PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Environmental uncertainty and the advantage of impulsive choice strategies

Diana C. Burk, Bruno B. Averbeck*

Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, United States of America

* averbeckbb@mail.nih.gov

OPEN ACCESS

Citation: Burk DC, Averbeck BB (2023) Environmental uncertainty and the advantage of impulsive choice strategies. PLoS Comput Biol 19(1): e1010873. https://doi.org/10.1371/journal.pcbi.1010873

Editor: Alireza Soltani, Dartmouth College, UNITED STATES

Received: August 15, 2022

Accepted: January 15, 2023

Published: January 30, 2023

Data Availability Statement: All code used to generate the results in this manuscript can be accessed on GitHub here: https://github.com/dcb4p/impulsive_choice_code.

Funding: This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA MH002928 (BA)). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have no competing interests.

Abstract

Choice impulsivity is characterized by the choice of immediate, smaller reward options over future, larger reward options, and is often thought to be associated with negative life outcomes. However, some environments make future rewards more uncertain, and in these environments impulsive choices can be beneficial. Here we examined the conditions under which impulsive vs. non-impulsive decision strategies would be advantageous. We used Markov Decision Processes (MDPs) to model three common decision-making tasks: Temporal Discounting, Information Sampling, and an Explore-Exploit task. We manipulated environmental variables to create circumstances where future outcomes were relatively uncertain. We then manipulated the discount factor of an MDP agent, which affects the value of immediate versus future rewards, to model impulsive and non-impulsive behavior. This allowed us to examine the performance of impulsive and non-impulsive agents in more or less predictable environments. In Temporal Discounting, we manipulated the transition probability to delayed rewards and found that the agent with the lower discount factor (i.e. the impulsive agent) collected more average reward than the agent with a higher discount factor (the non-impulsive agent) by selecting immediate reward options when the probability of receiving the future reward was low. In the Information Sampling task, we manipulated the amount of information obtained with each sample. When sampling led to small information gains, the impulsive MDP agent collected more average reward than the non-impulsive agent. Third, in the Explore-Exploit task, we manipulated the substitution rate for novel options. When the substitution rate was high, the impulsive agent again performed better than the non-impulsive agent, as it explored the novel options less and instead exploited options with known reward values. The results of these analyses show that impulsivity can be advantageous in environments that are unexpectedly uncertain.

Author summary

Impulsive choice behavior, or valuing immediate smaller rewards over larger, delayed rewards, is typically considered to be detrimental in decision-making. In this study, we use Markov Decision Processes (MDPs) to demonstrate that impulsive choices can be beneficial in three common decision-making tasks: Temporal Discounting, Information

Sampling and an Explore-Exploit task. Specifically, we found that when a task environment is more uncertain than expected, an impulsive agent can collect more average reward than a non-impulsive agent. Our work suggests that impulsivity is not inherently negative. Valuing immediate rewards over delayed rewards can be an adaptive strategy when faced with uncertainty.

Introduction

Impulsive decision making is frequently defined as disadvantageous. It has many descriptive definitions, including "choosing a smaller-sooner option when a larger-later option produces a better outcome,"[1] "swift action without forethought or conscious judgment,"[2] and "actions that are poorly conceived, prematurely expressed, unduly risky, or inappropriate to the situation and that often result in undesirable outcomes" [3]. Impulsivity is also considered a component of many clinical conditions, including gambling disorder and other behavioral addictions [4–6], substance-abuse [7–9], attention deficit/hyperactivity disorder [10,11], and other psychiatric disorders [2,12–14]. Taken together, these definitions and clinical manifestations suggest that favoring immediate rewards over delayed rewards leads to suboptimal outcomes [2,3,15–17]. Because impulsivity has carried this negative characterization, many studies have focused on impulsivity as maladaptive. However, there has been some investigation that suggests that impulsive choice behavior might be due to adaptation to the statistics of certain environments [18–22].

Impulsivity is measured with a variety of self-report questionnaires and laboratory tasks in human and animal subjects (For a review, see [23]). There are roughly 25 commonly used self-report questionnaires that measure impulsivity [15,24–27]. Laboratory tasks have also been designed to assess several dimensions of impulsivity, including motor impulsivity (for a review see [28]), attention impulsivity [29–31], risk preference [32–35], and impulsive choice behavior [36]. Choice impulsivity tasks, which we consider in the present manuscript, were developed to assess the weighting of immediate vs. future rewards. One commonly used choice task is Temporal Discounting [37–39], which measures preference for a smaller immediate reward or a larger future reward. Impulsive participants, by definition, favor the smaller, immediate rewards over the delayed, larger rewards. Information sampling tasks, such as the Beads task, are also used to measure the tradeoff between collecting more information or committing to a choice [40–45]. And N-armed bandit tasks that periodically introduce novel options have been used to assess the tendency for subjects to explore new options versus exploiting known options [21,46–50].

In this paper, we used a Markov decision process (MDP) framework to compare the behavior of impulsive and non-impulsive agents in three common decision-making tasks where current choices affect future rewards. The MDP framework models decisions of an agent in an environment where the current state affects the immediate reward an agent can obtain, as well as the probabilities of transitioning to future states [51,52] (**Fig 1**). If it is assumed that the agent is maximizing the expected reward, the MDP provides insight into the optimal strategy (that is, to maximize over state-action values), in a decision-making task.

Within the MDP framework, action values, $Q(s_t, a)$, are the sum of immediate and discounted future expected rewards:

$$Q(s_t, a) = r(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j)$$

where the first term $r(s_t, a)$ is the immediate expected reward in state s at time t if action a is

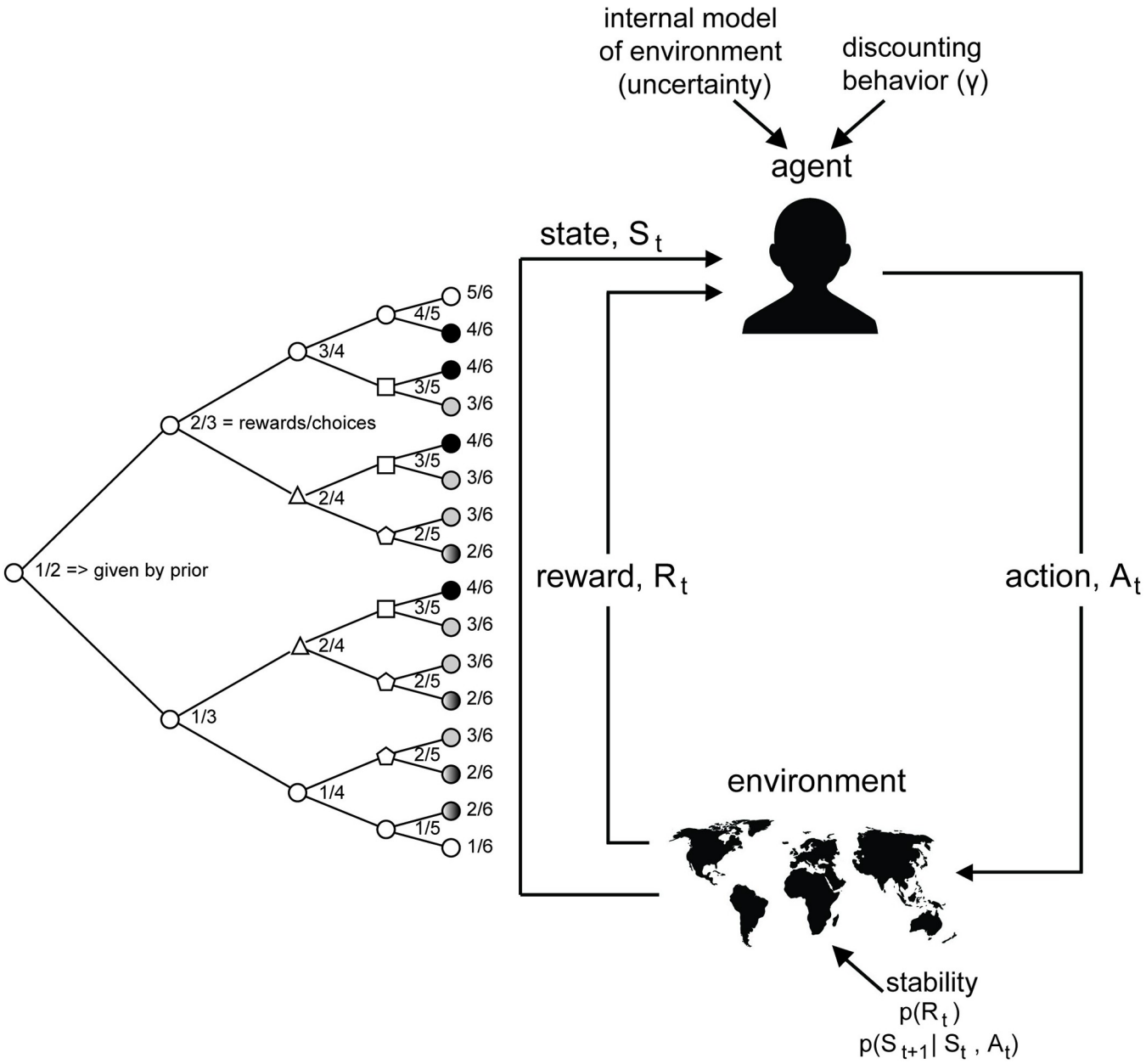

Fig 1. Agent and Environment Interactions in Reinforcement Learning (RL), Markov Decision Process (MDP) framework. A schematic of how an agent interacts with the environment and learns to maximize rewards in a MDP framework. An agent selects actions, A_t, which lead to changes in state, S_t and rewards, R_t, where t indicates the trial or time point. The agent's internal model of the environment and weighting of future rewards, or discount factor, γ, affect the actions taken. The stability of the environment is captured by transition probabilities to future states p ($S_{t+1}|S_t$, A_t) as well as the probability of receiving reward p (R_t); these also affect reward outcomes. An example reward distribution tree for a binomial bandit is shown on the left for a bandit option in a choice task. As an agent selects the option that gives a probabilistic reward, it traverses the tree based on outcomes. Each node in the tree represents a choice point where that option was chosen. The node shape and shading indicate whether a node represents a unique state. Circle nodes are unique. Other shapes or shading of nodes indicate duplicate states that have multiple choice paths to them. While MDPs are independent of time and history, these factors often affect decision-making behavior. Each upper branch in the state space tree represents a rewarded choice, and each lower branch represents an unrewarded choice. Thus, the number at each node indicates the posterior over the number of rewards and the number of times the option has been chosen. Traversal through this tree leads to the accumulation of evidence for whether an option is highly rewarding or not rewarding, which in turn affects the agent's future actions. Image credit: Wikimedia commons (bust image); Openclipart.org (map image).

https://doi.org/10.1371/journal.pcbi.1010873.g001

taken, and the second term, $\gamma \sum_{j \in S} p(j s_t, a) u_{t+1}(j)$, estimates the discounted future expected value (FEV) of rewards. The second term, therefore, quantifies the future values of actions taken in the present, i.e. delayed rewards. This second term is the product of the discount factor, γ, and an expectation over future utilities, $u_{t+1}(j)$, with the expectation taken over the transition function, which is the conditional distribution of futures states, $p(j|s_t, a)$. Thus, the equation can also be framed as:

$$Q(s_t, a) = IEV + \gamma * FEV$$

where IEV is the immediate expected value and FEV is the future expected value. For the (mostly) episodic tasks we will consider, the maximum average reward per episode would be obtained by an agent with a discount factor, γ, of 1.0 and the transition function given by the environment or the task. Algorithmically, discount factors are important for fitting infinite horizon models [53] but play a smaller role in fitting finite horizon, episodic models, unless episodes are very long. Discount factors are traits of agents, artificial or biological, and are not part of the environment. Naturally, if the discount factor, γ, is low, then the FEV affects the action value less.

Here we demonstrate parameter regimes where impulsive agents can perform better than non-impulsive agents; this effect is strongest when there is a mismatch between the agent's expectation and the environment. In laboratory experiments, the question becomes whether reductions in weighted FEV occur due to a change in discount factor (γ) or due to a change in the transition function ($\sum_{j \in S} p(j|s_t, a) u_{t+1}(j)$). The transition function is not always given (e.g. in temporal discounting tasks), or, when it is given, it may not be accurately approximated by subjects [54], and this misestimation can be mathematically indistinguishable from a change in discount factor. For example, participants may assume that environments are less predictable than is suggested by the experimenter (i.e. that the entropy of $p(j|s_t, a)$ is higher than stated), because participants have adapted to unstable environments outside the lab. This could result in an overall adjustment of discounting through the discount factor or flattening of the probability distribution affecting transitions to future states. In either case, the FEV is reduced, and the participant is more likely to choose immediate rewards. More formally, in unpredictable environments the conditional distribution, $p(j|s_t, a)$, has higher entropy, meaning that one cannot make choices that lead to desired future states, j. If some future states are rewarding and some are not, unpredictability means that the expectation over future utilities, $\sum_{j \in S} p(j|s_t, a) u_{t+1}(j)$ will be smaller, or even negative. Because the value of delayed rewards is the product of the discount factor and the expectation over future utilities, subjects that do not value delayed rewards may be doing so because they have a lower discount factor, or because they assume environments have unpredictable transition functions. In laboratory experiments this is usually assumed to load on the discount factor, but these effects can also be captured by increasing the uncertainty of the transition function [43]. In this manuscript, we demonstrate that when the discount factor is low, this reduces the impact of the FEV and any related uncertainty caused by changes in $p(j|s_t, a)$. In cases where $p(j|s_t, a)$ is lower than expected, and future rewards are less likely, an impulsive agent can fare better than a non-impulsive agent.

In the present study we examined the tradeoff between the discount factor and uncertainty in three decision-making tasks that can be related to each other through the discount factor and MDP framework. We show that when task environments are more uncertain than an MDP model expects, agents with smaller discount factors outperform agents with higher discount factors in tasks where discount factors of 1 would be optimal if the transition function was accurately approximated by the agent. This correspondingly implies that agents, and possibly human subjects, that are adapted to relatively uncertain environments can outperform

agents not adapted to uncertainty. While this second point follows directly from the models, it leads to an interpretation of impulsive choice strategies as optimal adaptations to environments with substantial uncertainty, rather than pathological deficits in decision making.

