The curse of optimism: a persistent distraction by novelty

Alireza Modirshanechi[1,2,†,*], He A. Xu[1,†], Wei-Hsiang Lin[1], Michael H. Herzog[1], and Wulfram Gerstner[1,2]

[1] Brain-Mind Institute, School of Life Sciences, EPFL, Lausanne, Switzerland
[2] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
[†] AM and HAX contributed equally to this work.
[*] Corresponding author: alireza.modirshanechi@epfl.ch

Abstract

Human curiosity has been interpreted as a drive for exploration and modeled by intrinsically motivated reinforcement learning algorithms. An unresolved challenge in machine learning is that these algorithms are prone to distraction by reward-independent stochastic stimuli. We ask whether humans exhibit the same distraction pattern in their behavior as the algorithms. To answer this question, we design a multi-step decision-making paradigm containing an unknown number of states in a stochastic part of the environment, combined with a reward-manipulation that varies optimism of participants regarding the availability of rewards. We show that (i) participants exhibit a persistent distraction by novelty in the stochastic part; (ii) reward optimism increases this distraction; and (iii) the persistent distraction is consistent with the predictions of algorithms driven by novelty but not with 'optimal' algorithms driven by information-gain. Our results suggest that humans use sup-optimal but computationally cheap curiosity-driven policies for exploration in complex environments.

Introduction

Curiosity drives humans and animals to explore their environments[1–3] and search for potentially more valuable sources of reward (e.g., more nutritious foods or better-paid jobs) than those currently available[4,5]. In computational neuroscience and psychology, intrinsically motivated reinforcement learning (RL) algorithms[6–8] have been proposed as models of curiosity-driven behavior[9–12] with novelty, surprise, or information-gain as intrinsic motivation in addition to the extrinsic motivation by nutritious or monetary reward[11]. These algorithms have been successful not only in explaining aspects of exploration in humans and animals[13–19] but also in solving complex machine learning tasks with sparse (extrinsic) rewards[20–24]. Despite their successes, these algorithms face an important challenge: Intrinsically motivated agents are prone to distraction by reward-independent stochasticity[25], i.e., they are attracted to novel, surprising, or just noisy states independent of whether or not these states are rewarding[25–27].

The extent of distraction varies between different algorithms, and designing efficient noise-robust algorithms is an ongoing line of research in machine learning[25]. In particular, it is well-known that agents seeking information-gain eventually lose their interest in stochasticity when they realize that there is no information to gain, whereas agents seeking surprise or novelty exhibit a persistent attraction by stochasticity[24,25,28,29]. 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 proceed in 3 steps. First, we design an experimental paradigm that brings ideas from machine learning[25–27] to behavioral neuroscience: We add a reward-independent stochastic part to an experimental paradigm where intrinsically motivated algorithms were previously successful to describe human behavior[19]. Second, we simulate three different intrinsically motivated algorithms (i.e., driven by novelty, surprise, and information-gain) and show that they exhibit different levels of distraction by the stochastic part. Third, the models predict that, after the 1st encounter of a reward, the remaining degree of distraction correlates with the relative importance of intrinsic motivation versus extrinsic reward in action-selection. We hypothesize that this balance is governed in humans by how optimistic they are about finding a higher reward. To test this hypothesis, we introduce a novel reward-manipulation in our experiment that influences the degree of 'reward optimism' in human participants. With this manipulation, we show that the behavior of human participants is both qualitatively and quantitatively consistent with that of novelty-driven RL agents: Human participants exhibit a persistent distraction by novelty, and the degree of this distraction correlates with their degree of reward optimism.

Our results provide evidence for (i) novelty-driven RL algorithms as models of human curiosity even when novelty is distracting and (ii) the influence of reward optimism on the relative importance of novelty-seeking versus reward-seeking in human decision making.

Results

Experimental Paradigm

We first design an experimental paradigm for human participnats that allows us dissociate predictions of different intrinsically motivated RL algorithms. We employ a sequential decision-making paradigm[30–32] for navigation in an environment with 58 states plus three goal states (Fig. 1A-B).

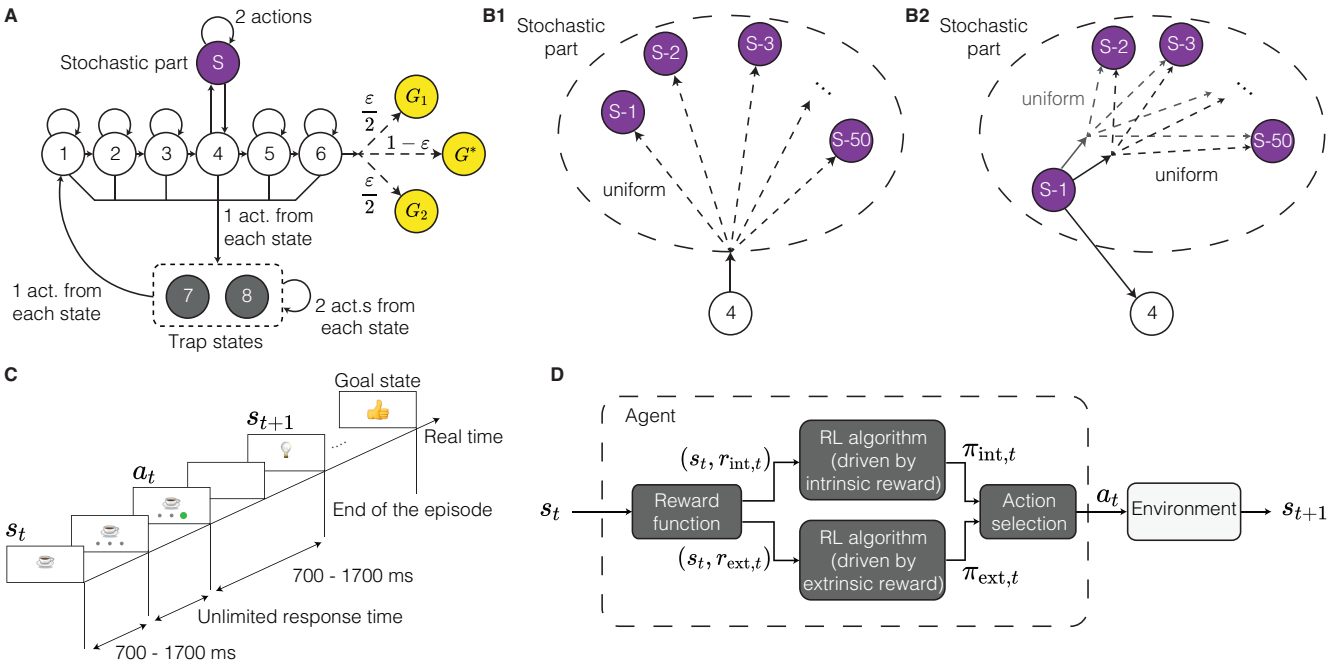

Figure 1: **Experimental paradigm and computational model. A.** Structure of the environment with the stochastic states merged together (purple circle). Each circle represents a state and each arrow an action. All actions except for the ones to the stochastic part or to the goal states are deterministic. Dashed arrows indicate random transitions – values (e.g., $1-\varepsilon$) show probabilities of each transition. We choose $\varepsilon \ll 1$ (c.f. Methods). **B.** Structure of the stochastic part of the environment (states S-1 to S-50). **B1.** In state 4, one action takes agents randomly (with uniform distribution) to one of the stochastic states. **B2.** In each stochastic state, 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., the thumb image in this example) is seen. **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, surprise, or information-gain) and the extrinsic reward $r_{\text{ext},t}$ (i.e., the extrinsic 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 next action a_t is made by a combination of both policies. See Methods for details.

Three actions are available in each non-goal state, and 'agents' can move from one state to another by choosing these actions (arrows in Fig. 1A-B), where we use the term 'agents' to refer to either human participants or participants simulated using RL algorithms. In the human experiments, states are represented by images on a computer screen and actions by three disks below each image (Fig. 1C); for simulated participants, both states and actions are abstract entities (i.e., we consider RL in a tabular setting[33]). 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 Fig. 1A), and their task is to move to any of the three 'goal' states 5 times. Importantly, neither human participants nor simulated participants 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 Fig. 1A), trap states (7 and 8 in Fig. 1A), and stochastic states (S-1 to S-50 in Fig. 1B). In each progressing state, one action ('progressing' action) takes agents one step closer to the goals and another 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 at 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 (Fig. 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 (Fig. 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.

Except for the stochastic states and the multiple goal states, the environment is similar to an earlier paradigm[19] in which a novelty-seeking RL algorithm explains the exploratory behavior of human participants better than alternative algorithms. The stochastic part of the environment is inspired by the work of Burda et al. (2019)[26] on curiosity-driven RL algorithms in machine learning. Without the stochastic part, intrinsic motivation helps agents to avoid the trap states and find the goal[19], hence it helps exploration before and does not harm exploitation after finding a goal. With the new stochastic part, we want to quantify (i) whether initial exploration before finding a goal is biased towards the stochastic part and (ii) whether exploitation of the discovered goal is influenced by a remaining 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 chose a small enough ε (Fig. 1A) to safely assume that all agents visit only G^* while being aware that G_1 and G_2 exist (Methods).

Simulating intrinsically motivated agents

To formulate predictions for human behavior, we simulate three intrinsically motivated RL algorithms. Agents simulated by each algorithm are able to navigate in an environment with an unknown number of states by seeking a combination of extrinsic and intrinsic rewards (Fig. 1D; c.f. Methods for details): At each time t, the 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[34–36] to propose a policy $\pi_{\text{ext},t}$ that maximizes future extrinsic rewards and $\pi_{\text{int},t}$ that maximizes future intrinsic rewards[19,25], respectively. The two policies are combined into a hybrid policy for taking the next action a_t, controlled by a set of free parameters (c.f. Methods). The degree of exploration is high if $\pi_{\text{int},t}$ dominates $\pi_{\text{ext},t}$ during action-selection.

