The curse of optimism: a persistent distraction by novelty

Alireza Modirshanechi[1,2,*], Wei-Hsiang Lin[1], He A. Xu[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
[*] 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 several of these algorithms get distracted by reward-independent stochastic stimuli. Here, we ask whether humans get distracted by the same stimuli as the algorithms. We design an experimental paradigm where human participants search for rewarding states in an environment with a highly 'stochastic' but reward-free sub-region. We show that (i) participants get repeatedly and persistently distracted by novelty in the stochastic part of the environment; (ii) optimism about the availability of other rewards increases this distraction; and (iii) the observed distraction pattern 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 suboptimal but computationally cheap curiosity-driven policies for exploration in complex environments.

Introduction

Curiosity drives humans and animals to explore their environments[1–3] and to 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[6,7] have been proposed as models of curiosity-driven behavior[8–11] 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[12–18] but also in solving complex machine learning tasks with sparse or even no (extrinsic) rewards such as in computer games[19–21] and high-dimensional control problems[22–25]. Despite their successes, these algorithms face a serious challenge: Intrinsically motivated agents are prone to distraction by reward-independent stochasticity (the so-called 'noisy TV' problem)[26,27], i.e., they are attracted to novel, surprising, or just noisy states independently of whether or not these states are rewarding[28].

The extent of distraction varies between different algorithms, and designing efficient noise-robust algorithms is an ongoing line of research in machine learning[29–32]. In particular, 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 attraction by stochasticity[26,27,33]. 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[26,27] to behavioral neuroscience and design a novel experimental paradigm with a highly stochastic part. We test the predictions of three different intrinsically motivated 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: Human participants exhibit a persistent distraction by novelty, and the degree of this distraction correlates with their degree of 'reward optimism', where reward optimism is defined by the experimental procedure. Our results provide evidence for (i) novelty-driven RL algorithms as models of human curiosity even when novelty-seeking is suboptimal 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 participants that allows us dissociate predictions of different intrinsically motivated RL algorithms. We employ a sequential decision-making paradigm[34–36] for navigation in an environment with 58 states plus three goal states (Fig. 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 Fig. 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 (Fig. 1C); for simulated participants, both states and actions are abstract entities (i.e., we consider RL in a tabular setting[37]). The assignment of images to states and disks to actions is random but fixed

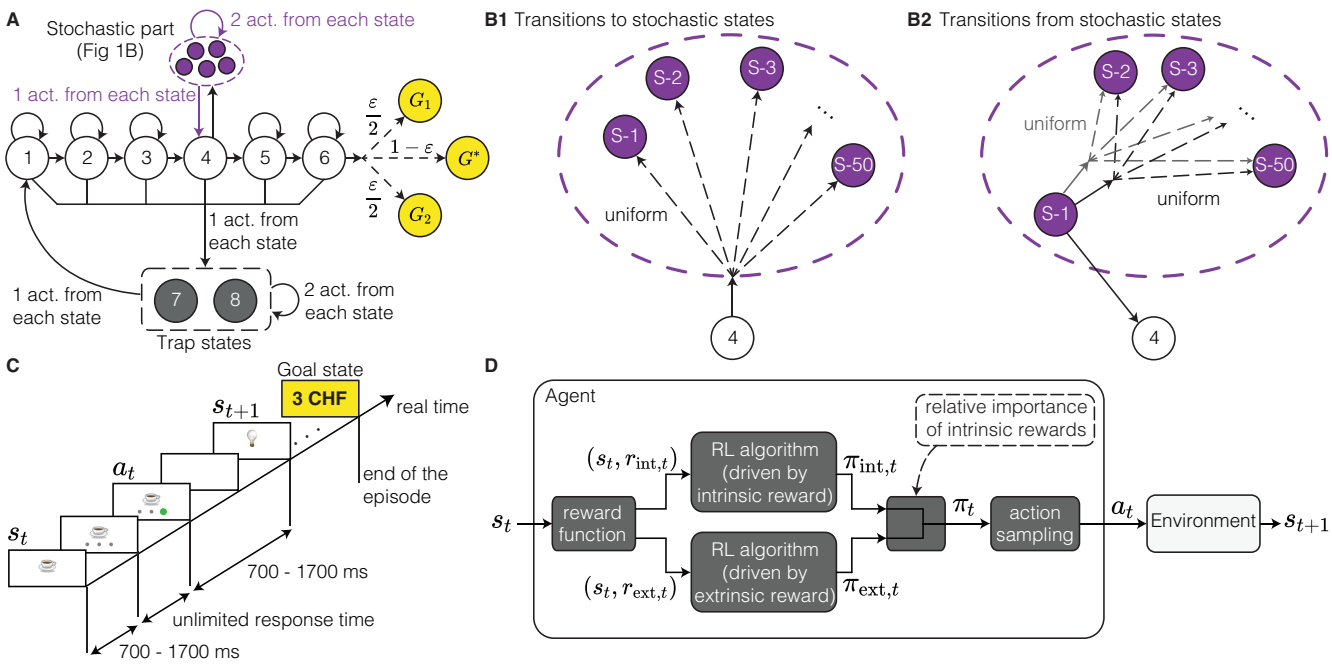

Figure 1: **Experimental paradigm and computational model. A.** Structure of the environment with the stochastic states merged together (dashed oval; see B). Each circle represents a state and each solid arrow an action. All actions except for the ones to the stochastic part or to the goal states are deterministic. Dashed arrows indicate random transitions; values (e.g., $1-\varepsilon$) show 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.