Results

The goal of this study was to examine the hypothesis that impulsive choice strategies, defined as a relative preference for immediate over future rewards through the discount factor, can perform better than non-impulsive choice strategies, when environments are more uncertain than expected. More specifically, when agents are not able to make choices that lead to preferred future states, due to environmental variability, choice strategies that favor immediate rewards can be superior. We combined models of three decision-making tasks: Temporal Discounting, Beads, and Explore-Exploit into a single MDP framework and related the tasks to each other through the discount factor, which has been previously used to operationalize impulsive choice behavior [55,56]. In all three tasks, we dissociated the expectations of the agent from the true uncertainty in the environment, to establish the conditions under which an impulsive choice strategy would be beneficial. For each task, we varied the parameters to simulate uncertain and certain environments, to test whether impulsive and non-impulsive agents would fare better. In the certain environments, future rewards were more likely than agent's expectations, and in the uncertain environments, less likely. To model impulsive and non-impulsive agents, we varied the discount factor, which captures the value of future rewards, and computed action values in the model. Thus, impulsive agents have lower discount factors ($\gamma_{Impulsive}$) and weight immediate rewards more, and non-impulsive agents have higher discount factors ($\gamma_{Non-Impulsive}$) and weight future rewards relatively more than immediate rewards. Although the statistics assumed by the agent vs. those that characterize the environment can be dissociated, only agents have discount factors.

In the Temporal Discounting task, the agents were given pairs of options with varying reward magnitudes and delays. Without manipulation of the future reward probability, the agent with the higher discount factor (i.e. less discounting) will collect more reward for choosing the larger, delayed rewards. However, we demonstrate that when the future reward is more uncertain than expected, the impulsive agent collects more average reward. In the Information Sampling task, the impulsive and non-impulsive agents are given bead draw sequences that are more or less informative about the majority color than expected. We demonstrate that when the bead information is less informative than expected, the impulsive agent collects more average reward by avoiding excessive draw costs for low value information. In the Explore-Exploit Task, the impulsive and non-impulsive agents choose between three bandits to learn which is the most rewarding option. Periodically, one of the bandits is replaced with a novel bandit. We demonstrate that when the substitution rate is high, the impulsive agent collects more average reward by not exploring the novel options. Thus, across three decision-making tasks, we show that when future rewards are more uncertain than expected, impulsive choices can lead to more reward.

Impulsive agents benefit from choosing immediate rewards in a Temporal Discounting task

The Temporal Discounting task was based on the Kirby delayed discounting questionnaire, which is typically used to evaluate how human participants value immediate and delayed rewards [38,43,57,58]. In this task and similar temporal discounting tasks, participants are presented with a set of choices between smaller immediate monetary rewards and larger, delayed monetary rewards (**Fig 2A**). Previous work has shown that delayed rewards are typically

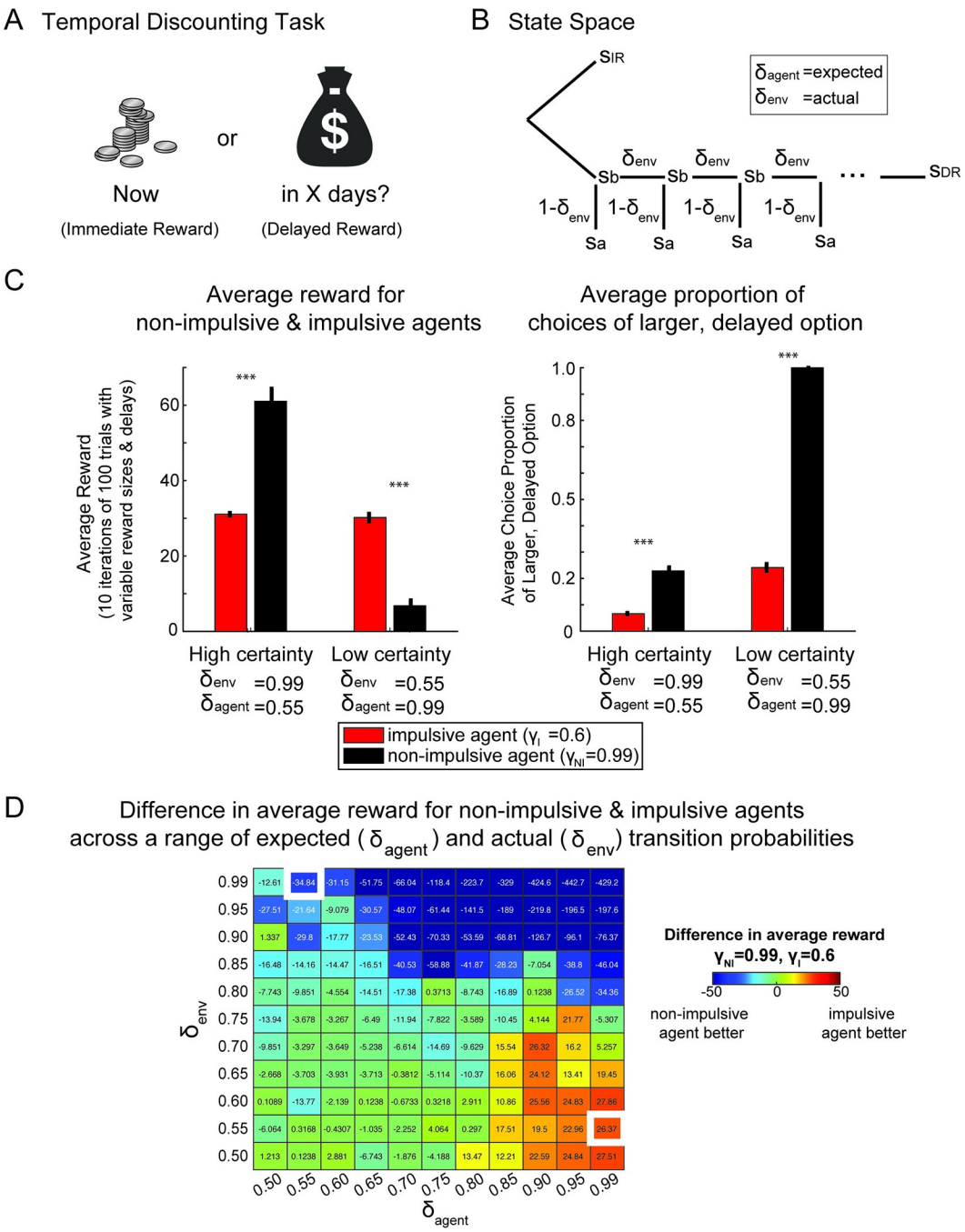

Fig 2. Temporal Discounting task and performance of impulsive and non-impulsive agents in different task environments. A) Task schematic for the Temporal Discounting task. Participants or agents are given a series of questions with two offers, one for a small immediate reward and the other for a larger, delayed reward. **B)** The state space tree for one pair of options in the task. The agent starts on the far left with a choice between the immediate reward or delayed reward. If the immediate reward is chosen, the agent proceeds on the upper branch to the immediate reward state (s_{IR}) and always collects the immediate reward. If the agent chooses the delayed reward, the agent proceeds through the lower branch towards the delayed reward state (s_{DR}). Along this branch are a sequence of intermediate transition states (s_b) which the agent progresses through with probability δ. At each transition state, the agent might proceed to a terminal, non-rewarding state (s_a) with probability 1-δ. The number of transition states is defined by the delay to the larger reward. **C)** Average reward collected and choice behavior across simulated trials in certain and uncertain task environments for impulsive and non-impulsive agents in the Temporal Discounting task. "High certainty" is when $\delta_{env} > \delta_{agent}$ and "low certainty" is when $\delta_{agent} < \delta_{env}$. The non-impulsive agent (black) has a discount factor of γ = 0.99 and the impulsive agent (red) has a discount factor of γ = 0.6. Left: Average reward collected for the two agents. Right: Average proportion of trials in which an agent

selected the larger, delayed option. Error bars are s.e.m. across 10 iterations of 100 trials using variable reward sizes and delays. *** indicates p<0.0001 paired t-test. **D)** Difference in average reward across a range of δ_{env} and δ_{agent} values. The heatmap shows domains where the non-impulsive agent performs better (more blue), the impulsive agent performs better (more red) or there are marginal differences between the two agents (red). The values shown in each box on the heatmap is the difference in average reward for the two agents. The white boxes indicate the task regimes shown in Fig 2C. See S1 Fig for other discount factors. Image credit: Openclipart.org (coins image, money image).

https://doi.org/10.1371/journal.pcbi.1010873.g002

valued less than immediate rewards of the same size. However, it remains unclear why future rewards are discounted, and there exist multiple possible mechanisms [59]. Here we examined the performance of impulsive (low discount factor, $\gamma = 0.6$) and non-impulsive (high discount factor, $\gamma = 0.99$) agents that also assumed different state-transition probabilities to future rewards. We examined the performance of these agents in environments where the actual transitions to future rewards were stochastic, such that future rewards were not always collected. The link between uncertainty and the performance of impulsive and non-impulsive agents is straightforward in this task. However, it illustrates the point that we generalize in subsequent tasks.

The state space for this task consists of two branches, one representing the smaller, immediate reward, and the other representing the larger, delayed reward (**Fig 2B**). If the immediate reward is chosen, then progression to the terminal, rewarding state, s_{IR} is guaranteed and the reward is collected. If the delayed reward is chosen, the agent proceeds through a sequence of states representing the passage of time. The agent progresses through transition states (s_b) towards the final delayed reward state (s_{DR}) with probability δ or terminates at intermediate, non-rewarding states (s_a) with probability 1−δ at each intermediate timestep. The transition states represent the passage of time and uncertainty about the delayed reward. The only decision is whether to take the immediate reward, or to pursue the future, larger reward.

In the model, the future expected value (FEV) of the delayed option, from the initial choice state, is calculated by discretizing the delay to the larger reward into steps with a probability of transitioning to each step with δ_{env}. When the transition probability, δ_{env}, to the delayed reward is high, the FEV of the delayed option is higher than the immediate reward. On the contrary, when the transition probability is low, the FEV of the delayed option is small.

Expanding upon the idea that in some cases, the value of the immediate option can be larger than the FEV of the delayed option, we examined whether an agent that discounted future rewards (i.e. impulsive) might fare better on average when the certainty of the delayed reward in the environment was worse than expected. In this case an agent expects a transition probability that is higher than the actual transition probability in the environment. We tested agents with two different discount factors (impulsive and non-impulsive) and two different transition probability assumptions, in two different environments. Specifically, we tested impulsive and non-impulsive agents under conditions in which the probability of transitioning to the delayed reward in the environment is higher than expected by the agent ($\delta_{env} = 0.99$, $\delta_{agent} = 0.55$) and under conditions in which the probability of transitioning to the delayed reward in the environment is lower than expected by the agent ($\delta_{env} = 0.55$, $\delta_{agent} = 0.99$). We simulated batches of trials with various sizes of rewards and delays. We then used the discount factor, γ, to model variable levels of discounting to reflect impulsive or non-impulsive behavior (γ_I and γ_{NI}, respectively).

The results from testing these two agents in the high and low certainty environments shows that in the high certainty environment, the impulsive agent collects less average reward than the non-impulsive agent (**Fig 2C left**; paired sample t-test, $t(9) = -20.92$, $p<0.001$, $d = -3.66$, power > 0.99). In the low certainty environment, the impulsive agent fares better than the non-impulsive agent, by collecting more average reward (paired sample t-test, $t(9) = 12.84$,

$p<0.001$, $d = 6.06$, power > 0.99). This outcome is driven by the frequency with which each agent selects the larger, delayed option in each environment. In both environments, the non-impulsive agent selects the larger, delayed option more often (**Fig 2C, right**). In the high certainty environment, when the transition probability is higher than expected ($\delta_{agent} = 0.55$, $\delta_{env} = 0.99$), the non-impulsive agent selects the delayed option more than the impulsive agent, due to the higher discount factor of the agent (paired sample t-test, $t(9) = -21.10$, $p<0.001$, $d = -5.76$, power > 0.99). However, the non-impulsive agent only selects the larger, delayed reward, about 30% of the time, due to the expectation of a low transition probability to delated rewards, as $\delta_{agent} = 0.55$. In the low certainty environment ($\delta_{agent} = 0.99$, $\delta_{env} = 0.55$), the non-impulsive agent selects the delayed option every time and significantly more than the impulsive agent (paired sample t-test, $t(9) = -52.25$, $p<0.001$, $d = -27.91$, power > 0.99), due to the expectation of a high transition probability to the delayed reward. The impulsive agent selects the delayed option less often in both environments. There are combinations of transition probabilities for which the impulsive agent collects more reward, less reward, or roughly equal reward to the non-impulsive agent (Fig 2D). In general, when $\delta_{agent}<\delta_{env}$, the non-impulsive agent collects more average reward and when $\delta_{agent}>\delta_{env}$, the impulsive agent collects more average reward. When δ_{agent} and δ_{env} are both high (approximately > 0.8), the non-impulsive agent collects more average reward. Note that when both δ_{agent} and δ_{env} are very low (i.e. 0.55 and 0.5), the impulsive agent can collect at least as much or marginally more reward than the non-impulsive agent, showing that the main effects are driven by the mismatch between expected transition probability and actual transition probability. Furthermore, the effect sizes between the pairs of agents decrease as γ_I becomes closer to γ_{NI}, as expected, but these relationships between δ_{agent}, δ_{env}, and reward remain the same (S1 Fig). Power analyses were conducted to make recommendations for an experiment with human subjects. Assuming an allocation ratio of 1.0 (i.e. equal number of subjects for each group), minimum power of 0.8, and alpha of 0.05, an experimenter would only need 3 participants with 100 completed trials, in each group to find a significant difference in average reward collected. However, given that the variability of human participants would be higher than that of our simulated agent behavior derived with a single discount factor, this is a low estimate of the number of subjects needed to run an experiment and see effects. For smaller effects in mean reward, (e.g. in the domain of $\delta_{agent}<0.75$ & $\delta_{env} = 0.65$), power analyses suggest that an average of 200 participants would be required to detect statistical differences in mean reward.

In summary, choosing the immediate option in the Temporal Discounting task is advantageous when the larger, delayed reward is more uncertain than expected. This suggests that in a more complex task, it might be possible to find a regime where choosing immediate rewards is also beneficial. We discuss examples of such tasks in the following two sections.