For the simulated agents, all three goals have the same reward of $r_{\text{ext},t} = 1$ (c.f. Methods). For the intrinsic reward $r_{\text{int},t}$, we choose one option from each of the three main categories of intrinsic motivation in machine learning[10,25]: (i) novelty[20–22] quantifies how infrequent the state s_t has been

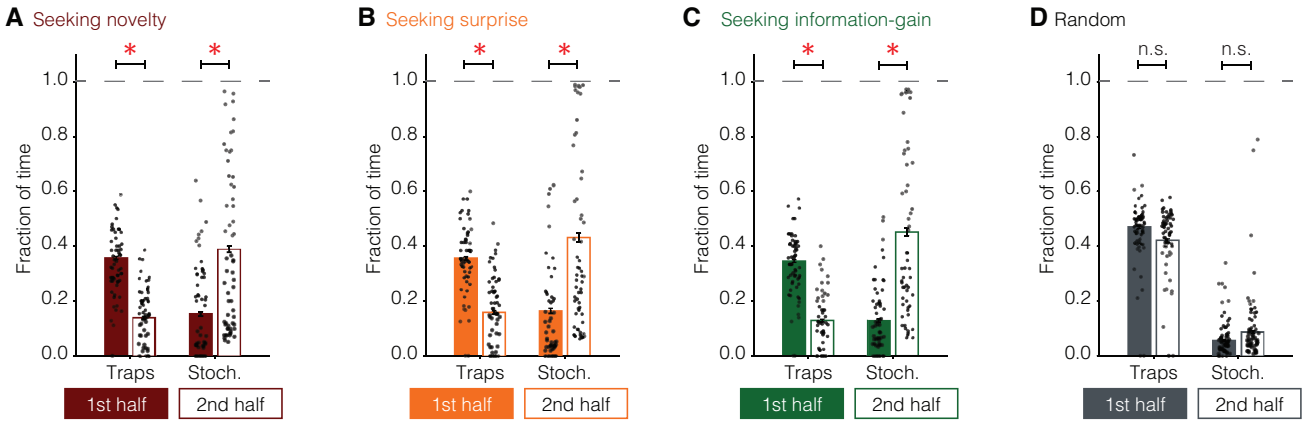

Figure 2: **During episode 1, simulated participants learn to avoid traps and get attracted to the stochastic part of the environment.** Fraction of time spent in the trap states and stochastic part of the environment (including state 4) during the 1st and 2nd half of episode 1 for seeking novelty (**A**), seeking surprise (**B**), seeking information-gain (**C**), and random exploration (**D**). Error bars show the standard error of the mean (SEM) over 500 simulations for each algorithm and single dots the data of 60 (randomly chosen out of 500) individual simulations. Red asterisks: significant differences (p-value ≤ 0.05) with a statistical power $\geq 80\%$, given a sample-size of 60 participants. The differences between the 1st and 2nd half are significant for intrinsically motivated algorithms but not random exploration.

until time t; (ii) surprise[23,26] quantifies how unexpected it is to observe state s_t after taking action a_{t-1} at state s_{t-1}; and (iii) information-gain[37–39] 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.

The three different intrinsic reward signals lead to three intrinsically motivated algorithms and to three groups of simulated participants: those (i) seeking novelty, (ii) seeking surprise, and (iii) seeking information-gain. We also run a random action-selection policy (randomly choosing each action with 1/3 probability) for control. In the following three sections, we formulate the predictions of different intrinsically motivated RL algorithms for human behavior by characterizing the behavior of these simulated participants in two different phases of the experiment, i.e., before (episode 1) and after (episode 2-5) finding a goal.

During episode 1, simulated participants learn to avoid traps and get attracted by stochasticity

We first focus on the behavior of simulated participants in episode 1 (Fig. 2). We optimize the parameters of each algorithm to have the lowest number of actions in episode 1 (i.e., most efficient exploration) and use these parameters to simulate 500 participants for each algorithm (Methods).

For each simulated participant, we split episode 1 into two equal parts and compare the number of state visits in the 1st and 2nd half (Fig. 2): In the 2nd half of episode 1, all three groups of intrinsically motivated simulated participants spend on average less time in the trap states and more time in the stochastic part than in the 1st half (Fig. 2A-C). For simulated participants with

random exploration, no change can be observed; an agent that randomly explores the environment spends only a small proportion of its time in the stochastic part and nearly 50% of time in the trap states (Fig. 2D).

Despite using the same set of parameters within each group of 500 simulated participants, we observe a substantial variability between different simulations (single dots in Fig. 2). The observed variability is due to stochastic action-selection by each agent and stochastic transitions in the environment. When considering the behavior of simulated participants as qualitative predictions for human behavior, we use the observed variability in simulations as an estimate of the variability in the behavior of human participants. We note, however, that there are additional sources of between-participant randomness in humans (e.g., different learning speed or different fatigue levels) that are not captured by our simulations; hence our estimate should be interpreted as a *lower*

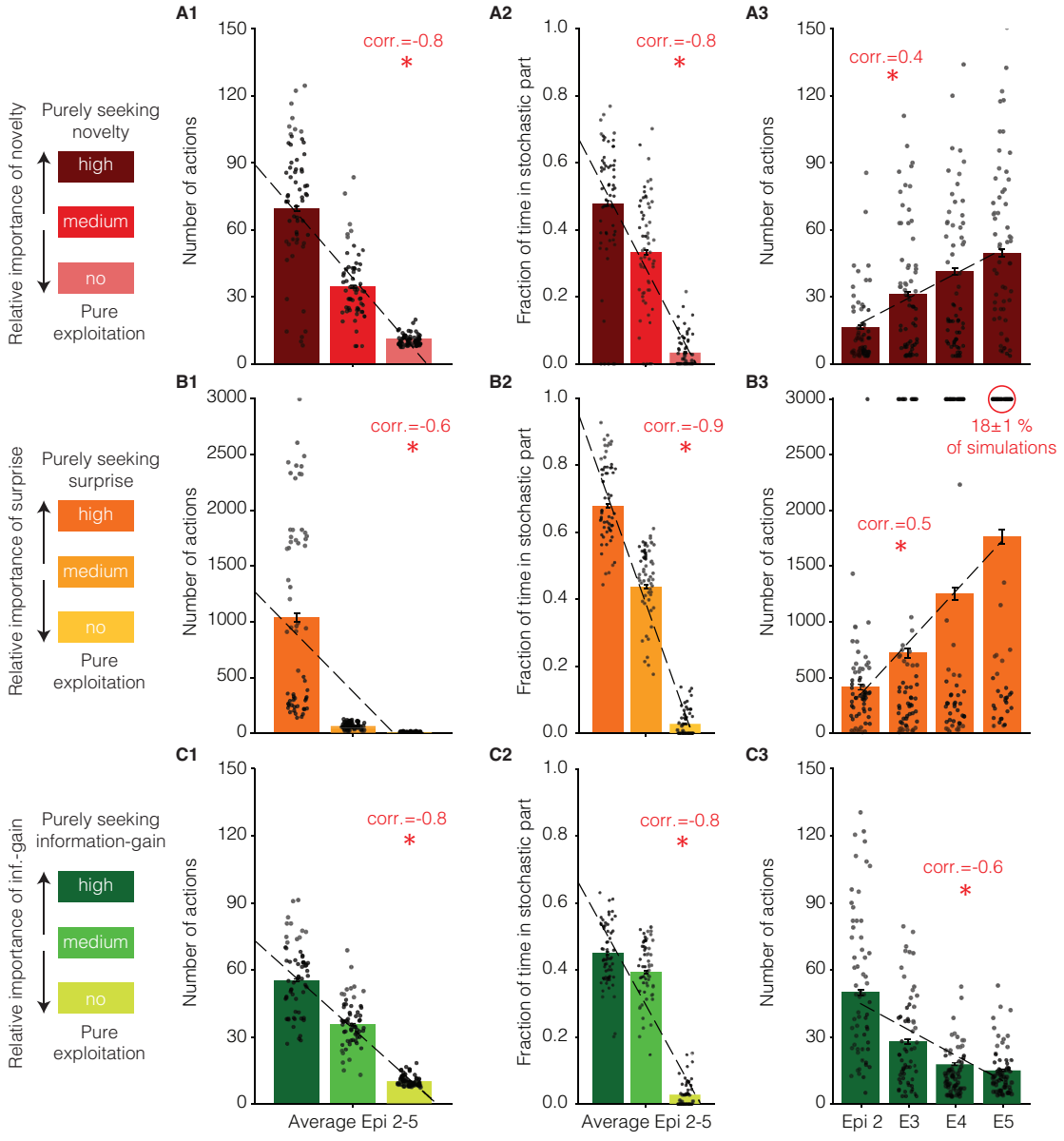

Figure 3: (Caption next page.)

