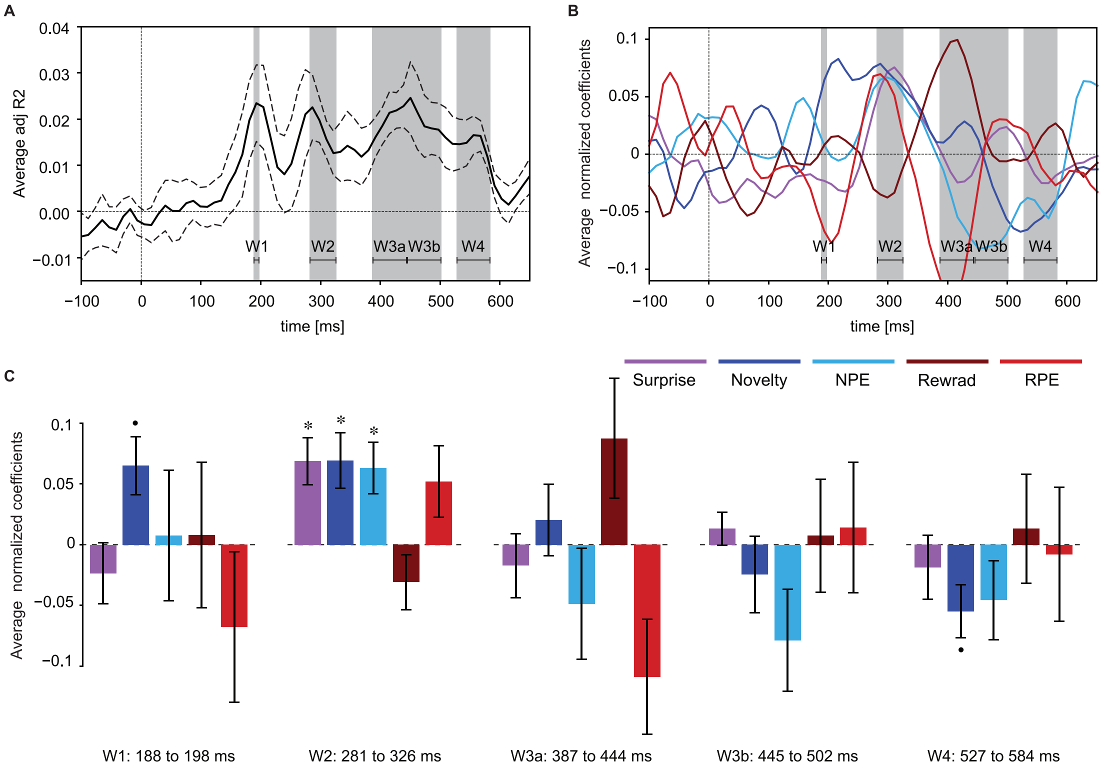


Figure B.9: **ERP variations explained by trial-by-trial and participant-by-participant multivariate linear regression analysis.** Figure B.9 uses a simplified preprocessing pipeline without orthogonalization but is otherwise analogous to Figure 3.10 in the main text. Surprise (magenta), Novelty (dark blue), NEP (light blue), Reward (brown) and RPE (red) were used as explanatory variables, and the ERP amplitude at each time point was considered as the response variable. **A.** Encoding power (adjusted R-squared values) averaged over 10 participants (dashed lines show the standard error of the mean) at each time point. Shaded areas and horizontal lines indicate four time intervals (W1, ..., W4) of significant encoding power (FDR controlled by 0.1, one-sample t-test, only for the time-points after the baseline). The 3rd time interval has been split into two time windows of equal length for the analysis in C. **B.** Values of the regression coefficients (averaged over participants) for Surprise, Novelty, NEP, Reward, and RPE as a function of time. Errors are not shown to simplify the illustration. **C.** In each of the 5 time windows, the regression coefficients plotted in B have been averaged over time. Error bars show the standard error of the mean (across participants). Asterisks show significantly non-zero values (FDR controlled by 0.1 for each time window, one-sample t-test). The Novelty coefficients in the 1st and the last time windows (dot) have p-values of 0.03 and 0.04, respectively, which are not significant after FDR correction. In the second time window, Surprise, Novelty, and NEP have significantly positive coefficients. See section B.2 for details.

C Appendix to chapter 5

C.1 Supplementary results for simulated efficient agents

C.1.1 Simulated efficient agents behave similarly in episode 1

All intrinsically motivated simulated efficient agents find the goal states faster than simulated agents that randomly explore the environment (Figure C.1A1 vs. Figure C.1A2-4). Importantly, in the 2nd half of episode 1, all intrinsically motivated simulated efficient agents spend on average less time in the trap states and more time in the stochastic part than in the 1st half (Figure C.1B2-4; compare it to Figure C.2B). For simulated agents with random exploration, no change can be observed (Figure C.2B1). We conclude that seeking different intrinsic rewards cannot be *qualitatively* distinguished during episode 1, at least given based on the data statistics of simulated efficient agents in Figure C.1B2-B4. This is expected as we fine-tuned the parameters of the simulated efficient agents to have the most efficient exploration during episode 1.

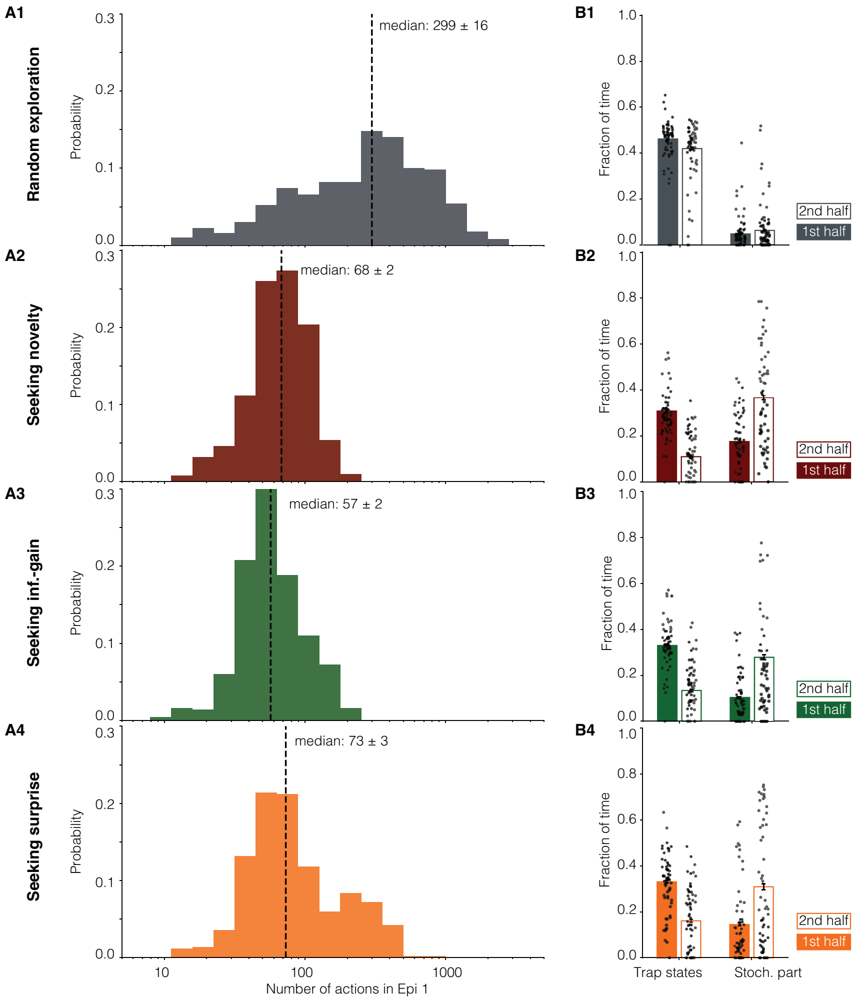


Figure C.1: **Optimized performance of intrinsically motivated algorithms in episode 1 (supplementary to Fig. 2 in the main text). A.** Distribution of number of actions (1500 simulations for each algorithm). SEMed is evaluated by bootstrapping. **B.** Fraction of time spent in the trap states and in the stochastic part during the 1st and the 2nd half of episode 1. Error bars show the SEMean and single dots the data of (60 out of 1500) individual simulations.

C.2 Supplementary results for human participants

C.2.1 Human participants in episode 1

As expected, human participants are on average slower than the intrinsically motivated simulated efficient agents in finding a goal state in episode 1 (Figure C.2A versus Figure C.1A2-A4). However, they show the same qualitative pattern of behavior as the intrinsically motivated simulated efficient agents: In the 2nd half of episode 1, all human participants spend on average less time in the trap states and more time in the stochastic part than in the 1st half (Figure C.2B versus Figure C.1B). These results show that human participants exhibit patterns of directed exploration also in episode 1 – consistent with the results of Xu et al. (2021).

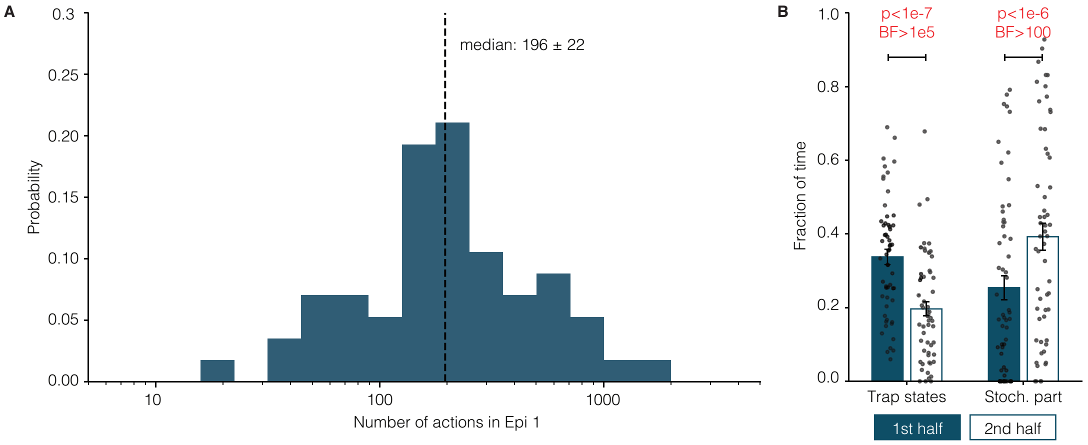


Figure C.2: **Human behavior in episode 1 (supplementary to Fig. 3 in the main text). A.** Distribution of number of actions taken by all 57 participants in episode 1. SEMed is evaluated by bootstrapping. **B.** Fraction of time spent in the trap states and in the stochastic part during the 1st and the 2nd half of episode 1. Error bars: SEMean. Red p-values: Significant effects with False Discovery Rate controlled at 0.05 for all tests in the figure. Red Bayes Factors (BF): Significant evidence in favor of the alternative hypothesis (BF$\geq$ 3).

C.2.2 Action frequencies in progressing states other than states 4

During episode 2, the action frequencies of human participants have the same characteristics in states 1, 2, 3, 5, and 6: A high preference for progressing actions and a low preference for self-looping actions. This implies that all three groups of human participants learned the ‘good’ action in the progressing states during episode 1.

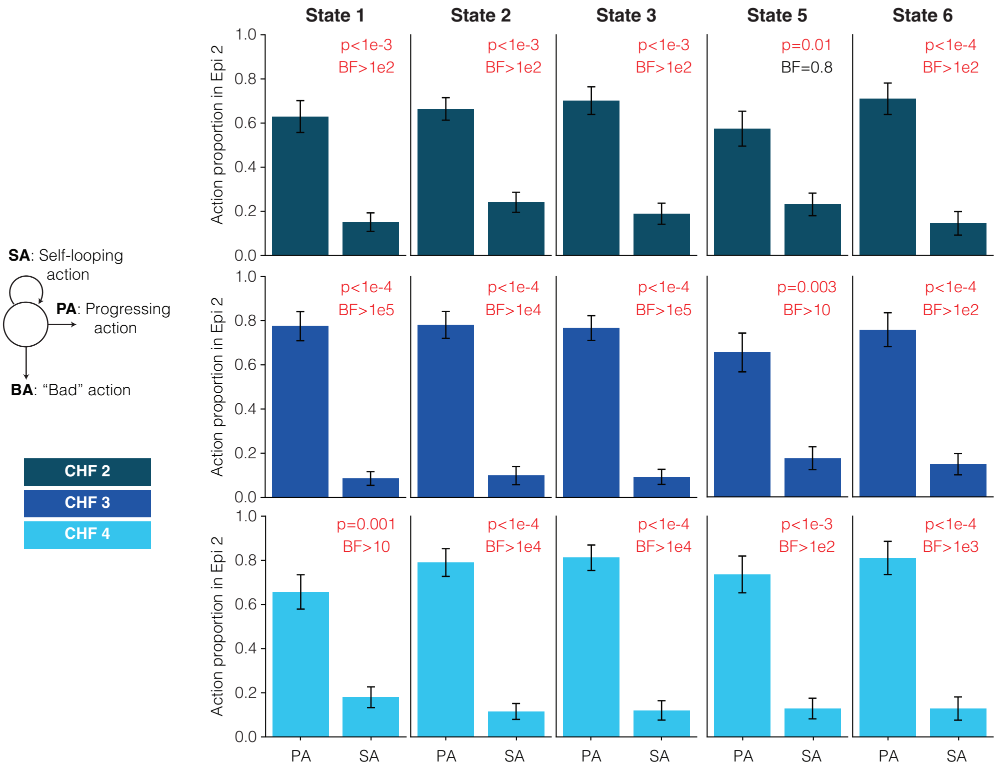


Figure C.3: **Action frequencies of human participants during episode 2 (supplementary to Fig. 3 in the main text).** Fraction of time taking the progressing action (PA) and the self-looping action (SA) when encountering state 1, 2, 3, 5, and 6 during episode 2 for three different groups of participants (different colors) – supplementary to Fig. 5C in the main text. Error bars show the SEMean. Red p-values: Significant effects with False Discovery Rate controlled at 0.05 for all tests in the figure. Red Bayes Factors (BF): Significant evidence in favor of the alternative hypothesis (BF$\geq$ 3).

C.2.3 The relative importance of model-free in action-selection

Similarly to how we quantified the relative importance of novelty in action-selection by ω_{i2e} (see Methods in the main text), we can quantify the relative importance of the model-free branch in action selection by

$$\omega_{\text{MF2MB}} = \frac{\Delta \bar{Q}_{\text{MF}}}{\Delta \bar{Q}_{\text{MB}} + \Delta \bar{Q}_{\text{MF}}}. \tag{C.1}$$

Consistent with the results of Xu et al. (2021), we observe a persistent dominance of the model-free branch throughout the experiment (Figure C.4A-B). Parameter recovery confirms the validity of this observation (Figure C.4C). The dominance of the model-free policy in explaining the action choices of human participants is another source of suboptimality in human behavior (in addition to exploration driven by novelty instead of information gain).

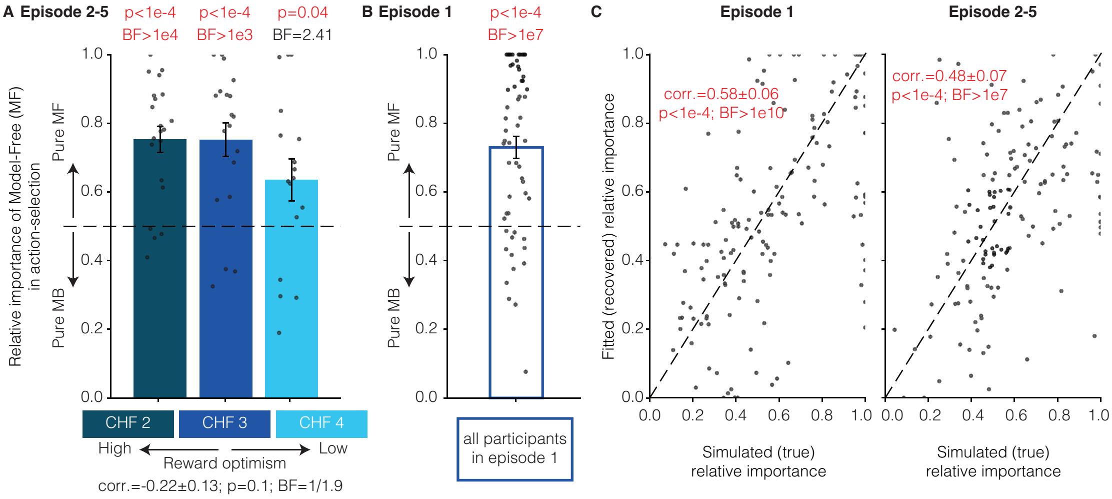


Figure C.4: **The relative importance of model-free policy in action-selection (supplementary to Fig. 6 in the main text). A-B.** The relative importance of the model-free policy in episodes 2-5 (A) and episode 1 (B) is computed for each participant after fitting the model to data (similar to Fig. 6A-B in the main text). Error bars show the SEMean and single dots the data of individual participants. **C.** Parameter-recovery using the action-choices of 150 (= 50 per group) simulated participants seeking novelty (see Methods in the main text). The comparison between the true contribution of the model-free policy to action-selection (computed with the parameters used for simulations) and the recovered contribution (computed with the parameters fitted to the simulated action-choices) shows that the relative importance of the model-free policy is on average identifiable in our experimental paradigm: Red p-values: Significant effects with False Discovery Rate controlled at 0.05 for all tests in the figure. Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3).