Impulsive agents benefit from guessing sooner in an Information Sampling (Beads) task

In information sampling tasks, the objective is to collect information and make an informed decision based on accumulated evidence. We used the previously developed Beads task [41,44,45,60] to examine information sampling behavior. In the Beads task, the objective is to correctly guess the color of the majority of beads in an urn with two colors of beads (Fig 3A). To collect information about the proportions of colors, participants must draw one bead at a time, and incur a cost for each draw. Thus, at each step in the task, participants either choose to draw a bead or guess the majority bead color in the urn. If they guess correctly, they receive a reward (+10) and if they guess incorrectly, they receive a penalty (-12). This decision-making sequence can be represented with a state-space tree (Fig 3B). In this diagram, each node

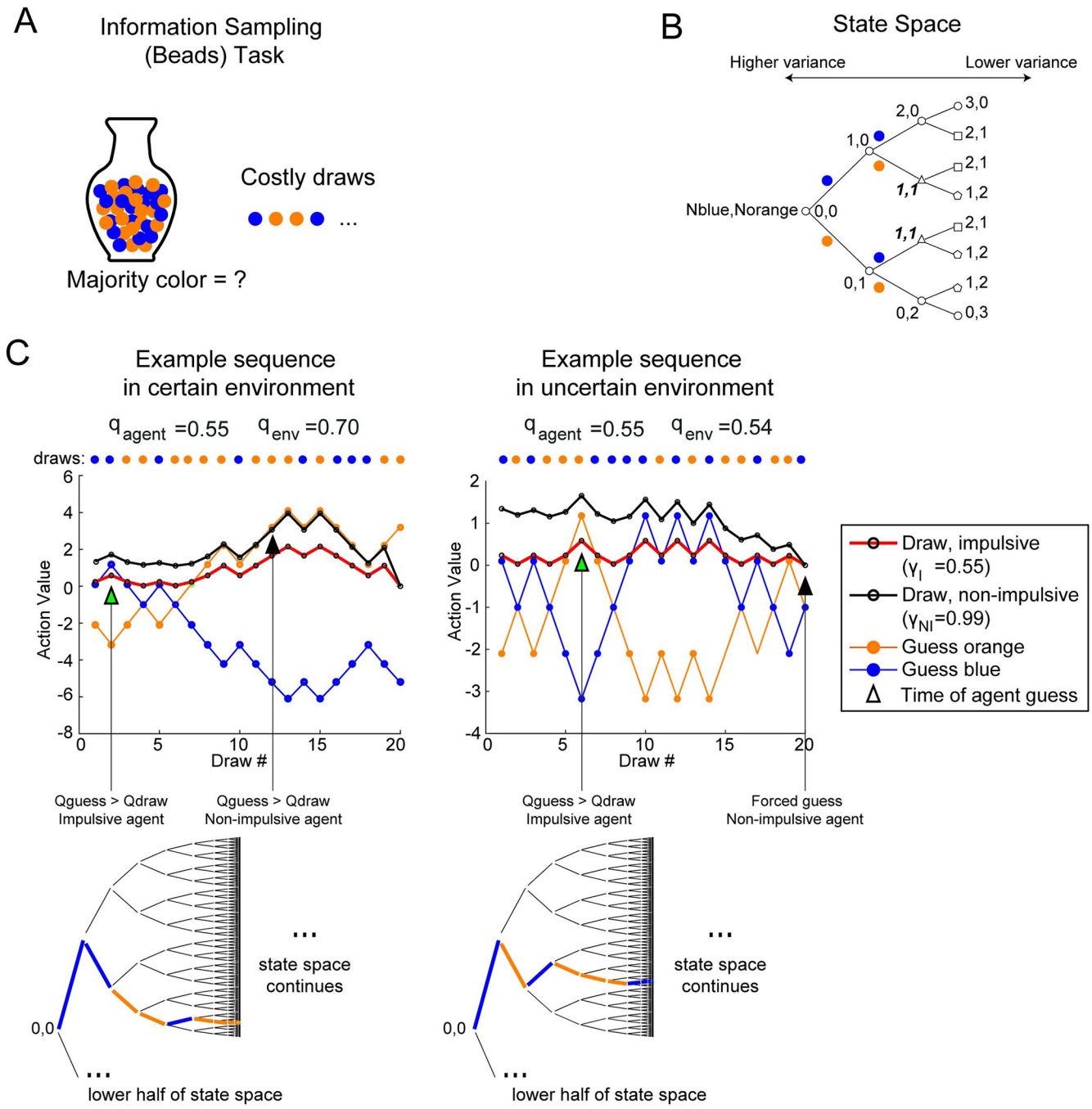

Fig 3. Information Sampling (Beads) Task and examples of agent performance in high and low certainty environments. A) Task schematic for the Beads task. In this task, the objective is to correctly guess the majority color of beads (e.g. orange or blue) in the urn. The participant or agent has the option to draw one bead at a time (for a cost, e.g. $0.10) to accumulate evidence. The agent's goal is to accumulate sufficient evidence to make a confident guess without incurring maximum draw cost. Once the maximum number of draws is reached, the agent is forced to guess a color. An agent receives a reward for a correct guess (e.g. $10) or a cost for an incorrect guess (e.g. -$12). **B)** The state space tree for the beads task up to 3 draws. Each node represents the number of orange and blue beads that have been drawn thus far and a decision point, where the agent can either draw again, guess orange, or guess blue. If the agent draws another bead, they stochastically transition to the next state according to a binomial probability. At the start of the tree, the variance in the probability distribution over the majority probability is highest and decreases with increasing numbers of draws. Note that the states with the same number of orange and blue beads after 3 draws are the same state. We draw repeated states as separate for clarity,. Repeated states are illustrated by the shape of the node. Circular nodes are unique, nodes of other shapes indicate a repeated state. **C)** Two example bead draw sequences in certain and uncertain task environments and the behavior of impulsive and non-impulsive agents. On the left, a sequence of 20 draws is shown from a set of task parameters that creates an environment where there is high certainty the majority color is orange (q_{agent} = 0.55, q_{env} = 0.7, C_{draw} = 0.10, $R_{correct}$ = 10, $R_{incorrect}$ = -12). The plot shows the action values for guessing orange and guessing blue which are identical for both agents. The plot also shows the corresponding action values for drawing a

bead for the non-impulsive agent (black) and the impulsive agent (red). Because the agents always select the largest action value on each time step, the agents only guess a color when the action value for guessing blue or orange surpasses the action value to draw another bead. In the case on the left, the non-impulsive agent guesses orange correctly after 11 draws (black arrow), whereas the impulsive agent guesses blue incorrectly after the first draw (green arrow). In the uncertain case (right), the task parameters create an environment where there is low certainty about the majority color ($q_{agent} = 0.55$, $q_{env} = 0.54$, $C_{draw} = 0.10$, $R_{correct} = 10$, $R_{incorrect} = -12$). The same traces for the action values are shown. The non-impulsive agent draws until it is forced to guess and incurs maximum draw cost (black arrow). The impulsive agent guesses correctly after 5 draws (green arrow). Below each plot of action values are the corresponding truncated state space trees, showing traversal through the state space for the example bead sequences. Only the top half of the state space tree is expanded through the first 10 bead draws.

https://doi.org/10.1371/journal.pcbi.1010873.g003

represents a decision point to either draw a bead or guess the color of the urn. The state is given by the number of blue beads and the number of orange beads that have been drawn. At the start of the tree, (0,0), there are no beads of either color. As we proceed deeper into the tree, the variance of the binomial distribution over the proportion of beads of each color gets lower as we accumulate information through bead draws, and the estimates of the fraction of beads of each color is more accurate. If the urn fraction is low, e.g. 60%/40%, the uncertainty around the correct guess decreases slowly.

We hypothesized that an impulsive agent might fare better than a non-impulsive agent when the majority fraction of beads in the urn is lower than expected by the agent, and therefore bead draws are less informative than expected. To test this, we examined a condition in which the agents believed the majority color in the urn (q_{agent}) was not far above chance (e.g. $q_{agent} = 0.55$). We then compared performance of impulsive ($\gamma_I = 0.55$) and non-impulsive ($\gamma_{NI} = 0.99$) agents in situations where the environment was more or less certain than expected (**Fig 3C**). The agent for this task has three actions available at each step: draw, guess blue, and guess orange, and at each step the agent picks the action with the highest value. In the more certain environment, the true underlying bead majority (q_{env}) is 70% orange, and the action value for guessing orange continues to increase as draws are made and evidence accumulates that orange is the majority color (**Fig 3C, left**). When the action value for guessing one of the two colors surpasses the action value for drawing a bead, an agent will stop and guess that color. For the impulsive agent in the certain environment, the agent guesses a color after the first draw, which leads to an incorrect choice of blue. On the contrary, the action value for drawing a bead for the non-impulsive agent starts out high and the action value for guessing orange surpasses the action value for drawing only after 11 draws. The agent has accumulated evidence that the majority is likely orange and chooses orange correctly. The non-impulsive agent with the higher discount factor values future rewards and is driven to go further into the state space to reduce uncertainty about the majority color. The choice to guess a color terminates the sequence and therefore does not depend on the discount factor, as there are no future possible states that can be reached after a guess, and the discount factor only affects future state values.

On the other hand, in the uncertain environment, the action values for guessing each color do not diverge as clearly because subsequent draws are not consistently of one color (**Fig 3C, right**). In this example, the beads are sampled from an urn with 54% orange beads. This low majority drives the action value to draw a bead for the non-impulsive agent to stay higher than the action values for guessing the two colors until the max number of draws allowed, at which time the agent is forced to guess a color. In contrast, the impulsive agent makes multiple draws, but fewer than the non-impulsive agent, and guesses orange correctly. The impulsive agent was able to guess without accruing as much cost from the charges for additional bead draws. The partial state spaces for these examples show that the bead sequences start out identically for the first two draws of each of the bead sequences, but then the draws in the certain environment (left) quickly dive towards the lower edge of the subtree, reflecting increased

probability of a majority color (**Fig 3C, bottom**). The path through the state space in the uncertain environment meanders towards the middle branches of the state space tree. On average, the closer to chance the majority color fraction, the less consistent the path through the state-space will be across trials of bead sequences.

To compare the average performance and choice behavior of the two agents in these environments, we simulated batches of bead sequences and choices using agents with the two discount factors. In the certain environment, where the majority fraction of beads is high, the non-impulsive agent with the higher discount factor ($\gamma_{NI} = 0.99$, black) collects more average reward (paired sample t-test, $t(99) = -20.70$, $p<0.001$, $d = -2.93$, power > 0.99) (**Fig 4A**). In the uncertain environment, the impulsive agent ($\gamma_I = 0.55$, red) collects more reward, despite both agents collecting less reward than in the certain environment (paired sample t-test, $t(99) = 4.16$, $p<0.005$, $d = 0.59$, power > 0.99). The reason for this is illustrated by the average number of draws each agent takes before guessing the color of the majority (**Fig 4A, right panel**). In both task environments, the non-impulsive agent makes more average draws before making a choice (paired sample t-test, $t(99) = -104.76$, $p<0.001$, $d = -14.82$, power > 0.99 for certain environment, $t(99) = -139.74$, $p<0.001$, $d = -19.76$, power > 0.99 for uncertain environment). This leads to a more informed choice in the certain environment, but in the uncertain environment, this only leads to a small improvement in guessing accuracy, and on average accrues more cost. The impulsive agent on the other hand, does not make as many draws before making a guess about the majority color, and thus avoids accruing additional draw costs for draws that do not improve the accuracy of the guess. The bead information in the uncertain environment is not only less reliable than expected, as the actual fraction of beads of one color (q_{env}) is lower than what is expected by the agent (q_{agent}), but also less informative, as the environment majority fraction (q_{env}) is closer to 0.5. The bead information in the certain environment is also unreliable, in the sense that it does not reflect the expected majority fraction (q_{agent}), but is more informative, as it provides a better estimate of the actual majority color.

Power analyses suggest that to observe these effects in an experiment with human subjects, (assuming minimum power of 0.8, alpha 0.05), a minimum of two subjects would be required to observe differences in choice behavior. However, approximately 20 participants in each group, impulsive and non-impulsive, would be the minimum number of subjects to observe differences in average reward collected in the uncertain environment as shown in Fig 4A. This number is a low estimate, as an experiment would have to account for the variability in discounting across subjects in the participant pools.

We also examined relative performance across a wider space of parameters, including the majority fraction of beads in the urn that is used to generate the bead draw sequences (q_{env}), the agent's belief about the majority fraction of beads (q_{agent}), the draw cost (C_{draw}), and the model discount factor (γ). We varied these parameters for multiple impulsive agents ($\gamma = 0.55$, 0.6, 065) and compared the average reward collected by the impulsive agents and the non-impulsive agents ($\gamma = 0.99$) across these task conditions (**Fig 4B**). As q_{agent} increases, the area of the parameter domain in which the non-impulsive agent fares better, expands. There exists a range of task parameters where an impulsive agent can collect more reward than a non-impulsive agent. For all task conditions, $R_{correct}$ was +10 and R_{error} was -12 to encourage more than one draw from the impulsive agent. However, there also exist parameter domains where impulsive agents can perform better than non-impulsive agents when $R_{correct} = |R_{error}|$ (see **S2 Fig**). Thus, in an information sampling task, impulsive behavior can be beneficial when the information that is accumulated is less informative than expected and is associated with a growing cost.