Figure 3: (Previous page.) **In episodes 2-5, seeking information-gain is most and seeking surprise is least robust against distraction by stochasticity.** We consider three relative levels of importance for intrinsic motivation: high (dark colors), medium (shaded colors), and no (light colors). For each level, we run 500 simulations of each algorithm. **A1-C1.** Average number of actions in episode 2-5 for seeking novelty (A1), seeking surprise (B1), and seeking information-gain (C1). Error bars show the SEM and single dots the data of 60 (randomly chosen out of 500) individual simulations. Simulations stopped after 3000 actions, even if a goal state was not reached. The Pearson correlation between the search duration and the relative importance of the intrinsic motivation is negative. **A2-C2.** Average fraction of time spent in the stochastic part (including state 4) of the environment during episodes 2-5. We find a negative Pearson correlation between the fraction of time spent in the stochastic part and the relative importance of the intrinsic motivation. **A3-C3.** Average number of actions in episode 2-5 for simulated participants *purely* seeking novelty (A3), seeking surprise (B3), and seeking information-gain (C3). The Pearson correlation between the search duration and episode number is positive for seeking novelty or surprise but is negative for seeking information-gain. Red asterisks: significant correlations (p-value ≤ 0.05) with a statistical power $\geq 80\%$, given a sample size of 20 participants for each importance level of intrinsic motivation.

bound for the variability in the behavior of human participants. We use this lower bound for power analysis[40] (red asterisks in Fig. 2).

In episodes 2-5, seeking information-gain is most and seeking surprise is least robust against distraction by stochasticity

After finding the goal G^* for the 1st time, an agent in episode 2-5 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 $\pi_{\text{ext},t}$ (Fig. 1D): pure exploitation (no intrinsic motivation), pure exploration (no reward-seeking), and a mixture of both (different shades of each color in Fig. 3). Since the action-choices with the random policy do not vary from one episode to another and remain the same as in Fig. 2D, we exclude random exploration from our further analyses.

For all three groups of intrinsically motivated simulated participants, decreasing the relative importance of intrinsic motivation decreases both the search duration (Fig. 3A1-C1) and the fraction of time spent in the stochastic part (Fig. 3A2-C2). This observation implies that intrinsically motivated exploration leads to an attraction to the stochastic part, effectively keeping the simulated participants away from the goal region beyond state 6 (Fig. 1A). Our results thus confirm earlier findings in machine learning[25,26] that intrinsically motivated agents get distracted by noisy reward-independent stimuli.

While all three groups of simulated participants get distracted by the stochastic part, their degree of distraction is different (different colors in Fig. 3A3-C3). For simulated participants that purely seek information-gain (i.e., no exploitation), the time spent in the stochastic part decreases over episodes (Fig. 3C3), whereas we observe the opposite pattern for simulated participants that purely seek novelty (Fig. 3A3) or surprise (Fig. 3B3). In particular, simulated participants that purely

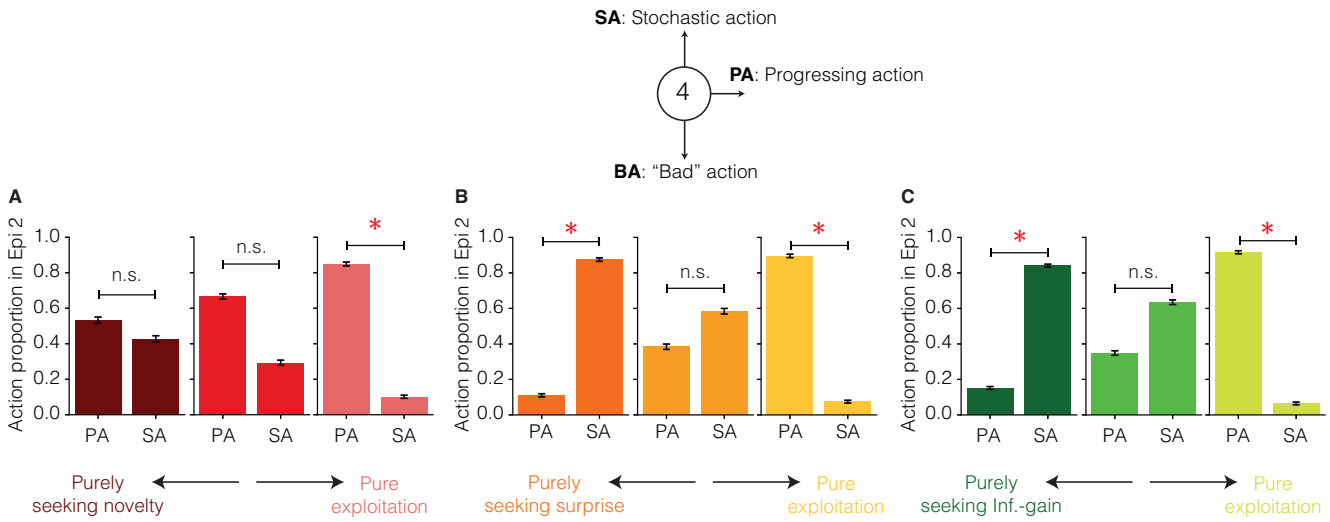

Figure 4: **In episode 2, seeking novelty is more robust against distraction by stochasticity than other intrinsically motivated algorithms.** Fraction of time taking the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2 for three different importance levels of intrinsic motivation – same data, color code, and shading of bars as in Fig. 3. Purely seeking novelty (**A**) shows a comparable preference for SA and PA in state 4, whereas purely seeking surprise (**B**) or information-gain (**C**) leads to a higher preference for SA than PA. Error bars show the SEM. Red asterisks: significant differences (p-value ≤ 0.05) with a statistical power $\geq 80\%$, given a sample size of 20 participants for each importance level of intrinsic motivation.

seek surprise get often (i.e., in $\sim 20\%$ of simulations in episode 5) stuck in the stochastic part and do not escape it within 3000 actions (Fig. 3B3). These observations confirm the inefficiency of seeking surprise and the efficiency of seeking information gain in dealing with noise[25,29].

In episode 2, novelty-seeking selects fewer stochastic actions than other algorithms

We next focus on episode 2 (Fig. 4): In order to further dissociate predictions of different algorithms, we analyze the action preferences of simulated participants in state 4.

For all three groups of simulated participants, increasing the relative importance of intrinsic motivation increases their preference for the stochastic action (Fig. 4). Moreover, for the highest level of intrinsic motivation, the probability of choosing the progressing action is lower than that of choosing the stochastic action for those simulated participants seeking surprise or information-gain (Fig. 4B-C) but not for those seeking novelty (Fig. 4A).

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; c.f. 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 a novelty-seeking agent) compared to most stochastic states, whereas there are many actions between the stochastic states that have potentially never been chosen and are, thus, attractive for an agent seeking surprise or information-gain.

Human participants

To test the predictions of intrinsically motivated algorithms, we compare the behavior of human participants with that of simulated participants. For simulated participants, the relative importance of the intrinsic motivation for action-selection (Fig. 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 motivation compared to extrinsic reward. Inspired by earlier studies[41–43], 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 relative importance of intrinsic motivation is positively correlated with the degree of 'reward optimism', where we define reward optimism as the expectancy of finding a goal of higher value than those already discovered (c.f. ref.[44] for a review of studies on optimism in psychology). Based on this hypothesis, we introduce 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 (c.f. 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 Fig. 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 explore further in episodes 2-5 (= a low degree of reward optimism), whereas participants who are assigned to an environment with 2 CHF reward value for G^* are likely to keep searching for more valuable goals in episodes 2-5 (= a high degree of reward optimism). Therefore, we have three different groups of participants with three different levels of reward optimism in episodes 2-5.

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 paper, 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^*. 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).

Human participants exhibit a persistent distraction by the stochastic part

We performed the same series of analyses on the behavior of human participants as on that of simulated participants (Fig. 5): Human participants exhibit the same qualitative characteristics in their behavior as those observed for simulated participants seeking novelty (compare Fig. 5 with Fig. 2-Fig. 4).

First, in episode 1, human participants spend less time in the trap states during the 2nd than during the 1st half of episode 1 and more time in the stochastic part of the environment (Fig. 5A). This behavior is consistent with the predictions of intrinsically motivated algorithms but not with those of random exploration (Fig. 2).

Second, in episodes 2-5, increasing the degree of reward optimism influences the behavior of human