C.2.4 Fitted parameters of novelty-seeking

Figure C.5 shows the fitted parameters of novelty-seeking algorithm for all 57 participants – see Methods in the main text for the details of parameter fitting.

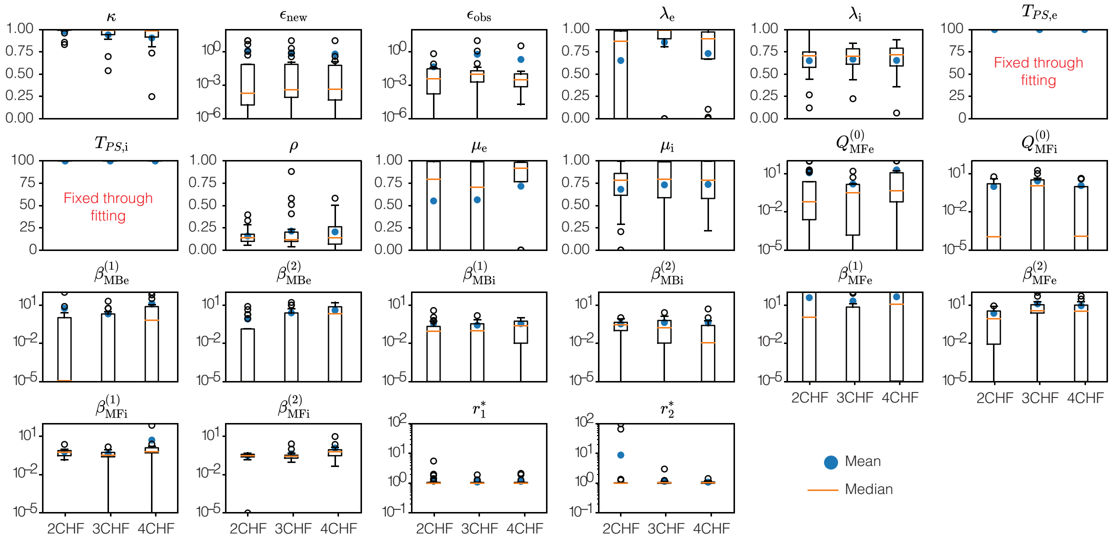


Figure C.5: **Participant-by-participant fitted parameters for novelty-seeking.** Blue circles show group averages and yellow lines group medians.

C.3 Supplementary results for simulated participants

Our comparison of the two most discriminating data statistics of simulated participants with those of human participants in the main text (Fig. 5) helped us gain insights about why algorithms driven by information gain or surprise fail to explain the action-choices of human participants in Bayesian model-selection (see Fig. 4 in the main text). In this section, we present supplementary results on the data of simulated participants as further evidence for our interpretations in the main text.

In subsection C.3.1, we compare all data statistics of human participants reported in Fig. 3 in the main text with the corresponding data statistics of simulated participants. In subsection C.3.2, we go beyond analyzing a handful of data statistics and train a model-agnostic classifier to separate different groups of simulated participants based on their patterns of state visitation. The results in both subsections are consistent with our model-selection results (Fig. 4 in the main text).

C.3.1 Systematic comparisons of data statistics

We analyzed 19 data statistics of human participants in Fig. 3 in the main text, but only two of these data statistics were compared with the data statistics of simulated participants (Fig. 5 in the main text). In this section, we compare all the 19 data statistics of human participants with those of simulated participants (see Figure C.6). Overall, the behavior of simulated participants seeking novelty seems to be the closest to the behavior of human participants (Figure C.6A-B), followed by simulated participants seeking information gain (Figure C.6C-D) and surprise (Figure C.6E-F), respectively. This order is consistent with the results of our Bayesian model-selection in Fig. 4 in the main text. Importantly, novelty-seeking simulated participants can precisely reproduce all *main* data statistics of the human participants based on which we made our conclusions: Correlations in Figure C.6A and the equal action preferences of the 2 CHF group for PA and SA in Figure C.6B. We emphasize that these results are found despite the fact that all the three algorithms have the same number of free parameters that were optimized to maximize the likelihood of action-choices of human participants.

The results shown in Figure C.6 are overall consistent with the interpretations provided in the main text on why algorithms driven by information gain and surprise fail to explain the action-choices of human participants. An additional interesting observation is that simulated participants seeking information gain or surprise can reproduce the between group differences in the median number of actions taken by human participants (inset of Figure C.6C1 and E1) but cannot reproduce the between group differences observed in their fraction of time spent in the stochastic part (inset of Figure C.6C2 and E2). This is due to the fact that the search duration can be, in general, manipulated by the level of randomness in action-selection (controlled by the inverse temperatures of the Softmax policy; see Alg. 5 and Methods in the main text), i.e., increasing random exploration increases the median number of actions even in the absence of any intrinsic reward. However, increasing randomness does not make simulated participants attracted to the stochastic part. Therefore, the failure of algorithms driven by information gain and surprise are essentially due to their inaccuracy in explaining directed exploration. Novelty-seeking can successfully explain this aspect of directed exploration in human

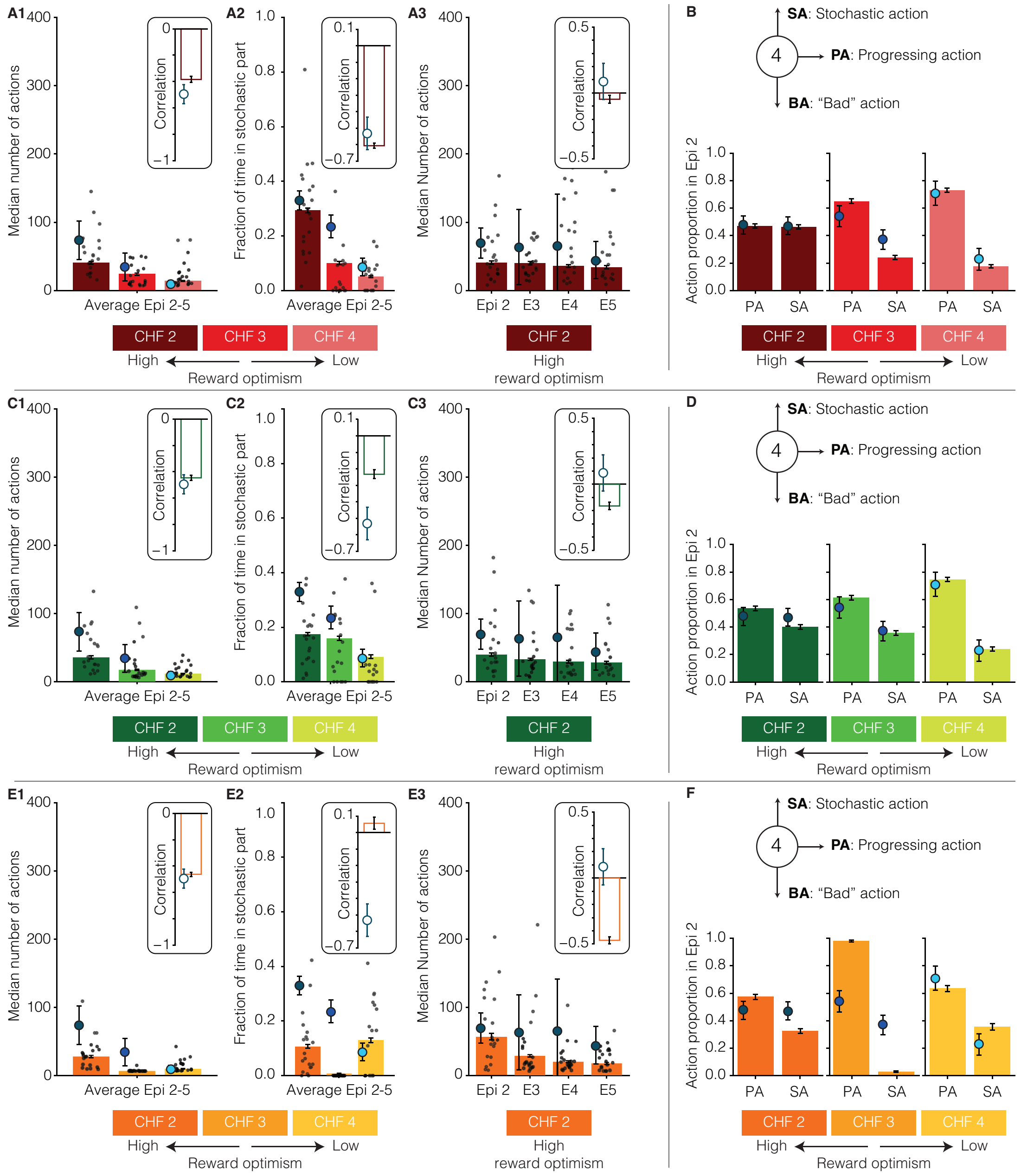


Figure C.6: (Caption next page.)

Figure C.6: (Previous page.) **Data statistics of simulated participants (supplementary to Fig. 5 in the main text).** Red (A-B), green (C-D), and orange bars (E-F) show data statistics of simulated participants seeking novelty, information gain, and surprise, respectively. Data statistics of human participants are shown by blue disks for comparison (same data as in Fig. 3 in the main text). Different shades of color indicate different reward groups: 2 CHF (dark), 3 CHF (medium), and 4 CHF (light). **A1, C1, and E1.** Median number of actions over episodes 2-5: Error bars show the SEMed (evaluated by bootstrapping) and single dots the data of 20 individual participants. Inset: Correlation between the search duration and the goal value of the environment (error estimated by bootstrapping). **A2, C2, and E2.** Average fraction of time spent in the stochastic part of the environment during episodes 2-5. Error bars show the SEMean and single dots the data of 20 individual participants. Inset: Correlation between the fraction of time and the goal value of the environment (error estimated by bootstrapping). **A3, C3, and E3.** Median number of actions in episodes 2-5 for the 2 CHF group. Error bars show the SEMed and single dots the data of 20 individual participants. Inset: Correlation between the search duration and episode number (error estimated by bootstrapping). **B, D, and F.** Fraction of time choosing the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2. Error bars show the SEMean.

participants (inset of Figure C.6A2).

C.3.2 Model-selection based on state visitation patterns

The results in Figure C.6 (and Fig. 5 in the main text) are based on a handful of data statistics that we had picked to dissociate different intrinsically motivated algorithms. In this section, we automatize the process of choosing discriminating data statistics by training a classifier on state visitation patterns of simulated participants (Figure C.7).

Data representation: We divide each episode into 5 equal time-windows and, for each simulated participant, compute the fraction of time it spends in different parts of the environment during each of these time-windows. We average the state-visitation patterns of episode 2-5 to summarize data into two 5×8 matrices (for 5 time-windows and 8 different parts of the environment), one for episode 1 and another for episodes 2-5 (Figure C.7A). By doing so, we summarize and represent the action-choices of each simulated participants by a 80-dimensional vector of state visitations. The 80-dimensional vector can be seen as a more general version of the fraction of time shown in Figure C.6A2, C2, and E2.

Training procedure and classifier evaluation on testing sets: To make a large enough dataset for accurate training of a classifier, we simulate 1500 participants for each of the 3 reward groups of each of the 3 intrinsically motivated algorithms (resulting in $3 \times 3 \times 1'500 = 13'500$ simulated participants), following the procedure described in 'Posterior Predictive Checks' (PPC) section of Methods in the main text. We note that due to stochastic action selection, the sequence of actions choices is different even when the same parameter set is used for several simulated participants. Here, 57 parameter sets (corresponding to 57 individual human participants) is used several hundred times to generate 1500 simulated participants with different action sequences and hence different state visitation patterns (see 'Posterior Predictive Checks' section of Methods in the main text). After removing the outliers (see Methods in the main text) and balancing the number of remaining simulated participants for each algorithm, we train a neural network classifier on the data of simulated participants: The input to the classifier is the 80-dimensional representation of the action-choices of a participant, and the output of the classifier is 3 values for the probability that the participant's action-choices were simulated by seeking novelty, information gain, or surprise, respectively (Figure C.7B1); note that we know the true intrinsic reward corresponding to each simulated participant as we ran the simulations ourselves. We repeat the training procedure 100 times to evaluate the robustness of our results (Figure C.7B1), where, for each training repetition, we use a different random split of data to the training (90%) and testing sets (10%), a different bootstrapping of the training and testing sets, and a different random initialization of the network's weights. The classifier robustly achieves an accuracy rate much higher than chance level (Figure C.7B2); this observation is found despite the facts (i) that the classifiers do not have access to the action-choices of simulated participants and only receive the state visitation patterns, (ii) that the state visitation patterns for many simulated participants (e.g., the 4 CHF group with high level exploitation) may look very similar to each other, and (iii) that the classifiers receive information about different algorithms only through simulations without any explicit access to the their likelihood functions.

Application of the trained classifier to human data: We apply the trained classifiers to the data of human participants represented by the 80-dimensional vectors as in Figure C.7A-B. The output of classifiers applied to the data of a human participant shows the posterior probability that the human participant's action-choices were *generated* by seeking novelty, information gain, or surprise, respectively. The average (over 100 training repetition) classifier output confirms the results of our Bayesian model-selection in Fig. 4 in the main text (Figure C.7C). In summary, our results show that state-visitation patterns of human participants are most similar to those of simulated participants seeking novelty, followed by those seeking information gain and surprise, respectively.

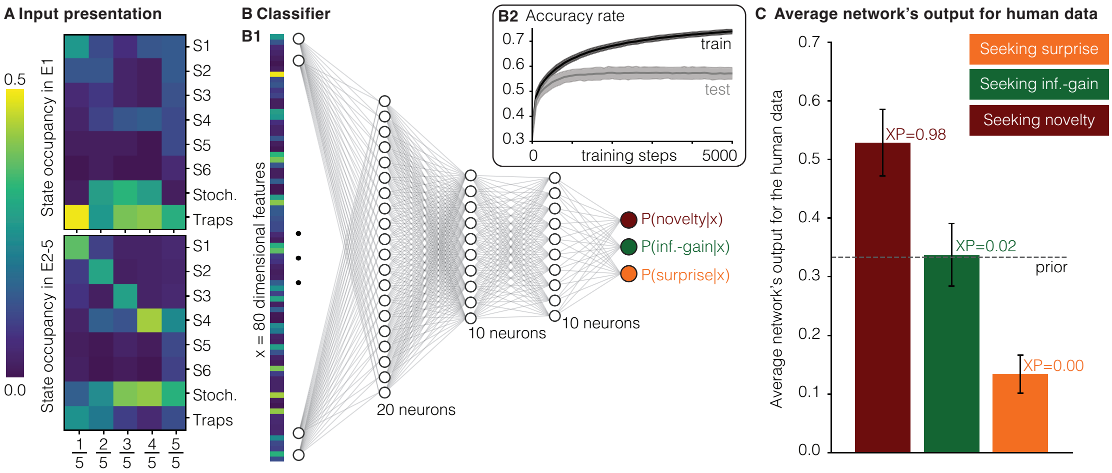


Figure C.7: **A model-agnostic classifier trained on the data statistics of simulated participants confirms the model-selection results in the main text (supplementary to Fig. 4 and Fig. 5 in the main text).** **A**. We divide each episode into 5 equal time-windows and summarize the action-choices of a simulated or human participant into two matrices of state visitations during these 5 time-windows, one for during episode 1 (top) and one averaged over episodes 2-5 (bottom). Color-code indicates the fraction of time spent in different parts of the environment (rows) in a given time-window (columns); the sum of the values in each column is equal to 1. **B**. Given the representation in A, data of each human or simulated participant can be represented by a 80-dimensional vector. **B1**. We consider a neural network with 3 hidden layers as a classifier that receives, as input, the 80-dimensional representation of the action choices of a participant and delivers, as output, the probability that the action-choices of this participant was generated by an algorithm driven by novelty, information gain, or surprise. Each neuron includes a batch-normalization unit and a Rectified Linear Unit (ReLu). **B2**. We train the neural network in B1 for 100 times on the data of simulated participants, each time with a different random split of data into the training and testing sets, a different bootstrapping of the training and testing sets, and a different random initialization of the network's weights (using cross-entropy loss function). The average accuracy rate at the beginning of the training is equal to the chance level (i.e., 33.3%) for both training (black) and testing sets (grey). At the end of the training, the network is able to correctly classify more than 50% of the testing samples. Shaded areas: Standard deviation across the 100 training repetitions. **C**. We apply the trained classifiers to the action-choices of human participants (represented as in A). Average (across the 100 training repetitions; error bars: SEMean) output of the classifiers matches the results of our Bayesian model-selection in the main text. XP stands for the exceedance probability computed as the ratio of the 100 training repetitions that resulted in a higher average P(novelty|x) than P(inf.-gain|x) and P(surprise|x) for human participants. We emphasize that the action-choices of human participants were never presented to the classifier throughout the training.

C.4 Supplementary methods

C.4.1 Model-building in an environment of unknown size

In this section, we use ideas from non-parametric Bayesian inference (Gershman and Blei, 2012; Ghahramani, 2013) and Dirichlet processes (Teh, 2010) to derive a Bayesian estimate $p^{(t)}(s'|s,a)$ of the transition probabilities in an environment of unknown size.