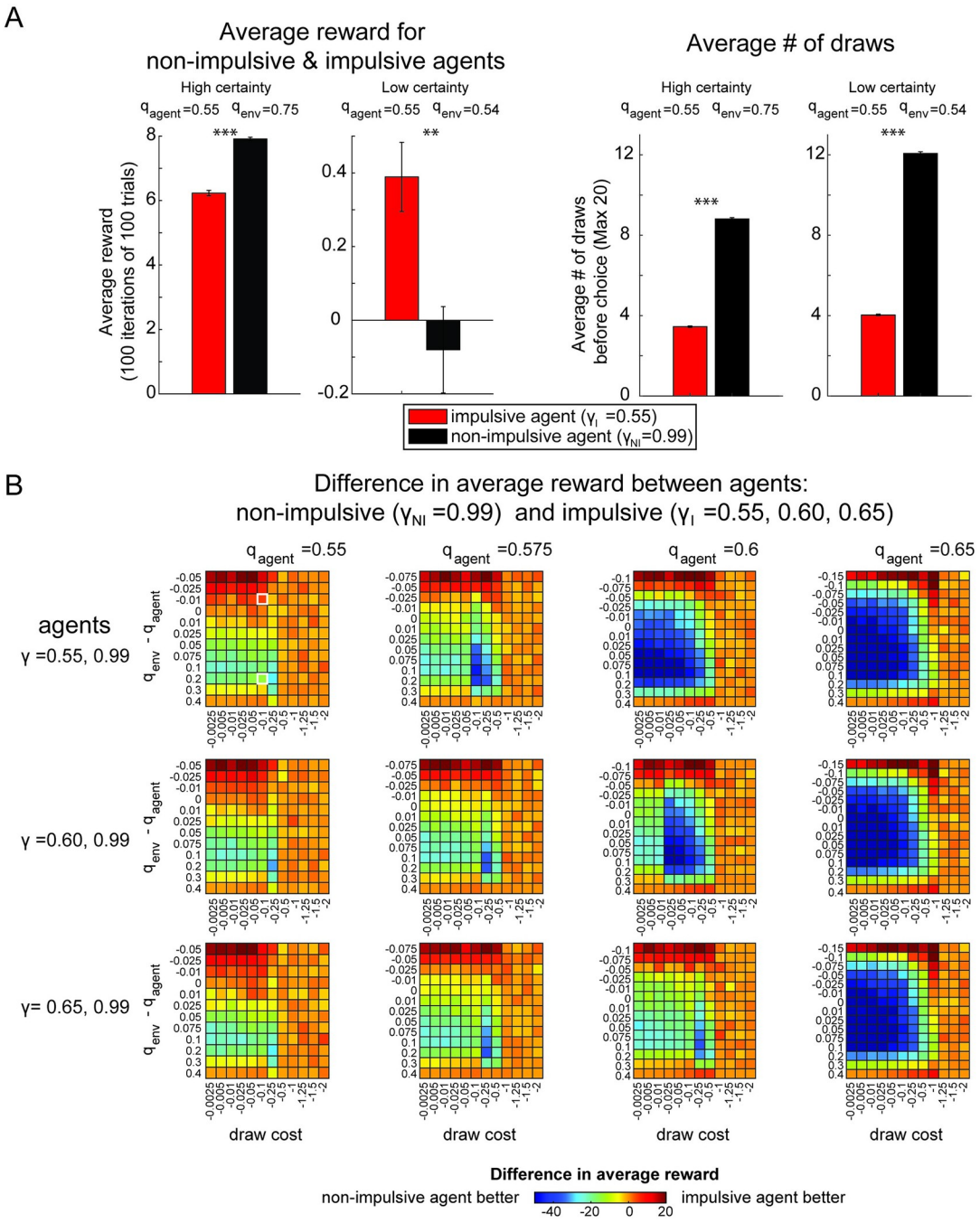

Fig 4. Performance and choice behavior of impulsive vs. non-impulsive agents in the Information Sampling (Beads) Task. A) Left: Average reward collected across simulated trials in certain and uncertain task environments for impulsive and non-impulsive agents in the Beads task. The non-impulsive agent (black) has a discount factor of γ = 0.99 and the impulsive agent (red) has a discount factor of γ = 0.55. Error bars are s.e.m. across 100 iterations of 100 trials using bead sequences from two different task parameters (q_{env} = 0.75 certain environment, q_{env} = 0.54 uncertain environment, q_{agent} = 0.55. C_{draw} = 0.1). **Right:** Average number of bead draws before guessing a color for each model in each task environment. In both task environments, the impulsive agent (red) draws similarly often, but significantly less than the non-impulsive agent (black). *** indicates $p<0.0001$ ** indicates $p<0.001$. **B)** Model performance across a range of parameter values. Each panel is a heatmap showing the differences in average reward for a pair of non-impulsive and impulsive agents, indicated by the discount factors on the far left. Each column has a set of heatmaps for the expected majority fraction of beads, q_{agent}. Each row has a set of heatmaps for a pair of discount factors (impulsive & non-impulsive). The x-axis of each heatmap is the draw cost and the y-axis is the difference between the model input q_{agent} and the majority fraction used to generate the bead draws, q_{env}. The color of the heatmap indicates whether the impulsive agent (red) or non-impulsive agent (black) collected more reward. More blue values indicate

the non-impulsive agent collected more average reward and more red values indicate the impulsive agent collected more reward. As q_{agent} increases (left to right), the domain in which the non-impulsive agent performs better expands. The white boxes in the heatmap in the top left panel highlight the data used to create the bar plots Fig 4A (left). All heatmaps were generated using $R_{correct} = 10$, $R_{incorrect} = -12$. See S1 Fig for (B) with $R_{correct} = 10$, $R_{incorrect} = -10$.

https://doi.org/10.1371/journal.pcbi.1010873.g004

Impulsive agents benefit by exploring novel options less in an Explore-Exploit task

In the Explore-Exploit task, there are three options that pay off with an equal, fixed reward, but with variable reward probabilities. An agent must learn which option is most valuable by selecting the options and experiencing reward. The bandits are stationary, in that the reward rate for each option remains fixed. However, novel choice options replace familiar options at stochastic intervals. When this happens the agent must choose between exploring the novel option, which has an expected value of 0.5 before it is sampled, and exploiting familiar options, for which the agent has an estimated reward probability (**Fig 5A**). In this example series of trials, three choices (A, B, and C) are shown. Through exploration of these options, the agent learns the approximate reward rate of each of the options, and should learn to pick A more often, as it is the most valuable. In the last panel in the series, a novel option is introduced to replace option A. The value of the novel option on the first trial is not known. The rate of replacing an option with a novel option is parametrized with the substitution rate of the environment (p_{env}). The higher the substitution rate, the more volatile the environment. Substitution with novel options affects where an agent is in the state space.

The state space for this task can be represented with one binomial tree for each option. As an option is chosen, the agent traverses that particular tree towards the upper half of the tree if the option is rewarded, and lower half if the option is unrewarded. Introduction of novel options resets the tree for the option which was replaced to the root node. For example, consider an agent that selects among three options (A, B, and C) and makes a sequence of three choices: C, A, B, A, A, B. After these choices, a novel option is introduced to replace option A (**Fig 5B**). In this example, assume A was chosen 3 times and rewarded 3 times, then the position in the tree would be along the uppermost branch of the state space tree for option A. If B was chosen twice, and rewarded and then not, rewarded, the tree for option B would appear as shown. C was chosen once, and not rewarded, and the position in the state space would be one step along the lowermost branch. When a novel option is introduced, as in this case, after three choices for option A (N^A), the agent's position in the tree for option A jumps back to the start, because now nothing is known about the new option. The positions in the other choice trees (B and C) remain the same. As the substitution rate gets higher, the agent rarely reaches deep nodes in the tree, which reflects more accurate reward probability estimates, for any of the options, because they are replaced before the agent can reach an accurate estimate of the value of an option. We hypothesized that in this context, an agent that discounts future rewards might fare better by exploiting known options as long as possible, rather than exploring novel options. Exploring novel options is a time investment that only pays off, on average, in the future, and therefore exploration in this context is more valuable with higher time-constants, because the relative value of exploration is obtained in the future.

To test the hypothesis that an impulsive strategy would fare better than a non-impulsive strategy when the novel substitution rate was high, we varied the discount factor ($\gamma_I = 0.65$, $\gamma_{NI} = 0.99$) and novel option substitution rate (p_{env}). Similar to the beads task, we examined a situation in which the agent believed the substitution rate (i.e. the probability) per trial was 0.08, and the actual substitution rate was above (0.2) or below (0.02) that value (i.e. $p_{agent} = 0.08$, $p_{env} = 0.02$ or 0.2). When the substitution rate is higher, the environment is more uncertain

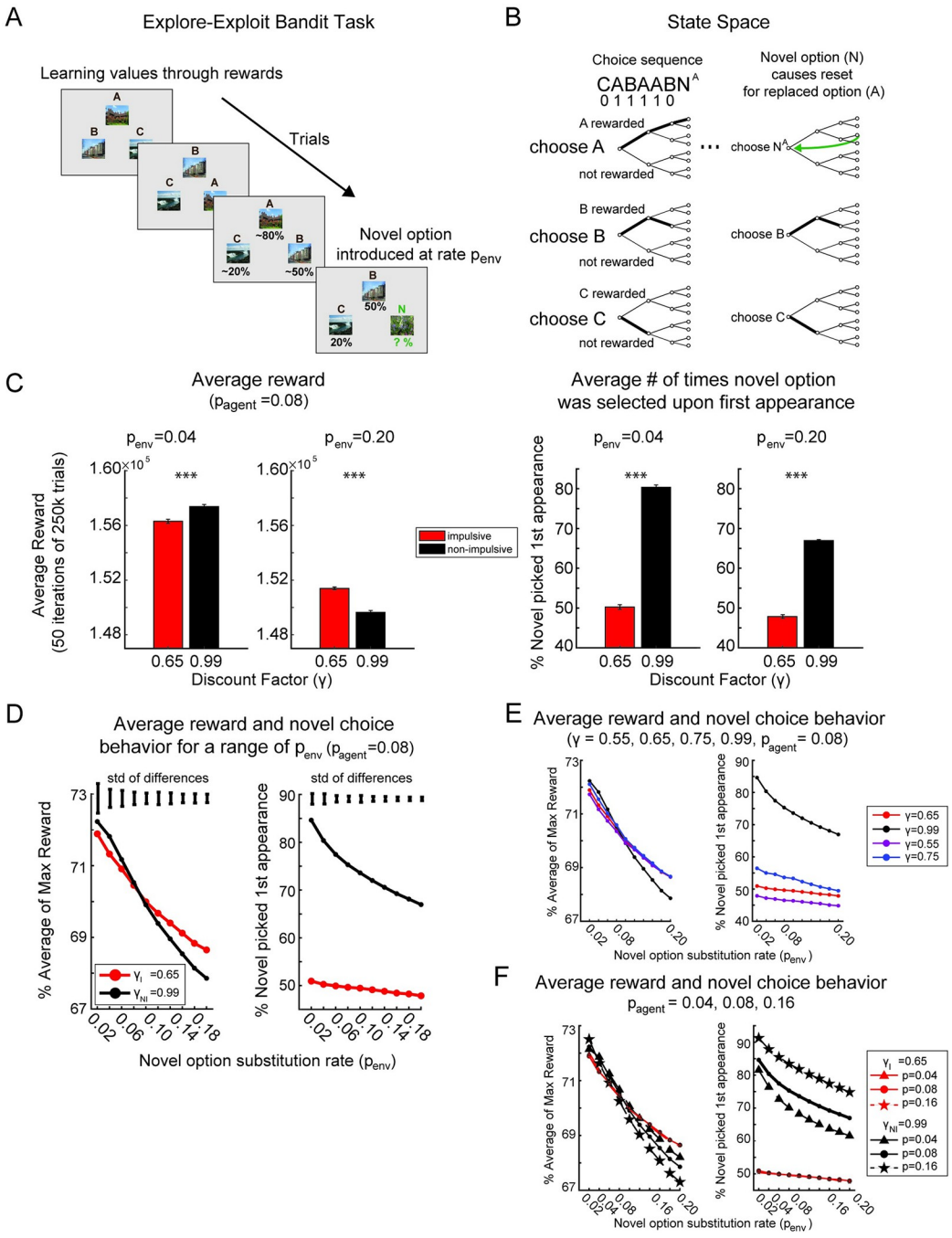

Fig 5. Explore-Exploit bandit task with novelty. A) Example sequence of trials in the Explore-Exploit task. In this example, each option (A, B, C) is a picture with an underlying reward rate. The agent or participant must learn the values of the three options through experience of choosing the options and receiving or not receiving a reward on each trial. In this example, the agent has learned the approximate values of the options over the course of multiple trials (not all are shown) and then a novel option is substituted for one of the options (option A). The novel option substitution rate (p_{env}) affects the number of trials the agent has to learn about an option. When p_{env} is high, it is harder for the agent to learn about the underlying values of the options. **B)** State space representation of the Explore-Exploit task. Each option can be represented with a separate subtree. Thus, for an example sequence of choices such as C, A, B, A, A, B, the agent progresses through 1 step of the tree for C, three steps for A, and two steps for B. The agent progresses to an upper branch or lower branch depending on whether the choice was rewarded. Rewards are shown for this example sequence as 0s or 1s. Thus, the agent progresses through the uppermost branches of the subtree for option A, as it was rewarded all three times it was chosen. The introduction of a novel option causes the position in the subtree for that option to reset. When the novel option is presented at the end of this sequence and replaces option A, the agent jumps back to the start for that option, as

reward history no longer represents the underlying value of the novel option. **C)** Bar plots of average raw reward (left) and average selection of the novel option upon first appearance (right) in low and high certainty environments. On the left, average reward for the non-impulsive (black) and impulsive (red) agents at p_{env} values of 0.04 and 0.20 are shown. On the right, average selection of the novel option upon first appearance is shown for the same values of p_{env}. *** indicates $p<0.0001$. Error bars on plots s.e.m. across iterations. Error bars above plots represent the standard deviation of the differences between the mean values for the non-impulsive and impulsive agents. **D)** Average % of maximum possible reward and average novel option choice behavior across a range of novel option substitution rates for both the non-impulsive (black) and impulsive agent (red). On the left, the plot of average reward shows a that when the novel option substitution rate (p_{env}) is low, the non-impulsive agent collects more reward than the impulsive agent, but when p_{env} is high (greater than 0.1), the impulsive agent collects more reward than the non-impulsive agent. The plot of novel choice behavior shows that for all novel option substitution rates tested, the non-impulsive agent selects the novel option significantly more often than the impulsive agent on the first trial it appears. Error bars above the graphs represent the standard deviation of the differences between the mean values for non-impulsive and impulsive agents. **E)** Average % of maximum reward and choice behavior for a range of discount factors and $p_{agent} = 0.08$. **F)** Average % of maximum reward and choice behavior for discount factors shown in (C) and (D) with $p_{agent} = 0.04, 0.08, 0.16$. Image credit: Wikimedia Commons (scene images).

https://doi.org/10.1371/journal.pcbi.1010873.g005

because options are frequently replaced, and when the substitution rate is lower the environment is less uncertain. In the case where the substitution rate in the environment was lower than the agent expected, the non-impulsive agent collected more average reward (**Fig 5C, left**) (paired sample t-test, $t(49) = -13.74$, $p < 0.0001$, $d = -1.94$, power > 0.99). In the case where the substitution rate in the environment was higher than the agent expected, the impulsive agent collected more average reward (**Fig 5C, left**) (paired sample t-test, $t(49) = 47.74$, $p < 0.0001$, $d = 6.75$, power > 0.99). This corresponded to a difference in choice behavior of the novel options. In both sets of task conditions, the impulsive agent selected the novel option less often (**Fig 5C, right**) (paired sample t-test, $t(49) = -202.23$, $p<0.0001$, $d = -28.60$, power > 0.99 for certain environment, $t(49) = -379.45$, $p<0.0001$, $d = -53.66$, power > 0.99 for uncertain environment). As the substitution rate increased, both agents selected the novel option less often.