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 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 Fig. 1A), trap states (7 and 8 in Fig. 1A), and stochastic states (S-1 to S-50 in Fig. 1B, shown as a dashed oval in Fig. 1A). 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.

The stochastic part of the environment – which is inspired by the machine-learning literature and mimics the main features of a 'noisy TV'[28] – is a crucial difference to existing paradigms in the literature of behavioral neuroscience[18,38,39]. Without the stochastic part, intrinsic motivation helps agents to avoid the trap states and find the goal[18], hence it helps exploration before and does not harm exploitation after finding a goal. By adding the stochastic part, we aim to quantify how much exploitation of the discovered goal is reduced because of the distraction by the stochastic states.

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 (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 with efficient algorithms

To formulate qualitative predictions for human behavior, we simulate three intrinsically motivated RL algorithms. Intrinsic motivation is described in each algorithm by 'intrinsic rewards' that agents give to themselves upon visiting 'novel', 'surprising', or 'informative' states (see Methods for details). Extrinsic rewards, on the other hand, are received only when visiting the three goal states. 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): 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[40,41] to propose a policy $\pi_{\text{ext},t}$ that maximizes future extrinsic rewards and $\pi_{\text{int},t}$ that maximizes future intrinsic rewards[18,26], 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[26,27]: (i) novelty[18–20] quantifies how infrequent the state s_t has been until time t; (ii) information-gain[23,25,42,43] 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[21,28,44] 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 focus on episodes 2-5 and formulate the *qualitative* predictions of different intrinsically motivated RL algorithms for human behavior by characterizing the behavior of these simulated efficient agents. We note that these predictions are made by using efficient RL algorithms with perfect memory and high computational power. Thus these efficient agents are a starting point and used only (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; a more realistic simulation of human behavior is presented in a later section.

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 a similar performance during episode 1 and find the goal G^* almost equally fast (Supplementary Materials). 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 (Fig. 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 $\pi_{\text{ext},t}$ (Fig. 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 Fig. 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 (Fig. 2A1, C1, and E1) and the fraction of time spent in the stochastic part (Fig. 2A2, C2, and E2). This observation implies that intrinsically motivated exploration leads to an attraction to the stochastic part of the environment, effectively keeping the simulated efficient agents away from the goal region beyond state 6 (Fig. 1A). Our results thus confirm earlier findings in machine learning[26,28] that intrinsically motivated agents get distracted by noisy 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 Fig. 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 (Fig. 2C3), whereas we observe the opposite pattern for efficient agents that purely seek novelty (Fig. 2A3) or surprise (Fig. 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 (Fig. 2E3). These observations confirm the inefficiency

of seeking surprise and the efficiency of seeking information-gain in dealing with noise[26].

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 (Fig. 2B, D, and F). For all three