Time dependent base distribution as the expected prior

Consider the 1st time an agent takes action a at state s. Which is the state s' where the agent is expected to land given that it has zero experience for taking action a at state s? There are two possibilities: (i) s' is one of the already known state, i.e., $s' \in \mathcal{S}^{(t)}$, and (ii) s' is one of the infinitely many imaginable states $\mathcal{S}$ that the agent has not observed yet, i.e., $s' \notin \mathcal{S}^{(t)}$. We assume that the agent considers different weights for these two possibilities even in the prior distribution. We give a precise definition of this prior distribution in the next subsection, but we first need to give a definition for our *time dependent base distribution* (Teh, 2010) which we will used later.

We define the probability measure H as a continuous probability distribution (i.e. without any atom) on the space of all imaginable states $\mathcal{S}$ – e.g., the space of all images that can appear on the computer screen. Our results are independent of the exact shape of H – as long as it is a *continuous* probability distribution. We then define the time-dependent base distribution on $\mathcal{S}$ as

$$H^{(t)} = \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} H + \frac{\epsilon_{\text{obs}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} \sum_{s \in \mathcal{S}^{(t)}} \delta_s, \tag{C.2}$$

where δ_s is the Dirac measure at s, ϵ_{obs} and ϵ_{new} are the weights for combining the two possibilities of (i) transiting to a known state $s' \in \mathcal{S}^{(t)}$ and (ii) transiting to a new and unknown state $s' \in \mathcal{S}$. In the next section, we use this base distribution in a way to have $p^{(t)}(.|s,a) = H^{(t)}$ for any state-action pair (s,a) that has not been experienced before.

Derivation of the world-model

We indicate the matrix of transition probabilities as a parameter Θ that fully summarizes the environment. Then, given underlying $\Theta = \theta : \mathcal{S} \times \mathcal{A} \to \text{Measures}[\mathcal{S}]$, we have

$$\mathbb{P}(S_{t+1} = s'|S_t = s, A_t = a, \Theta = \theta) = \theta_{s,a}(s') \tag{C.3}$$

for any s and $s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Given the sequence of states $S_{1:t} = s_{1:t}$ and actions $A_{1:t-1} = a_{1:t-1}$, an agent's belief about the transition matrix θ is defined as the posterior

$$b^{(t)}(\theta) = \mathbb{P}_t(\theta | s_{1:t}, a_{1:t-1}) \propto \mathbb{P}_t(\theta)\mathbb{P}_t(s_{2:t}|\theta, a_{1:t-1}, s_1) = \mathbb{P}_t(\theta) \prod_{t'=1}^{t-1} \theta_{s_{t'},a_{t'}}(s_{t'+1}), \tag{C.4}$$

where the prior $\mathbb{P}_t$ is a time-dependent prior distribution over transition probabilities. We assume that $\Theta_{s,a}$s are a priory i.i.d. samples of a Dirichlet process prior (Teh, 2010) with the base distribution $H^{(t)}$ and a time-dependent concentration parameter $\alpha^{(t)}$, that is, for any finite and countable $\mathcal{S}' \subseteq \mathcal{S}$,

$$\mathbb{P}_t\big(\{\theta_{s,a} : s \in \mathcal{S}', a \in \mathcal{A}\}\big) = \prod_{s \in \mathcal{S}', a \in \mathcal{A}} \mathrm{DP}(\theta_{s,a}; \alpha^{(t)}, H^{(t)}), \tag{C.5}$$

where DP stands for Dirichlet Process. $H^{(t)}$ is the prior expected value of Θ and can be seen as a prior estimate of transition probabilities, and $\alpha^{(t)}$ shows how many samples this estimate is worth (Efron and Hastie, 2016; Teh, 2010). Putting a weight of ϵ_{obs} for each known state and ϵ_{new} for all unknown state (Equation C.2), we end up with $\alpha^{(t)} = \epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|$ as the number of samples that $H^{(t)}$ is worth.

It is straightforward to show that the posterior distribution $b^{(t)}$ has a the same form as the prior (Teh, 2010), that is, for any finite and countable $\mathcal{S}' \subseteq \mathcal{S}$,

$$\begin{aligned} &b^{(t)}\big(\{\theta_{s,a} : s \in \mathcal{S}', a \in \mathcal{A}\}\big) = \\ &\prod_{s \in \mathcal{S}', a \in \mathcal{A}} \mathrm{DP}\Big(\theta_{s,a}\,;\, \alpha^{(t)} + C^{(t)}_{s,a}\,,\, \frac{\alpha^{(t)}}{\alpha^{(t)} + C^{(t)}_{s,a}} H^{(t)} + \frac{1}{\alpha^{(t)} + C^{(t)}_{s,a}} \sum_{s' \in \mathcal{S}^{(t)}} C^{(t)}_{s,a,s'} \delta_{s'}\Big), \end{aligned} \tag{C.6}$$

where $C^{(t)}_{s,a}$ is the number of times action a has been taken at state s' until time t, and $C^{(t)}_{s,a,s'}$ is the number of times transition $(s,a) \to s'$ has been experienced. We consider the posterior expected value of Θ as an estimate of the world-model (Teh, 2010)

$$\begin{aligned} p^{(t)}(s'|s,a) &= \hat{\theta}^{(t)}_{s,a}(s') = \mathbb{E}_{b^{(t)}}[\Theta_{s,a}(s')] \\ &= \frac{\alpha^{(t)}}{\alpha^{(t)} + C^{(t)}_{s,a}} H^{(t)}(s') + \frac{1}{\alpha^{(t)} + C^{(t)}_{s,a}} \sum_{s'' \in \mathcal{S}^{(t)}} C^{(t)}_{s,a,s''} \delta_{s''}(s') \\ &= \frac{\alpha^{(t)}(1 - c_t)}{\alpha^{(t)} + C^{(t)}_{s,a}} H(s') + \frac{1}{\alpha^{(t)} + C^{(t)}_{s,a}} \sum_{s'' \in \mathcal{S}^{(t)}} \Big(\frac{\alpha^{(t)} c_t}{|\mathcal{S}^{(t)}|} + C^{(t)}_{s,a,s''}\Big) \delta_{s''}(s'), \end{aligned} \tag{C.7}$$

where we used $c_t = \frac{\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}$ to shorten the notation. Equation C.7 can be simplified and written as

$$p^{(t)}(s'|s,a) = \hat{\theta}^{(t)}_{s,a}(s') = \begin{cases} \frac{\epsilon_{\text{obs}} + C^{(t)}_{s,a,s'}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + C^{(t)}_{s,a}} & \text{if} \quad s' \in \mathcal{S}^{(t)}\,, \\ \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + C^{(t)}_{s,a}} & \text{if} \quad s' = s_{\text{new}}\,. \end{cases} \tag{C.8}$$

where by $s' = s_{\text{new}}$ we mean $s' \notin \mathcal{S}^{(t)}$, i.e.,

$$\begin{aligned}\hat{\theta}^{(t)}_{s,a}(s_{\text{new}}) &= \mathbb{E}_{\Theta_{s,a}\sim b^{(t)}}\Big[\mathbb{E}_{S'\sim\Theta_{s,a}}[\mathcal{I}_{S'\notin\mathcal{S}^{(t)}}]\Big] \\ &= \frac{\alpha^{(t)}(1-c_t)}{\alpha^{(t)}+C^{(t)}_{s,a}}\int_{s'\notin\mathcal{S}^{(t)}} H(s')ds' = \frac{\alpha^{(t)}(1-c_t)}{\alpha^{(t)}+C^{(t)}_{s,a}}.\end{aligned} \tag{C.9}$$

For the case of $\epsilon_{\text{new}} = 0$, $\epsilon_{\text{obs}} = \epsilon$, and $\mathcal{S}^{(t)} = \mathcal{S}$ being a finite and countable set, the transition matrix is the same as the transition matrix conventionally used for finite state-spaces (Xu et al., 2021). For the case of $\epsilon_{\text{obs}} = 0$, the transition matrix has the form of a Chinese restaurant process (Blei and Frazier, 2011; Teh, 2010).

To account for imperfect model-building, we use leaky counts $\tilde{C}^{(t)}_{s,a,s'}$ and $\tilde{C}^{(t)}_{s,a} = \sum_{s'}\tilde{C}^{(t)}_{s,a,s'}$ instead of $C^{(t)}_{s,a}$ and $C^{(t)}_{s,a,s'}$, where $\tilde{C}^{(t)}_{s,a,s'}$ is recursively updated via

$$\tilde{C}^{(t+1)}_{s,a,s'} = \begin{cases}\kappa\tilde{C}^{(t)}_{s,a,s'} + \delta_{s',s_{t+1}} & \text{if}\, s = s_t\,,\, a = a_t \\ \tilde{C}^{(t)}_{s,a,s'} & \text{otherwise,}\end{cases} \tag{C.10}$$

where δ is the Kronecker delta function, $\tilde{C}^{(0)}_{s,a,s'} = 0$, and $\kappa \in [0,1]$ is the leak parameter (Liakoni et al., 2021; Meyniel et al., 2016; Yu and Cohen, 2009). If $\kappa = 1$, then $\tilde{C}^{(t+1)}_{s,a,s'} = C^{(t+1)}_{s,a,s'}$.

One may argue that the whole Bayesian formulation could be avoided by considering Equation C.8 as the starting point – similar to how we present the model in Methods in the main text. However, as we will see in the next two sections, Equation C.8 without the Bayesian formulation of this section is not enough for deriving (i) the update rule for the model-based branch and (ii) computation of the information gain.

C.4.2 Prioritized sweeping for updating the MB Q-values

Given a reward function R (i.e., R_{ext} for the extrinsic reward or $R_{\text{int},t}$ for the intrinsic reward) the Bellman equations (Sutton and Barto, 2018) are

$$Q^{(t)}(s,a) = \mathbb{E}_{S'\sim\hat{\theta}^{(t)}_{s,a}}\Big[R_{s,a}(S') + \lambda\max_{a'\in\mathcal{A}} Q^{(t)}(S',a')\Big], \tag{C.11}$$

where $Q^{(t)}$ is the Q-value (i.e., $Q^{(t)}_{\text{MB,ext}}$ for the extrinsic reward or $Q^{(t)}_{\text{MB,int}}$ for the intrinsic reward), $\lambda \in [0,1)$ is the discount factor (i.e., λ_{ext} for the extrinsic reward or λ_{int} for the intrinsic reward), and we show $R(s, a \to s')$ by $R_{s,a}(s')$ to shorten the notation.

We assume that $R_{s,a}(s')$ is the same for all $s' \notin \mathcal{S}^{(t)}$. Then, given the fact that $\hat{\theta}^{(t)}_{s,a}(s')$ is also the same for all $s' \notin \mathcal{S}^{(t)}$ or $s \notin \mathcal{S}^{(t)}$ (according to Equation C.8), we can deduce that $Q^{(t)}(s,a)$ is the same for all $s \notin \mathcal{S}^{(t)}$ and for all actions $a \in \mathcal{A}$. Then, Equation C.11

can be re-written as

$$
\begin{aligned}
Q^{(t)}(s,a) = \sum_{s' \in \mathcal{S}^{(t)}} & \hat{\theta}^{(t)}_{s,a}(s')\Big(R_{s,a}(s') + \lambda V^{(t)}(s')\Big) + \\
& \hat{\theta}^{(t)}_{s,a}(s_{\text{new}})\Big(R_{s,a}(s_{\text{new}}) + \lambda V^{(t)}(s_{\text{new}})\Big),
\end{aligned}
\tag{C.12}
$$

where $V^{(t)}(s') := \max_{a' \in \mathcal{A}} Q^{(t)}(s', a')$. We use the fact that $V^{(t)}(s_{\text{new}}) = Q^{(t)}(s_{\text{new}}, a)$ is independent of a and find $V^{(t)}(s_{\text{new}})$ by solving

$$
\begin{aligned}
V^{(t)}(s_{\text{new}}) &= \sum_{s' \in \mathcal{S}^{(t)}} \hat{\theta}^{(t)}_{s_{\text{new}}}(s')\Big(R_{s_{\text{new}}}(s') + \lambda V^{(t)}(s')\Big) + \\
&\qquad \hat{\theta}^{(t)}_{s_{\text{new}}}(s_{\text{new}})\Big(R_{s_{\text{new}}}(s_{\text{new}}) + \lambda V^{(t)}(s_{\text{new}})\Big) \\
&= \frac{\epsilon_{\text{obs}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} \sum_{s' \in \mathcal{S}^{(t)}} \Big(R_{s_{\text{new}}}(s') + \lambda V^{(t)}(s')\Big) + \\
&\quad \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}\Big(R_{s_{\text{new}}}(s_{\text{new}}) + \lambda V^{(t)}(s_{\text{new}})\Big),
\end{aligned}
\tag{C.13}
$$

where we used the fact that $R_{s_{\text{new}}}(s') := R_{s_{\text{new}},a}(s')$ is independent of a. The solution to Equation C.13 is given by

$$
\begin{aligned}
V^{(t)}(s_{\text{new}}) =& \frac{\epsilon_{\text{obs}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} \sum_{s' \in \mathcal{S}^{(t)}} \Big(R_{s_{\text{new}}}(s') + \lambda V^{(t)}(s')\Big) + \\
& \frac{\epsilon_{\text{new}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} R_{s_{\text{new}}}(s_{\text{new}}) \\
=& W^{(t)}_{\text{obs}} \sum_{s' \in \mathcal{S}^{(t)}} \Big(R_{s_{\text{new}}}(s') + \lambda V^{(t)}(s')\Big) + W^{(t)}_{\text{new}} R_{s_{\text{new}}}(s_{\text{new}}),
\end{aligned}
\tag{C.14}
$$

where in the last line we shortened the notation by defining constants

$$
W^{(t)}_{\text{obs}} := \frac{\epsilon_{\text{obs}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|} \quad \text{and} \quad W^{(t)}_{\text{new}} := \frac{\epsilon_{\text{new}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}.
\tag{C.15}
$$

We can combine Equation C.14 and Equation C.12 and found a set of equations for the Q-values of states only in $\mathcal{S}^{(t)}$:

$$
\begin{aligned}
Q^{(t)}(s,a) =& \lambda \sum_{s' \in \mathcal{S}^{(t)}} \Big(\hat{\theta}^{(t)}_{s,a}(s') + \lambda \hat{\theta}^{(t)}_{s,a}(s_{\text{new}}) W^{(t)}_{\text{obs}}\Big) V^{(t)}(s') + \sum_{s' \in \mathcal{S}^{(t)}} \hat{\theta}^{(t)}_{s,a}(s') R_{s,a}(s') + \\
& \hat{\theta}^{(t)}_{s,a}(s_{\text{new}})\Big(R_{s,a}(s_{\text{new}}) + \lambda W^{(t)}_{\text{new}} R_{s_{\text{new}}}(s_{\text{new}}) + \lambda W^{(t)}_{\text{obs}} \sum_{s' \in \mathcal{S}^{(t)}} R_{s_{\text{new}}}(s')\Big).
\end{aligned}
\tag{C.16}
$$

To solve this set of equations, we use prioritized sweeping (Brea, 2017; Sutton and Barto, 2018; Van Seijen and Sutton, 2013) with some modifications (similar to Xu et al. (2021)). The modified algorithm is presented in Alg. 7 in which we use the QUpdate operator defined as

$$\begin{aligned} &Q\text{Update}(s,a;\lambda,\hat{\theta},V,R,W_{\text{obs}},W_{\text{new}},\tilde{\mathcal{S}}) \\ &\quad:=\lambda\sum_{s'\in\tilde{\mathcal{S}}}\Big(\hat{\theta}_{s,a}(s')+\lambda\hat{\theta}_{s,a}(s_{\text{new}})W_{\text{obs}}\Big)V(s')+\sum_{s'\in\tilde{\mathcal{S}}}\hat{\theta}_{s,a}(s')R_{s,a}(s')+ \\ &\quad\quad\hat{\theta}_{s,a}(s_{\text{new}})\Big(R_{s,a}(s_{\text{new}})+\lambda W_{\text{new}}R_{s_{\text{new}}}(s_{\text{new}})+\lambda W_{\text{obs}}\sum_{s'\in\tilde{\mathcal{S}}}R_{s_{\text{new}}}(s')\Big). \end{aligned} \tag{C.17}$$

C.4.3 Derivation of information gain

information gain-seeking algorithms (Mobin et al., 2014; Schmidhuber, 2010) consider the intrinsic reward as the amount of change in the world-model $\hat{\theta}^{(t)}_{s,a}$ upon observing the transition $(s,a)\to s'$, defined as

$$R_{\text{int},t}(s,a\to s')=IG^{(t)}(s,a\to s')=\text{D}_{\text{KL}}\Big[\ \hat{\theta}^{(t)}_{s,a}\ ||\ \hat{\theta}^{(t+1)}_{s,a\to s'}\ \Big], \tag{C.18}$$

where $\hat{\theta}^{(t+1)}_{s,a\to s'}$ is $\hat{\theta}^{(t+1)}_{s,a}$ if $S_{t+1}=s'$, and D_{KL} is the Kullback-Leibler divergence (Cover, 1999). In different contexts, $IG^{(t)}(s,a\to s')$ is also called Postdictive surprise (Kolossa et al., 2015), but it has a fundamentally different behavior from the *prediction* surprise $-\log\hat{\theta}^{(t)}_{s,a}(s')$ (Modirshanechi et al., 2022) that we used for our surprise-seeking algorithm (see Methods in the main text).