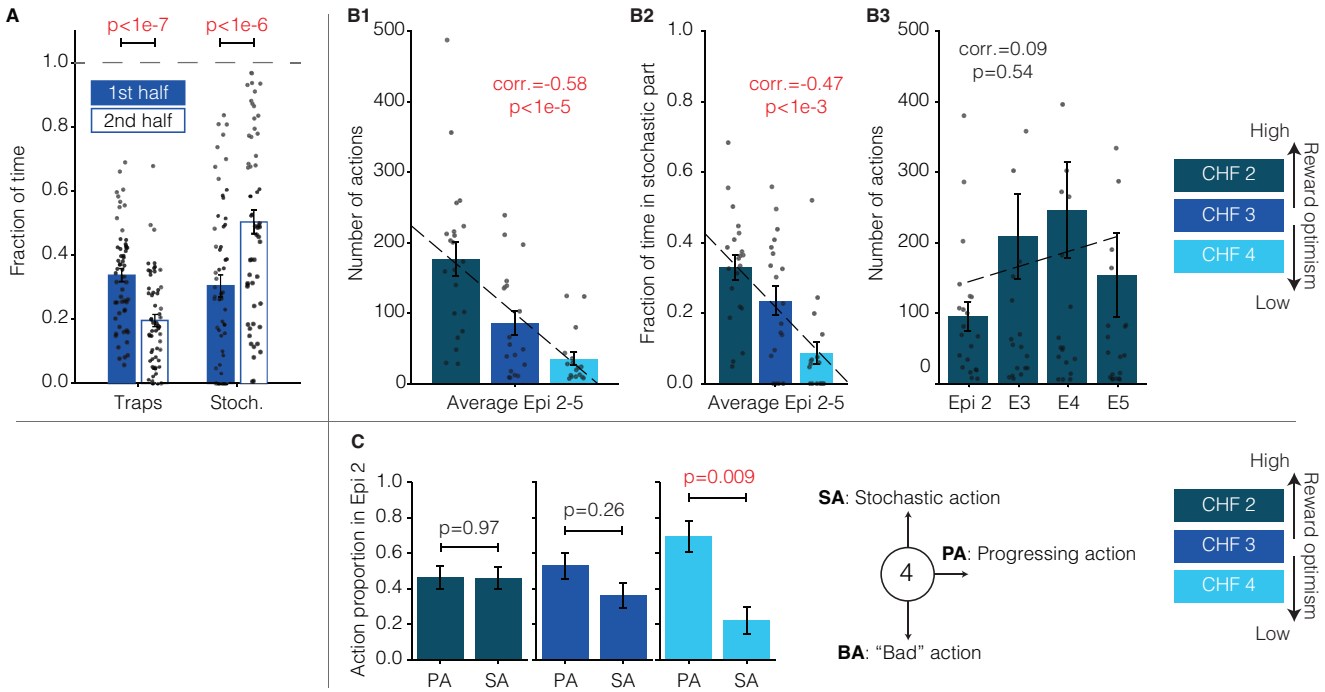

Figure 5: **The behavior of human participants shows characteristics similar to that of simulated participants seeking novelty. A.** Average fraction of time spent in different parts of the environment, separately in the 1st (filled bars) and the 2nd (open bars) half of episode 1 (c.f. Fig. 2). Since there is no difference between the three groups of participants in episode 1, all statistics are computed over all 57 participants. **B.** Search duration in episodes 2-5. **B1.** Average number of actions in episode 2-5 for the three different groups – dark: 2 CHF, medium: 3 CHF, light: 4 CHF (c.f. Fig. 3A1-C1). **B2.** Average fraction of time spent in the stochastic part of the environment (including state 4) during episodes 2-5 (c.f. Fig. 3A2-C2). **B3.** Average number of actions in episode 2-5 for the 2 CHF group (c.f. Fig. 3A3-C3). No significant correlation was observed between the search duration and episode number. **C.** Fraction of time choosing the progressing action (PA) and the stochastic action (SA) when encountering state 4 during episode 2 (c.f. Fig. 4); c.f. Supplementary Materials for other progressing states. Error bars show the SEM and single dots the data of individual participants. Red p-values: significant differences and correlations (controlled for multiple testing at a False Discovery Rate of 0.05[45]); see Table 1 in Methods for detailed statistics.

participants in the same way as increasing the relative importance of intrinsic motivation influences the behavior of simulated participants (Fig. 5B vs. Fig. 3): The search duration (Fig. 5B1) and the fraction of time spent in the stochastic part (Fig. 5B2) in episodes 2-5 are decreasing functions of the goal value, e.g., the 2 CHF group has longer search duration and spends more time in the stochastic part than the other two groups. This supports our hypothesis that increasing the degree of reward optimism increases the relative importance of intrinsic motivation against extrinsic reward. Importantly, we observe for the 2 CHF group only a non-significant increase of the search duration over episodes 2-5 (Fig. 5B3). This implies that the 2 CHF group (i.e., the most optimistic group of participants) persistently explore the stochastic part. This behavior is qualitatively more similar to the behavior of simulated participants seeking novelty or surprise than the behavior of those seeking information-gain (Fig. 3A3-B3).

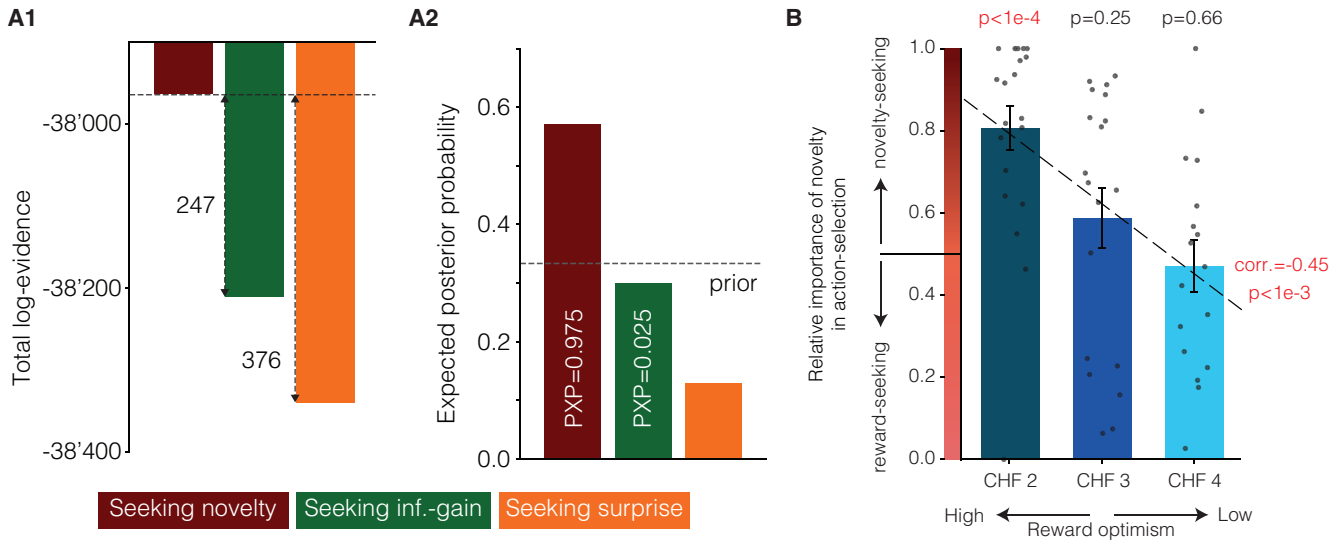

Figure 6: **Fitting intrinsically motivated RL algorithms to the behavioral data. A.** Human participants' action choices are best explained by novelty-seeking. **A1.** Model log-evidence (computed using the Schwarz approximation[46]) summed over all participants (i.e., assuming that different participants have the same exploration strategy but can have different parameters[47]) is significantly higher for seeking novelty than seeking information-gain or surprise. High values indicate good performance, and differences greater than 3 or 10 are traditionally[48] considered as significant or strongly significant, respectively. **A2.** The expected posterior model probability with random effects assumption[49] (i.e., assuming that different participants can have different exploration strategies and different parameters) given the data of all participants. PXP stands for Protected Exceedance Probability[50], i.e., the probability of one model being more probable than the others. **B.** The relative contribution of novelty-seeking (i.e., dominance of $\pi_{\text{int},t}$ over $\pi_{\text{ext},t}$; c.f. Fig. 1D) in episode 2-5 is computed for each participant after fitting the model to data – c.f. Eq. 15 in Methods. Color-bar on the left shows how the y-axis relates to the relative importance of novelty in Fig. 3-Fig. 4. Error bars show the SEM and single dots the data of individual participants. P-values on top of each bar corresponds to a one-sample t-test comparing the average with 0.5. Red p-values: significant differences and correlations (controlled for multiple testing at a False Discovery Rate of 0.05[45]); see Table 1 in Methods for detailed statistics.

Third, in episode 2, decreasing the degree of reward optimism increases the preference of human participants for the progressing action in state 4 (Fig. 5C). Importantly, the 2 CHF group chooses the progressing and the stochastic actions about equally often (Fig. 5C). This observation is consistent with the behavior of simulated participants seeking novelty but inconsistent with that of simulated participants seeking surprise or information-gain (Fig. 5C vs. Fig. 4).

Novelty-seeking is the most probable model of human exploration

In the previous section, we observed that human participants exhibit the same *qualitative* patterns of behavior as novelty-seeking simulated participants. We now use Bayesian model-comparison[47,49] to *quantitatively* test whether human behavior is explained better by seeking novelty than seeking surprise or information-gain (Methods).

To have an accurate model-comparison, we use, in the following, a more realistic model of behavior than that of simulations in Fig. 2-Fig. 4: We use a hybrid RL model[19,32,51,52] combining model-based planning[36] and model-free habit-formation[53] and allow our algorithms to assign different extrinsic reward values to different goal states (Methods). Our model-comparison results show that seeking novelty is the most probable model for the majority ($\sim$ 60%) of human participants, followed by seeking information-gain as the 2nd most probable model for $\sim$ 30% of human participants (Fig. 6A) – c.f. Supplementary Materials for model-recovery[54].

We next use our winning novelty-seeking algorithm with its parameters fitted to behavioral data and simulate new participants – i.e., we perform posterior predictive checks[54,55]. The newly simulated participants (c.f. Supplementary Materials) exhibit the same *qualitative* pattern of behavior as human participants (Fig. 5) and efficient novelty-seeking simulated participants (Fig. 2-Fig. 4). Importantly, the search duration of the newly simulated participants with highest level of reward optimism does not change over episodes 2-5. This behavior more closely matches the behavior of human participants (Fig. 5B3) compared to that of efficient simulated participants (Fig. 3A3). Moreover, the estimated parameters of our novelty-seeking algorithms confirms reliance of human participants on habit-formation instead of planning (Supplementary Materials), consistent with previous works on similar experimental paradigms[19,32].

Taken together, our results provide strong evidence for novelty-seeking combined with habit-formation as a model of human behavior in our experiment.

Reward optimism correlates with relative importance of novelty in episodes 2-5

Since novelty-seeking turns out to be the most probable model of human behavior, we can now explicitly test our hypothesis that reward optimism increases the relative importance of intrinsic motivation. 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 (Fig. 6B). Moreover, the participants in the 2 CHF group continue with a almost fully exploratory policy in episodes 2-5 indicating that they have only a small bias towards exploiting the small but known reward (Fig. 6B). See Supplementary Materials for parameter recovery[54].