The average reward collected, and novel choice behavior, differ between the impulsive and non-impulsive agent depending on the novel option substitution rate (**Fig 5D**). The impulsive agent collects less reward than the non-impulsive agent when the novel option substitution rate is lower than 0.06, and more than the non-impulsive agent when the substitution rate is higher than 0.14 (**Fig 5D, left**). Note that at 0.02, only 2 out of every 100 trials are novel option trials, and thus the agents perform similarly due to limited encounters with novel options. On average, the impulsive agent chooses the novel option less upon first appearance across all substitution rates (**Fig 5D, right**). A two-way ANOVA was performed to analyze the effect of discount factor (γ) and novel option substitution rate (p_{env}) on average reward. As substitution rate increased, average reward decreased (main effect: substitution rate, $F(9,980) = 2176.12$, $p<0.001$). At low substitution rates less than 0.10, the non-impulsive agent ($\gamma_{NI} = 0.99$) collected more reward than the impulsive agent ($\gamma_I = 0.65$) and at high substitution rates, greater than 0.10, this effect reversed such that the impulsive agent collected more average reward than the non-impulsive agent (main effect: discount factor $F(1,980) = 91.01$, $p<0.001$, interaction: substitution rate x discount factor $F(9,980) = 65.16$, $p<0.001$). Similarly, a two-way ANOVA was performed to assess the effect of discount factor and substitution rate on novel choice behavior. As the substitution rate increased, selection of the novel choice option upon first appearance decreased for both agents (main effect: substitution rate, $F(9,980) = 2298.43$, $p<0.0001$). Furthermore, the non-impulsive agent selected the novel option significantly more often than the impulsive agent across all substitution rates (main effect: discount factor, $F(1,980) = 316014.59$, $p<0.0001$). A change in environment substitution rate had a larger effect on novel choice behavior for the non-impulsive agent, as the non-impulsive agent had a ~20% decrease in selection of the novel option from the lowest substitution rate ($p_{env} = 0.02$) to the highest substitution rate ($p_{env} = 0.20$) and the non-impulsive agent only selected the novel

option 45–50% of the time across all substitution rates (interaction: substitution rate x discount factor, $F(9,980) = 1170.38$, $p<0.0001$).

We also examined relative performance across a range of agent substitution rates and discount factors. First, we varied the impulsive agent discount factor (γ_I) while keeping the agent's substitution rate, $p_{agent} = 0.08$, constant (**Fig 5E**). As the discount factor of the impulsive agent (γ_I) became closer to the discount factor of the non-impulsive agent γ_{NI}, the differences between the agents across a range of p_{env} decreased. Next, we varied the agent's substitution rate (p_{agent}) while keeping the discount factors constant ($\gamma_I = 0.65$, $\gamma_{NI} = 0.99$) (**Fig 5F**). Changing the agent's substitution rate had negligible effects on average reward and novel choice behavior for the impulsive agent, however, changing p_{agent} for the non-impulsive agent affected both average reward collected and novel option choice behavior (Fig 5F). When the trained substitution rate was the highest ($p_{agent} = 0.16$), the non-impulsive agent collected the least average reward when p_{env} was greater than 0.10. These results suggest that the discount factor has a larger effect on choice behavior than the trained substitution rate. In all cases, there were no differences in average reward between the impulsive and non-impulsive agents at the agents' trained substitution rates. Thus, it was the differences in discount factors and mismatch between expected substitution rate by the agent (p_{agent}) and the actual substitution rate (p_{env}) that were responsible for differences in average reward and choice behavior between impulsive and non-impulsive agents.

Power analyses showed that to observe effects like those shown in Fig 5C, only 7 iterations would be required to observe the smallest effect. However, each of these iterations included 250,000 trials. To provide guidance for readers with regards to effect sizes that might be observed in experiments using human participants, we ran simulations using a more reasonable number of trials per iteration that might be possible in an experiment, while keeping the number of iterations, 50, fixed. Simulations with 5000 trials for each iteration (or theoretical human participant) produced results that were still significant for some substitution rates, but much weaker and not over the entire range of substitution rates (two-way ANOVA, main effect: substitution rate $F(9,980) = 1.11$, $p = 0.2919$, main effect: discount factor $F(9,980) = 45.46$, $p < 0.0001$, interaction substitution rate x discount factor $F(9,980) = 2.36$, $p < 0.05$). In particular, following the example in Fig 5C, at $p_{sub} = 0.04$ differences in mean reward between agents were not statistically significant (paired sample t-test, $t(49) = -0.79$, $p = 0.43$, $d = -0.11$) but differences at $p_{sub} = 0.20$ remained significant (paired sample t-test, $t(49) = 6.58$, $p<0.0001$, $d = 0.93$). This is because with only 5,000 total trials, $p_{sub} = 0.04$ results in only 200 novel option trials. Thus, if someone were interested in pursuing an experiment with human subjects, it would be possible to titrate the number of trials with available participants in each subject group to observe benefits of impulsive behavior at high substitution rates.

In summary, we have shown that across three common decision-making tasks, that non-impulsive choice strategies can be beneficial. In particular, this is true when task variables create an environment where future rewards are less certain than expected.

Discussion

We used Markov decision process models to examine the trade-off between environmental uncertainty and the advantages of impulsive choice strategies. We found, across three tasks, that when the environment was more uncertain than expected, agents with impulsive choice strategies that favored immediate over future rewards were more effective than agents with less impulsive choice strategies. In Temporal Discounting, an agent that selects an immediate, smaller, certain option, earns more rewards than an agent that selects future, larger, uncertain options. This finding extends to other tasks that have been used to measure impulsivity. In the

Information Sampling task, when subjects draw beads (at a cost) to improve their ability to guess the correct urn color, deciding early is advantageous when beads are less informative than expected. This is particularly true when incorrect choices lead to large losses. Finally, in an Explore-Exploit task in which novel options are periodically introduced, exploration of novel options is only beneficial when they will be available for exploitation in the future. Therefore, when the available options turn over more frequently than expected, exploration is less valuable and impulsive strategies that select options with higher immediate expected values are more advantageous. Our results show that an impulsive choice strategy, which is often considered maladaptive, can be advantageous when environments are consistently more uncertain than expected.

The value of future rewards depends on the ability of an agent to execute a sequence of choices that lead to future states that deliver those rewards. They also rely on the subjective weighting of future rewards. When environments are uncertain, actions will not necessarily lead to desired future states. This leads to decreased future expected values (FEV) and increased relative value of immediate reward. If the conditional distribution of future states, $p(j|s_t, a)$, is broad (i.e. high entropy), conditioned on actions and states, an agent cannot control its transition to a future state because many states are likely to occur. Stated another way, agents have limited ability to control future outcomes. If only a few future states have high utility, and particularly if some futures states have negative utility, this lack of control will significantly decrease the values of the future expected reward term in the action value equation. Thus, agents that are adapted to uncertain environments should learn to consistently reduce future expected values. Here, we have simulated this reduction by manipulating the discount factor in situations where the agent had a different expectation for $p(j|s_t, a)$ than given by the task and showed that having a low discount factor can be beneficial.

Laboratory decision-making tasks used to measure impulsivity assess subjects under the assumption that all subjects will assume the same transition probabilities, which are often given by task instructions, or left implicit. If, however, participants have adapted to different levels of uncertainty in the environments in which they live, they may make choices with different implicit levels of uncertainty in the distributions of conditional state transitions. As these are assumed fixed by experiments, differences in behavior will be attributed to differences in discount factors. However, it is also possible that subjects have poor estimates of transition probabilities, and it is not straightforward to dissociate an agent's discount factor from the uncertainty in the state transition function they bring to a task, and both can decrease future expected values. In other words, task performance is always optimal when the statistics of the environment are accurately modeled. However, if an agent has different expectations than the true environment statistics, as was the case in this study, then discounting future rewards can be beneficial to task performance. We chose to model situations in which the transition probabilities of the environment were more uncertain than expected by the agent, because this led to an advantage for smaller discount factors. However, we could also have matched the discount factors, and shown that in uncertain environments, agents that better approximated that uncertainty would do better than agents that thought the environment was more certain. Either reducing the discount factor or increasing estimates of environmental uncertainty decreases the value of future rewards, and therefore makes immediate rewards relatively more valuable.

The Temporal Discounting task in our study was modeled after the KDD behavioral assessment, which is a questionnaire used to assess subject specific preferences for smaller, immediate rewards relative to larger, future rewards [58]. We simulated stochastic environments, such that future rewards were not always delivered. Importantly, we modeled the delay to the larger reward with a transition probability and used the MDP, a utility based model, to compute

action values when the transition probabilities were higher than expected (certain environment) and lower than expected (uncertain environment), which made the TD simulations risky intertemporal choices. We found that impulsive agents performed worse when transition probabilities to larger, delayed rewards were higher than expected, similar to previous findings using probabilistic future rewards (for reviews, see [61,62]). However, when the transition probabilities to delayed rewards in the environment were lower than expected, the impulsive agent with the lower discount factor collected more average reward than the non-impulsive agent that chose larger, future rewards that were not delivered. The success of the impulsive agent was amplified by the mismatch between the expectation about future reward and the underlying probability of reaching that reward. As previously described, impulsivity is frequently given a negative interpretation. In contrast, we demonstrate that choosing a smaller, immediate reward can be beneficial in some cases, in this case, risky intertemporal choice. It remains an open debate whether attribute-comparison (i.e. time vs. time and probability vs. probability) or utility based models are more appropriate for capturing intertermporal choice behavior and neural representation, and there are many kinds of intertemporal choices based on combinations of attributes [62]. Here we demonstrate an example where an impulsive agent can perform better than a non-impulsive agent, and this example could be extended to other kinds of intertemporal choices by using mismatched expectations across a range of attributes. Recent work with related discounting tasks used to assess weighting of immediate and future rewards, such as the Marshmallow Task [63], have also shown that preference for immediate rewards can be related to the perceived reliability of the experimenters, and trust, rather than trait impulsivity, which suggests that the accuracy of expectations can affect choice behavior [64]. Other work has suggested that immediate choices in the Marshmallow task are rational adaptation to time delays rather than failures of self-control [20]. Thus, although patients with substance use disorders and some psychiatric disorders can exhibit higher impulsive choice in behavioral tasks [65], and this is given as a possible dimensional explanation of their disorder, favoring immediate, smaller rewards, can be beneficial when the task environment makes future rewards less likely than expected.

Information sampling tasks have also been used to assess impulsivity [5,42,66]. Variations on these tasks include random dot motion perceptual inference [67], perceptual-motor inference [68], and sequential sampling paradigms [44,66,69–71]. We modeled choices in the Beads task, which has also been used to assess discrete information sampling with sampling cost [41,44,45]. In this task, participants are asked to guess the majority color of beads in an urn. In each trial, they can draw an additional bead from the urn for a small cost or guess the majority color. Drawing additional beads, therefore, improves accuracy, but at a cost. Past work has used a variety of models to capture both reaction time and choice behavior in perceptual inference tasks, including the well-known drift diffusion framework [72,73] and variants [67,74], including full POMDP developments [67], similar to what we have used. The drift diffusion framework captures the decision to terminate information sampling with a threshold crossing. Here we modeled the decision without the need to fit a threshold by quantifying the action values for continuing to accrue information (i.e. draw a bead) versus making a choice based on previously gathered information (i.e. guessing a color)[41,44]. We manipulated the probability distribution for the actual bead draws such that they were higher or lower than the majority fraction the agent expected. We also made the cost for guessing incorrectly larger than the cost for guessing correctly, to encourage drawing behavior from the impulsive agent. When the majority fraction of beads in the urn was lower than the agent's expectation, the impulsive agent accumulated more reward than the non-impulsive agent, because the non-impulsive agent accumulated costs for draws that were less informative than expected.

We showed that an impulsive agent can perform better in conditions in which we manipulated cost and uncertainty, but the effect is strengthened when the cost for guessing incorrectly is larger than the reward for guessing incorrectly. There has been some past work with the Beads task and asymmetric reward structure, but to our knowledge, only small and large rewards, not cost for guessing [75]. It would be interesting to explore asymmetric payouts in future work. Based on previous modeling of cognitive resources during information sampling in the Beads task, we would predict that a loss context would inhibit guessing for human participants in a way that reflects general risk preference, rather than a precise, online computation that would be cognitively demanding [76]. Past work has also shown that manipulating sampling costs can lead to changes in sampling, such that participants can be driven to oversample when sampling costs are low [68,77]. Sampling can also be affected by perseverative behaviors, not just information seeking, particularly in impulsive subjects. In one study, subjects were asked to report their estimate of the probability of the majority color in a variant of the Beads task, and subsequent analyses showed that schizophrenic patients, characterized with impulsive behavior, had persistent drawing that correlated with the frequency of clinical delusions. However, when delusions were controlled for in analyses, the same patients exhibited decreased information seeking compared to healthy individuals, suggesting that perseverative drawing is sometimes unrelated to the goal of information seeking [78].

Our results show that not only the cost to sample, but also the expected utility of the information sampled, can affect sampling and overall performance. However, the simulations here do not account for perseverative actions, which can be a feature of impulsivity and drive what appears to be perseverative information seeking. In our simulations, the impulsive agent benefitted from sampling less when the information gained from sampling was less informative. Future experiments involving impulsive human subjects could test both the effects of this loss context and also incorporate a separate term in the model for perseverative drawing that is independent from drawing related to information seeking.

Impulsive choice has also been shown to be related to novelty-seeking in clinical disorders and substance abuse [79–83]. However, these studies frequently use self-report questionnaires that measure sensation seeking as a metric for novelty seeking behavior. Our measure of novelty seeking is related to the explore-exploit trade-off, and operationalizes an investment in learning about a novel option (i.e. exploration) because the investment may pay off in the future (exploitation), in a well-characterized bandit task with novel options [44,46,84–86]. In the Explore-Exploit task, we manipulated the substitution rate of novel options. When the substitution rate was higher than expected, the impulsive agent collected more reward on average by not exploring the novel option as often. This was advantageous because the novel options were replaced more often than expected, and thus had short time horizons, and therefore could not be exploited in the future. When environments are unstable, or time horizons are short, exploration does not pay off, because the options are not available in the future, and an impulsive strategy that prioritizes immediate rewards is more beneficial. Direct manipulation of the time horizon of available choices has shown a similar result and has shown that human subjects can adapt to the time horizon for options during an explore-exploit task [21]. However, past work investigating novelty-seeking in clinical groups has shown mixed outcomes. Clinical groups that rank high on impulsivity on self-report questionnaires have been shown to exhibit risk-seeking and novelty-seeking behaviors, but not in all cases [87], and in some patient populations, novelty-seeking and impulsivity are largely separable behaviors [88,89]. Past work with the Explore-Exploit task as we have simulated it here, has shown that as the discount factor increases in this model, the novelty bonus increases [44,85]. This novelty bonus can account for high rates of choosing the novel option among other options [84,85]. While the results here show less novelty-seeking for impulsive agents, the framework would allow for

experiments that decouple these two features of decision-making. For example, we would predict that some clinical groups labeled as impulsive would perform similar to our computational impulsive agents and perform better than healthy controls in high substitution rate environments, while others would choose the novel options more often, which might hurt overall performance. By manipulating the task parameters, it would be possible to shed light on the interactions between impulsivity in clinical populations and novelty-seeking, which we have defined as exploration of options with unknown reward rates.

In all three tasks presented, we modeled impulsive choice behavior in the context of misestimation of the task environment and manipulated the discount factor that weights the value of future rewards. However, an individual in a laboratory task might exhibit preference for a smaller, more certain option either because it will come sooner (time preference) or because it is certain (risk preference). Past work has shown that individual attitudes toward risk might play an independent role from time preference in estimating the discount factor [90,91]. While we do not dissociate these two factors in our models, past work has incorporated preferences for time and risk into the discount factor term to improve estimates of discounting in human subjects [92].