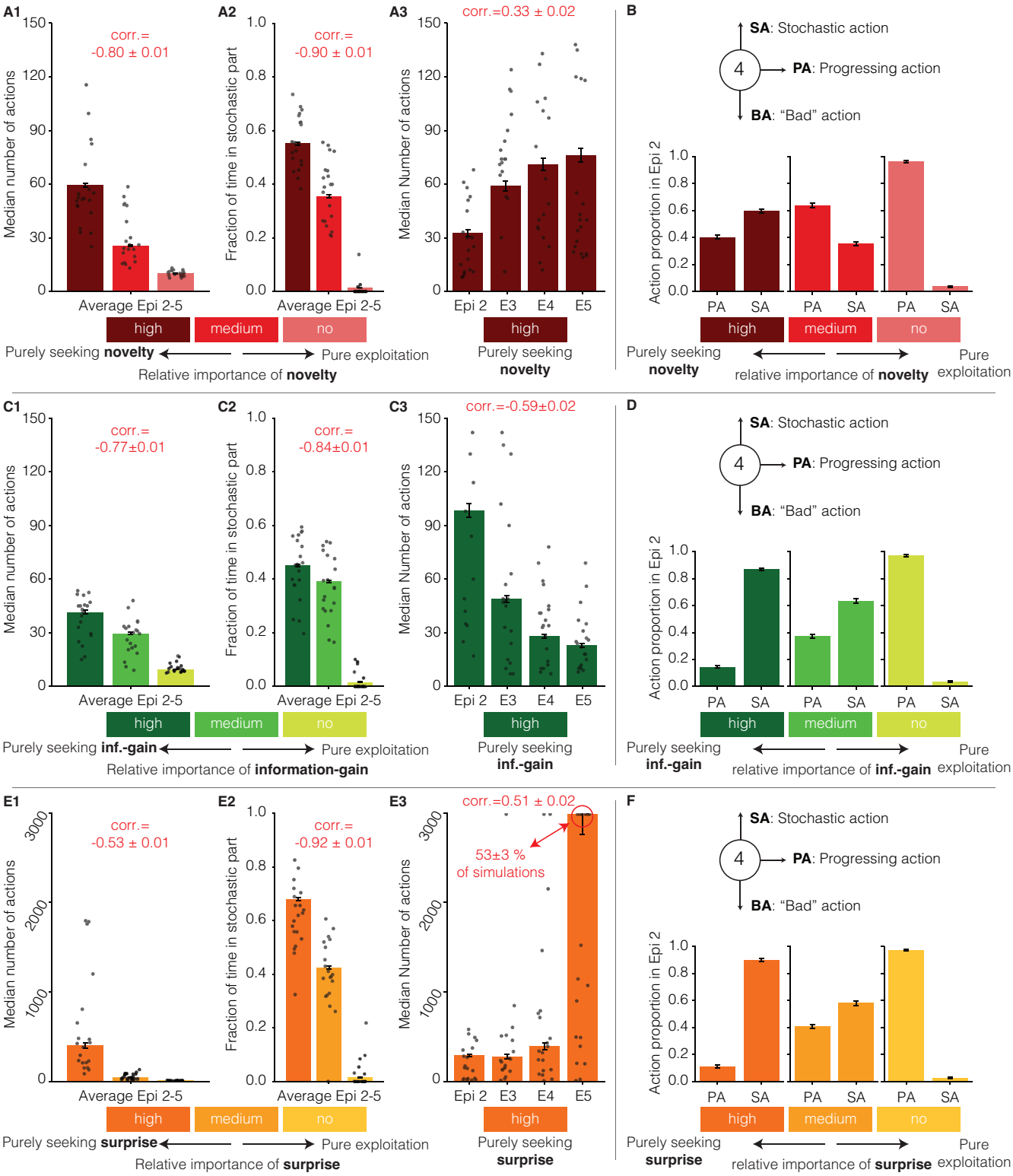

Figure 2: (Caption next page.)

Figure 2: (Previous page.) **Qualitative predictions for episodes 2-5 with efficient algorithms seeking novelty (A-B), surprise (C-D), and information-gain (E-F) as intrinsic rewards.** We consider three levels of importance for intrinsic rewards (Fig. 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.

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%; Fig. 2D and F), whereas this difference is much smaller for seeking novelty (40% vs. 60%; Fig. 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.

Human participants

To test the predictions of intrinsically motivated algorithms, 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 (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 rewards compared to extrinsic rewards. Inspired by earlier studies[45–47], 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 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 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 further explore 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; see Methods for how this information is incorporated in the RL algorithms. 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[48].

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 Fig. 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).

Human participants exhibit a persistent distraction by stochasticity

We perform the same series of analyses on the behavior of human participants as those performed on the behavior of simulated efficient agents (Fig. 3). In episodes 2-5, the search duration of human participants (Fig. 3A1) and the fraction of time they spend in the stochastic part (Fig. 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 (Fig. 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 Fig. 3A1, A2, and B with Fig. 2A1, A2, and B, respectively).

The behavior of the 2 CHF group is particularly interesting since they are the most optimistic group of participants. The 2 CHF group exhibits a constant search duration over episodes 2-5 (zero correlation accepted by Bayesian hypothesis testing[49]; Fig. 3A3). This implies that they

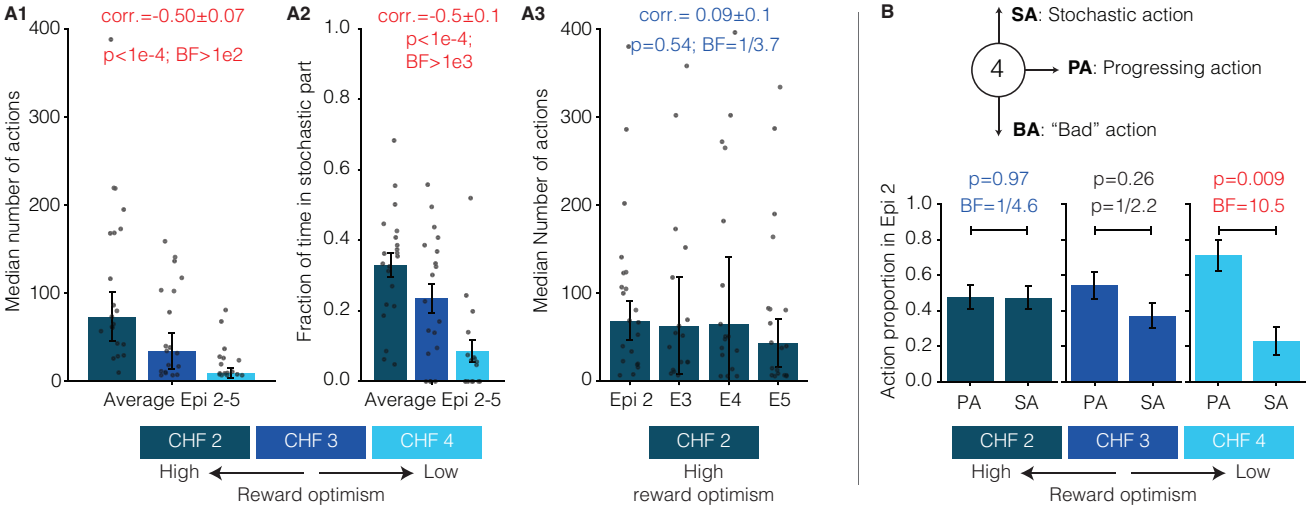

Figure 3: **Human participants persistently explore the stochastic part if they are highly optimistic. A.** Search duration in episodes 2-5. **A1.** Median number of actions over episodes 2-5 for the three different groups: 2 CHF (dark), 3 CHF (medium), and 4 CHF (light). Error bars show the SEMed (evaluated 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[49] 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 Supplementary Materials 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[49] 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[50] (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).