If $s'\notin\mathcal{S}_t$, the naïve definition of D_{KL} cannot be used in Equation C.18 because $\hat{\theta}^{(t)}_{s,a}$ and $\hat{\theta}^{(t+1)}_{s,a\to s'}$ has different supports for their atoms. To resolve this issue, Mobin et al. (2014) propose a padding mechanism as a heuristic solution. We use a more general definitions of D_{KL} as the expected Radon–Nikodym derivative of $\hat{\theta}^{(t)}_{s,a}$ with respect to $\hat{\theta}^{(t+1)}_{s,a\to s'}$ that is well-defined in our Bayesian framework:

$$\text{D}_{\text{KL}}\Big[\ \hat{\theta}^{(t)}_{s,a}\ ||\ \hat{\theta}^{(t+1)}_{s,a\to s'}\ \Big]=\mathbb{E}_{S''\sim\hat{\theta}^{(t)}_{s,a}}\left[\frac{d\hat{\theta}^{(t)}_{s,a}}{d\hat{\theta}^{(t+1)}_{s,a\to s'}}(S'')\right], \tag{C.19}$$

where $\frac{d\hat{\theta}^{(t)}_{s,a}}{d\hat{\theta}^{(t+1)}_{s,a\to s'}}(S'')$ is the Radon–Nikodym derivative of $\hat{\theta}^{(t)}_{s,a}$ with respect to $\hat{\theta}^{(t+1)}_{s,a\to s'}$ at S'' – note that $\hat{\theta}^{(t)}_{s,a}$ is always absolutely continuous with respect to $\hat{\theta}^{(t+1)}_{s,a\to s'}$. Whenever $s'\in\mathcal{S}^{(t)}$, the Radon–Nikodym derivative is

$$\frac{d\hat{\theta}^{(t)}_{s,a}}{d\hat{\theta}^{(t+1)}_{s,a\to s'}}(s'')=\begin{cases}\frac{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}+1}{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}} & \text{if}\quad s''\neq s'\,, \\ \frac{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}+1}{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}}\,\frac{\epsilon_{\text{obs}}+C^{(t)}_{s,a,s'}}{\epsilon_{\text{obs}}+C^{(t)}_{s,a,s'}+1} & \text{if}\quad s''=s'\,.\end{cases} \tag{C.20}$$

and whenever $s' \notin \mathcal{S}^{(t)}$, the Radon–Nikodym derivative is

$$\frac{d\hat{\theta}^{(t)}_{s,a}}{d\hat{\theta}^{(t+1)}_{s,a\to s'}}(s'') = \begin{cases} \frac{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+\epsilon_{\text{obs}}+C^{(t)}_{s,a}+1}{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}} & \text{if} \quad s'' \neq s' \,, \\ 0 & \text{if} \quad s'' = s' \,. \end{cases} \tag{C.21}$$

As a result, the information gain in Equation C.18 can be calculated as

$$R_{\text{int},t}(s, a \to s') = \begin{cases} \log \frac{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}+1}{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}} + \hat{\theta}^{(t)}_{s,a}(s') \log \frac{\epsilon_{\text{obs}}+C^{(t)}_{s,a,s'}}{\epsilon_{\text{obs}}+C^{(t)}_{s,a,s'}+1} & \text{if} \quad s' \in \mathcal{S}^{(t)} \,, \\ \log \frac{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+\epsilon_{\text{obs}}+C^{(t)}_{s,a}+1}{\epsilon_{\text{new}}+\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|+C^{(t)}_{s,a}} & \text{if} \quad s' \notin \mathcal{S}^{(t)} \,. \end{cases} \tag{C.22}$$

If $\epsilon_{\text{new}} \to 0$, then the momentary average gain in information after taking action a in state s can be written as

$$\begin{aligned} \bar{IG}^{(t)}(s,a) &:= \mathbb{E}_{S' \sim \hat{\theta}^{(t)}_{s,a}} \Big[IG^{(t)}(s, a \to S') \Big] \\ &= \log \Big[1 + \frac{1}{B^{(t)}_{s,a}} \Big] - \sum_{s' \in \mathcal{S}^{(t)}} \Big(\hat{\theta}^{(t)}_{s,a}(s') \Big)^2 \log \Big[1 + \frac{1}{B^{(t)}_{s,a} \hat{\theta}^{(t)}_{s,a}(s')} \Big], \end{aligned} \tag{C.23}$$

where we defined $B^{(t)}_{s,a} := \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + C^{(t)}_{s,a}$. With a few line of algebra, we can show

$$\frac{\partial \bar{IG}^{(t)}(s,a)}{\partial C^{(t)}_{s,a}} = -\frac{1}{B^{(t)}_{s,a}(1+B^{(t)}_{s,a})} \Big[1 - \sum_{s' \in \mathcal{S}^{(t)}} \Big(\hat{\theta}^{(t)}_{s,a}(s') \Big)^2 \frac{1+B^{(t)}_{s,a}}{1+\hat{\theta}^{(t)}_{s,a}(s') B^{(t)}_{s,a}} \Big] \leq 0, \tag{C.24}$$

where the equality holds if an only if $\hat{\theta}^{(t)}_{s,a}(s') = 1$ for some s'. Hence, $\bar{IG}^{(t)}(s,a)$ is a decreasing function of the count $C^{(t)}_{s,a}$ of the state-action pair (s,a), i.e., the more action a is taken in state s, the less informative it becomes.

C.4.4 Analysis of the MB optimistic initialization in episode 2

To theoretically analyze the influence of the MB optimistic initialization in episode 2, we make a few simplistic assumptions:

1. ϵ_{new} in Equation C.8 is negligible.
2. All transition probabilities expect for the ones between the stochastic states and the progressing action in state 6 (because of the only one time experience) have been learned with certainty during the 1st episode, i.e., the progressing actions have been identified to lead with probability 1 to the next progressing state.
3. The counts for the actions in the stochastic part are roughly the same for all states and actions and is denoted by $\bar{C}^{(t)}$, i.e., for any state s_s in the stochastic part, we

assume that $C^{(t)}_{s_s,a} = \bar{C}^{(t)}$ for every action a.

Given these assumptions, the Q-values in stochastic part are the same for all states (due to the symmetry). In all following equations, we use s_s to denote a representative state in the stochastic part, use a_p to refer to the progressing actions and a_s to the stochastic/self-looping actions, denote state 4 by $s4$ and state 6 by $s6$, use r_{G^*} to denote the reward value of the already discovered goal and define

$$\bar{R} := \frac{1}{|\mathcal{S}^{(t)}|}(1 + r_1^* + r_2^*) \quad \text{and} \quad \bar{V}^{(t)} := \frac{1}{|\mathcal{S}^{(t)}|}\sum_{s'} V^{(t)}_{\text{MB,ext}}(s').$$

Using these notations and assumptions as well as Equation C.8 and Equation C.11, we have

$$\begin{aligned} Q^{(t)}_{\text{MB,ext}}(s_s, a_p) &= p_s^{(t)}(\bar{R} + \lambda_{\text{ext}}\bar{V}^{(t)}) + \lambda_{\text{ext}}\Big(1 - p_s^{(t)}\Big)V^{(t)}_{\text{MB,ext}}(s4) \\ Q^{(t)}_{\text{MB,ext}}(s_s, a_s) &= p_s^{(t)}(\bar{R} + \lambda_{\text{ext}}\bar{V}^{(t)}) + \lambda_{\text{ext}}\Big(1 - p_s^{(t)}\Big)V^{(t)}_{\text{MB,ext}}(s_s), \end{aligned} \tag{C.25}$$

where

$$p_s^{(t)} = \frac{\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}{\epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \bar{C}^{(t)}}. \tag{C.26}$$

Note that $\frac{p_s^{(t)}}{|\mathcal{S}^{(t)}|}$ is equal to the probability of transition to any state s' for which $C_{s_s,a,s'} = 0$ (see Equation C.8).

If the optimal policy is to leave the stochastic part and go to the already discovered goal state, then we must have

$$\text{Condition 1: } Q^{(t)}_{\text{MB,ext}}(s_s, a_s) < Q^{(t)}_{\text{MB,ext}}(s_s, a_p) = V^{(t)}_{\text{MB,ext}}(s_s). \tag{C.27}$$

According Equation C.25, Condition 1 is equivalent to $Q^{(t)}_{\text{MB,ext}}(s_s, a_p) = V^{(t)}_{\text{MB,ext}}(s_s) \leq V^{(t)}_{\text{MB,ext}}(s4)$, which, by using Equation C.25 again and after a few lines of algebra, can be written as

$$\text{Condition 1} \equiv \frac{p_s^{(t)}}{1 - \lambda_{\text{ext}}(1 - p_s^{(t)})}\Big[\bar{R} + \lambda_{\text{ext}}\bar{V}^{(t)}\Big] < V^{(t)}_{\text{MB,ext}}(s4). \tag{C.28}$$

Given that the optimal policy under Condition 1 is to leave the stochastic part and go to the already discovered goal state, we can write the value of state 4 as

$$\begin{aligned} V^{(t)}_{\text{MB,ext}}(s4) &= Q^{(t)}_{\text{MB,ext}}(s4, a_p) \\ &= \lambda^2_{\text{ext}} Q^{(t)}_{\text{MB,ext}}(s6, a_p) = \lambda^2_{\text{ext}}\Big[\tilde{p}_g^{(t)}(\bar{R} + \lambda_{\text{ext}}\bar{V}^{(t)}) + (1 - p_g^{(t)})r_{G^*}\Big], \end{aligned} \tag{C.29}$$

where

$$p_g^{(t)} = \frac{\epsilon_{\text{obs}}|\mathcal{S}^{(t)}|}{\epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + 1} \quad \text{and} \quad \tilde{p}_g^{(t)} = p_g^{(t)} + \lambda_{\text{ext}}(1 - p_g^{(t)}). \tag{C.30}$$

Note that $\frac{p_g^{(t)}}{|\mathcal{S}^{(t)}|}$ is equal to the probability of transition to any state s' for which $C_{s6,a_p,s'} = 0$ (see Equation C.8). Using Equation C.29, we can simplify Equation C.28 as

$$\text{Condition 1} \equiv f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) < r_{G^*}, \tag{C.31}$$

with

$$\begin{aligned} f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, &\epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) := \\ &\frac{1}{\lambda_{\text{ext}}^2(1 - p_g^{(t)})}\Big[\frac{p_s^{(t)}}{1 - \lambda_{\text{ext}}(1 - p_s^{(t)})} - \lambda_{\text{ext}}^2 \tilde{p}_g^{(t)}\Big]\Big[\bar{R} + \lambda_{\text{ext}}\bar{V}^{(t)}\Big]. \end{aligned} \tag{C.32}$$

The variable $R_{\text{Stoch.}}^{(t)}$ that is discussed in the Method section of the main text is equal to $\lambda_{\text{ext}}^2 f_{\text{C1}}$.

An important observation is that, independently of the choice of the model free parameters, we have

$$\lim_{\bar{C}^{(t)} \to \infty} f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) < 0.$$

This implies that, for any value of r_2^* and $r_{G^*} > 0$, increasing $\bar{C}^{(t)}$ would eventually result in a preference for leaving the stochastic part and going towards the already discovered goal (Condition 1 is satisfied): After a sufficiently long and unsuccessful exploration phase, agents will eventually give up exploration. This is essentially what makes the MB optimistic initialization similar to exploration driven by information gain.

Moreover, by analyzing f_{C1}, we can gain further insights about how the model free parameters influence exploration based on the MB optimistic initialization:

1. For any value of r_2^* and r_{G^*}, we have

   $$\lim_{\lambda_{\text{ext}} \to 0} f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) = \infty.$$

   This implies that decreasing the discount factor to put a small weight on the future rewards would make the agent stay in the stochastic part (Condition 1 is violated).

2. If $r_{G^*} < r_2^*$ (i.e., the agent knows that there exists a goal state with a reward higher than the one already discovered) and $\lambda_{\text{ext}}^2 \tilde{p}_g^{(t)} < \frac{p_s^{(t)}}{1-\lambda_{\text{ext}}(1-p_s^{(t)})}$ (i.e., the discount factor is small enough; see point 1), then we have

   $$\lim_{r_2^* \to \infty} f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) > r_{G^*}.$$

This implies that, if $r_{G^*} < r_2^*$, then increasing r_2^* would eventually result in a preference for staying in the stochastic part (Condition 1 is violated): If the reward value of a goal state is much greater than the reward value of the discovered goal state, then the agent prefers to keep exploring the stochastic part.

3. For any value of r_2^* and r_{G^*}, we have

$$\lim_{\epsilon_{\text{obs}} \to 0} f_{\text{C1}}(r_1^*, r_2^*, \lambda_{\text{ext}}, \epsilon_{\text{obs}}, \bar{C}^{(t)}, |\mathcal{S}^{(t)}|) < 0.$$

This implies that, independently of the reward value of the discovered goal state, if the agent assigns a very small prior probability to the unseen transitions, then the agent always prefer to leave the stochastic part and go to the already discovered goal state (i.e., Condition 1 is satisfied).

C.5 Algorithmic implementation

C.5.1 Initialization

For Epi > 1, $\mathcal{S}^{(0)}$, $\tilde{C}^{(0)}$, $U^{(0)}_{\text{ext}}$, $U^{(0)}_{\text{int}}$, $Q^{(0)}_{\text{MB,ext}}$, $Q^{(0)}_{\text{MB,int}}$, $Q^{(0)}_{\text{MF,ext}}$, and $Q^{(0)}_{\text{MF,int}}$ are initialized by their latest value in the previous episode.

For Epi $= 1$, the initial values are as follows:

$$\begin{aligned}
\mathcal{S}^{(0)} &= \{G_0, G_1, G_2\},\\
\tilde{C}^{(0)} &= 0,\\
Q^{(0)}_{\text{MF,ext}}(s,a) &= Q^{(0)}_{\text{MF,ext}},\\
Q^{(0)}_{\text{MF,int}}(s,a) &= Q^{(0)}_{\text{MF,int}}.
\end{aligned} \tag{C.33}$$

For the model-based Q-values, we can analytically solve the bellman equations at time $t = 0$, resulting in

$$\begin{aligned}
&U^{(0)}_{\text{ext}}(s) = Q^{(0)}_{\text{MB,ext}}(s,a) = \frac{\hat{\theta}_{\text{obs}} + \lambda_{\text{ext}}\hat{\theta}_{\text{new}}W_{\text{obs}}}{1 - \lambda_{\text{ext}}|\mathcal{S}^{(0)}|\Big(\hat{\theta}_{\text{obs}} + \lambda_{\text{ext}}\hat{\theta}_{\text{new}}W_{\text{obs}}\Big)}(1 + r_1 + r_2),\\
&U^{(0)}_{\text{int}}(s) = Q^{(0)}_{\text{MB,int}}(s,a)\\
&= \frac{\hat{\theta}_{\text{obs}}|\mathcal{S}^{(0)}|R^{(\text{int})}_{\text{obs}}(s_{\text{obs}}) + \hat{\theta}_{\text{new}}\Big(R^{(\text{int})}_{\text{obs}}(s_{\text{new}}) + \lambda_{\text{int}}W_{\text{new}}R^{(\text{int})}_{\text{new}}(s_{\text{new}}) + |\mathcal{S}^{(0)}|\lambda_{\text{int}}W_{\text{obs}}R^{(\text{int})}_{\text{new}}(s_{\text{obs}})\Big)}{1 - \lambda_{\text{int}}|\mathcal{S}^{(0)}|\Big(\hat{\theta}_{\text{obs}} + \lambda_{\text{int}}\hat{\theta}_{\text{new}}W_{\text{obs}}\Big)},
\end{aligned} \tag{C.34}$$

with

$$\begin{aligned}
W_{\text{obs}} &= \frac{\epsilon_{\text{obs}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(0)}|} \quad , \quad & W_{\text{new}} &= \frac{\epsilon_{\text{new}}}{(1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(0)}|},\\
\hat{\theta}_{\text{obs}} &= \frac{\epsilon_{\text{obs}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(0)}|} \quad , \quad & \hat{\theta}_{\text{new}} &= \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(0)}|}
\end{aligned} \tag{C.35}$$

and

$$\begin{aligned}
R^{(\text{int})}_{\text{new}}(s_{\text{obs}}) &= R_{\text{int},0}(s_{\text{new}}, a \to s) \quad , \quad & R^{(\text{int})}_{\text{new}}(s_{\text{new}}) &= R_{\text{int},0}(s_{\text{new}}, a \to s_{\text{new}})\\
R^{(\text{int})}_{\text{obs}}(s_{\text{obs}}) &= R_{\text{int},0}(s, a \to s) \quad , \quad & R^{(\text{int})}_{\text{obs}}(s_{\text{new}}) &= R_{\text{int},0}(s, a \to s_{\text{new}})
\end{aligned} \tag{C.36}$$

for any $a \in \mathcal{A}$ and $s \in \mathcal{S}^{(0)}$.

However, the final (after learning the transition probabilities) values for $Q_{\text{MB,ext}}(s,a)$ are much smaller than the analytic solution to Bellman equation at $t = 0$ – due to the sparse connections and the fact that there is only one path to one goal state. We, therefore, use a heuristic and put $U^{(0)}_{\text{ext}}(s) = Q^{(0)}_{\text{MB,ext}}(s,a) = 0$.

C.5.2 Pseudocode

See Algorithms 5, 6, 8, and 7 for . Note that, in all pseudocode, we use an alternative shorter notation by defining $R^{(\mathrm{int},t)}_{s,a}(s') := R_{\mathrm{int},t}(s, a \to s')$ and $R^{(\mathrm{ext})}_{s,a}(s') := R_{\mathrm{ext}}(s, a \to s')$.