Discussion

We designed a novel experimental paradigm to study human curiosity-driven exploration in the presence of stochasticity and to test and compare predictions of three different intrinsically motivated RL algorithms as models of this exploration. We made three main observations: (i) Human participants exhibit a persistent distraction by the novelty signal in the stochastic part of the environment; (ii) seeking novelty explains this persistent distraction of human participants better than seeking surprise or information-gain, even though seeking information-gain is more robust in dealing with noise; and (iii) optimism about finding a higher value of reward in the environment increases the level of this distraction in humans in the same way as the relative importance of novelty increases the level of distraction in RL algorithms.

How humans deal with the exploration-exploitation trade-off has been a long-lasting question in neuroscience and psychology[56,57]. Experimental studies have shown that humans use a combination of random and directed exploration[14,17], potentially linked to different neural mechanisms[58–60]. However, there are multiple distinct theoretical models to describe directed exploration[5,12,25,61–63], and it has been debated which one is best suited to explain human behavior. Here, we focused on the class of intrinsically motivated RL algorithms and designed an experimental paradigm that enabled us to dissociate predictions of three main models of intrinsic motivation in neuroscience[13,18,19,64,65] and psychology[16,17,66,67]. Our 1st and 2nd observations above support novelty-driven RL algorithms as models of human curiosity-driven exploration even in situations where seeking novelty is not optimal for exploration – visible as a persistent distraction by reward-independent stochasticity.

Our 2nd observation implies that humans are not optimal in dealing with stochasticity and challenges the idea of defining curiosity as a normative solution to the exploration-exploitation trade-off[3,7,66]. However, we note that computing novelty is computationally much cheaper than computing information-gain (c.f. Methods). This implies that a potentially higher level of distraction by noise in humans may be just the price of spending less computational power – an idea consistent with the 'resource rationality' hypothesis[68,69]. This observation also implies that finding more robust exploration algorithms in machine learning does not necessarily help us to find better models of human exploration.

The core assumption of using intrinsically motivated algorithms as models of human curiosity is the existence of an intrinsic reward parallel to the extrinsic reward. There are, however, multiple ways to incorporate intrinsic rewards into the RL framework[25]. The common practice is to use a weighted sum of the intrinsic and the extrinsic reward as a single scalar reward signal driving action-selection[9,20,26]. An alternative approach is to treat different reward signals in parallel and compute separate action-policies which are combined to drive action-selection later in the processing stream[19,70,71]. The latter approach provides higher flexibility to arbitrate between different policies and is hence more compatible with the rapid change of behavior observed in our human participants that is triggered by the reward value they find in episode 1 (Fig. 5B1-2 and Supplementary Materials); this approach is also more consistent with experimental evidence for partially separate neural pathways of novelty- and reward-induced behaviors[19,65,72–75]. Moreover, our 3rd observation shows that the relative importance of different policies is regulated by the degree of reward optimism – in line with the known influence of environmental variables on the preference for novelty[41,42,74] and curiosity-driven behavior[76,77]. A potential future direction to further investigate how reward optimism influences the interplay of novelty and reward is to study the behavior of human participants in our environment with other reward distributions (e.g., including negative reward values) or even when extrinsic reward is independent of participants' action choices (similar to ref.[16,65,78]).

Optimism in psychology has been defined as a 'variable that reflects the extent to which people hold generalized favorable expectancies for their future'[44] and has been linked to several neural and behavioral characteristics[44,79,80]. While the traditional approach to measure optimism is through self-tests[81], more recently statistical inference using RL[82] and Bayesian[83,84] models of behavior have been proposed to quantify variables correlated with traditional measurements. In the machine learning community, on the other hand, seeking intrinsic rewards has been interpreted as an 'optimistic policy' for exploration[61]. Our 3rd observation shows that the preference for an intrinsic reward is indeed correlated with a notion of optimism – i.e., the expectancy of finding a goal

of higher value in episodes 2-5 (different colors in Fig. 6B). We speculate that the within-group differences in the preference for novelty among our participants (different points for each bar in Fig. 6B) may be explained by the individual-specific level of optimism that could be quantified by self-tests. Moreover, we note that persistent distraction by the stochastic part of the environment observed in our participants (Fig. 5B3) is consistent with the known phenomena of optimism bias[85] and optimistic belief updating in humans[79,86,87].

An alternative type of intrinsic motivation that we did not systematically study here is 'empowerment', a measure of controllability over the environment[63]. The empowerment theory predicts that, in the environment of our experiment (Fig. 1A-B), humans would avoid the stochastic part where they do not have any control over the outcome of their actions. This prediction is inconsistent with our 1st observation. We note, however, that our results do not exclude the possibility that seeking empowerment explains well the behavior of humans in environments different from ours or for tasks with a goal other than searching for a rewarding state[88].

In conclusion, our results show (i) that human decision-making is influenced by an evident interplay of intrinsic motivation with extrinsic rewards that is controlled by reward optimism and (ii) that novelty-seeking RL algorithms can successfully model this interplay.

Methods

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).

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 naive 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[89].

Before starting the experiment, the participants were informed that they need to find either one of the 3 goal states for 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'. At each trial, participants were presented an image (state) and three grey disks below the image (Fig. 1C). Clicking on a disk (action) led participants to a subsequent image which was chosen based on the underlying graph of the environment in Fig. 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 (Fig. 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 (Fig. 1B1) and the other two disks randomly to the bad and progressing actions, respectively (Fig. 1A). With this assignment, we made sure that all human participants would visit the stochastic part at least once during episode 1.

Before the start of 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 Fig. 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 assignment 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. We considered ε to be around machine precision 10^{-8}, so we have $(1-\varepsilon)^{5\times 61} \approx 1 - 10^{-5} \approx 1$, meaning that all 61 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 for 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.

	Test	Statistic	95% CI	DF	P value
Fig. 5A - Traps	one-sample t-test	$t = 6.36$	$(0.097, 0.186)$	56	$< 10^{-7}$
Fig. 5A - Stoch.	one-sample t-test	$t = -5.61$	$(-0.27, -0.13)$	56	$< 10^{-6}$
Fig. 5B1	corr. test	$t = -5.3$	$(-0.73, -0.39)$	55	$< 10^{-5}$
Fig. 5B2	corr. test	$t = -3.98$	$(-0.65, -0.24)$	55	0.0002
Fig. 5B3	one-sample t-test on individual corr.s	$t = 0.63$	$(-0.20, 0.37)$	20	0.54
Fig. 5C - 2 CHF	one-sample t-test	$t = 0.039$	$(-0.25, 0.26)$	20	0.97
Fig. 5C - 3 CHF	one-sample t-test	$t = 1.17$	$(-0.13, 0.47)$	18	0.26
Fig. 5C - 4 CHF	one-sample t-test	$t = 2.99$	$(0.14, 0.81)$	16	0.009
Fig. 6B - Corr.	corr. test	$t = -3.77$	$(-0.64, -0.22)$	55	0.0004
Fig. 6B - 2 CHF	one-sample t-test	$t = 5.67$	$(0.69, 0.92)$	20	$< 10^{-4}$
Fig. 6B - 3 CHF	one-sample t-test	$t = 1.18$	$(0.43, 0.74)$	18	0.25
Fig. 6B - 4 CHF	one-sample t-test	$t = -0.45$	$(0.34, 0.61)$	16	0.66

Table 1: **Detailed statistics for null hypothesis testing in Fig. 5 and Fig. 6**. All tests are two-sided, and all correlations are Pearson correlation; we found similar results and the same conclusions for the Spearman correlation. P-value threshold for significant effects by controlling the False Discovery Rate at 0.05[45] = 0.029. CI = Confidence Interval. DF = Degrees of freedom.

The details of statistical tests in Fig. 5 and Fig. 6 are reported in Table 1.

Full hybrid model

We first present the most general case of our algorithm as visualized in Fig. 1D and then explain the special cases used for Fig. 2-Fig. 4 and Fig. 6. We used ideas from non-parametric Bayesian inference[90–92] to generalize an earlier algorithm[19] to the case that the total number of states is unknown. We present the final results here and present the derivations and pseudo-code in Supplementary Materials.

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

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

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. In the most general case, $\{\tilde{G}_0, \tilde{G}_1, \tilde{G}_2\}$ 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

$$R_{\text{ext}}(s, a \to s') = \delta_{s',\tilde{G}_0} + r_1^* \delta_{s',\tilde{G}_1} + r_2^* \delta_{s',\tilde{G}_2}, \tag{2}$$

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. We discuss choices of $R_{\text{int},t}$ in the next section.