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[49]; Fig. 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 Fig. 3B against Fig. 2B, D, and F).

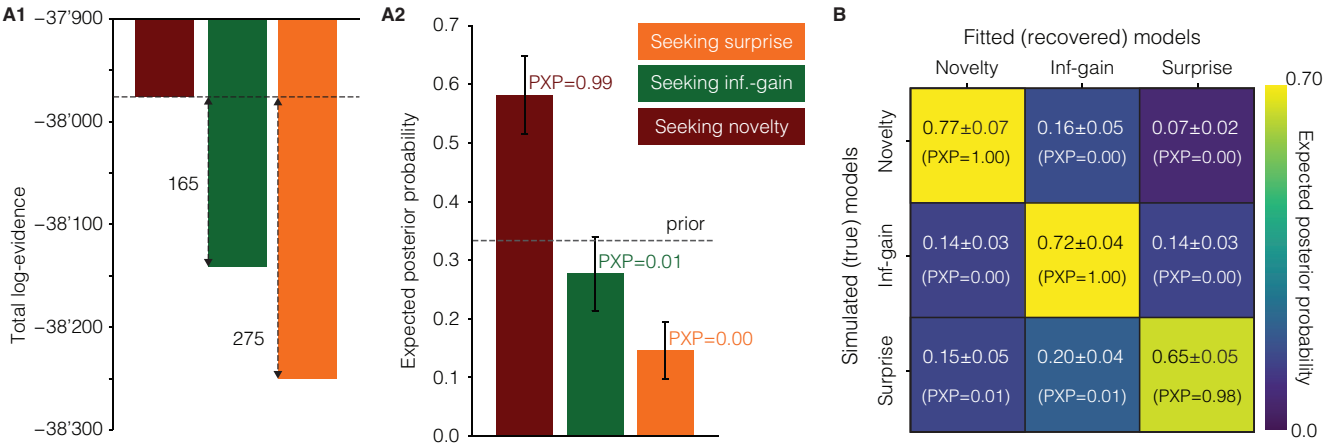

Figure 4: **Novelty-seeking is the most probable model of human behavior. A.** Human participants' action-choices are best explained by novelty-seeking (see Methods for details). **A1.** Model log-evidence summed over all participants (i.e., assuming that different participants have the same exploration strategy but can have different parameters[53]) is significantly higher for seeking novelty than seeking information-gain or surprise. High values indicate good performance, and differences greater than 10 are traditionally[50] considered as strongly significant. **A2.** The expected posterior model probability with random effects assumption[54] (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[54], 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).

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 to those of novelty-seeking simulated efficient agents. However, the qualitative predictions in Fig. 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 Fig. 2: In order to model the behavior of human participants, we use a hybrid RL model[18,36,38,51] combining model-based planning[41] and model-free habit-formation[52], 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[53,54] to *quantitatively* test whether human behavior is explained better by seeking novelty than seeking information-gain or surprise

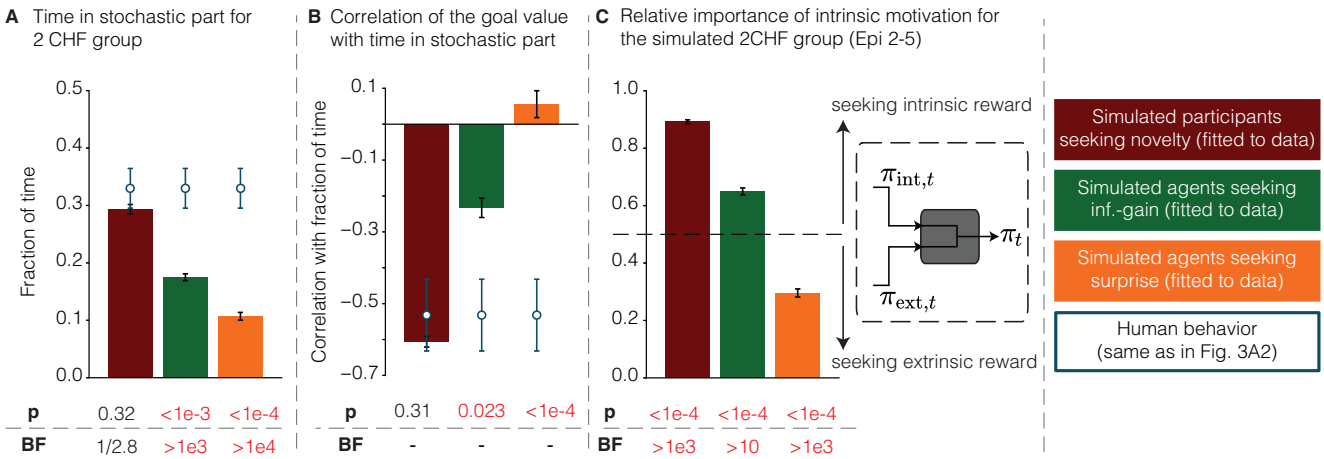

Figure 5: **Seeking novelty, but not surprise or information-gain, can reproduce data statistics.** For each of the three intrinsic rewards, we run 1500 simulations of algorithms with parameters fitted to individual human participants; random seeds are different in each simulation. We divide the simulated 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 Fig. 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 Fig. 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., dominance of $\pi_{\text{int},t}$ over $\pi_{\text{ext},t}$; Eq. 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[50] (see Methods). Red BFs: Significant evidence in favor of the alternative hypothesis (BF≥ 3).