Furthermore, it remains an open question whether individual preferences for immediate rewards are due to attitudes towards risk or due to an inability to learn transition probabilities to future rewards. While beyond the scope of this study, it is worth acknowledging the possibility that impulsive choices could arise from poor planning ability, or from a conscious devaluing of future expected values. However, recent work suggests that deficits in planning or goal pursuit might be separable from impulsive choice behavior, as human subjects labeled as impulsive can also exhibit goal-oriented behaviors that require extensive planning [93].

In summary, previous work suggests that impulsive decision-making in clinical groups is maladaptive [94,95]. In contrast, our results across the three tasks suggest that impulsive behavior is not inherently negative and can be beneficial when an environment is more volatile than expected. Therefore, impulsive choice patterns can be adaptively optimal. It is not the agent that is suboptimal, but the match between the environment to which an agent is adapted, and the environment in which an agent is being tested. Furthermore, the framework here makes predictions about how human subjects, labeled impulsive by self-report or other means, might perform better in a variety of decision-making tasks. While past work has suggested that delay and risk are not necessarily equitable or represented as a single construct at the neural level [37,96], past literature has operationalized impulsivity through discounting of future rewards and the discount factor [55,56,97]. By combining these three tasks into a single framework, united by the discount factor, it becomes possible to validate the consistency of the discount factor for human participants. We have demonstrated parameter regimes where impulsive agents could fare better than non-impulsive agents that could be used to test human participants. For example, if "impulsive" human participants exhibit impulsive choice in TD and Beads, but choose novel options much more than non-impulsive agents in the Explore-Exploit task, this would suggest that the discount factor should be reconsidered as a way to operationalize impulsive choice in the context of novelty.

There is a growing literature on how experience in resource-poor environments and early-life stress can lead to changes in decision-making behavior and to favoring immediate over future rewards [98–103], which suggests impulsive choice behavior might be an adaptation to environmental instability. Furthermore, accurate assessment of environmental controllability has been shown to improve with development and age, suggesting that some impulsive choice behavior might arise from a dysfunction during development [104]. Although impulsivity is often assumed to be a trait, it may be a state, perhaps slowly changing, and impulsive choice behavior might reflect the environment to which an agent has adapted. Future work should investigate the flexibility of patients to adapt to impulsive task environments. The computational framework presented here opens a variety of possibilities to understand impulsive

choice behavior as a gradient, rather than a binary label, and to better understand how human subjects weigh immediate and future rewards in the contexts of monetary discounting, information sampling, and novelty seeking. We believe this framework allows for quantification of impulsive choice behavior in a new light that will be useful to clinicians and researchers investigating factors that lead to impulsive choices.

Methods

All simulations and analyses described below were conducted using MATLAB.

General algorithm

We first discuss aspects of the algorithm that are consistent across all tasks. Similar methods were used to analyze patient data in these same three tasks [43]. In the present manuscript we are carrying out theoretical analyses to simulate behavior preferences of different agents. Simulations of two of the tasks (Information sampling & 3-armed Bandit) were previously described [44]. We first summarize the basic framework, which is described in more detail in the two previous studies. We then describe the specifics of each task and the manipulations of the agent and the environment used to achieve varied levels of uncertainty to answer the question posed in this study.

All tasks involved considering immediate rewards and future rewards at each step without consideration of previous steps. Thus, all tasks can be modeled as Markov Decision Processes (MDP) or Partially Observed MDPs (POMDP). The MDP framework models the utility, u, of a state, s, at time t as

$$u_t(s_t) = \max_{a \in A_{s_t}} \{Q(s_t, a)\} \quad (1)$$

where A_{s_t} is the set of available actions in state s at time t, a is an action, and $Q(s_t, a)$ is the action value. The action value is the combination of immediate reward, possible cost, and discounted expected future rewards:

$$Q(s_t, a) = r(s_t, a) + C(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j) \quad (2)$$

where $r(s_t, a)$ is the immediate reward received in state s at time t if action a is taken and $C(s_t, a)$ is the cost to sample. These quantities make up the immediate expected value (IEV), which is the reward (cost) that will be received in the current time step when an action is taken. The future expected value (FEV) is the discounted expected future rewards, given an action. The expectation is taken over all possible future states, S, at time $t + 1$. Each transition probability, $p(j|s_t, a)$, is the probability of transitioning to a particular state, j, from the current state if one takes action a. The discount factor, g, defines the discounting of future rewards and takes on values between 0 and 1. Thus the utility equation is a maximization across all possible actions to find the most valuable action to take.

For discrete state, finite horizon models with tractable state spaces (e.g. Temporal discounting & Information sampling), utility estimates can be calculated by backward induction [44,53,105]. Because there is a termination of the sequence of choices in these tasks and a defined final reward (outcome), we can start by defining the utilities at the final state(s). We can then work backward to define the utilities of the previous states. If N is the final state:

1. Set t = N

$$u_N(s_N) = r(s_N) \text{ for all } s_N \in N \quad (3)$$

2. Substitute *t-1* and compute the utility:

$$u_t(s_t) = \max_{a \in A_{s_t}} \{r(s_t, a) + C(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j)\} \quad (4)$$

Then set:

$$A^*_{s_{t,t}} = \text{argmax}_{a \in A_{s_t}} \{r(s_t, a) + C(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j)\} \quad (5)$$

3. If $t = 1$, stop, otherwise return to 2.

The set $A^*_{s_{t,t}}$ contains all actions, a, which maximize the utility.

The Explore-Exploit task was modeled as an infinite horizon POMDP. Utilities were fit using the value iteration algorithm [44,53]. The algorithm starts by initializing a vector of utilities across states,u^0, to random values, and then computing:

$$u(s_t)^{n+1} = \max_{a \in A_{s_t}} \{r(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u^n(j)\} \quad (6)$$

Because the state-space of the task was intractable over useful horizons, we used a B-spline basis function approximation [44] to estimate the utilities:

$$\hat{u}(s) = \sum_{i=1}^{M} b_i \phi_i(s) \quad (7)$$

where $\hat{u}(s)$ is the approximation of the utility, b_i are the basis coefficients, and $\phi_i(s)$ are the basis functions. We then calculated a projection matrix, H, and the approximation:

$$\hat{u} = Hu \quad (8)$$

The approximation was plugged into the righthand side of Eq (6) in place of $u^n(j)$. Approximations to the new values were iteratively calculated until convergence:

$$\hat{u}^{n+1} = Hu^{n+1} \quad (9)$$

Manipulation of uncertainty

Agents built on MDPs optimize expected reward when they are matched to the statistics of the environment, where matched means that the parameters of the probability model on which the agent is built are the parameters of the environment from which the agent samples in the simulations [53]. Therefore, an impulsive MDP agent will outperform a non-impulsive agent when the non-impulsive agent is not as well matched to the statistics of the environment. Here we were interested in the trade-off between immediate and future expected value, as this is the trade-off assessed with experimental measures of impulsivity. Impulsive subjects overweight IEVs, relative to FEVs, because they prefer immediate to delayed rewards. Therefore, we considered mismatches between agents and environments in FEVs, which are products of the uncertainty of state transitions, $p(j|s_t, a)$, and discount factor, γ.

One way to approach this would be to show that when transition probabilities in the environment are more uncertain, i.e., when $p(j|s_t, a)$ in the environment is high entropy, agents that assume $p(j|s_t, a)$ is low entropy will do worse than agents that have the proper environmental model. However, this would not show differences in the discount factor as this would be true with matched discount factors. Behavioral measures of impulsivity used in the laboratory and descriptive definitions of impulsivity often use discount factors to characterize impulsive choices. Therefore, we chose an approach that would show that having a shorter time horizon, characterized by a smaller discount factor, can be beneficial when environment and

agent expectations are not matched. Specifically, when environments are more uncertain than expected, impulsive choice strategies can be beneficial. After the description of each decision-making task and model, we describe how we modified the parameters that described the agent's expectation and the parameters that described the environment in which the agent made choices to achieve a mismatch between the agent's expectations and the actual environment. Thus, we modeled MDP agents using assumed uncertainty values, and subsequently used these agents to make choices in environments that had mismatched uncertainty values. We use subscripts of "agent" for MDP model parameters, and subscripts of "env" (to indicate environment), to refer to the statistics used to generate the actual outcomes on each trial. Thus, agents were not matched to their environments, and we examined the effect of this mismatch, and different discount factors, on the number of rewards received.

Manipulation of impulsivity

Across all tasks, we use the discount factor, γ, to model impulsive and non-impulsive choice strategies. Impulsive agents are characterized by a low discount factor $\gamma_{Impulsive\ (I)}<0.7$. Non-impulsive agents have a high discount factor $\gamma_{Non-impulsive\ (NI)} = 0.99$.

Statistical analyses

To compare mean reward and choice behavior of pairs of agents, paired t-tests and paired sample t-tests were used as noted in the results. For the Explore-Exploit task, a two-way ANOVA was used to determine main effects of discount factor and substitution rate and interaction effects on mean reward and choice behavior. To compute effect sizes, we used Cohen's d. When agents were given identical trials, we used Cohen's d effect size for paired samples x1 and x2:

$$d = (\mu_1 - \mu_2)/\sqrt{var(\mu_1 - \mu_2)} \tag{10}$$

and when agents were given different trials, we used

$$d = (\mu_1 - \mu_2)/\sqrt{(s_1^2 - s_2^2)/2} \tag{11}$$

where μ_1 and μ_2 were the mean values and s_1 and s_2 were the sample standard deviations for reward or choice behavior for each agent. To provide guidance for using these tasks with human participants, we calculated the number of iterations (i.e. sample size) required to ensure that a comparison has a specified power, given the effect size observed. we used a power of $\beta = 0.80$, significance level $\alpha = 0.05$ [106].

Temporal Discounting task

In the Temporal Discounting task, an agent is given a choice between a smaller, immediate reward (R_1) and a larger, delayed (and possibly probabilistic) reward (R_2). The task comes in several variants. For example, the Kirby delayed discounting questionnaire includes questions like, "Would you prefer \$54 today, or \$55 in 117 days?" and "Would you prefer \$55 today or \$75 in 61 days?"[38]. Replies to these questions are used in decision making models to estimate discount factors. Extensive work has shown that reward value decreases with delay to reward [38,107–109]. Furthermore, even when an experiment suggests that delayed rewards will be certain, human participants select options with lower expected values more often when outcomes are immediate rather than delayed. When both options are offered with a delay, participants choose the option with the larger expected value, even if that delay is larger. Experiments combining manipulations of uncertainty (through probabilistic reward offers) and time delays

show that manipulating uncertainty directly has little effect on the preferences for delayed rewards. These experiments suggest that human participants attribute uncertainty to delayed rewards [37].

To model this task, we used a previously published, quasi-hyperbolic discounting model [43,44,109]. We assume a state space in which an action a (choose immediate reward or choose delayed reward) leads to the immediate reward state (s_{IR}) or a sequence of transition states (s_b). Each transition state leads to the subsequent transition state, an intermediate terminal state (s_a) that terminates the episode and results in no reward, or if it is the final transition state, the final reward state (s_{DR}) in which R_2 is received. The sequence of unrewarded states models the temporal delay to the second option and the uncertainty around one's ability to reach the terminal delayed reward state (s_{DR}). The transition probabilities are defined by two parameters: β, which parameterizes the transition probability of the first step at t = 0, and δ, which is the discretized transition probability between the sequential s_b transition states. Thus, the model implements the progression through the state space with the following probabilities:

The probability of moving to the next intermediate transition state at the start is:

$$p(s_1 = s_b) = \beta\delta \tag{12}$$

The probability of terminating in an exit state at the start is:

$$p(s_1 = s_a) = 1 - \beta\delta \tag{13}$$

The probability of moving to the next intermediate transition state given that we are in an intermediate transition state:

$$p(s_{t+1} = s_b | s_t = s_b) = \delta \tag{14}$$

and the probability of terminating at an exit state, given that one is in an intermediate transition state is:

$$p(s_{t+1} = s_a | s_t = s_b) = 1 - \delta \tag{15}$$

The value of the immediate reward is R_1 and the value of delayed reward is Q(a = choose R_2 at delay N) = $R_2\beta\delta^N$. For the modeling in this study, $\beta = 1$ for all conditions, which makes the quasi-hyperbolic model equivalent to an exponential model. While MDPs inherently discount future rewards exponentially, past work has suggested that human behavior can be fit better by hyperbolic discounting [110–112]), and a value of $\beta<1$ would likely be more appropriate for fitting human behavioral data but would not affect the interpretation of the results presented here.

Manipulation of uncertainty in the Temporal Discounting task

For the Temporal Discounting task, the transition probability, δ, was used to manipulate uncertainty. If $\delta = 0.5$, exiting at an intermediate, unrewarded state is as likely as moving one step closer to the final reward state, and if $\delta = 0.9$, progression to the next state happens 90% of the time. Uncertainty in the environment was modeled as a smaller value of δ for the expected transition probability to the delayed reward than the one used to calculate the state-action value for choosing the delayed reward in the agent's model. The state-action value, $Q(s_t, a)$, was computed using δ_{agent} and the true outcomes were simulated using δ_{env}, where $\delta_{agent}<\delta_{env}$ (certain environment) and $\delta_{agent}>\delta_{env}$ (uncertain environment). Thus, each δ, δ_{agent} and δ_{env}, had two possible values of 0.55 and 0.99, although results are not contingent on these exact values. We compared the performance of two agents, one with a low discount factor ($\gamma_{Impulsive}$ = 0.6) and one with a high discount factor ($\gamma_{Non-Impulsive}$ = 0.99), to model impulsive and non-impulsive behavior, respectively.

To simulate outcomes across multiple trials, trials were generated using a range of unitless small reward sizes (R_1 = 1: 0.5: 51) and large reward sizes (R_2 = 50: 10: 1050), and unitless time delays (N = 1:20). For each trial, the action value $Q(s_t, a)$ was computed for the two options using the discount factor, γ_{agent}, and δ_{agent} such that $Q(Choose\ R_1) = R_1$, and $Q(Choose\ R_2) = R_2\ \gamma_{agent}{}^N \delta_{agent}{}^N$. The agent then picked the larger action value which determined whether they received R_1 or proceeded through the simulation of transition states towards R_2 on that trial. To simulate transition states to the delayed reward, the series of probabilistic states were simulated using δ_{env} and N, such that the agent effectively proceeded through N Bernoulli trials with p = δ_{env} to determine whether R_2 was received on that trial when R_2 was selected, or no reward was received. Average reward was calculated across 10 iterations of 100 trials for each agent in each environment. We then compared the average reward received and frequency of choosing the larger, delayed option when $\delta_{agent} < \delta_{env}$ and when $\delta_{agent} > \delta_{env}$ for both agents.