Algorithm 5 General pseudocode for algorithm

```
# Setting specification
Specify
    $\Phi = \{r_1^*, r_2^*, Q^{(0)}_{\mathrm{MF,ext}}, Q^{(0)}_{\mathrm{MF,int}}, \lambda_{\mathrm{ext}}, \lambda_{\mathrm{int}}, \mu_{\mathrm{ext}}, \mu_{\mathrm{int}}, \rho, \kappa, \epsilon_{\mathrm{new}}, \epsilon_{\mathrm{obs}}, T_{PS,\mathrm{ext}}, T_{PS,\mathrm{int}},$
    $\beta^{(1)}_{\mathrm{MB,ext}}, \beta^{(2)}_{\mathrm{MB,ext}}, \beta^{(1)}_{\mathrm{MB,int}}, \beta^{(2)}_{\mathrm{MB,int}}, \beta^{(1)}_{\mathrm{MF,ext}}, \beta^{(2)}_{\mathrm{MF,ext}}, \beta^{(1)}_{\mathrm{MF,int}}, \beta^{(2)}_{\mathrm{MF,int}}\}.$
Specify the intrinsic reward function $R^{(\mathrm{int},t)}_{s,a}(s')$.
Specify Episode (Epi) and the set of possible actions $\mathcal{A}$.
if Epi $= 1$ then
    Put $\beta_{\mathrm{MB,ext}} = \beta^{(1)}_{\mathrm{MB,ext}}$, $\beta_{\mathrm{MF,ext}} = \beta^{(1)}_{\mathrm{MF,ext}}$, $\beta_{\mathrm{MB,int}} = \beta^{(1)}_{\mathrm{MB,int}}$, and $\beta_{\mathrm{MF,int}} = \beta^{(1)}_{\mathrm{MF,int}}$.
else
    Put $\beta_{\mathrm{MB,ext}} = \beta^{(2)}_{\mathrm{MB,ext}}$, $\beta_{\mathrm{MF,ext}} = \beta^{(2)}_{\mathrm{MF,ext}}$, $\beta_{\mathrm{MB,int}} = \beta^{(2)}_{\mathrm{MB,int}}$, and $\beta_{\mathrm{MF,int}} = \beta^{(2)}_{\mathrm{MF,int}}$.
end if
# Initialization (all variables are defined only for $s \in \mathcal{S}^{(0)}$)
Initialize $\mathcal{S}^{(0)}$, $\tilde{C}^{(0)}$, $U^{(0)}_{\mathrm{ext}}$, $U^{(0)}_{\mathrm{int}}$, $Q^{(0)}_{\mathrm{MB,ext}}$, $Q^{(0)}_{\mathrm{MB,int}}$, $Q^{(0)}_{\mathrm{MF,ext}}$, and $Q^{(0)}_{\mathrm{MF,int}}$ (cf. subsection C.5.1).
# 1st observation
$t = 0$
Initialize state $S_1 = s_1$ and update $\tilde{C}^{(1)}_s = \kappa \tilde{C}^{(0)}_s + \delta_{s,s_1}$ and $\mathcal{S}^{(1)} = \mathcal{S}^{(0)} \cup \{s_1\}$.
Put $\tilde{C}^{(1)}_{s,a,s'} = \tilde{C}^{(0)}_{s,a,s'}$ for all $s$ and $s' \in \mathcal{S}^{(0)}$.
Put $Q^{(1)}_{\mathrm{MF,ext}}(s,a) = Q^{(0)}_{\mathrm{MF,ext}}(s,a)$ and $Q^{(1)}_{\mathrm{MF,int}}(s,a) = Q^{(0)}_{\mathrm{MF,int}}(s,a)$ for all $s \in \mathcal{S}^{(0)}$.
Put $e^{(1)}_{\mathrm{ext}} = 0$ and $e^{(1)}_{\mathrm{int}} = 0$.
# Extensions of variables for $s_1 \notin \mathcal{S}^{(0)}$
Put $\tilde{C}^{(1)}_{s,a,s'} = 0$ if $s = s_1$ or $s' = s_1$ and $s_1 \notin \mathcal{S}^{(0)}$.
Put $Q^{(1)}_{\mathrm{MF,ext}}(s_1,a) = Q^{(0)}_{\mathrm{MF,ext}}$ and $Q^{(1)}_{\mathrm{MF,int}}(s_1,a) = Q^{(0)}_{\mathrm{MF,int}}$ if $s_1 \notin \mathcal{S}^{(0)}$.
Update $U^{(1)}_{\mathrm{ext}}$, $U^{(1)}_{\mathrm{int}}$, $Q^{(1)}_{\mathrm{MB,ext}}$, and $Q^{(1)}_{\mathrm{MB,int}}$ using the model-based branch in Alg. 6.
# Going through the task
$t \leftarrow 1$.
while $s_t \neq G_i$ for $i \in \{0,1,2\}$ do
    # Making action
    Compute $Q^{(t)}_{\mathrm{MF}}(s,a) = \beta_{\mathrm{MF,ext}} Q^{(t)}_{\mathrm{MF,ext}}(s,a) + \beta_{\mathrm{MF,int}} Q^{(t)}_{\mathrm{MF,int}}(s,a)$.
    Compute $Q^{(t)}_{\mathrm{MB}}(s,a) = \beta_{\mathrm{MB,ext}} Q^{(t)}_{\mathrm{MB,ext}}(s,a) + \beta_{\mathrm{MB,int}} Q^{(t)}_{\mathrm{MB,int}}(s,a)$.
    Sample $a_t$ from $\pi(a_t|s_t) \propto \exp\Big\{Q^{(t)}_{\mathrm{MF}}(s_t,a_t) \, + \, Q^{(t)}_{\mathrm{MB}}(s_t,a_t)\Big\}$.
    Observe $S_{t+1} = s_{t+1}$.
    # Updating internal variables
    $\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} \cup \{s_{t+1}\}$.
    Update counts $\tilde{C}^{(t+1)}_s = \kappa \tilde{C}^{(t)}_s + \delta_{s,s_{t+1}}$ and $\tilde{C}^{(t+1)}_{s_t,a_t,s'} = \kappa \tilde{C}^{(t)}_{s_t,a_t,s'} + \delta_{s',s_{t+1}}$.
    Put $\tilde{C}^{(t+1)}_{s,a,s'} = \tilde{C}^{(t)}_{s,a,s'}$ if $s \neq s_t$ or $a \neq a_t$.
    Update $U^{(t+1)}_{\mathrm{ext}}$, $U^{(t+1)}_{\mathrm{int}}$, $Q^{(t+1)}_{\mathrm{MB,ext}}$, and $Q^{(t+1)}_{\mathrm{MB,int}}$ using the model-based branch in Alg. 6.
    Update $e^{(t+1)}_{\mathrm{ext}}$, $e^{(t+1)}_{\mathrm{int}}$, $Q^{(t+1)}_{\mathrm{MF,ext}}$, and $Q^{(t+1)}_{\mathrm{MF,int}}$ using the model-free branch in Alg. 8.
    # Going to the next step
    $t \leftarrow t + 1$.
end while
```

Algorithm 6 Pseudocode for the model-based branch

```
Put $\tilde{C}_{s,a}^{(t+1)} = \sum_{s'} \tilde{C}_{s,a,s'}^{(t+1)}$.
# Updating the world model
Update $\hat{\theta}_{s,a}^{(t+1)}(s') = \big(\epsilon_{\text{obs}} + \tilde{C}_{s,a,s'}^{(t+1)}\big) / \big(\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t+1)}| + \tilde{C}_{s,a}^{(t+1)}\big)$ for $s' \in \mathcal{S}^{(t+1)}$.
Update $\hat{\theta}_{s,a}^{(t+1)}(s_{\text{new}}) = \big(\epsilon_{\text{new}}\big) / \big(\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t+1)}| + \tilde{C}_{s,a}^{(t+1)}\big)$.
# Updating the values
Update $Q_{\text{MB,int}}^{(t+1)}$ and $U_{\text{int}}^{(t+1)}$ using Alg. 7 and $R^{(\text{int},t+1)}$ as rewards.
Update $Q_{\text{MB,ext}}^{(t+1)}$ and $U_{\text{ext}}^{(t+1)}$ using Alg. 7 and $R^{(\text{ext})}$ as rewards.
```

Algorithm 7 Pseudocode for the modified version of Prioritized Sweeping Algorithm for one time-step at time $t+1$

```
# Specifying whether the update is for the intrinsic or the extrinsic reward
Put $\lambda = \lambda_{\text{ext}}$ for extrinsic and $\lambda = \lambda_{\text{int}}$ for intrinsic reward.
Put $Q^{(t)} = Q_{\text{MB,ext}}^{(t)}$, $U^{(t)} = U_{\text{ext}}^{(t)}$, and $R = R^{(\text{ext})}$ for extrinsic, and put $Q^{(t)} = Q_{\text{MB,int}}^{(t)}$, $U^{(t)} = U_{\text{int}}^{(t)}$, and $R = R^{(\text{int},t+1)}$ for intrinsic reward.
# Extending $U$-values
Compute $W_{\text{obs}} = \epsilon_{\text{obs}} / \big((1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t+1)}|\big)$ and $W_{\text{new}} = \epsilon_{\text{new}} / \big((1-\lambda)\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t+1)}|\big)$
if $s_{t+1} \notin \mathcal{S}^{(t)}$ then
    Put $U^{(t)}(s_{t+1}) = W_{\text{obs}} \sum_{s' \in \mathcal{S}^{(t)}} \Big(R_{s_{\text{new}}}(s') + \lambda U^{(t)}(s')\Big) + W_{\text{new}} R_{s_{\text{new}}}(s_{\text{new}})$.
end if
# Applying the effect of the latest observation on $Q$-values using previous $U$-values
for $(s,a) \in \mathcal{S}^{(t+1)} \times \mathcal{A}$ do
    $Q^{(t+1)}(s,a) = Q\text{Update}(s,a;\lambda,\hat{\theta}^{(t+1)},U^{(t)},R,W_{\text{obs}},W_{\text{new}},\mathcal{S}^{(t+1)})$ defined in Eq. C.17.
end for
# Making the priority queue
for $s \in \mathcal{S}^{(t+1)}$ do
    $U^{(t+1)}(s) = U^{(t)}(s)$
    $\text{PriorityQueue}(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$
end for
# Updating $U$-values for $T_{\text{PS}}$ steps
for $T_{\text{PS}}$ iterations do
    $s' = \arg\max_{s \in \mathcal{S}^{(t+1)}} \text{PriorityQueue}(s)$
    $\Delta V = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a) - U^{(t+1)}(s')$
    $U^{(t+1)}(s') = \max_{a \in \mathcal{A}} Q^{(t+1)}(s',a)$
    # Applying the effect of the update of $U$-values on $Q$-values
    for $(s,a) \in \mathcal{S}^{(t+1)} \times \mathcal{A}$ do
        $Q^{(t+1)}(s,a) \leftarrow Q^{(t+1)}(s,a) + \lambda\Big(\hat{\theta}_{s,a}^{(t+1)}(s') + \lambda\hat{\theta}_{s,a}^{(t+1)}(s_{\text{new}})W_{\text{obs}}\Big)\Delta V$
    end for
    # Updating the priority queue
    for $s \in \mathcal{S}^{(t+1)}$ do
        $\text{PriorityQueue}(s) = |U^{(t+1)}(s) - \max_{a \in \mathcal{A}} Q^{(t+1)}(s,a)|$
    end for
end for
```

Algorithm 8 Pseudocode for the model-free branch

```
# Prediction errors
Compute $RPE_{\text{ext},t+1} = R^{(\text{ext})}_{s_t,a_t}(s_{t+1}) + \lambda_{\text{ext}} \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,ext}}(s_{t+1}, a') - Q^{(t)}_{\text{MF,ext}}(s_t, a_t)$.
Compute $RPE_{\text{int},t+1} = R^{(\text{int},t)}_{s_t,a_t}(s_{t+1}) + \lambda_{\text{int}} \max_{a' \in \mathcal{A}} Q^{(t)}_{\text{MF,int}}(s_{t+1}, a') - Q^{(t)}_{\text{MF,int}}(s_t, a_t)$.
# Update of the eligibility traces
Update $e^{(t+1)}_{\text{ext}}(s_t, a_t) = 1$, and $e^{(t+1)}_{\text{ext}}(s, a) = \lambda_{\text{ext}}\mu_{\text{ext}}e^{(t)}_{\text{ext}}(s, a)$, for all $s \neq s_t$ and $a \neq a_t$.
Update $e^{(t+1)}_{\text{int}}(s_t, a_t) = 1$, and $e^{(t+1)}_{\text{int}}(s, a) = \lambda_{\text{int}}\mu_{\text{int}}e^{(t)}_{\text{int}}(s, a)$, for all $s \neq s_t$ and $a \neq a_t$.
# TD-learners
Update $Q^{(t+1)}_{\text{MF,ext}}(s, a) = Q^{(t)}_{\text{MF,ext}}(s, a) + \rho e^{(t+1)}_{\text{ext}}(s, a) RPE_{\text{ext},t+1}$, $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$.
Update $Q^{(t+1)}_{\text{MF,int}}(s, a) = Q^{(t)}_{\text{MF,int}}(s, a) + \rho e^{(t+1)}_{\text{int}}(s, a) RPE_{\text{int},t+1}$, $\forall s \in \mathcal{S}$ and $a \in \mathcal{A}$.
```

D Learning in volatile environments with the Bayes Factor surprise

This appendix is a pointer to the author's paper in Neural Computation (Liakoni et al., 2021).

Authors: Vasiliki Liakoni*, Alireza **Modirshanechi***, Wulfram Gerstner, and Johanni Brea

*: V Liakoni and A Modirshanechi are joint first authors.

Abstract: Surprise-based learning allows agents to rapidly adapt to nonstationary stochastic environments characterized by sudden changes. We show that exact Bayesian inference in a hierarchical model gives rise to a surprise-modulated trade-off between forgetting old observations and integrating them with the new ones. The modulation depends on a probability ratio, which we call the Bayes Factor Surprise, that tests the prior belief against the current belief. We demonstrate that in several existing approximate algorithms, the Bayes Factor Surprise modulates the rate of adaptation to new observations. We derive three novel surprise-based algorithms, one in the family of particle filters, one in the family of variational learning, and one in the family of message passing, that have constant scaling in observation sequence length and particularly simple update dynamics for any distribution in the exponential family. Empirical results show that these surprise-based algorithms estimate parameters better than alternative approximate approaches and reach levels of performance comparable to computationally more expensive algorithms. The Bayes Factor Surprise is related to but different from the Shannon Surprise. In two hypothetical experiments, we make testable predictions for physiological indicators that dissociate the Bayes Factor Surprise from the Shannon Surprise. The theoretical insight of casting various approaches as surprise-based learning, as well as the proposed online algorithms, may be applied to the analysis of animal and human behavior and to reinforcement learning in nonstationary environments.

Funding: This research was supported by the Swiss National Science Foundation (200020 184615) and the European Union Horizon 2020 Framework Program under grant agreement 785907 (Human Brain Project, SGA2).

Author contribution: VL, AM, and JB conceived and designed the project. AM defined the Bayes Factor Surprise and worked out the surprise-based interpretation of the exact Bayesian inference, with the help of VL and JB. VL, AM, and JL developed the biologically plausible algorithms and worked out the surprise-based interpretation of the approximate algorithms. AM and WG conceived and worked out the experimental predictions. VL wrote the code for the algorithms, the simulations, the experimental predictions, and the visualization, with the help and feedback of AM and JB. All authors interpreted the results and wrote the manuscripts.

E Curiosity-driven exploration: foundations in neuroscience and computational modeling

This appendix is a pointer to the author's paper in Trends in Neurosciences (Modirshanechi et al., 2023b).

Authors: Alireza **Modirshanechi**, Kacper Kondrakiewicz, Wulfram Gerstner, and Sebastian Haesler

Abstract: Curiosity refers to the intrinsic desire of humans and animals to explore the unknown, even when there is no apparent reason to do so. Thus far, no single, widely accepted definition or framework for curiosity has emerged, but there is growing consensus that curious behavior is not goal-directed but related to seeking or reacting to information. In this review, we take a phenomenological approach and group behavioral and neurophysiological studies which meet these criteria into three categories according to the type of information seeking observed. We then review recent computational models of curiosity from the field of machine learning and discuss how they enable integrating different types of information seeking into one theoretical framework. Combinations of behavioral and neurophysiological studies along with computational modeling will be instrumental in demystifying the notion of curiosity.

Funding: This work was supported by the Swiss National Science Foundation No. 200020 207426 (AM and WG), KU Leuven C14/21/111 (SH), Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO-Flanders) postdoctoral fellowship 1276122N (KK) and grant G097022N (SH).

Author contribution: All authors contributed to the conceptualization of the study. AM and SH did the visualization and wrote the original draft. All authors revised the text.

F Pointers to other publications as a contributing author

This appendix chapter consists of pointers to the other publications of the author as a contributing author, during his doctoral study.

F.1 Rapid suppression and sustained activation of distinct cortical regions for a delayed sensory-triggered motor response

This appendix section is a pointer to the author's paper in Neuron (Esmaeili et al., 2021).

Authors: Vahid Esmaeili*, Keita Tamura*, Samuel P. Muscinelli, Alireza **Modirshanechi**, Marta Boscaglia, Ashley B. Lee, Anastasiia Oryshchuk, Georgios Foustoukos, Yanqi Liu, Sylvain Crochet, Wulfram Gerstner, and Carl C.H. Petersen

*: V Esmaeili and K Tamura are joint first authors.