(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 (Fig. 4A; Protected Exceedance Probability[54] $= 0.99$ and 0.01 for seeking novelty and information-gain, respectively). 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)[55,56]. As opposed to the simulations in Fig. 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 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[55] 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; Fig. 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[54] ≥ 0.98 for the true model in Fig. 4B). 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 Fig. 5A-B and a systematic analysis in Supplementary Materials). Our results show that simulated participants using novelty as intrinsic rewards reproduce all data statistics (including the zero correlation observed in Fig. 3A3; see Supplementary Materials), 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 (Fig. 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 (Fig. 5B). We emphasize that both shortcomings are observed despite the fact that 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 Fig. 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; Fig. 5C), which in turn cannot explain the overall high level of exploration observed in the 2 CHF group of human participants (Fig. 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 Fig. 2D with Fig. 3B). With such reduced importance of information-gain, the hybrid algorithm cannot, however, explain the specific behavioral features in Fig. 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-grain.

Taken together our results with simulated participants provide strong quantitative evidence for novelty-seeking as a model of human exploration in our experiment.

Reward optimism correlates with 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 the relative importance of intrinsic rewards.

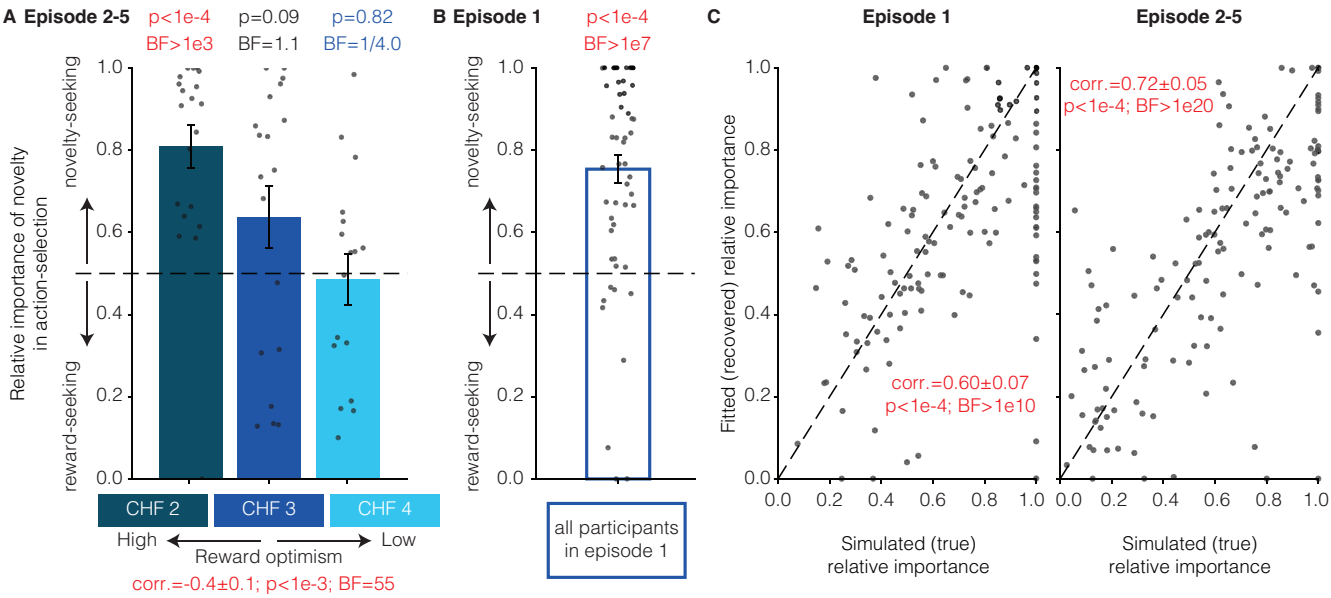

Figure 6: **Reward optimism increases the relative importance of novelty in action-selection.** **A.** The relative importance of novelty-seeking in episodes 2-5 is computed for each participant after fitting the model to data (similar to Fig. 5C but using action-choices of human participants instead 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[49] 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[55] 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[50] (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$\leq 1/3$).

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. 6A; parameter-recovery[55] in Fig. 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 (Fig. 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[37] 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 (Fig. 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.

Discussion

We designed a novel experimental paradigm to study human curiosity-driven exploration in the presence of stochasticity. We made two main observations: (i) Human participants who are optimistic about finding higher rewards than those already discovered are persistently distracted by stochasticity; and (ii) this persistent pattern of distraction 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 deal with the exploration-exploitation trade-off has been a long-lasting question in neuroscience and psychology[57,58]. Experimental studies have shown that humans use a combination of random and directed exploration[13,16], potentially linked to different neural mechanisms[59–61]. However, there are multiple distinct theoretical models to describe directed exploration[5,10,26,62–65], and it has been debated which one is best suited to explain human behavior. In a general setting, human exploration is driven by multiple motivational signals[15,65], but it has been also shown that a particular signal can dominate exploration in specific tasks[3,17,45,66–68]. In an earlier work[18], we have shown that novelty signals dominantly drive human exploration in situations where one needs to search for rewarding states in unknown but deterministic environments. Observations (i) and (ii) above provide further evidence for novelty as the dominant drive of human search strategy even in situations where seeking novelty is not optimal and leads to distraction by reward-independent stochasticity. Further experimental studies are needed to investigate the role novelty in other types of human exploratory behavior.