As a general choice for the RL algorithm in Fig. 1D, we consider a hybrid of model-based and model-free policy[19,32,52,53]. 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 Fig. 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[33], thus the MF component of our algorithm remains similar to that of previous studies[19,31]: 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:

$$\begin{aligned} RPE_{\text{ext},t+1} &= r_{\text{ext},t+1} + \lambda_{\text{ext}} V^{(t)}_{\text{MF,ext}}(s_{t+1}) - Q^{(t)}_{\text{MF,ext}}(s_t, a_t) \\ RPE_{\text{int},t+1} &= r_{\text{int},t+1} + \lambda_{\text{int}} V^{(t)}_{\text{MF,int}}(s_{t+1}) - Q^{(t)}_{\text{MF,int}}(s_t, a_t), \end{aligned} \tag{3}$$

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 $V^{(t)}_{\text{MF,int}}(s_{t+1}) = \max_{a'} Q^{(t)}_{\text{MF,int}}(s_{t+1}, a')$ are the extrinsic and intrinsic V-values[33] of the state s_{t+1}, respectively. We use two separate eligibility traces for the update of Q-values, one for extrinsic reward $e_{\text{ext},t}$ and one for intrinsic reward $e_{\text{int},t}$, 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

$$\begin{aligned} e_{\text{ext},t+1}(s, a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_{\text{ext}} \mu_{\text{ext}} e_{\text{ext},t}(s, a) & \text{otherwise} \end{cases} \\ e_{\text{int},t+1}(s, a) &= \begin{cases} 1 & \text{if} \quad s = s_t, a = a_t \\ \lambda_{\text{int}} \mu_{\text{int}} e_{\text{int},t}(s, a) & \text{otherwise}\,, \end{cases} \end{aligned} \tag{4}$$

where λ_{ext} and λ_{int} are the discount factors defined above, and μ_{ext} and $\mu_{\text{int}} \in [0, 1]$ are the decay factors of the eligibility traces[33] for the extrinsic and intrinsic rewards, respectively. The update rule is then $\Delta Q^{(t+1)}_{\text{MF}}(s, a) = \rho e_{t+1}(s, a) RPE_{t+1}$, where e_{t+1} is the eligibility trace (i.e., either $e_{\text{ext},t+1}$ or $e_{\text{int},t+1}$), RPE_{t+1} is the reward prediction error (i.e., either $RPE_{\text{ext},t+1}$ or $RPE_{\text{int},t+1}$), 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[93,94]:

$$\tilde{C}^{(t+1)}_{s,a,s'} = \begin{cases} \kappa \tilde{C}^{(t)}_{s,a,s'} + \delta_{s',s_{t+1}} & \text{if } s = s_t \,,\; a = a_t \\ \tilde{C}^{(t)}_{s,a,s'} & \text{otherwise,} \end{cases} \tag{5}$$

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 model-building in humans. If $\kappa = 1$, then $\tilde{C}^{(t+1)}_{s,a,s'}$ is the exact count of transition $(s,a) \to s'$. Then, these counts are used to estimate the transition probabilities

$$p^{(t)}(s'|s,a) = \begin{cases} \frac{\epsilon_{\text{obs}} + \tilde{C}^{(t)}_{s,a,s'}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if} \quad s' \in \mathcal{S}^{(t)} \,, \\ \frac{\epsilon_{\text{new}}}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}} & \text{if} \quad s' = s_{\text{new}} \,, \end{cases} \tag{6}$$

where $\tilde{C}^{(t)}_{s,a} = \sum_{s'} \tilde{C}^{(t)}_{s,a,s'}$ is the counts of taking action a at state s, $\epsilon_{\text{obs}} \in \mathbb{R}^+$ is a free parameter for the prior probability of transition to a known state (i.e., states in $\mathcal{S}^{(t)}$), and $\epsilon_{\text{new}} \in \mathbb{R}^+$ is a free parameter for the prior probability of transition to a new state (i.e., states not in $\mathcal{S}^{(t)}$) – c.f. Supplementary Materials for derivations. Choosing $\epsilon_{\text{new}} = 0$ is equivalent to assuming there is no unknown states in the environment, for which the estimate in Eq. 6 is reduced to the classic Bayesian estimate of transition probabilities in bounded discrete environments[19,32]. The transition probabilities are then used in a novel variant of prioritized sweeping[33–35] 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[33] 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**[33]: The probability of taking action a in state s at time t is

$$\pi_t(a|s) \propto \exp\Big[\beta_{\text{MB,ext}} Q^{(t)}_{\text{MB,ext}}(s,a) \,+ \beta_{\text{MF,ext}} Q^{(t)}_{\text{MF,ext}}(s,a) \,+ \beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) \,+ \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)\Big], \tag{7}$$

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[33]) expressing the contribution of each Q-value to action-selection. For Fig. 1D, we defined

$$\begin{aligned} \pi_{\text{ext},t}(a|s) &\propto \exp\Big[\beta_{\text{MB,ext}} Q^{(t)}_{\text{MB,ext}}(s,a) \,+ \beta_{\text{MF,ext}} Q^{(t)}_{\text{MF,ext}}(s,a)\Big] \\ \pi_{\text{int},t}(a|s) &\propto \exp\Big[\beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) \,+ \beta_{\text{MF,int}} Q^{(t)}_{\text{MF,int}}(s,a)\Big], \end{aligned} \tag{8}$$

and as a result $\pi_t \propto \pi_{\text{ext},t} \cdot \pi_{\text{int},t}$.

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[53,95], cognitive load[96], and whether the location of reward is known[19]. 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:

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.

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 four independent free parameters.

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.

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

$$\begin{aligned}\Phi = \{&r^*_1, r^*_2, Q^{(0)}_{\text{MF,ext}}, Q^{(0)}_{\text{MF,int}}, \lambda_{\text{ext}}, \lambda_{\text{int}}, \mu_{\text{ext}}, \mu_{\text{int}}, \rho, \kappa, \epsilon_{\text{new}}, \epsilon_{\text{obs}}, T_{PS,\text{ext}}, T_{PS,\text{int}},\\ &\beta^{(1)}_{\text{MB,ext}}, \beta^{(2)}_{\text{MB,ext}}, \beta^{(1)}_{\text{MB,int}}, \beta^{(2)}_{\text{MB,int}}, \beta^{(1)}_{\text{MF,ext}}, \beta^{(2)}_{\text{MF,ext}}, \beta^{(1)}_{\text{MF,int}}, \beta^{(2)}_{\text{MF,int}}\},\end{aligned} \tag{9}$$

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 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.

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, e.g., $r_{\text{int},t+1} = R_{\text{int},t}(s_t, a_t \to s_{t+1})$. In this section, we present our 3 choices of $R_{\text{int},t}$.

Novelty: For an agent seeking novelty (red in Fig. 2-Fig. 4 and Fig. 6A), we define the intrinsic reward function as

$$R_{\text{int},t}(s, a \to s') = -\log p^{(t)}_N(s'), \tag{10}$$

where $p^{(t)}_N(s') = \frac{1+C^{(t)}_{s'}}{1+|\mathcal{S}^{(t)}|+t}$ is the state frequency with $C^{(t)}_{s'} = \left|\left\{t' \in \{1, ..., t\} : s_{t'} = s'\right\}\right|$ the number of encounters of state s' up to time t. With this definition, that generalizes earlier works[19] to the case where the number of states is unknown, the least novel states are those that have been encountered most (i.e., with highest $C^{(t)}_{s'}$). Moreover, novelty is at its highest value for the unobserved states as we have $C^{(t)}_{s'} = 0$ for any unobserved state $s' \notin \mathcal{S}^{(t)}$. Similar intrinsic rewards have been used in machine learning[20–22].

Surprise: For an agent seeking surprise (orange in Fig. 2-Fig. 4 and Fig. 6A), we define the intrinsic reward function as the Shannon surprise[29]

$$R_{\text{int},t}(s, a \to s') = -\log p^{(t)}(s'|s, a), \tag{11}$$

where $p^{(t)}(s'|s, a)$ is defined in Eq. 6. With this definition, the expected intrinsic reward of taking action a at state s is equal to the entropy of the distribution $p^{(t)}(.|s, a)$[97]. 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[23,26].

Information-gain: For an agent seeking information-gain (green in Fig. 2-Fig. 4 and Fig. 6A), we define the intrinsic reward function as

$$R_{\text{int},t}(s, a \to s') = \mathrm{D}_{\mathrm{KL}}\Big[p^{(t)}(.|s,a)||p^{(t+1)}(.|s,a)\Big], \tag{12}$$

where D_{KL} is the Kullback-Leibler divergence[97], and $p^{(t+1)}$ is the updated world-model upon observing $(s,a) \to s'$. Note that if $s' \notin \mathcal{S}^{(t)}$, then there are some technical problems in the naive computation of D_{KL} – since $p^{(t)}$ and $p^{(t+1)}$ has different supports. We deal with these problems using a more fundamental definition of D_{KL} using the Radon–Nikodym derivative; c.f. Supplementary Materials for derivations and see ref.[37] for an alternative heuristic solution. Note that the information-gain in Eq. 12 has been also interpreted as a measure of surprise (called 'Postdictive surprise'[98]), but it has a radically different behavior from the Shannon surprise in Eq. 11 – see ref.[29] for an elaborate treatment of the topic. Similar intrinsic rewards have been used in machine learning[24,28,37,39,99].

Efficient model-based planning for simulated participants

For simulations in Fig. 2-Fig. 4, we set $\varepsilon = 0$ (c.f. Fig. 1A) and used a pure MB version of our algorithm with 13 parameters:

$$\{r_1^*, r_2^*, \lambda_{\text{ext}}, \lambda_{\text{int}}, \kappa, \epsilon_{\text{new}}, \epsilon_{\text{obs}}, T_{PS,\text{ext}}, T_{PS,\text{int}}, \beta^{(1)}_{\text{MB,ext}}, \beta^{(2)}_{\text{MB,ext}}, \beta^{(1)}_{\text{MB,int}}, \beta^{(2)}_{\text{MB,int}}\}. \tag{13}$$

We considered perfect model-building by assuming $\kappa = 1$ and almost perfect planning by assuming $T_{PS,\text{ext}} = T_{PS,\text{int}} = 100$. 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)[19]: $\lambda_{\text{ext}} = 0.95$, $\lambda_{\text{int}} = 0.70$, $\epsilon_{\text{new}} = 10^{-5}$ and $\epsilon_{\text{obs}} = 10^{-4}$. To separate the effect of optimistic initialization[33] from seeking intrinsic reward, we assumed the same value of reward for all goals, i.e., $r_1^* = r_2^* = 1$. Finally, we considered $\beta^{(1)}_{\text{MB,ext}} = 0$ to have pure intrinsic reward seeking in episode 1.