Abstract: The neuronal mechanisms generating a delayed motor response initiated by a sensory cue remain elusive. Here, we tracked the precise sequence of cortical activity in mice transforming a brief whisker stimulus into delayed licking using wide-field calcium imaging, multiregion high-density electrophysiology, and time-resolved optogenetic manipulation. Rapid activity evoked by whisker deflection acquired two prominent features for task performance: (1) an enhanced excitation of secondary whisker motor cortex, suggesting its important role connecting whisker sensory processing to lick motor planning; and (2) a transient reduction of activity in orofacial sensorimotor cortex, which contributed to suppressing premature licking. Subsequent widespread cortical activity during the delay period largely correlated with anticipatory movements, but when these were accounted for, a focal sustained activity remained in frontal cortex, which was causally essential for licking in the response period. Our results demonstrate key cortical nodes for motor plan generation and timely execution in delayed goal-directed licking.

Funding: This work was supported by the Swiss National Science Foundation (grants 310030B 166595, 31003A 182010, and CRSII5 177237 to CCHP and grant 200020 165538 to SPM), the European Research Council (grant ERC-2011-ADG 293660 to CCHP), European Union's Marie Skłodowska-Curie Actions (grants 665667 and 798617), the Research Foundation for Opto-Science and Technology, the Brain Science Foundation, the Japan Society for the Promotion of Science, and the Ichiro Kanehara Foundation (KT).

Author contribution: VE, KT, SC, and CCHP conceptualized the study; VE and KT developed neural and behavioral experiment setups; VE, KT, and MB obtained neural and behavioral data; VE, KT, ABL, and AO obtained histological data; GF and YL built the two-photon tomography system; VE, KT, SPM, and AM analyzed the data; WG advised clustering and fitting of neuronal data; VE, KT, SC, and CCHP wrote the manuscript; all authors discussed and edited the manuscript; and CCHP provided overall supervision.

F.2 Fitting summary statistics of neural data with a differentiable spiking network simulator

This appendix section is a pointer to the author's paper in NeurIPS 2021 (Bellec et al., 2021).

Authors: Guillaume Bellec*, Shuqi Wang*, Alireza **Modirshanechi**, Johanni Brea**, and Wulfram Gerstner**

*: G Bellec and S Wang are joint first authors.
**: J Brea and W Gerstner are joint senior authors.

Abstract: Fitting network models to neural activity is an important tool in neuroscience. A popular approach is to model a brain area with a probabilistic recurrent spiking network whose parameters maximize the likelihood of the recorded activity. Although this is widely used, we show that the resulting model does not produce realistic neural activity. To correct for this, we suggest to augment the log-likelihood with terms that measure the dissimilarity between simulated and recorded activity. This dissimilarity is defined via summary statistics commonly used in neuroscience and the optimization is efficient because it relies on back-propagation through the stochastically simulated spike trains. We analyze this method theoretically and show empirically that it generates more realistic activity statistics. We find that it improves upon other fitting algorithms for spiking network models like GLMs (Generalized Linear Models) which do not usually rely on back-propagation. This new fitting algorithm also enables the consideration of hidden neurons which is otherwise notoriously hard, and we show that it can be crucial when trying to infer the network connectivity from spike recordings.

Funding: This research was supported by the Swiss National Science Foundation (no. 200020 184615) and the Intel Neuromorphic Research Lab.

Author contribution: GB and SW conceived the project initially. SW did most of the simulations under the supervision of GB. All authors contributed significantly to the theory and the writing.

F.3 Brain signals of a Surprise-Actor-Critic model: Evidence for multiple learning modules in human decision making

This appendix section is a pointer to the author's paper in NeuroImage (Liakoni et al., 2022).

Authors: Vasiliki Liakoni*, Marco P. Lehmann*, Alireza **Modirshanechi**, Johanni Brea, Antoine Lutti, Wulfram Gerstner**, and Kerstin Preuschoff**

*: V Liakoni and MP Lehmann are joint first authors.

**: W Gerstner and K Preuschoff are joint senior authors.

Abstract: Learning how to reach a reward over long series of actions is a remarkable capability of humans, and potentially guided by multiple parallel learning modules. Current brain imaging of learning modules is limited by (i) simple experimental paradigms, (ii) entanglement of brain signals of different learning modules, and (iii) a limited number of computational models considered as candidates for explaining behavior. Here, we address these three limitations and (i) introduce a complex sequential decision making task with surprising events that allows us to (ii) dissociate correlates of reward prediction errors from those of surprise in functional magnetic resonance imaging (fMRI); and (iii) we test behavior against a large repertoire of model-free, model-based, and hybrid reinforcement learning algorithms, including a novel surprise-modulated actor-critic algorithm. Surprise, derived from an approximate Bayesian approach for learning the world-model, is extracted in our algorithm from a state prediction error. Surprise is then used to modulate the learning rate of a model-free actor, which itself learns via the reward prediction error from model-free value estimation by the critic. We find that action choices are well explained by pure model-free policy gradient, but reaction times and neural data are not. We identify signatures of both model-free and surprise-based learning signals in blood oxygen level dependent (BOLD) responses, supporting the existence of multiple parallel learning modules in the brain. Our results extend previous fMRI findings to a multi-step setting and emphasize the role of policy gradient and surprise signalling in human learning.

Funding: This research was supported by the Swiss National Science Foundation (CRSII2-147636 to KP and WG and 200020 184615 to WG) as well as by the European Union Horizon 2020 Framework Program No.785907 (Human Brain Project, SGA2).AL was supported by the ROGER DE SPOELBERCH foundation and the Swiss National Science Foundation (320030 184784). The experimental work was carried out on the MRI platform of the Département des Neurosciences Cliniques - Centre Hospitalier Universitaire Vaudois (CHUV), which is generously supported by the ROGER DE SPOELBERCH and Partridge Foundations.

Author contribution: Conceptualization: VL, MPL, AM, JB, AL, WG, and KP. Methodology: VL, MPL, AM, JB, AL, WG, and KP. Software: VL and MPL. Validation: VL, MPL, AM, and AL. Formal analysis: VL and MPL. Investigation: VL, MPL, AM, JB, and AL. Writing – original draft: VL, MPL, and WG. Writing – review & editing: VL, MPL, AM, JB, WG, and KP. Visualization: VL and MPL. Resources: AL. Supervision: AL, WG, and KP. Project administration and Funding acquisition: WG and KP.

F.4 Distributed and specific encoding of sensory, motor, and decision information in the mouse neocortex during goal-directed behavior

This appendix section is a pointer to the author's paper in Cell Reports (Oryshchuk et al., 2024).

Authors: Anastasiia Oryshchuk, Christos Sourmpis, Julie Weverbergh, Reza Asri, Vahid Esmaeili, Alireza **Modirshanechi**, Wulfram Gerstner, Carl C.H. Petersen*, and Sylvain Crochet*

**: CCH Petersen and S Crochet are joint senior authors.

Abstract: Goal-directed behaviors involve coordinated activity in many cortical areas, but whether the encoding of task variables is distributed across areas or is more specifically represented in distinct areas remains unclear. Here, we compared representations of sensory, motor, and decision information in the whisker primary somatosensory cortex, medial prefrontal cortex, and tongue-jaw primary motor cortex in mice trained to lick in response to a whisker stimulus with mice that were not taught this association. Irrespective of learning, properties of the sensory stimulus were best encoded in the sensory cortex, whereas fine movement kinematics were best represented in the motor cortex. However, movement initiation and the decision to lick in response to the whisker stimulus were represented in all three areas, with decision neurons in the medial prefrontal cortex being more selective, showing minimal sensory responses in miss trials and motor responses during spontaneous licks. Our results reconcile previous studies indicating highly specific vs. highly distributed sensorimotor processing.

Funding: This work was supported by the Swiss National Science Foundation: grants 31003A_182010 (CCHP), 310030_146252 (CCHP), and TMAG-3_209271 (CCHP).

Author contribution: AO, CCHP, and SC conceptualized the study; AO performed the experiments; AO obtained histological data; AO, CS, JW, and RA analyzed the data; VE, AM, and WG advised on clustering and decoding of neuronal data; AO, CCHP, and SC wrote the manuscript; all authors discussed and edited the manuscript; and CCHP and SC provided overall supervision.

F.5 Remembering the "When": Hebbian Memory Models for the Time of Past Events

This appendix section is a pointer to the author's pre-print on bioRxiv (Brea et al., 2023).

Authors: Johanni Brea, Alireza **Modirshanechi**, Georgios Iatropoulos, and Wulfram Gerstner

Abstract: Humans and animals can remember how long ago specific events happened. In contrast to interval-timing on the order of seconds and minutes, little is known about the neural mechanisms that enable remembering the "when" of autobiographical memories stored in the episodic memory system. Based on a systematic exploration of neural coding, association and retrieval schemes, we develop a family of hypotheses about the reconstruction of the time of past events, consistent with Hebbian plasticity in neural networks. We compare several plausible candidate mechanism in simulated experiments and, accordingly, propose how combined behavioral and physiological experiments can be used to pin down the actual neural implementation of the memory for the time of past events.

Funding: This research was supported by Swiss National Science Foundation (no. 200020 184615).

Author contribution: JB: Conceptualization, Simulations, Figures, Writing: original draft, review, editing. AM: Conceptualization, Writing: review, editing. GI: Conceptualization, Writing: review, editing. WG: Conceptualization, Writing: review, editing, Funding acquisition.

Bibliography

Achiam, J. and Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

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

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

Ahmadlou, M., Houba, J. H. W., van Vierbergen, J. F. M., Giannouli, M., Gimenez, G.-A., van Weeghel, C., Darbanfouladi, M., Shirazi, M. Y., Dziubek, J., Kacem, M., de Winter, F., and Heimel, J. A. (2021). A cell type-specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science*, 372(6543):eabe9681.

Akam, T., Costa, R., and Dayan, P. (2015). Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS Computational Biology*, 11(12):e1004648.

Akiti, K., Tsutsui-Kimura, I., Xie, Y., Mathis, A., Markowitz, J. E., Anyoha, R., Datta, S. R., Mathis, M. W., Uchida, N., and Watabe-Uchida, M. (2022). Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron*, 110(22):3789–3804.e9.

Alamia, A., VanRullen, R., Pasqualotto, E., Mouraux, A., and Zenon, A. (2019). Pupil-linked arousal responds to unconscious surprisal. *Journal of Neuroscience*, 39(27):5369–5376.

Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., and Norman, K. A. (2021). Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, 109(2):377–390.e7.

Aubret, A., Matignon, L., and Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*.

Baldi, P. (2002). *A Computational Theory of Surprise*, pages 1–25. Springer US, Boston, MA.

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

Barascud, N., Pearce, M. T., Griffiths, T. D., Friston, K. J., and Chait, M. (2016). Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences*, 113(5):E616–E625.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning.* Cambridge University Press.

Barry, M. L. L. R. and Gerstner, W. (2022). Fast adaptation to rule switching using neuronal surprise. *bioRxiv*, pages 2022–09.

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

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

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

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences*, 106(5):1672–1677.

Bellec, G., Wang, S., Modirshanechi, A., Brea, J., and Gerstner, W. (2021). Fitting summary statistics of neural data with a differentiable spiking network simulator. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18552–18563. Curran Associates, Inc.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying count-based exploration and intrinsic motivation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Berlemont, K. and Nadal, J.-P. (2022). Confidence-controlled Hebbian learning efficiently extracts category membership from stimuli encoded in view of a categorization task. *Neural Computation*, 1(34):45–77.

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology. General Section*, 41(1-2):68–80.

Bhui, R., Lai, L., and Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41:15–21.

Binz, M. and Schulz, E. (2022). Modeling human exploration through resource-rational reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(74):2461–2488.

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

Brainard, D. H. and Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10(4):433–436.

Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., and Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*.

Brea, J. (2017). Is prioritized sweeping the better episodic control? *arXiv preprint arXiv:1711.06677*.

Brea, J., Modirshanechi, A., Iatropoulos, G., and Gerstner, W. (2023). Remembering the "when": Hebbian memory models for the time of past events. *bioRxiv*.

Bromberg-Martin, E. S., Feng, Y.-Y., Ogasawara, T., White, J. K., Zhang, K., and Monosov, I. E. (2024). A neural mechanism for conserved value computations integrating information and rewards. *Nature Neuroscience*, 27:159–175.

Brändle, F., Binz, M., and Schulz, E. (2022). *Exploration Beyond Bandits*, page 147–168. Cambridge University Press.

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

Carver, C. S., Scheier, M. F., and Segerstrom, S. C. (2010). Optimism. *Clinical Psychology Review*, 30(7):879–889.

Chang, H. (2021). Operationalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

Chentanez, N., Barto, A., and Singh, S. (2005). Intrinsically motivated reinforcement learning. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Cockburn, J., Man, V., Cunningham, W. A., and O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, 110(16):2691–2702.

Cogliati Dezza, I., Cleeremans, A., and Alexander, W. H. (2022). Independent and interacting value systems for reward and information in the human brain. *eLife*, 11:e66358.

Cogliati Dezza, I., Noel, X., Cleeremans, A., and Yu, A. J. (2021). Distinct motivations to seek out information in healthy individuals and problem gamblers. *Translational Psychiatry*, 11:408.

Cohen, J. D., McClure, S. M., and Yu, A. J. (2007). Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942.

Corder, G. W. and 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.

Cubit, L. S., Canale, R., Handsman, R., Kidd, C., and Bennetto, L. (2021). Visual attention preference for intermediate predictability in young children. *Child development*, 92(2):691–703.

da Silva, C. F. and Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, 4:1053–1066.

Daddaoua, N., Lopes, M., and Gottlieb, J. (2016). Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates. *Scientific reports*, 6:20202.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput Biol*, 10(1):e1003441.

Daw, N. (2011). Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, 23(1).

Daw, N. and 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., and Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Daw, N., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8:1704–1711.

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668.

Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21.

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

Dubey, R. and 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., and Griffiths, T. (2021). Aha! moments correspond to meta-cognitive prediction errors. *PsyArXiv*.

Dubois, M., Habicht, J., Michely, J., Moran, R., Dolan, R. J., and Hauser, T. U. (2021). Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *eLife*, 10:e59907.

Duncan-Johnson, C. C. and Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, 14(5):456–467.

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

English, G., Ghasemi Nejad, N., Sommerfelt, M., Yanik, M. F., and von der Behrens, W. (2023). Bayesian surprise shapes neural responses in somatosensory cortical circuits. *Cell Reports*, 42(2):112009.

Esmaeili, V., Tamura, K., Muscinelli, S. P., Modirshanechi, A., Boscaglia, M., Lee, A. B., Oryshchuk, A., Foustoukos, G., Liu, Y., Crochet, S., Gerstner, W., and Petersen, C. C. (2021). Rapid suppression and sustained activation of distinct cortical regions for a delayed sensory-triggered motor response. *Neuron*, 109(13):2183–2201.e9.

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

Fearnhead, P. and 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., and Koechlin, E. (2021). Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, 5:99–112.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., and Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, 19(12):1658–1664.

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

Fong, E. and Holmes, C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.

Foucault, C. and Meyniel, F. (2023). Two determinants of dynamic adaptive learning for magnitudes and probabilities. *bioRxiv*.

Fox, L., Dan, O., and Loewenstein, Y. (2023). On the computational principles underlying human exploration. *eLife*, 12:RP90684.

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

Frey, B. S. and Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5):589–611.

Friedman, D., Hakerem, G., Sutton, S., and Fleiss, J. L. (1973). Effect of stimulus uncertainty on the pupillary dilation response and the vertex evoked potential. *Electroencephalography and clinical neurophysiology*, 34(5):475–484.

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., and Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, 29(1):1–49.

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., and Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science*, 305(5688):1258–1264.

Garrett, M., Groblewski, P., Piet, A., Ollerenshaw, D., Najafi, F., Yavorska, I., Amster, A., Bennett, C., Buice, M., Caldejon, S., et al. (2023). Stimulus novelty uncovers coding diversity in visual cortical circuits. *bioRxiv*, pages 2023–02.

Garrett, M., Manavi, S., Roll, K., Ollerenshaw, D. R., Groblewski, P. A., Ponvert, N. D., Kiggins, J. T., Casal, L., Mace, K., Williford, A., Leon, A., Jia, X., Ledochowitsch, P., Buice, M. A., Wakeman, W., Mihalas, S., and Olsen, S. R. (2020). Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife*, 9:e50340.

Garrett, N. and Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and Cognition*, 50:12–22.

Garrido, M. I., Sahani, M., and Dolan, R. J. (2013). Outlier responses reflect sensitivity to statistical structure in the human brain. *PLoS computational biology*, 9(3):e1002999.

Garrido, M. I., Teng, C. L. J., Taylor, J. A., Rowe, E. G., and Mattingley, J. B. (2016). Surprise responses in the human brain demonstrate statistical learning under high concurrent cognitive demand. *npj Science of Learning*, 1:16006.

Gershman, S. J. (2019a). Uncertainty and exploration. *Decision*, 6(3):277.

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

Gershman, S. J. (2021). Just looking: The innocent eye in neuroscience. *Neuron*, 109(14):2220–2223.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

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

Gershman, S. J. and Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science*, 7(3):391–415.

Gershman, S. J., Radulescu, A., Norman, K. A., and Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, 10:1–13.

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

Gesiarz, F., Cahill, D., and Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLoS Computational Biology*, 15(6):1–15.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553.