Information Sampling (Beads) task

In the information sampling task, participants are asked to guess the majority color of beads in an urn (one of two colors, for example, blue and green). Evidence for the majority bead color is accumulated one bead at a time, with a small cost for each bead drawn. At each time step, there are three possible actions: (a) guess green (b) guess blue or (c) draw another bead to gather more information. The state, s_t, is given by the number of draws (n_d) and the number of accumulated blue beads $s_t = \{n_d, n_b\}$. Each bead draw incurs a cost, $C_{draw}(s_t, a)$, and there is a maximum number of allowable draws. This allowed us to model the task using a finite-horizon, finite state, POMDP [45]. Additional parameters include the true fraction of beads in the majority urn (q), the reward for guessing correctly ($R_{correct}$) and the cost for guessing incorrectly (R_{error}).

For a given trial, a bead draw sequence (of length max draws) was generated using the fraction of majority beads q. State-action values were calculated for each possible action for each step to determine when the agent should stop drawing and guess a majority color.

For guessing that the urn is majority blue:

$$r(s_t, a = guess\ blue) = R_{error} p_g + R_{correct} p_b \tag{16}$$

where p_b is the probability the urn is majority blue, given by:

$$p_b = \left[1 + \left(\frac{q}{1-q}\right)^{(n_d - 2n_b)}\right]^{-1} \tag{17}$$

and p_g is the probability the urn is majority green, given by $p_g = 1 - p_b$. For guessing an urn color, the second term in the MDP utility equation that represents the FEV is 0, as choosing an urn terminates the sequence of actions.

For drawing again, a = draw, we have:

$$Q_t(s_t, a = draw) = C_{draw} + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j) \tag{18}$$

From a given state, s_t, if the agent draws again, the two possible next states are $s_{t+1} = \{n_d+1, n_b+1\}$ if a blue bead is drawn, or $s_{t+1} = \{n_d+1, n_b\}$ if a green bead is drawn. The corresponding transition probabilities are:

$$p(n_{d+1} + 1, n_b + 1 | s_t = [n_d, n_b], a = draw) = q p_b + (1 - q) p_g \tag{19}$$

and

$$p(n_{d+1}+1, n_b | s_t = [n_d, n_b], a = draw) = (1-q)p_b + qp_g \quad (20)$$

The action taken on each step was the one with the highest value. When the action value for guessing blue or green was higher than the action value for draw, the corresponding urn was chosen and total reward (whether the guess was correct or incorrect, and how many draws were taken) was computed. To model average agent behavior, 100 batches of 100 draw sequences were generated for each set of task parameters. Action values were computed for each step of each bead draw sequence and the agent picked the action associated with the largest action value at each step. Once the agent picked a color or reached the maximum number of draws (20 in these simulations), the reward collected and draw cost incurred was calculated and the number of draws before choice was recorded. This was conducted across all simulated sequences in a batch and average reward and average number of draws was calculated across the batches of bead draws. This was repeated for each discount factor across all task parameter sets.

Manipulation of uncertainty in the Information Sampling (Beads) task

To vary the level of uncertainty in the beads task, three parameters were modified to create parametric environments where either a non-impulsive agent (higher discount factor, $\gamma_{NI} = 0.99$) or impulsive agent (lower discount factor, $\gamma_I = 0.55$) would obtain more overall reward. First, the fraction of majority beads, q_{env}, used to generate the bead draw sequences was either higher or lower than the majority fraction used to calculate the agent's state-action values, q_{agent}. For example, if q_{env}, used to generate the bead draws, was lower than q_{agent}, then the agent would expect more information from each bead draw was present in the actual sequences. The second parameter modified was the cost to draw a bead (C_{draw}). Varying C_{draw} affected whether the impulsive or non-impulsive agent collected more reward on average. Third, the cost of guessing incorrectly (R_{error}) was set larger than the reward for guessing correctly ($R_{correct}$). While there exists a parameter range where $|R_{error}| = |R_{correct}|$ and the impulsive agent can collect more average reward, in this domain the agent typically only makes one draw before the action value for guessing one of the colors becomes greater than the action value for drawing a bead. If $|R_{error}| > |R_{correct}|$, then this encourages multiple draws from the impulsive agent, leading to a richer behavioral output.

Explore-Exploit task

The Explore-Exploit task is a 3-armed bandit task in which one option is replaced with a novel option at a parametrized, stochastic rate. The size of the reward is the same for each option, but the probability of receiving a reward from each option differs. The agent must learn the value of each option through experience. After the agent experiences the three available options for a period, one of the options is randomly selected and replaced by a novel option. The agent must then decide whether to choose the novel option (explore) or select (exploit) one of the remaining two options with which the agent has more experience. The replacements are not known in advance and happen stochastically, so there is no way to plan for an option being replaced.

In the model states are defined by the number of times each option has been chosen and the number of times it has been rewarded $s_t = \{R_1, C_1, R_2, C_2, R_3, C_3\}$. The immediate reward estimate is given by:

$$r(s_t, a = choose\ option\ i) = \frac{R_i + 1}{C_i + 2} \quad (21)$$

The numerator and denominator include the assumption of a beta(1,1) prior, reflecting an a-priori reward probability of 0.5. The set of possible next states is given by the chosen target, whether it was rewarded, and whether one of the options was replaced with a novel option. The probability of a novel substitution, h was a parameter and $q_i = r(s_t, a = i)$. The transition probability to a state without a novel choice substitution and no reward is given by:

$$p(\ldots, C_i + 1, R_i, \ldots | s_t = [\ldots, C_i, R_i, \ldots], a = \textit{choose option } i) = (1 - q_i)(1 - h)$$

and if the chosen target was rewarded and there was still no novel option:

$$p(\ldots, C_i + 1, R_i + 1, \ldots | s_t = [\ldots, C_i, R_i, \ldots], a = \textit{choose option } i) = q_i(1 - h)$$

When a novel option was introduced, it could replace the chosen option or a different option. If the chosen target, i, was not rewarded and a different target, j, was replaced, the transition probability is:

$$p(\ldots, C_i + 1, R_i, C_j = 0, R_j = 0 | s_t = [\ldots, C_i, R_i, \ldots], a = \textit{choose option } i) = (1 - q_i)h/3$$

As long as the chosen target was not rewarded, the transition probability is the same, even if the chosen target, i, was replaced instead. Correspondingly, if the chosen target, i, was rewarded, and a different target, j, was replaced, the transition probability is given by:

$$p2(\ldots, C_i + 1, R_i + 1, C_j = 0, R_j = 0 | s_t = [\ldots, C_i, R_i, \ldots], a = \textit{choose option } i) = q_i h/3$$

and is the same following reward and replacement of the chosen target, i.

Manipulation of uncertainty in the Explore-Exploit task

To manipulate uncertainty in the Explore-Exploit task, we varied the substitution rate of the novel option. Similar to the mismatch method in the Information Sampling task, the agents had a single substitution rate ($p_{agent} = 0.08$) and the novel option substitution rate in the environment was varied from $p_{env} = 0.02$ to $p_{env} = 0.2$. Thus, the agents expected a substitution rate of 0.08, but in each experimental condition, the substitution rate in the environment was either higher than, lower than, or equal to the expected substitution rate. The low substitution rate represents a certain environment, where the values of the three options are stable for long periods. The high substitution rate represents an uncertain environment because there are frequent introductions of novel options and therefore any single option cannot be exploited for long periods.

To compare average reward collected for impulsive and non-impulsive agents, we varied the discount factor (γ) used to compute the action values for each of the three options. We simulated 50 iterations of 250,000 trials of three options. The underlying reward rates could be 0.8, 0.5 or 0.3, and when novel options were introduced their reward rate was assigned randomly. The agent had to explore novel options to learn their reward rates. Sets of available options could include any combination of these three reward probabilities. The novel options replaced one of the options at rate p_{env}. We used the model to generate the action values for these trials. Choices were generated by selecting the largest action value for each trial. Rewards were calculated based on choosing these options and their underlying reward rates. To compare agents with different discount factors, identical sequences of trials were given to the two agents for each substitution rate.

To compare the balance between exploitation and exploration of the novel option, we calculated how often different agents selected the novel option on first appearance. This was calculated using the same choice data used to calculate average reward.

Supporting information

S1 Fig. Heatmaps of differences in average reward for non-impulsive and impulsive agents across a range of expected and actual transition probabilities in the Temporal Discounting task. Each panel is a heatmap showing the differences in average reward for a pair of non-impulsive and impulsive agents for a range of transition probabilities. δ_{agent} (x-axis) is the transition probability fed to the model and δ_{env} (y-axis) is the actual transition probability used to calculate the future expected values of the delayed rewards.
(PDF)

S2 Fig. Model behavior across a range of Beads task parameters with even outcomes for correct and incorrect guesses ($R_{correct}$ = 10, $R_{incorrect}$ = -10). Each panel is a heatmap showing the differences in average reward for a pair of non-impulsive and impulsive agents, indicated by the discount factors on the far left. Each column has a set of heatmaps for the models' expected majority fraction of beads, q_{agent}. Each row has a set of heatmaps for a pair of discount factors (impulsive & non-impulsive). The x-axis of each heatmap is the draw cost and the y-axis is the difference between the model input q_{agent} and the majority fraction used to generate the bead draws, q_{env}. More blue values indicate the non-impulsive agent collected more average reward and more red values indicate the impulsive agent collected more reward. As q_{agent} increases (left to right), the domain in which the non-impulsive agent performs better expands.
(PDF)

Acknowledgments

We would like to thank Dr. Silvia Lopez Guzman for reviewing an early version of this manuscript.

Author Contributions

Conceptualization: Diana C. Burk, Bruno B. Averbeck.

Formal analysis: Diana C. Burk, Bruno B. Averbeck.

Funding acquisition: Bruno B. Averbeck.

Investigation: Diana C. Burk.

Methodology: Diana C. Burk, Bruno B. Averbeck.

Project administration: Bruno B. Averbeck.

Resources: Bruno B. Averbeck.

Software: Diana C. Burk, Bruno B. Averbeck.

Supervision: Bruno B. Averbeck.

Visualization: Diana C. Burk.

Writing – original draft: Diana C. Burk, Bruno B. Averbeck.

Writing – review & editing: Diana C. Burk, Bruno B. Averbeck.

References

1. Stevens JR, Stephens DW. The adaptive nature of impulsivity. Impulsivity: The behavioral and neurological science of discounting. 2010; 361–387. https://doi.org/10.1037/12069-013

2. Moeller FG, Barratt ES, Dougherty DM, Schmitz JM, Swann AC. Psychiatric aspects of impulsivity. Am J Psychiatry. 2001; 158: 1783–1793. https://doi.org/10.1176/appi.ajp.158.11.1783 PMID: 11691682
3. Evenden JL. Varieties of impulsivity. Psychopharmacology (Berl). 1999; 146: 348–361. https://doi.org/10.1007/pl00005481 PMID: 10550486
4. Ioannidis K, Hook R, Wickham K, Grant JE, Chamberlain SR. Impulsivity in Gambling Disorder and problem gambling: a meta-analysis. Neuropsychopharmacology 2019 44:8. 2019; 44: 1354–1361. https://doi.org/10.1038/s41386-019-0393-9 PMID: 30986818
5. Clark L, Robbins TW, Ersche KD, Sahakian BJ. Reflection Impulsivity in Current and Former Substance Users. Biol Psychiatry. 2006; 60: 515–522. https://doi.org/10.1016/j.biopsych.2005.11.007 PMID: 16448627
6. Rogers RD, Moeller FG, Swann AC, Clark L. Recent Research on Impulsivity in Individuals With Drug Use and Mental Health Disorders: Implications for Alcoholism. Alcohol Clin Exp Res. 2010; 34: 1319–1333. https://doi.org/10.1111/j.1530-0277.2010.01216.x PMID: 20528825
7. Mitchell SH. Measuring impulsivity and modeling its association with cigarette smoking. Behav Cogn Neurosci Rev. 2004; 3: 261–275. https://doi.org/10.1177/1534582305276838 PMID: 15812110
8. Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience 2005 8:11. 2005; 8: 1481–1489. https://doi.org/10.1038/nn1579 PMID: 16251991
9. Robbins TW, Gillan CM, Smith DG, Wit S de, Ersche KD. Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. Trends Cogn Sci. 2012; 16: 81–91. https://doi.org/10.1016/j.tics.2011.11.009 PMID: 22155014
10. Barkley RA. Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. Psychol Bull. 1997; 121: 65. https://doi.org/10.1037/0033-2909.121.1.65 PMID: 9000892
11. Dekkers TJ, de Water E, Scheres A. Impulsive and risky decision-making in adolescents with attention-deficit/hyperactivity disorder (ADHD): The need for a developmental perspective. Curr Opin Psychol. 2022; 44: 330–336. https://doi.org/10.1016/J.COPSYC.2021.11.002 PMID: 34953445
12. Peluso MAM, Hatch JP, Glahn DC, Monkul ES, Sanches M, Najt P, et al. Trait impulsivity in patients with mood disorders. J Affect Disord. 2007; 100: 227–231. https://doi.org/10.1016/j.jad.2006.09.037 PMID: 17097740
13. Swann AC, Dougherty DM, Pazzaglia PJ, Pham M, Moeller FG. Impulsivity: a link between bipolar disorder and substance abuse. Bipolar Disord. 2004; 6: 204–212. https://doi.org/10.1111/j.1399-5618.2004.00110.x PMID: 15117399
14. Reddy LF, Lee J, Davis MC, Altshuler L, Glahn DC, Miklowitz DJ, et al. Impulsivity and Risk Taking in Bipolar Disorder and Schizophrenia. Neuropsychopharmacology. 2014; 39: 456. https://doi.org/10.1038/npp.2013.218 PMID: 23963117
15. Barratt ES. Factor Analysis of Some Psychometric Measures of Impulsiveness and Anxiety. Psychol Rep. 1965; 16: 547–554. https://doi.org/10.2466/pr0.1965.16.2.547 PMID: 14285869
16. Eysenck SBG, Eysenck HJ. The place of impulsiveness in a dimensional system of personality description. British Journal of Social and Clinical Psychology. 1977; 16: 57–68. https://doi.org/10.1111/j.2044-8260.1977.tb01003.x PMID: 843784
17. Dalley JW, Robbins TW. Fractionating impulsivity: neuropsychiatric implications. Nature Reviews Neuroscience 2017 18:3. 2017; 18: 158–171. https://doi.org/10.1038/nrn.2017.8 PMID: 28209979
18. Otto AR, Markman AB, Love BC. Taking More, Now: The Optimality of Impulsive Choice Hinges on Environment Structure. Soc Psychol Personal Sci. 2012; 3: 131–138. https://doi.org/10.1177/1948550611411311 PMID: 22348180
19. Raio CM, Konova AB, Otto AR. Trait impulsivity and acute stress interact to influence choice and decision speed during multi-stage decision-making. Sci Rep. 2020; 10: 1–12. https://doi.org/10.1038/s41598-020-64540-0 PMID: 32385327
20. McGuire JT, Kable JW. Rational temporal predictions can underlie apparent failures to delay gratification. Psychol Rev. 2013; 120: 395–410. https://doi.org/10.1037/a0031910 PMID: 23458085
21. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. Humans use directed and random exploration to solve the explore-exploit dilemma. J Exp Psychol Gen. 2014; 143: 2074–2081. https://doi.org/10.1037/a0038199 PMID: 25347535
22. Haynes JM, Willis-Moore ME, Perez D, Cousins DJ, Odum AL. Temporal expectations in delay of gratification. J Exp Anal Behav. 2022 [cited 8 Dec 2022]. https://doi.org/10.1002/JEAB.814 PMID: 36477783