Observation (ii) is particularly surprising as it has been believed that humans are not prone to the 'noisy TV' problem[4,31,33]. Our results with human participants challenge the idea of defining curiosity as a normative solution to the exploration-exploitation trade-off[3,6]; hence, algorithmic advances in machine learning do not necessarily help 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 noise 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[69–71].

The core assumption of using intrinsically motivated algorithms as models of human curiosity is the existence of an intrinsic exploration reward parallel to the extrinsic reward. There are, however, multiple ways to incorporate intrinsic rewards into the RL framework[26]. 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[8,19,28]. 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[18,72,73]. The latter approach provides higher flexibility to arbitrate between

different policies based on changes in the relative importance of intrinsic versus extrinsic rewards because it does not need re-planning or re-evaluation of already learned policies. Parallel processing paths are compatible with the rapid change of behavior observed in our human participants after finding a goal at the end of episode 1 (Fig. 3B1-2) and also consistent with experimental evidence for partially separate neural pathways of novelty- and reward-induced behaviors[18,74–78].

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[45,46,76]. In particular, theories of 'motivation crowding effect'[79] and 'undermining effect'[80,81] suggest that the absolute value of extrinsic reward might contribute, in addition to the reward optimism, to the observed negative correlation in Fig. 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 to investigate the interplay of novelty and reward is to study human behavior in various environments with different reward distributions and different 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'[48] and has been linked to several neural and behavioral characteristics[48,82,83]. While the traditional approach to measure optimism is through self-tests[84], more recently statistical inference using RL[85] and Bayesian[86,87] 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[62]. 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 Fig. 6A). Moreover, the persistent exploration of the stochastic part of our environment observed in the behavior of human participants (Fig. 3B3) is consistent with the known phenomena of optimism bias[88] and optimistic belief updating in humans[82,89,90].

Even though notions of 'novelty', 'surprise', and 'information-gain' are frequently used in neuroscience[18,91,92], psychology[93,94], and machine learning[26,27,33], there is no consensus on the precise definitions of these notions as scientific terms[44,95]. 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 with respect to the precise choice of definitions as long as (i) novelty quantifies infrequency of *states*[18], e.g., defined based on density models in machine learning[19,20]; (ii) surprise quantifies mismatches between observations and agents' expectations, where the expectations are made based on the previous *state-action* pair, including all measure of prediction surprise[44] and typical measures of prediction error in machine learning[21,28]; and (iii) information-gain quantifies improvements in the agents' *world-model* and vanishes by accumulation of experience, e.g., including Bayesian[91] and Postdictive surprise[92] and measures of disagreement and progress-rate in machine learning[23–25,29].

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.

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 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[96].

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 (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 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. 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 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 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 Fig. 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 Fig. 2A, C, and E as well as for the differences for the highest importance of intrinsic rewards in Fig. 2D and F.

The correction for multiple hypotheses testing was done by controlling the False Discovery Rate at 0.05[50] over all 22 null hypotheses that are tested in Fig. 3, Fig. 5, and Fig. 6 (p-value threshold: 0.034). All Bayes Factors (abbreviated BF in the figures) were evaluated using Schwartz approximation[49] to avoid any assumptions on the prior distribution. We note that evaluating the Bayes Factors using priors suggested by ref.[97,98] does not change our conclusions. We also note that using the Spearman correlation instead of the Pearson correlation in Fig. 2A, C, and E, Fig. 3A, and Fig. 6A does not change our conclusions.

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 simulating efficient agents (Fig. 2) and for modeling human behavior (Fig. 4-Fig. 6). We used ideas from non-parametric Bayesian inference[99] 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 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. 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

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

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 Fig. 1D, we consider a hybrid of model-based and model-free policy[18,36,38,52]. 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[37], thus the MF component of our algorithm remains similar to that of previous studies[18,35]: 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[37] of the state s_{t+1}, respectively. We use two separate eligibility traces[35,37] 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 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[100,101]:

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

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 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

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

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 $\mathcal{S}^{(t)}$) – see Supplementary Materials for derivations. Choosing $\epsilon_{\text{new}} = 0$ is equivalent to assuming there is no unknown state in the environment, for which the estimate in Eq. 6 is reduced to the classic Bayesian estimate of transition probabilities in bounded discrete environments[18,36]. The transition probabilities are then used in a novel variant of prioritized sweeping[37,40] 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[37] 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**[37]: The probability of taking action a in state s at time t is

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

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

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

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[39,52], cognitive load[102], and whether the location of reward is known[18]. 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 (Fig. 5C and Fig. 6).

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 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.

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 Fig. 2, Fig. 4, and Fig. 5), we define the intrinsic reward function as

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

where $p_N^{(t)}(s') = \frac{1+\tilde{C}^{(t)}_{s'}}{1+|\mathcal{S}^{(t)}|+\sum_{s''}\tilde{C}^{(t)}_{s''}}$ is the state frequency with $\tilde{C}^{(t)}_{s'}$ the pseudo-count of encounters of state s' up to time t (similar to Eq. 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[18] 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[19,20].