Ghavamzadeh, M., Mannor, S., Pineau, J., and Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Found. Trends Mach. Learn.*, 8(5–6):359–483.

Ghazizadeh, A., Fakharian, M. A., Amini, A., Griggs, W., Leopold, D. A., and Hikosaka, O. (2020). Brain Networks Sensitive to Object Novelty, Value, and Their Combination. *Cerebral Cortex Communications*, 1(1).

Ghazizadeh, A., Griggs, W., and Hikosaka, O. (2016). Ecological origins of object salience: Reward, uncertainty, aversiveness, and novelty. *Frontiers in Neuroscience*, 10:378.

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

Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., and Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, 7:1955–1967.

Gläscher, J., Daw, N., Dayan, P., and 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., Filipowicz, A. L., Kable, J. W., Balasubramanian, V., and Gold, J. I. (2018). A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, 2(3):213–224.

Glaze, C. M., Kable, J. W., and Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *eLife*, 4:e08825.

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

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–593.

Grossman, C. D., Bari, B. A., and Cohen, J. Y. (2022). Serotonin neurons modulate learning rate through uncertainty. *Current Biology*, 32(3):586–599.

Grundei, M., Schröder, P., Gijsen, S., and Blankenburg, F. (2023). EEG mismatch responses in a multimodal roving stimulus paradigm provide evidence for probabilistic inference across audition, somatosensation, and vision. *Human Brain Mapping*, 44(9):3644–3668.

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F., and Yamins, D. L. (2018). Learning to play with intrinsically-motivated, self-aware agents. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., Kennedy, R. T., Aragona, B. J., and Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature neuroscience*, 19(1):117–126.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., and 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., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119.

Heilbron, M. and Chait, M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, 389:54–73. Sensory Sequence Processing in the Brain.

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

Henry, G. H., Dreher, B., and Bishop, P. (1974). Orientation specificity of cells in cat striate cortex. *Journal of neurophysiology*, 37(6):1394–1409.

Hershenhoren, I., Taaseh, N., Antunes, F. M., and Nelken, I. (2014). Intracellular correlates of stimulus-specific adaptation. *Journal of Neuroscience*, 34(9):3303–3319.

Hodapp, A. and Rabovsky, M. (2021). The n400 erp component reflects an error-based implicit learning signal during language comprehension. *European Journal of Neuroscience*, 54(9):7125–7140.

Holroyd, C. B. and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4):679.

Homann, J., Koay, S. A., Chen, K. S., Tank, D. W., and Berry, M. J. (2022). Novel stimuli evoke excess activity in the mouse primary visual cortex. *Proceedings of the National Academy of Sciences*, 119(5):e2108882119.

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

Horvitz, J. C., Stewart, T., and Jacobs, B. L. (1997). Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research*, 759(2):251–258.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243.

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

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

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., and 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:e18073.

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

Ito, M. and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21(3):368–373.

Itti, L. and Baldi, P. (2006). Bayesian surprise attracts human attention. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Jaegle, A., Mehrpour, V., and Rust, N. (2019). Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Current Opinion in Neurobiology*, 58:167–174.

Jarrett, D., Tallec, C., Altché, F., Mesnard, T., Munos, R., and Valko, M. (2022). Curiosity in hindsight. In *Deep Reinforcement Learning Workshop NeurIPS 2022*.

Jordan, R. (2023). The locus coeruleus as a global model failure system. *Trends in Neurosciences*.

Jordan, R. and Keller, G. B. (2023). The locus coeruleus broadcasts prediction errors across the cortex to promote sensorimotor plasticity. *eLife*, 12:RP85111.

Juechems, K. and Summerfield, C. (2019). Where does value come from? *Trends in Cognitive Sciences*, 23(10):836–850.

Kahneman, D. (2013). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559.

Kao, C.-H., Khambhati, A. N., Bassett, D. S., Nassar, M. R., McGuire, J. T., Gold, J. I., and Kable, J. W. (2020a). Functional brain network reconfiguration during learning in a dynamic environment. *Nature communications*, 11(1):1682.

Kao, C.-H., Lee, S., Gold, J. I., and Kable, J. W. (2020b). Neural encoding of task-dependent errors during adaptive learning. *eLife*, 9:e58809.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kelly, C., Sharot, T., et al. (2021). Individual differences in information-seeking. *Nature Communications*, 12:7062.

Kidd, C. and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, 88(3):449–460.

Kidd, C., Piantadosi, S. T., and Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, 7(5):e36399.

Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li, Y., Watabe-Uchida, M., Gershman, S. J., et al. (2020a). A unified framework for dopamine signals across timescales. *Cell*, 183(6):1600–1616.

Kim, K., Sano, M., De Freitas, J., Haber, N., and Yamins, D. (2020b). Active world model learning with progress curiosity. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5306–5315. PMLR.

Klyubin, A., Polani, D., and Nehaniv, C. (2005). Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135 Vol.1.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M., and Gottlieb, J. (2019). Diverse motives for human curiosity. *Nature Human Behaviour*, 3:587–595.

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

Kolter, J. Z. and Ng, A. Y. (2009). Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 513–520, New York, NY, USA. Association for Computing Machinery.

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

Kool, W., Cushman, F. A., and Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology*, 12(8):e1005090.

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

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

Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., and Norman, K. A. (2023). Bayesian surprise predicts human event segmentation in story listening. *Cognitive Science*, 47(10):e13343.

Ladosz, P., Weng, L., Kim, M., and Oh, H. (2022). Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22.

Lecaignard, F., Bertrand, O., Caclin, A., and Mattout, J. (2022). Neurocomputational underpinnings of expected surprise. *Journal of Neuroscience*, 42(3):474–486.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4):1–9.

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

Li, Y. S., Nassar, M. R., Kable, J. W., and Gold, J. I. (2019). Individual neurons in the cingulate cortex encode action monitoring, not selection, during adaptive decision-making. *Journal of Neuroscience*, 39(34):6668–6683.

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

Liakoni, V., Modirshanechi, A., Gerstner, W., and 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., and Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Comput Biol*, 9(2):e1002911.

Lieder, F. and Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1.

Lindborg, A., Musiolek, L., Ostwald, D., and Rabovsky, M. (2023). Semantic surprise predicts the n400 brain potential. *Neuroimage: Reports*, 3(1):100161.

Little, D. Y.-J. and Sommer, F. T. (2013). Learning and exploration in action-perception loops. *Frontiers in Neural Circuits*, 7:37.

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

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Macedo, L. and Cardoso, A. (2019). A contrast-based computational model of surprise and its applications. *Topics in Cognitive Science*, 11(1):88–102.

Macedo, L., Reisezein, R., and Cardoso, A. (2004). Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

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

Maguire, R., Maguire, P., and Keane, M. T. (2011). Making sense of surprise: an investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1):176–186.

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

Makeig, S., Delorme, A., Westerfield, M., Jung, T.-P., Townsend, J., Courchesne, E., and Sejnowski, T. J. (2004). Electroencephalographic brain dynamics following manually responded visual targets. *PLoS Biol*, 2(6):e176.

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

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., and 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.

Martin, J., Narayanan, S. S., Everitt, T., and Hutter, M. (2017). Count-based exploration in feature space for reinforcement learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2471–2478. AAAI Press.

Masegosa, A., Nielsen, T. D., Langseth, H., Ramos-López, D., Salmerón, A., and Madsen, A. L. (2017). Bayesian models of data streams with hierarchical power priors. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2334–2343. JMLR. org.

Mathis, M. W., Mathis, A., and Uchida, N. (2017). Somatosensory cortex plays an essential role in forelimb motor adaptation in mice. *Neuron*, 93(6):1493–1503.

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

Mattar, M. G. and Lengyel, M. (2022). Planning in the brain. *Neuron*, 110(6):914–934.

Mavor-Parker, A., Young, K., Barry, C., and Griffin, L. (2022). How to stay curious while avoiding noisy TVs using aleatoric uncertainty estimation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15220–15240. PMLR.

McGuire, J. T., Nassar, M. R., Gold, J. I., and Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, 84(4):870–881.

Meder, B. and Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, 7(2):119–148.

Mehrpour, V., Meyer, T., Simoncelli, E. P., and Rust, N. C. (2021). Pinpointing the neural signatures of single-exposure visual recognition memory. *Proceedings of the National Academy of Sciences*, 118(18):e2021660118.

Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D., and Pathak, D. (2021). Discovering and achieving goals via world models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24379–24391. Curran Associates, Inc.

Meyer, T. and Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences*, 108(48):19401–19406.

Meyer, T., Ramachandran, S., and Olson, C. R. (2014). Statistical learning of serial visual transitions by neurons in monkey inferotemporal cortex. *Journal of Neuroscience*, 34(28):9332–9337.

Meyer, T. and Rust, N. C. (2018). Single-exposure visual memory judgments are reflected in inferotemporal cortex. *eLife*, 7:e32259.

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

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

Miles, J. (2005). *R-Squared, Adjusted R-Squared.* American Cancer Society.

Mobin, S. A., Arnemann, J. A., and Sommer, F. (2014). Information-based learning by agents in unbounded state spaces. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Modirshanechi, A., Becker, S., Brea, J., and Gerstner, W. (2023a). Surprise and novelty in the brain. *Current Opinion in Neurobiology*, 82:102758.

Modirshanechi, A., Brea, J., and Gerstner, W. (2022). A taxonomy of surprise definitions. *Journal of Mathematical Psychology*, 110:102712.

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

Modirshanechi, A., Kondrakiewicz, K., Gerstner, W., and Haesler, S. (2023b). Curiosity-driven exploration: foundations in neuroscience and computational modeling. *Trends in Neurosciences*, 46(12):1054–1066.

Modirshanechi, A., Lin, W.-H., Xu, H. A., Herzog, M. H., and Gerstner, W. (2023c). The curse of optimism: a persistent distraction by novelty. *bioRxiv*.

Moens, V. and Zénon, A. (2019). Learning and forgetting using reinforced Bayesian change detection. *PLoS Computational Biology*, 15(4):e1006713.

Montag, C., Lachmann, B., Herrlich, M., and Zweig, K. (2019). Addictive features of social media/messenger platforms and freemium games against the background of psychological and economic theories. *International journal of environmental research and public health*, 16(14):2612.

Montag, C., Yang, H., and Elhai, J. D. (2021). On the psychology of tiktok use: A first glimpse from empirical findings. *Frontiers in public health*, 9:641673.

Montgomery, K. (1953). Exploratory behavior as a function of "similarity" of stimulus situation. *Journal of Comparative and Physiological Psychology*, 46(2):129–133.

Montgomery, K. C. (1954). The role of the exploratory drive in learning. *Journal of Comparative and Physiological Psychology*, 47(1):60–64.

Morrens, J., Çağatay Aydin, Janse van Rensburg, A., Esquivelzeta Rabell, J., and Haesler, S. (2020). Cue-evoked dopamine promotes conditioned responding during learning. *Neuron*, 106(1):142–153.e7.

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

Murayama, K. (2022). A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review*, 129(1):175–198.

Murayama, K., Matsumoto, M., Izuma, K., and Matsumoto, K. (2010). Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences*, 107(49):20911–20916.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (mmn) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118(12):2544–2590.

Näätänen, R. and Picton, T. (1987). The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, 24(4):375–425.

Nassar, M. R. and Frank, M. J. (2016). Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, 11:49–54.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., and 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., and 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.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4):979–999.

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

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, 135(5):601–609.

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

Niv, Y. and Langdon, A. (2016). Reinforcement learning with marr. *Current Opinion in Behavioral Sciences*, 11:67–73.

Nour, M. M., Dahoun, T., Schwartenbeck, P., Adams, R. A., FitzGerald, T. H. B., Coello, C., Wall, M. B., Dolan, R. J., and 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'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.

Ogasawara, T., Sogukpinar, F., Zhang, K., Feng, Y.-Y., Pai, J., Jezzini, A., and Monosov, I. E. (2022). A primate temporal cortex–zona incerta pathway for novelty seeking. *Nature Neuroscience*, 25.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 34:171–175.

Oryshchuk, A., Sourmpis, C., Weverbergh, J., Asri, R., Esmaeili, V., Modirshanechi, A., Gerstner, W., Petersen, C. C., and Crochet, S. (2024). Distributed and specific encoding of sensory, motor, and decision information in the mouse neocortex during goal-directed behavior. *Cell Reports*, 43(1).

Ostrovski, G., Bellemare, M. G., van den Oord, A., and Munos, R. (2017). Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2721–2730. JMLR.org.

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

Oudeyer, P.-Y. (2018). Computational theories of curiosity-driven learning. *arXiv preprint arXiv:1802.10546*.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286.

Oxford-English-Dictionary (2021a). "novelty, n. and adj.". In *OED Online*. Oxford University Press.

Oxford-English-Dictionary (2021b). surprise, n. In *OED Online*. Oxford University Press.

Ozkan, E., Smidl, V., Saha, S., Lundquist, C., and Gustafsson, F. (2013). Marginalized adaptive particle filtering for nonlinear models with unknown time-varying noise parameters. *Automatica*, 49(6):1566–1575.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., and 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.

O'Toole, S. M., Oyibo, H. K., and Keller, G. B. (2023). Molecularly targetable cell types in mouse visual cortex have distinguishable prediction error responses. *Neuron*, 111(18):2918–2928.e8.

Paller, K. A., Zola-Morgan, S., Squire, L. R., and Hillyard, S. A. (1988). P3-like brain waves in normal monkeys and in monkeys with medial temporal lesions. *Behavioral neuroscience*, 102(5):714.

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

Palminteri, S. and Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, 26(7):607–621.

Pasturel, C., Montagnini, A., and Perrinet, L. U. (2020). Humans adapt their anticipatory eye movements to the volatility of visual motion properties. *PLoS Computational Biology*, 16(4):1–28.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2778–2787. JMLR.org.

Pathak, D., Gandhi, D., and Gupta, A. (2019). Self-supervised exploration via disagreement. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5062–5071. PMLR.

Pearce, J. M. and 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.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042.

Piloto, L. S., Weinstein, A., Battaglia, P., and Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6:1257–1267.

Piray, P. and Daw, N. D. (2021a). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications*, 12:4942.

Piray, P. and Daw, N. D. (2021b). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12:6587.

Poli, F., Meyer, M., Mars, R. B., and Hunnius, S. (2022). Contributions of expected learning progress and perceptual novelty to curiosity-driven exploration. *Cognition*, 225:105119.

Polich, J. (2007). Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology*, 118(10):2128–2148.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25.

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

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

Qiyuan, J., Richer, F., Wagoner, B. L., and Beatty, J. (1985). The pupil and stimulus probability. *Psychophysiology*, 22(5):530–534.

Ramachandran, S., Meyer, T., and Olson, C. R. (2017). Prediction suppression and surprise enhancement in monkey inferotemporal cortex. *Journal of Neurophysiology*, 118(1):374–382.

Reisenzein, R., Horstmann, G., and Schützwohl, A. (2019). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science*, 11(1):50–74.

Richter, D., Kietzmann, T. C., and de Lange, F. P. (2023). High-level prediction errors in low-level visual cortex. *bioRxiv*, pages 2023–08.

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage*, 84:971–985.

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

Rouder, J. N. and Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6):877–903.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, 16:225–237.

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

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

Rouhani, N., Norman, K. A., Niv, Y., and Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, 203:104269.

Rowan, T. H. (1990). *Functional stability analysis of numerical algorithms.* PhD thesis, The University of Texas at Austin.

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

Rust, R. T. and Schmittlein, D. C. (1985). A Bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. *Marketing Science*, 4(1):20–40.

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

SanMiguel, I., Costa-Faidella, J., Lugo, Z. R., Vilella, E., and Escera, C. (2021). Standard tone stability as a manipulation of precision in the oddball paradigm: Modulation of prediction error responses to fixed-probability deviants. *Frontiers in Human Neuroscience*, 15:577.

Santucci, V., Baldassarre, G., and Mirolli, M. (2013). Which is the best intrinsic motivation signal for learning multiple skills? *Frontiers in Neurorobotics*, 7.

Savinov, N., Raichuk, A., Vincent, D., Marinier, R., Pollefeys, M., Lillicrap, T., and Gelly, S. (2019). Episodic curiosity through reachability. In *International Conference on Learning Representations*.

Scheier, M. F., Carver, C. S., and Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, 67(6):1063–1078.

Schmidhuber, J. (2008). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on anticipatory behavior in adaptive learning systems*, pages 48–76. Springer.

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

Schomaker, J. and Meeter, M. (2015). Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. *Neuroscience & Biobehavioral Reviews*, 55:268–279.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schulz, E. and Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, 55:7–14.

Schützwohl, A. and Reisenzein, R. (2012). Facial expressions in response to a highly surprising event exceeding the field of vision: a test of Darwin's theory of surprise. *Evolution and Human Behavior*, 33(6):657–664.

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

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. (2020). Planning to explore via self-supervised world models. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8583–8592. PMLR.

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

Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23):R941–R945.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14(11):1475–1479.

Sheldon, A. B. (1969). Preference for familiar versus novel stimuli as a function of the familiarity of the environment. *Journal of Comparative and Physiological Psychology*, 67(4):516–521.

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

Sinclair, A. H., Manalili, G. M., Brunec, I. K., Adcock, R. A., and Barense, M. D. (2021). Prediction errors disrupt hippocampal representations and update episodic memories. *Proceedings of the National Academy of Sciences*, 118(51):e2117625118.