After fixing parameter values for 10 out of 13 parameters in Eq. 13, we fine-tuned $\beta^{(1)}_{\text{MB,int}}$ to minimize the average length of episode 1 (to find the goal as fast as possible). Fig. 2 show data of 500 simulations of each algorithm with the optimal choice of $\beta^{(1)}_{\text{MB,int}}$. For episodes 2-5, we first set $\beta^{(2)}_{\text{MB,int}} = 0$ and $\beta^{(2)}_{\text{MB,ext}} = 10$ to have a non-deterministic policy for purely seeking extrinsic reward after the 1st encounter of the goal (the lightest color in Fig. 3 and Fig. 4). Different shades of color in Fig. 3 and Fig. 4 corresponds to different choices of $\omega \in [0,1]$ for $\beta^{(2)}_{\text{MB,int}} = \omega\beta^{(1)}_{\text{MB,int}}$ and $\beta^{(2)}_{\text{MB,ext}} = (1-\omega)\cdot 10$. 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.

For all simulations, we used the same transition matrix (i.e., the same assignment of actions to transitions; Fig. 1A-B) independently of the action-choices of simulated participants. This choice, in contrary to the action-disk assignments in human experiments (c.f. Experimental procedure for human participants), made it possible for simulated participants to never visit the stochastic part during episode one (with a small probability). However, we found the same conclusions and similar results for our simulations even when we restricted our analyses in Fig. 2-Fig. 4 to the data of simulated participants who took the stochastic action at their first encounter of state 4.

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)[19] showing the negligible importance of this parameter. As a result, we had 20 free parameters for each of the three intrinsic motivations (i.e., novelty, surprise, and information-gain). For each intrinsic motivation $M \in \{\text{novelty}, \text{surprise}, \text{inf-gain}\}$ and for each participant $n \in \{1, ..., 57\}$, we estimated the algorithm's parameter by maximizing likelihood of data given parameters:

$$\hat{\Phi}_{n,M} = \arg\max_{\Phi} P(D_n|\Phi, M_n = M) \tag{14}$$

where D_n is the data of participant n, M_n is the intrinsic motivation used by participant n, $P(D_n|\Phi, M_n = M)$ is the probability of D_n being generated by our intrinsically motivated algorithm seeking $M_n = M$ with its parameter equal to Φ, and $\hat{\Phi}_{n,M}$ is the set of estimated parameters that maximizes that probability (c.f. Eq. 9). For optimization, we used Subplex algorithm[100] as implemented in Julia NLopt package[101].

Because all algorithms have the same number of parameters, we considered the maximum log-likelihood as the model log-evidence, i.e., for intrinsic motivation M and participant n, we consider $\log P(D_n|M_n = M) \approx \log P(D_n|\hat{\Phi}_{n,M}, M_n = M)$ – which is equal to a shifted Schwarz approximation of the model log-evidence (also called BIC)[46–48]. Fig. 6A1 shows the total log-evidence $\sum_n \log P(D_n|M_n = M)$. With the fixed effects assumption at the level of models (i.e., assuming that $M_1 = M_2 = ... = M_{57} = M^*$), the total log-evidence $\sum_n \log P(D_n|M_n = M)$ is equal to $\log P(M^* = M|D_{1:57}) + \text{constant}$, where $P(M^* = M|D_{1:57})$ is the posterior probability of M being the intrinsic motivation used by all participants. See ref.[47,54] for tutorials.

We then consider the Bayesian model selection method of ref.[49,50] with the random effects assumption, i.e., assuming that participant n uses the intrinsic motive $M_n = M$, which is not necessarily the same as the one used by other participants, with probability P_M. We performed Markov Chain Monte Carlo sampling (using Metropolis Hasting algorithm[48] with uniform prior and 20 chains of length 10′000) for inference (instead of usual variational inference methods[49,50]) and estimated the joint posterior distribution $P(M_{1:57}, P_{\text{novelty}}, P_{\text{surprise}}, P_{\text{inf-gain}}|D_{1:57})$. Fig. 6A2 shows the expected posterior probability $\mathbb{E}[P_M|D_{1:57}]$ as well as the protected exceedance probabilities $P(P_M > P_{M'} \text{ for all } M' \neq M|D_{1:57})$ computed by using the participant-wise log-evidences.

The boxplots of the fitted parameters of novelty-seeking are shown in Supplementary Materials. The same set of parameters were used to compute the relative importance of novelty in Fig. 6B (c.f. Relative importance of novelty in action-selection). See Supplementary Materials for (i) posterior predictive checks, (ii) model-recovery, and (iii) parameter-recovery[54].

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, independent of the values of the inverse-temperatures, the action is taken by a pure novelty-seeking policy – because the policy in Eq. 7 can be re-written as $\pi^{(t)}(a|s) \propto \exp\big[\beta_{\text{MB,int}} Q^{(t)}_{\text{MB,int}}(s,a) +$

$\beta_{\mathrm{MF,int}} Q^{(t)}_{\mathrm{MF,int}}(s,a)\big]$. 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 $\omega_{\mathrm{i2e}} \in [0,1]$ 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 $Q^{(t)}_{\mathrm{ext}}(s,a) = \beta_{\mathrm{MB,ext}} Q^{(t)}_{\mathrm{MB,ext}}(s,a) + \beta_{\mathrm{MF,ext}} Q^{(t)}_{\mathrm{MF,ext}}(s,a)$ and $Q^{(t)}_{\mathrm{int}}(s,a) = \beta_{\mathrm{MB,int}} Q^{(t)}_{\mathrm{MB,int}}(s,a) + \beta_{\mathrm{MF,int}} Q^{(t)}_{\mathrm{MF,int}}(s,a)$, respectively. We further define the state-dependent variations in Q-values as $\Delta Q^{(t)}_{\mathrm{ext}}(s) = \max_a Q^{(t)}_{\mathrm{ext}}(s,a) - \min_a Q^{(t)}_{\mathrm{ext}}(s,a)$ and $\Delta Q^{(t)}_{\mathrm{int}}(s) = \max_a Q^{(t)}_{\mathrm{int}}(s,a) - \min_a Q^{(t)}_{\mathrm{int}}(s,a)$ as well as their temporal average $\Delta \bar{Q}_{\mathrm{ext}} = \left\langle \Delta Q^{(t)}_{\mathrm{ext}}(s_t) \right\rangle$ and $\Delta \bar{Q}_{\mathrm{int}} = \left\langle \Delta Q^{(t)}_{\mathrm{int}}(s_t) \right\rangle$, where $\langle . \rangle$ shows the temporal average. $\Delta \bar{Q}_{\mathrm{ext}}$ and $\Delta \bar{Q}_{\mathrm{int}}$ 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

$$\omega_{\mathrm{i2e}} = \frac{\Delta \bar{Q}_{\mathrm{int}}}{\Delta \bar{Q}_{\mathrm{ext}} + \Delta \bar{Q}_{\mathrm{int}}}. \tag{15}$$

Fig. 6B shows the value ω_{i2e} in episode 2-5 computed for each human participant (dots) and averaged over different groups (bars). See Supplementary Materials for ω_{i2e} during episode 1 and the results of a similar approach for quantifying the relative importance of the MB and MF policies in action-selection.

Acknowledgement

AM thanks Vasiliki Liakoni, Johanni Brea, Sophia Becker, Martin Barry, and Valentin Schmutz for many useful discussions on relevant topics. This research was supported by 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 Contributions

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, ran the simulations, and analyzed the data. AM, MHH, and WG wrote the paper.

Competing Interests statement

The authors declare no competing interests.

Code and data availability

All code and data needed to reproduce the results reported in this manuscript will be made publicly available after publication acceptance.

References

1. Berlyne, D. E. A theory of human curiosity. *British Journal of Psychology. General Section* **45**, 180–191 (1954).
2. Berlyne, D. E. Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology. General Section* **41**, 68–80 (1950).
3. Dubey, R. & Griffiths, T. L. Understanding exploration in humans and machines by formalizing the function of curiosity. *Current Opinion in Behavioral Sciences* **35**, 118–124 (2020).
4. Gottlieb, J., Oudeyer, P.-Y., Lopes, M. & Baranes, A. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences* **17**, 585–593 (2013).
5. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology* **55**, 7–14 (2019).
6. Barto, A. G. *Intrinsic Motivation and Reinforcement Learning*, 17–47 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
7. Singh, S., Lewis, R. L., Barto, A. G. & Sorg, J. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development* **2**, 70–82 (2010).

8. Santucci, V., Baldassarre, G. & Mirolli, M. Which is the best intrinsic motivation signal for learning multiple skills? *Frontiers in Neurorobotics* **7** (2013).

9. Jaegle, A., Mehrpour, V. & Rust, N. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Current Opinion in Neurobiology* **58**, 167–174 (2019).

10. Oudeyer, P.-Y. & Kaplan, F. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics* **1**, 6 (2009).

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

12. Murayama, K. A reward-learning framework of knowledge acquisition: An integrated account of curiosity, interest, and intrinsic–extrinsic rewards. *Psychological Review* **129**, 175–198 (2022).

13. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Networks* **15**, 549–559 (2002).

14. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* **143**, 2074–2081 (2014).

15. Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J. & Dayan, P. The modulation of savouring by prediction error and its effects on choice. *eLife* **5**, e13747 (2016).