Surprise: For an agent seeking surprise (orange in Fig. 2, Fig. 4, and Fig. 5), we define the intrinsic reward function as the Shannon surprise (a.k.a. surprisal)[44]

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

where $p^{(t)}(s'|s, a)$ is defined in Eq. 6. With this definition, the expected (over s') intrinsic reward of taking action a at state s is equal to the entropy of the distribution $p^{(t)}(s'|s, a)$[103]. 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[21,28].

Information-gain: For an agent seeking information-gain (green in Fig. 2, Fig. 4, and Fig. 5),

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[103], and $p^{(t+1)}$ is the updated world-model upon observing $(s, a) \to s'$. The dots in Eq. 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 Supplementary Materials for derivations and see ref.[42] 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'[92]), but it has a behavior radically different from that of the Shannon surprise introduced above (Eq. 11) – see ref.[44] 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 Supplementary Materials for the proof). Similar intrinsic rewards have been used in machine learning[23,29,33,42].

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 (Fig. 2, Fig. 5, and Fig. 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. Eq. 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 (Eq. 2). Our results for the fitted relative importance of intrinsic reward across different groups of human participants (Fig. 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 Fig. 4 and Fig. 6 (particularly Fig. 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 Fig. 4 and the best fit would indicate a relative importance of 0 for novelty in Fig. 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

$$Q^{(t)}_{\text{MB,ext}}(s,a) = \bar{R}^{(t)}_{\text{ext}}(s,a) + \lambda_{\text{ext}} \sum_{s'} p^{(t)}(s'|s,a) \max_{a'} Q^{(t)}_{\text{MB,ext}}(s',a'), \tag{13}$$

where $p^{(t)}(s'|s,a)$ is estimated transition probability in Eq. 6, and

$$\begin{aligned} \bar{R}^{(t)}_{\text{ext}}(s,a) &= \sum_{s'} p^{(t)}(s'|s,a) R_{\text{ext}}(s,a \to s') \\ &= p^{(t)}(\tilde{G}_0|s,a) + r_1^* p^{(t)}(\tilde{G}_1|s,a) + r_2^* p^{(t)}(\tilde{G}_2|s,a) \end{aligned} \tag{14}$$

is the average immediate extrinsic reward expected to be collected by taking action a in state s (see Eq. 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

$$\bar{R}^{(t)}_{\text{ext}}(s,a) = \frac{\epsilon_{\text{obs}}(1 + r_1^* + r_2^*)}{\epsilon_{\text{new}} + \epsilon_{\text{obs}}|\mathcal{S}^{(t)}| + \tilde{C}^{(t)}_{s,a}}. \tag{15}$$

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 (Supplementary Materials). 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 (Supplementary Materials). 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 (Eq. 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[37]. 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 $Q^{(0)}_{\text{MF,ext}}$. 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., $\mu_{\text{ext}} = 1$), 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 $r_{G^*} < \frac{1}{\lambda^2_{\text{ext}}}(1 - \lambda_{\text{ext}})(1 + \lambda_{\text{ext}} + \lambda^2_{\text{ext}})Q^{(0)}_{\text{MF,ext}}$. 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 (Fig. 6), the effective reward optimism generated by optimistic initialization is not strong enough to explain human behavior.

Efficient model-based planning for simulated participants

For simulating efficient agents in Fig. 2, we set $\varepsilon = 0$ (see Fig. 1A) and used a pure MB version of our algorithm with 13 parameters:

$$\{r^*_1, r^*_2, \lambda_{\text{ext}}, \lambda_{\text{int}}, \kappa, \epsilon_{\text{new}}, \epsilon_{\text{obs}}, T_{PS,\text{ext}}, T_{PS,\text{int}}, \beta^{(1)}_{\text{MB,ext}}, \beta^{(2)}_{\text{MB,ext}}, \beta^{(1)}_{\text{MB,int}}, \beta^{(2)}_{\text{MB,int}}\}. \tag{16}$$

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)[18]: $\lambda_{\text{ext}} = 0.95$, $\lambda_{\text{int}} = 0.70$, $\epsilon_{\text{new}} = 10^{-5}$ and $\epsilon_{\text{obs}} = 10^{-4}$. To relatively separate the effect of optimistic initialization[37] 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 $\beta^{(1)}_{\text{MB,ext}} = 0$ to have pure intrinsic reward seeking in episode 1.

After fixing parameter values for 10 out of 13 parameters in Eq. 16, we fine-tuned $\beta^{(1)}_{\text{MB,int}}$ to minimize the average length of episode 1 (to find the goal as fast as possible; see Supplementary Materials). 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 shade of colors in Fig. 2). Different shades of color in Fig. 2 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.

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)[18] 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:

$$\hat{\Phi}_{n,R} = \arg\max_{\Phi} P(D_n|\Phi, R_n = R) \tag{17}$$

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 Eq. 9), and $\hat{\Phi}_{n,R}$ is the set of estimated parameters that maximizes that probability. For optimization, we used Subplex algorithm[104] as implemented in Julia NLopt package[105].

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)[49,53]. Fig. 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 ref.[53,55] for tutorials.

We also considered the Bayesian model selection method of ref.[54] 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[50] with uniform prior and 40 chains of length $10'000$) for inference and estimated the joint posterior distribution

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

Fig. 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 Supplementary Materials. The same set of parameters were used for model-recovery in Fig. 4B, posterior predictive checks in Fig. 5, computing the relative importance of novelty in Fig. 6A-B (see 'Relative importance of novelty in action-selection'), and parameter recovery in Fig. 6C.