Singh, S., Lewis, R. L., and Barto, A. G. (2010a). Where do rewards come from? In *Proceedings of the annual conference of the cognitive science society*, pages 2601–2606. Cognitive Science Society.

Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. (2010b). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82.

Solomon, S. S., Tang, H., Sussman, E., and Kohn, A. (2021). Limited Evidence for Sensory Prediction Error Responses in Visual Cortex of Macaques and Humans. *Cerebral Cortex*, 31(6):3136–3152.

Soltani, A. and 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., and Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, 193(4258):1142–1146.

Stahl, A. E. and Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230):91–94.

Stankevicius, A., Huys, Q. J. M., Kalra, A., and Seriès, P. (2014). Optimism as a prior belief about the probability of future reward. *PLoS Computational Biology*, 10(5):1–9.

Starkweather, C. K. and Uchida, N. (2021). Dopamine signals as temporal difference errors: recent advances. *Current Opinion in Neurobiology*, 67:95–105.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017.

Stojić, H., Schulz, E., P Analytis, P., and Speekenbrink, M. (2020). It's new, but is it good? how generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General*, 149(10):1878–1907.

Stone, K., Khaleghi, N., and Rabovsky, M. (2023). The n400 is elicited by meaning changes but not synonym substitutions: Evidence from persian phrasal verbs. *Cognitive Science*, 47(12):e13394.

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the international conference on artificial neural networks, Paris*, volume 2, pages 159–164. Citeseer.

Strunk, D. R., Lopez, H., and DeRubeis, R. J. (2006). Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behaviour Research and Therapy*, 44(6):861–882.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

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

Tartaglia, E. M., Clarke, A. M., and Herzog, M. H. (2017). What to choose next? a paradigm for testing human sequential decision making. *Frontiers in psychology*, 8:312.

Teh, Y. W. (2010). *Dirichlet Process*, pages 280–287. Springer US, Boston, MA.

Teigen, K. H. and Keren, G. (2003). Surprises: low probabilities or high contrasts? *Cognition*, 87(2):55–71.

Ten, A., Kaushik, P., Oudeyer, P.-Y., and Gottlieb, J. (2021). Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, 12:5972.

Tiitinen, H., May, P., Reinikainen, K., and Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature*, 372(6501):90–92.

Todorovic, A. and de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *Journal of Neuroscience*, 32(39):13389–13395.

Tomov, M. S., Truong, V. Q., Hundia, R. A., and Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature communications*, 11:2371.

Traner, M. R., Bromberg-Martin, E. S., and Monosov, I. E. (2021). How the value of the environment controls persistence in visual search. *PLoS Computational Biology*, 17(12):1–34.

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

Tueting, P., Sutton, S., and Zubin, J. (1970). Quantitative evoked potential correlates of the probability of events. *Psychophysiology*, 7(3):385–394.

Ulanovsky, N., Las, L., and Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature neuroscience*, 6(4):391–398.

Van Seijen, H. and Sutton, R. (2013). Planning by prioritized sweeping with small backups. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 361–369, Atlanta, Georgia, USA. PMLR.

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

Visalli, A., Capizzi, M., Ambrosini, E., Kopp, B., and Vallesi, A. (2023). P3-like signatures of temporal predictions: a computational EEG study. *Experimental Brain Research.*

Visalli, A., Capizzi, M., Ambrosini, E., Mazzonetto, I., and Vallesi, A. (2019). Bayesian modeling of temporal expectations in the human brain. *NeuroImage*, 202:116097.

Walsh, M. M. and Anderson, J. R. (2012). Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, 36(8):1870–1884.

Walz, J. M., Goldman, R. I., Carapezza, M., Muraskin, J., Brown, T. R., and Sajda, P. (2013). Simultaneous eeg-fmri reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. *Journal of Neuroscience*, 33(49):19212–19222.

Wang, W., Eldridge, M. A., and Richmond, B. J. (2022). Novelty seeking for novelty's sake. *Nature Neuroscience*, 25.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23.

Wilmes, K. A., Petrovici, M. A., Sachidhanandam, S., and Senn, W. (2023). Uncertainty-modulated prediction errors in cortical microcircuits. *bioRxiv*.

Wilson, R. C., Bonawitz, E., Costa, V. D., and Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38:49–56. Computational cognitive neuroscience.

Wilson, R. C. and Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074–2081.

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

Wittmann, B. C., Daw, N. D., Seymour, B., and Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, 58(6):967–973.

Wu, C. M., Schulz, E., and Gershman, S. J. (2020). Inference and search on graph-structured spaces. *Computational Brain & Behavior*, pages 1–23.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., and Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, 2(12):915–924.

Wu, S., Blanchard, T., Meschke, E., Aslin, R. N., Hayden, B. Y., and Kidd, C. (2022). Macaques preferentially attend to intermediately surprising information. *Biology letters*, 18(7):20220144.

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, 15(5):786.

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

Yabe, H., Tervaniemi, M., Reinikainen, K., and Näätänen, R. (1997). Temporal window of integration revealed by mmn to sound omission. *Neuroreport*, 8(8):1971–1974.

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1620.

Yaron, A., Hershenhoren, I., and Nelken, I. (2012). Sensitivity to complex statistical regularities in rat auditory cortex. *Neuron*, 76(3):603–615.

Yu, A. J. and Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

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

Zajkowski, W. K., Kossut, M., and Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife*, 6:e27430.

Zhang, K., Bromberg-Martin, E. S., Sogukpinar, F., Kocher, K., and Monosov, I. E. (2022). Surprise and recency in novelty detection in the primate brain. *Current Biology*, 32(10):2160–2173.e6.

Zhao, S., Chait, M., Dick, F., Dayan, P., Furukawa, S., and Liao, H.-I. (2019). Pupil-linked phasic arousal evoked by violation but not emergence of regularity within rapid sound sequences. *Nature Communications*, 10(1):4030.

Alireza Modirshanechi

email: alireza.modirshanechi@epfl.ch
webpage: https://sites.google.com/view/modirsha

Education

Ph.D. in Computer and Communication Sciences — *2018 – 2024*
EPFL, Lausanne, Switzerland
Thesis Advisor: Prof. Wulfram Gerstner

B.Sc. in Electrical Engineering (Communication) — *2014 – 2018*
Sharif University of Technology, Tehran, Iran — GPA: 19.12/20

Research Interests

- Computational Neuroscience; Cognitive Science; Reinforcement Learning; Statistical Inference

Honors and Awards

- EPFL special bonus as the recognition of **outstanding performance in 2020-2021** and **2021-2022**
- **Teaching Assistant Award**, School of Computer and Communication Sciences, EPFL, 2019
- Recipient of 2018-2019 **EPFL EDIC Fellowship**
- **Ranked 2nd** in the annual **distinguished bachelor thesis competition**, Electrical Engineering Department, Sharif University of Technology, 2018
- **Ranked 2nd in Cumulative GPA** among all entrants of 2014 (nearly 200 students), Electrical Engineering Department, Sharif University of Technology
- **Silver Medal**, National Electrical Engineering Olympiad, Iran, 2018
- **Ranked 1st** in the National Tournament of fMRI Analysis, National Brain Mapping Lab., Iran, 2017
- **Silver Medal**, 45th International Physics Olympiad (**IPHO**), Kazakhstan, 2014
- **Gold Medal**, 26th National Physics Olympiad, Iran, 2013

Research Experience

Ph.D. Student — 2018 –2024
Laboratory of Computational Neuroscience, EPFL — *Lausanne, Switzerland*

- Supervisor: Prof. Wulfram Gerstner
- Working on mathematical definitions of surprise and novelty (e.g., Modirshanechi et al. 2022 in J. Math. Psych.), their roles as intrinsic reward signals in human reinforcement learning (e.g., Modirshanechi et al. 2023 on bioRxiv), and their contribution to adaptive learning in volatile environments (e.g., Liakoni* and Modirshanechi* et al. 2021 in Neural Comp.).
- Also worked on collaborative projects on statistical analysis of physiological recordings and neuroimaging data (e.g., Bellec et al. in NeurIPS 2021, Esmaeili et al. 2021 in Neuron, and Liakoni et al. 2022 in NeuroImage).

Semester-Project Student (Lab rotation) — Feb. 2019 – June 2019
Laboratory of Mathematical Data Science, EPFL — *Lausanne, Switzerland*

- Supervisor: Prof. Emmanuel Abbe (in collaboration with Prof. George Coukos)
- Network data analysis in immuno-oncology, using gene expression to predict the efficiency of different treatments on lung cancer.

Research Assistant — 2016 – 2018
AIR Lab, Sharif University of Technology — *Tehran, Iran*

- Supervisor: Prof. Hamid Aghajan
- Decoding subjective surprise using EEG signals (Modirshanechi et al. 2019 in NeuroImage), decoding stimulus-type using fMRI signals, and predicting election outcomes using Twitter and Google Trends data (Kassraie*, Modirshanechi*, and Aghajan in DATA 2017).

Publications (first or co-first author)

Google Scholar profile: https://scholar.google.com/citations?user=ysDOYB8AAAAJ&hl=en

[1] **The curse of optimism: a persistent distraction by novelty**
A Modirshanechi, W Lin, HA Xu, MH Herzog, W Gerstner. (2023)
bioRxiv preprint: 2022.07.05.498835
https://doi.org/10.1101/2022.07.05.498835

[2] **Curiosity-driven exploration: foundations in neuroscience and computational modelling**
A Modirshanechi, K Kondrakiewicz, W Gerstner, S Haesler. (2023)
Trends in Neurosciences 46.
https://doi.org/10.1016/j.tins.2023.10.002

[3] **Surprise and novelty in the brain**
A Modirshanechi, S Becker, J Brea, W Gerstner. (2023)
Current Opinion in Neurobiology 82.
https://doi.org/10.1016/j.conb.2023.102758

[4] **A taxonomy of surprise definitions**
A Modirshanechi, J Brea, W Gerstner. (2022)
Journal of Mathematical Psychology 110.
https://doi.org/10.1016/j.jmp.2022.102712

[5] **Novelty is not Surprise:**
human exploratory and adaptive behavior in sequential decision-making
HA Xu*, **A Modirshanechi***, MP Lehmann, W Gerstner†, MH Herzog†. (2021)
PLOS Computational Biology 17 (6).
https://doi.org/10.1371/journal.pcbi.1009070

[6] **Learning in volatile environments with the Bayes Factor surprise**
V Liakoni*, **A Modirshanechi***, W Gerstner, J Brea. (2021)
Neural Computation 33 (2).
https://doi.org/10.1162/neco_a_01352

[7] **Trial-by-trial surprise-decoding model for visual and auditory binary oddball task**
A Modirshanechi, MM Kiani, H Aghajan. (2019)
NeuroImage, 196, 302-317.
https://doi.org/10.1016/j.neuroimage.2019.04.028

[8] **Election vote share prediction using a sentiment-based fusion of Twitter data with Google Trends and online polls**
P Kassraie*, **A Modirshanechi***, H Aghajan. (2017)
Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017).
https://doi.org/10.5220/0006484303630370

$*$: Joint first authors; †: Joint senior authors

Publications (collaborating author)

[1] **Remembering the "When": Hebbian memory models for the time of past events**
J Brea, **A Modirshanechi**, G. Iatropoulos, W Gerstner. (2023)
bioRxiv preprint: 2022.11.28.518209.
https://doi.org/10.1101/2022.11.28.518209

[2] **Distributed and specific encoding of sensory, motor, and decision information in the mouse neocortex during goal-directed behavior**
A Oryshchuk, C Sourmpis, ..., **A Modirshanechi**, W Gerstner, CCH Petersen†, S Crochet†. (2024)
Cell Report, 43
https://doi.org/10.1016/j.celrep.2023.113618

[3] **Brain signals of a Surprise-Actor-Critic model: evidence for multiple learning modules in human decision making**
V Liakoni*, MP Lehmann*, **A Modirshanechi**, J Brea, A Lutti, W Gerstner†, K Preuschoff†. (2022)
NeuroImage, 246.
https://doi.org/10.1016/j.neuroimage.2021.118780

[4] **Fitting summary statistics of neural data with a differentiable spiking network simulator**
G Bellec*, S Wang*, **A Modirshanechi**, J Brea†, W Gerstner†. (2021)
NeurIPS 2021.
https://arxiv.org/abs/2106.10064

[5] **Rapid suppression and sustained activation of distinct cortical regions for a delayed sensory-triggered motor response**
V Esmaeili*, K Tamuras*, SP Muscinelli, **A Modirshanechi**, ..., W Gerstner, CCH Petersen. (2021)
Neuron, 109.
https://doi.org/10.1016/j.neuron.2021.05.005

Conference Talks and Posters

Contributed Talks:

- **Novelty drives exploration, whereas surprise modulates learning**
  A Modirshanechi, HA Xu, , MP Lehmann, MH Herzog, W Gerstner
  Bernstein 2021

Selected Posters:

- **Novelty drives human exploration even when it is suboptimal**
  A Modirshanechi, HA Xu, W Lin, MH Herzog, W Gerstner
  Cosyne 2023, CCN 2023, & Bernstein 2023
- **Environmental similarity structure modulates neural and behavioral signatures of novelty**
  S Becker, **A Modirshanechi**, W Gerstner
  Bernstein 2023
- **Theories of surprise: definitions and predictions**
  A Modirshanechi, J Brea, W Gerstner
  Cosyne 2022
- **A bio-plausible implementation of novelty-guided exploration**
  S Becker, **A Modirshanechi**, W Gerstner
  RLDM 2022 & Bernstein 2022
- **Model-based surprise modulates model-free learning**
  A Modirshanechi, HA Xu, , MP Lehmann, MH Herzog, W Gerstner
  Cosyne 2021
- **Neural circuits for converting sensory information into a motor plan**
  V Esmaeili, K Tamura, SP Muscinelli, **A Modirshanechi**, ..., W Gerstner, CCH Petersen
  Cosyne 2021

Invited Talks

- The Computational Neuroscience Center (INT, Marseille), February 2024
- Computational Phenomenology Discussion Group (hosted by Maxwell Ramstead), December 2023
- Angelika Steger Lab (ETH Zurich, Switzerland), February 2023
- Active Inference Institute, January 2023
- Eric Schulz Lab (Max Planck Inst. for Bio. Cybernetics, Tübingen, Germany), November 2022
- Sebastian Haesler Lab (Katholieke Universiteit Leuven, Belgium), October 2022
- Ilya Monosov Lab (University of Washington), May 2022
- Marcelo Mattar Lab (UC San Diego), May 2022
- Montreal AI-Neuro Seminar Series, May 2022
- The Allen Institute, October 2021
- IPM Theoretical Neuroscience Journal Club, Tehran, Iran, March 2021
- EPFL Neuro Symposium, February 2021
- Sharif Neuro Event (Sharif University of Technology), July 2019

Teaching Experience

Teaching Assistant 2019 – 2023
EPFL *Lausanne, Switzerland*

- Artificial Neural Networks and Reinforcement Learning – Spring 2020, 2021, 2022, and 2023
    - Head Teaching Assistant in 2021, 2022, and 2023.
    - Taught a series of lectures on "Curiosity-Driven Exploration" in 2023; recorded videos on YouTube: `https://www.youtube.com/playlist?list=PL7SYVykTNxXZgpOTVM9WBalpHr7rU177Q`
- Markov Chains and Algorithmic Applications – Fall 2019, 2020, and 2021
- Probability and Statistics – Spring 2019

Teaching Assistant 2015 – 2018
Sharif University of Technology *Tehran, Iran*

- Computational Neuroscience – Spring 2018 (Head Teaching Assistant)
- Communication Systems – Fall 2017
- Engineering Mathematics – Fall 2016 and 2017 (Head Teaching Assistant)
- Signals and Systems – Spring 2017
- Electromagnetism – Fall 2015

Advanced Physics Teacher 2013 – 2017
National Organization for Development of Exceptional Talents *Several cities in Iran*

- Thermodynamics, Classical Mechanics, Electromagnetism, and Calculus

Mentorship, Service, and Volunteer Work

- Writing **educational articles** in Medium Publications (e.g., *Towards Data Science*); selected articles:
    - 'Why does the optimal policy exist?'**
    - 'Is correlation distance a metric?'**
    - 'What can we learn from posterior distributions?'*
    - 'The counter-intuitive nature of probabilistic relationships'*

  * : more than 1K views; ** : more than 10K views;

- Workshop organizer (jointly with Franziska Brändle) at the Bernstein conference 2022:
  Surprise in the brain: Theory and Experiments
- Adhoc **Reviewer** for Experimental Brain Research, Frontiers in Psychology, Nature Communications, PLOS Computational Biology, and Scientific Reports.
- **Supervisor** of 6 master students at EPFL (2019 – 2023):
  MS Halvagal, A Mocanu, A Jakob, L Gruaz, Y El Hassan, and A Poiroux.
- **Supervisor** of 5 undergrad research assistants at the Sharif University of Technology (2016 – 2018):
  E Khalaj, A Shirali, A Hoseini, A Shirzad, and M Shahbazi.
- **Volunteer** in the Cosyne 2024 Mentoring Forum to give feedback to other researchers on their Cosyne abstract.
- **Volunteer** in the EPFL Buddy program (every year during 2020-2024) to help 1st-year Ph.D. students at EPFL with starting their Ph.D (e.g., EPFL course, life in Lausanne, etc.).
- **Volunteer** in Review of Application Materials (RAMP) program at EPFL (2022-2023 and 2023-2024) to help prospect Ph.D. students with their applications (e.g., reviewing CVs, cover letters, etc.).