16. Kobayashi, K., Ravaioli, S., Baranès, A., Woodford, M. & Gottlieb, J. Diverse motives for human curiosity. *Nature human behaviour* **3**, 587–595 (2019).

17. Gershman, S. J. Uncertainty and exploration. *Decision* **6**, 277 (2019).

18. Horvath, L. *et al.* Human belief state-based exploration and exploitation in an information-selective symmetric reversal bandit task. *Computational Brain & Behavior* (2021).

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

20. Bellemare, M. *et al.* Unifying count-based exploration and intrinsic motivation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 29 (Curran Associates, Inc., 2016).

21. Ostrovski, G., Bellemare, M. G., van den Oord, A. & Munos, R. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2721–2730 (JMLR.org, 2017).

22. Tang, H. *et al.* #exploration: A study of count-based exploration for deep reinforcement learning. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

23. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2778–2787 (JMLR.org, 2017).

24. Pathak, D., Gandhi, D. & Gupta, A. Self-supervised exploration via disagreement. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 5062–5071 (PMLR, 2019).

25. Aubret, A., Matignon, L. & Hassas, S. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976* (2019).

26. Burda, Y. *et al.* Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations* (2019).

27. Savinov, N. *et al.* Episodic curiosity through reachability. In *International Conference on Learning Representations* (2019).

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

29. Modirshanechi, A., Brea, J. & Gerstner, W. Surprise: a unified theory and experimental predictions. *bioRxiv* (2021).

30. Tartaglia, E. M., Clarke, A. M. & Herzog, M. H. What to choose next? a paradigm for testing human sequential decision making. *Frontiers in Psychology* **8**, 312 (2017).

31. Lehmann, M. P. *et al.* One-shot learning and behavioral eligibility traces in sequential decision making. *eLife* **8**, e47463 (2019).

32. Liakoni, V. *et al.* Brain signals of a surprise-actor-critic model: Evidence for multiple learning modules in human decision making. *NeuroImage* **246**, 118780 (2022).

33. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

34. Van Seijen, H. & Sutton, R. Planning by prioritized sweeping with small backups. In Dasgupta, S. & McAllester, D. (eds.) *Proceedings of the 30th International Conference on Machine Learning*, vol. 28 of *Proceedings of Machine Learning Research*, 361–369 (PMLR, Atlanta, Georgia, USA, 2013).

35. Brea, J. Is prioritized sweeping the better episodic control? *arXiv preprint arXiv:1711.06677* (2017).

36. Mattar, M. G. & Lengyel, M. Planning in the brain. *Neuron* **110**, 914–934 (2022).

37. Mobin, S. A., Arnemann, J. A. & Sommer, F. Information-based learning by agents in unbounded state spaces. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014).

38. Little, D. Y.-J. & Sommer, F. T. Learning and exploration in action-perception loops. *Frontiers in Neural Circuits* **7**, 37 (2013).

39. Achiam, J. & Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732* (2017).

40. Wasserman, L. *All of statistics: a concise course in statistical inference* (Springer, 2004).

41. Gershman, S. J. & Niv, Y. Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science* **7**, 391–415 (2015).

42. Stojić, H., Schulz, E., P Analytis, P. & Speekenbrink, M. It's new, but is it good? how generalization and uncertainty guide the exploration of novel options. *Journal of Experimental Psychology: General* **149**, 1878–1907 (2020).

43. Traner, M. R., Bromberg-Martin, E. S. & Monosov, I. E. How the value of the environment controls persistence in visual search. *PLoS Computational Biology* **17**, 1–34 (2021).

44. Carver, C. S., Scheier, M. F. & Segerstrom, S. C. Optimism. *Clinical Psychology Review* **30**, 879–889 (2010). Positive Clinical Psychology.

45. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

46. Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464 (1978).

47. Daw, N. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII* **23** (2011).

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

49. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).

50. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *Neuroimage* **84**, 971–985 (2014).

51. Daw, N., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience* **8**, 1704–1711 (2005).

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

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

54. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019).

55. Nassar, M. R. & Frank, M. J. Taming the beast: extracting generalizable knowledge from computational models of cognition. *Current opinion in behavioral sciences* **11**, 49–54 (2016).

56. Cohen, J. D., McClure, S. M. & Yu, A. J. 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**, 933–942 (2007).

57. Wilson, R. C., Bonawitz, E., Costa, V. D. & Ebitz, R. B. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences* **38**, 49–56 (2021). Computational cognitive neuroscience.

58. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* **6**, e27430 (2017).

59. Dubois, M. *et al.* Human complex exploration strategies are enriched by noradrenaline-modulated heuristics. *eLife* **10**, e59907 (2021).

60. Wittmann, B. C., Daw, N. D., Seymour, B. & Dolan, R. J. Striatal activity underlies novelty-based choice in humans. *Neuron* **58**, 967–973 (2008).

61. Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A. Bayesian reinforcement learning: A survey. *Found. Trends Mach. Learn.* **8**, 359–483 (2015).

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

63. Klyubin, A., Polani, D. & Nehaniv, C. Empowerment: a universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, vol. 1, 128–135 Vol.1 (2005).

64. Itti, L. & Baldi, P. Bayesian surprise attracts human attention. *Vision Research* **49**, 1295–1306 (2009).

65. Ogasawara, T. *et al.* A primate temporal cortex–zona incerta pathway for novelty seeking. *Nature neuroscience* **25** (2022).

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

67. Chan, H. K. & Toyoizumi, T. A decision-making model with anticipation of surprise for explaining irrational economic behaviors. *arXiv preprint arXiv:2108.12347* (2021).

68. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* **43**, e1 (2020).

69. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Current Opinion in Behavioral Sciences* **41**, 15–21 (2021). Value based decision-making.

70. Kim, Y., Nam, W., Kim, H., Kim, J.-H. & Kim, G. Curiosity-bottleneck: Exploration by distilling task-specific novelty. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 3379–3388 (PMLR, 2019).

71. Cogliati Dezza, I., Cleeremans, A. & Alexander, W. H. Independent and interacting value systems for reward and information in the human brain. *eLife* **11**, e66358 (2022).

72. Ghazizadeh, A. *et al.* Brain Networks Sensitive to Object Novelty, Value, and Their Combination. *Cerebral Cortex Communications* **1** (2020). Tgaa034.

73. Menegas, W., Babayan, B. M., Uchida, N. & Watabe-Uchida, M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife* **6**, e21886 (2017).

74. Akiti, K. *et al.* Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *bioRxiv* (2021).

75. Tapper, A. R. & Molas, S. Midbrain circuits of novelty processing. *Neurobiology of Learning and Memory* **176**, 107323 (2020).

76. Deci, E. L., Koestner, R. & Ryan, R. M. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* **125**, 627–668 (1999).

77. Murayama, K., Matsumoto, M., Izuma, K. & Matsumoto, K. Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences* **107**, 20911–20916 (2010).

78. Ghazizadeh, A., Griggs, W. & Hikosaka, O. Ecological origins of object salience: Reward, uncertainty, aversiveness, and novelty. *Frontiers in Neuroscience* **10**, 378 (2016).

79. Sharot, T., Korn, C. W. & Dolan, R. J. How unrealistic optimism is maintained in the face of reality. *Nature neuroscience* **14**, 1475–1479 (2011).

80. Strunk, D. R., Lopez, H. & DeRubeis, R. J. Depressive symptoms are associated with unrealistic negative predictions of future life events. *Behaviour Research and Therapy* **44**, 861–882 (2006).

81. Scheier, M. F., Carver, C. S. & Bridges, M. W. 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**, 1063–1078 (1994).

82. Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S. & Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour* **1**, 1–9 (2017).

83. Stankevicius, A., Huys, Q. J. M., Kalra, A. & Seriès, P. Optimism as a prior belief about the probability of future reward. *PLoS Computational Biology* **10**, 1–9 (2014).

84. Gesiarz, F., Cahill, D. & Sharot, T. Evidence accumulation is biased by motivation: A computational account. *PLoS Computational Biology* **15**, 1–15 (2019).

85. Sharot, T. The optimism bias. *Current Biology* **21**, R941–R945 (2011).

86. Garrett, N. & Sharot, T. Optimistic update bias holds firm: Three tests of robustness following shah et al. *Consciousness and Cognition* **50**, 12–22 (2017). Unrealistic Optimism – Its Nature, Causes and Effects.

87. Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences* **26**, 607–621 (2022).

88. Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J. & Schulz, E. Intrinsically motivated exploration as empowerment. *PsyArXiv* (2022).

89. Brainard, D. H. & Vision, S. The psychophysics toolbox. *Spatial vision* **10**, 433–436 (1997).

90. Teh, Y. W. *Dirichlet Process*, 280–287 (Springer US, Boston, MA, 2010).

91. Ghahramani, Z. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 20110553 (2013).

92. Gershman, S. J. & Blei, D. M. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology* **56**, 1–12 (2012).

93. Yu, A. J. & Cohen, J. D. Sequential effects: Superstition or rational behavior? In Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 21 (Curran Associates, Inc., 2009).

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

95. Huys, Q. J. *et al.* Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* **112**, 3098–3103 (2015).

96. Piray, P. & Daw, N. D. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications* **12**, 4942 (2021).

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

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

99. Sun, Y., Gomez, F. & Schmidhuber, J. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In Schmidhuber, J., Thórisson, K. R. & Looks, M. (eds.) *Artificial General Intelligence*, 41–51 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

100. Rowan, T. H. *Functional stability analysis of numerical algorithms*. Ph.D. thesis, The University of Texas at Austin (1990).

101. Johnson, S. G. The nlopt nonlinear-optimization package. URL http://github.com/stevengj/nlopt.