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 (Fig. 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. Fig. 5 and its supplements in Supplementary Materials 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 ref.[54] to 5 randomly chosen sub-populations of these 150 simulated participants (each with 60 participants, i.e., 20 from each participant group). Fig. 4B shows the results of the model-comparison averaged over these 5 repetitions. Fig. 6C shows the relative importance of novelty in action-selection (see Eq. 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).

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 Eq. 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 $Q^{(t)}_{\text{int}}(s,a) =$

$\beta_{\mathrm{MB,int}} Q^{(t)}_{\mathrm{MB,int}}(s,a) + \beta_{\mathrm{MF,int}} Q^{(t)}_{\mathrm{MF,int}}(s,a)$, 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

$$\omega_{\mathrm{i2e}} = \frac{\Delta \bar{Q}_{\mathrm{int}}}{\Delta \bar{Q}_{\mathrm{ext}} + \Delta \bar{Q}_{\mathrm{int}}}. \tag{18}$$

Fig. 6A shows the value ω_{i2e} in episode 2-5 computed for each human participant (dots) and averaged over different groups (bars), and Fig. 6B shows the same for episode 1. Fig. 5C shows the value ω_{i2e} in episode 2-5 for the 2 CHF group of simulated participants. See Supplementary Materials for 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, Valentin Schmutz, and Guillaume Bellec for many useful discussions on relevant topics. This research was supported by 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 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, did the formal analyses, 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. Reconciling novelty and complexity through a rational analysis of curiosity. *Psychological Review* **127**, 455–476 (2019).
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. 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).
7. Santucci, V., Baldassarre, G. & Mirolli, M. Which is the best intrinsic motivation signal for learning multiple skills? *Frontiers in Neurorobotics* **7** (2013).
8. 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).

9. Oudeyer, P.-Y., Kaplan, F. & Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation* **11**, 265–286 (2007).

10. 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).

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

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

13. 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).

14. 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).

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

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

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

18. 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).

19. 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).

20. 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).

21. 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).

22. Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F. & Yamins, D. L. Learning to play with intrinsically-motivated, self-aware agents. In Bengio, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc., 2018).

23. Sekar, R. *et al.* Planning to explore via self-supervised world models. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 8583–8592 (PMLR, 2020).

24. Kim, K., Sano, M., De Freitas, J., Haber, N. & Yamins, D. Active world model learning with progress curiosity. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 5306–5315 (PMLR, 2020).

25. Mendonca, R., Rybkin, O., Daniilidis, K., Hafner, D. & Pathak, D. Discovering and achieving goals via world models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 24379–24391 (Curran Associates, Inc., 2021).

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

27. Ladosz, P., Weng, L., Kim, M. & Oh, H. Exploration in deep reinforcement learning: A survey. *Information Fusion* **85**, 1–22 (2022).

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

29. 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).

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

31. Mavor-Parker, A., Young, K., Barry, C. & Griffin, L. How to stay curious while avoiding noisy TVs using aleatoric uncertainty estimation. In Chaudhuri, K. *et al.* (eds.) *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, 15220–15240 (PMLR, 2022).

32. Jarrett, D. *et al.* Curiosity in hindsight. In *Deep Reinforcement Learning Workshop NeurIPS 2022* (2022).

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

34. 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).

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

36. 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).

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

38. 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).

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

40. 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).

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

42. 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).

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

44. Modirshanechi, A., Brea, J. & Gerstner, W. A taxonomy of surprise definitions. *Journal of Mathematical Psychology* **110**, 102712 (2022).

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

46. 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).

47. 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).

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

49. Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the american statistical association* **90**, 773–795 (1995).

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

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. 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).

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

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

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

56. 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).

57. 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).

58. 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.

59. 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).

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

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

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

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

64. 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).

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

66. Ten, A., Kaushik, P., Oudeyer, P.-Y. & Gottlieb, J. Humans monitor learning progress in curiosity-driven exploration. *Nature communications* **12**, 5972 (2021).

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

68. Meder, B. & Nelson, J. D. Information search with situation-specific reward functions. *Judgment and Decision Making* **7**, 119–148 (2012).

69. 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).

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

71. Binz, M. & Schulz, E. Modeling human exploration through resource-rational reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K. (eds.) *Advances in Neural Information Processing Systems* (2022).

72. 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).

73. 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).

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

75. 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).

76. Akiti, K. *et al.* Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron* **110**, 3789–3804.e9 (2022).

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

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

79. Frey, B. S. & Jegen, R. Motivation crowding theory. *Journal of Economic Surveys* **15**, 589–611 (2001).

80. 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).

81. 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).

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

83. 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).

84. 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).

85. 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).

86. 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).

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

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

89. 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.

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

91. Baldi, P. *A Computational Theory of Surprise*, 1–25 (Springer US, Boston, MA, 2002).

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

93. Reisenzein, R., Horstmann, G. & Schützwohl, A. The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science* **11**, 50–74 (2019).

94. Nelson, J. D. Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review* **112**, 979–999 (2005).

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

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

97. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review* **16**, 225–237 (2009).

98. Rouder, J. N. & Morey, R. D. Default bayes factors for model selection in regression. *Multivariate Behavioral Research* **47**, 877–903 (2012).

99. 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).

100. 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).

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

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

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

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

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