23. Esteves M, Moreira PS, Sousa N, Leite-Almeida H. Assessing Impulsivity in Humans and Rodents: Taking the Translational Road. Front Behav Neurosci. 2021; 15: 79. https://doi.org/10.3389/fnbeh.2021.647922 PMID: 34025369

24. Cyders MA, Littlefield AK, Coffey S, Karyadi KA. Examination of a short English version of the UPPS-P Impulsive Behavior Scale. Addictive behaviors. 2014; 39: 1372–1376. https://doi.org/10.1016/j.addbeh.2014.02.013 PMID: 24636739

25. Patton J, Stanford M, Barratt E. Factor structure of the Barratt impulsiveness scale. J Clin Psychol. 1995; 51: 768–774. https://doi.org/10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1 PMID: 8778124

26. Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH. Fifty years of the Barratt Impulsiveness Scale: An update and review. Pers Individ Dif. 2009; 47: 385–395. https://doi.org/10.1016/J.PAID.2009.04.008

27. Hook RW, Grant JE, Ioannidis K, Tiego J, Yücel M, Wilkinson P, et al. Trans-diagnostic measurement of impulsivity and compulsivity: A review of self-report tools. Neurosci Biobehav Rev. 2021; 120: 455–469. https://doi.org/10.1016/j.neubiorev.2020.10.007 PMID: 33115636

28. Chowdhury NS, Livesey EJ, Blaszczynski A, Harris JA. Pathological Gambling and Motor Impulsivity: A Systematic Review with Meta-Analysis. Journal of Gambling Studies 2017 33:4. 2017; 33: 1213–1239. https://doi.org/10.1007/s10899-017-9683-5 PMID: 28255940

29. Halperin J, Wolf L, Pascualvaca D, Newcorn J, Healey J, O'BRIEN JD, et al. Differential Assessment of Attention and Impulsivity in Children. J Am Acad Child Adolesc Psychiatry. 1988; 27: 326–329. https://doi.org/10.1097/00004583-198805000-00010 PMID: 3379014

30. Dickman SJ. Impulsivity, arousal and attention. Pers Individ Dif. 2000; 28: 563–581. https://doi.org/10.1016/S0191-8869(99)00120-8

31. Carr MR, De Vriesa TJ, Pattija T. Optogenetic and chemogenetic approaches to manipulate attention, impulsivity and behavioural flexibility in rodents. Behavioural Pharmacology. 2018; 29: 560–568. https://doi.org/10.1097/FBP.0000000000000425 PMID: 30169376

32. Romer D. Adolescent risk taking, impulsivity, and brain development: Implications for prevention. Dev Psychobiol. 2010; 52: 263–276. https://doi.org/10.1002/dev.20442 PMID: 20175097

33. Lauriola M, Panno A, Levin IP, Lejuez CW. Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task. J Behav Decis Mak. 2014; 27: 20–36. https://doi.org/10.1002/BDM.1784

34. Ramírez-Martín A, Ramos-Martín J, Mayoral-Cleries F, Moreno-Küstner B, Guzman-Parra J. Impulsivity, decision-making and risk-taking behaviour in bipolar disorder: a systematic review and meta-analysis. Psychol Med. 2020; 50: 2141–2153. https://doi.org/10.1017/S0033291720003086 PMID: 32878660

35. Rosenbaum GM, Hartley CA. Developmental perspectives on risky and impulsive choice. Philosophical Transactions of the Royal Society B. 2019; 374: 20180133. https://doi.org/10.1098/rstb.2018.0133 PMID: 30966918

36. Hamilton KR, Mitchell MR, Wing VC, Balodis IM, Bickel WK, Fillmore M, et al. Choice impulsivity: Definitions, measurement issues, and clinical implications. Personality Disorders: Theory, Research, and Treatment. 2015; 6: 182–198. https://doi.org/10.1037/per0000099 PMID: 25867841

37. Keren G, Roelofsma P. Immediacy and certainty in intertemporal choice. Organ Behav Hum Decis Process. 1995; 63: 287–297. https://doi.org/10.1006/OBHD.1995.1080

38. Kirby KN. Bidding on the Future: Evidence Against Normative Discounting of Delayed Rewards. J Exp Psychol Gen. 1997; 126: 54–70. https://doi.org/10.1037/0096-3445.126.1.54

39. Lopez-Guzman S, Konova AB, Glimcher PW. Computational psychiatry of impulsivity and risk: how risk and time preferences interact in health and disease. Philosophical Transactions of the Royal Society B. 2019;374. https://doi.org/10.1098/rstb.2018.0135 PMID: 30966919

40. Huq SF, Garety PA, Hemsley DR. Probabilistic judgements in deluded and non-deluded subjects. Q J Exp Psychol A. 1988; 40: 801–812. https://doi.org/10.1080/14640748808402300 PMID: 3212213

41. Furl N, Averbeck BB. Parietal Cortex and Insula Relate to Evidence Seeking Relevant to Reward-Related Decisions. Journal of Neuroscience. 2011; 31: 17572–17582. https://doi.org/10.1523/JNEUROSCI.4236-11.2011 PMID: 22131418

42. Djamshidian A, O'Sullivan SS, Sanotsky Y, Sharman S, Matviyenko Y, Foltynie T, et al. Decision-making, impulsivity and addictions: Do Parkinson's disease patients jump to conclusions? Mov Disord. 2012; 27: 1137. https://doi.org/10.1002/mds.25105 PMID: 22821557

43. Averbeck BB, Djamshidian A, O'Sullivan SS, Housden CR, Roiser JP, Lees AJ. Uncertainty about mapping future actions into rewards may underlie performance on multiple measures of impulsivity in

behavioral addiction: Evidence from Parkinson's disease. Behavioral Neuroscience. 2013; 127: 245–255. https://doi.org/10.1037/a0032079 PMID: 23565936

44. Averbeck BB. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. PLoS Comput Biol. 2015; 11. https://doi.org/10.1371/journal.pcbi.1004164 PMID: 25815510

45. Moutoussis M, Bentall RP, El-Deredy W, Dayan P. Bayesian modelling of Jumping-to-Conclusions bias in delusional patients. 2011; 16: 422–447. https://doi.org/10.1080/13546805.2010.548678 PMID: 21480015

46. Costa VD, Tran VL, Turchi J, Averbeck BB. Dopamine modulates novelty seeking behavior during decision making. Behavioral Neuroscience. 2014; 128: 556–566. https://doi.org/10.1037/a0037128 PMID: 24911320

47. Meyer RJ, Shi Y. Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. 1995; 41: 817–834. https://doi.org/10.1287/MNSC.41.5.817

48. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience 2009 12:8. 2009; 12: 1062–1068. https://doi.org/10.1038/nn.2342 PMID: 19620978

49. Steyvers M, Lee MD, Wagenmakers EJ. A Bayesian analysis of human decision-making on bandit problems. J Math Psychol. 2009; 53: 168–179. https://doi.org/10.1016/J.JMP.2008.11.002

50. Lee MD, Zhang S, Munro M, Steyvers M. Psychological models of human and optimal performance in bandit problems. Cogn Syst Res. 2011; 12: 164–174. https://doi.org/10.1016/J.COGSYS.2010.07.007

51. Sutton R.S., Barto AG. Reinforcement Learning, An Introduction. second edition. In: MIT press. 2018.

52. Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems. Nature Machine Intelligence. Nature Research; 2019. pp. 133–143. https://doi.org/10.1038/s42256-019-0025-4

53. Puterman ML. Markov decision processes: Discrete stochastic dynamic programming. xvii. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley; 1994. https://doi.org/10.1002/9780470316887

54. Gilovich T, Griffin D, Kahneman D. Heuristics and Biases. Heuristics and Biases: The psychology of intuitive judgment. Cambridge University Press; 2002. https://doi.org/10.1017/CBO9780511808098

55. Martinez E, Pasquereau B, Drui G, Saga Y, Météreau É, Tremblay L. Ventral striatum supports Methylphenidate therapeutic effects on impulsive choices expressed in temporal discounting task. Scientific Reports 2020 10:1. 2020; 10: 1–11. https://doi.org/10.1038/s41598-020-57595-6 PMID: 31959838

56. Onoda K, Okamoto Y, Kunisato Y, Aoyama S, Shishida K, Okada G, et al. Inter-individual discount factor differences in reward prediction are topographically associated with caudate activation. Exp Brain Res. 2011; 212: 593–601. https://doi.org/10.1007/s00221-011-2771-3 PMID: 21695536

57. Kirby KN, Petry NM, Bickel WK. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. J Exp Psychol Gen. 1999; 128: 78–87. https://doi.org/10.1037//0096-3445.128.1.78 PMID: 10100392

58. Kirby KN, Maraković NN. Delay-discounting probabilistic rewards: Rates decrease as amounts increase. Psychonomic Bulletin & Review 1996 3:1. 1996; 3: 100–104. https://doi.org/10.3758/BF03210748 PMID: 24214810

59. Scholten H, Scheres A, de Water E, Graf U, Granic I, Luijten M. Behavioral trainings and manipulations to reduce delay discounting: A systematic review. Psychon Bull Rev. 2019; 26: 1803–1849. https://doi.org/10.3758/s13423-019-01629-2 PMID: 31270766

60. Cisek P, Puskas GA, El-Murr S. Decisions in Changing Conditions: The Urgency-Gating Model. Journal of Neuroscience. 2009; 29: 11560–11571. https://doi.org/10.1523/JNEUROSCI.1844-09.2009 PMID: 19759303

61. Green L, Myerson J. A discounting framework for choice with delayed and probabilistic rewards. Psychol Bull. 2004; 130: 769–792. https://doi.org/10.1037/0033-2909.130.5.769 PMID: 15367080

62. Luckman A, Donkin C, Newell BR. An evaluation and comparison of models of risky intertemporal choice. Psychol Rev. 2020; 127: 1097–1138. https://doi.org/10.1037/rev0000223 PMID: 32700921

63. Mischel W, Shoda Y, Peake PK. The nature of adolescent competencies predicted by preschool delay of gratification. J Pers Soc Psychol. 1988; 54: 687–696. https://doi.org/10.1037//0022-3514.54.4.687 PMID: 3367285

64. Kidd C, Palmeri H, Aslin RN. Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. Cognition. 2013; 126: 109–114. https://doi.org/10.1016/j.cognition.2012.08.004 PMID: 23063236

65. Amlung M, Vedelago L, Acker J, Balodis I, MacKillop J. Steep delay discounting and addictive behavior: a meta-analysis of continuous associations. Addiction. 2017; 112: 51–62. https://doi.org/10.1111/add.13535 PMID: 27450931

66. Cardinale EM, Pagliaccio D, Swetlitz C, Grassie H, Abend R, Costa V, et al. Deliberative Choice Strategies in Youths: Relevance to Transdiagnostic Anxiety Symptoms: https://doi.org/101177/2167702621991805. 2021; 9: 979–989.

67. Drugowitsch J, Moreno-Bote RN, Churchland AK, Shadlen MN, Pouget A. The Cost of Accumulating Evidence in Perceptual Decision Making. Journal of Neuroscience. 2012; 32: 3612–3628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012 PMID: 22423085

68. Juni MZ, Gureckis TM, Maloney LT. Information sampling behavior with explicit sampling costs. Decision (Wash D C). 2016; 3: 147–168. https://doi.org/10.1037/dec0000045 PMID: 27429991

69. Bennett D, Oldham S, Dawson A, Parkes L, Murawski C, Yücel M. Systematic Overestimation of Reflection Impulsivity in the Information Sampling Task. Biol Psychiatry. 2017; 82: e29–e30. https://doi.org/10.1016/j.biopsych.2016.05.027 PMID: 27587264

70. Costa VD, Averbeck BB. Frontal-parietal and limbic-striatal activity underlies information sampling in the best choice problem. Cereb Cortex. 2015; 25: 972–982. https://doi.org/10.1093/cercor/bht286 PMID: 24142842

71. Furl N, Averbeck BB, McKay RT. Looking for Mr(s) Right: Decision bias can prevent us from finding the most attractive face. Cogn Psychol. 2019; 111: 1–14. https://doi.org/10.1016/j.cogpsych.2019.02.002 PMID: 30826584

72. Gold JI, Shadlen MN. Neural computations that underlie decisions about sensory stimuli. Trends Cogn Sci. 2001; 5: 10–16. https://doi.org/10.1016/s1364-6613(00)01567-9 PMID: 11164731

73. Ditterich J. Stochastic models of decisions about motion direction: behavior and physiology. Neural Netw. 2006; 19: 981–1012. https://doi.org/10.1016/j.neunet.2006.05.042 PMID: 16952441

74. Tickle H, Tsetsos K, Speekenbrink M, Summerfield C. Human optional stopping in a heteroscedastic world. Psychol Rev. 2021 [cited 26 Nov 2022]. https://doi.org/10.1037/REV0000315 PMID: 34570524

75. Kobayashi K, Lee S, Filipowicz ALS, McGaughey KD, Kable JW, Nassar MR. Dynamic Representation of the Subjective Value of Information. Journal of Neuroscience. 2021; 41: 8220–8232. https://doi.org/10.1523/JNEUROSCI.0423-21.2021 PMID: 34380761

76. Petitet P, Attaallah B, Manohar SG, Husain M. The computational cost of active information sampling before decision-making under uncertainty. Nature Human Behaviour 2021 5:7. 2021; 5: 935–946. https://doi.org/10.1038/s41562-021-01116-6 PMID: 34045719

77. Bowler A, Habicht J, Moses-Payne ME, Steinbeis N, Moutoussis M, Hauser TU. Children perform extensive information gathering when it is not costly. Cognition. 2021; 208: 104535. https://doi.org/10.1016/j.cognition.2020.104535 PMID: 33370652

78. Baker SC, Konova AB, Daw ND, Horga G. A distinct inferential mechanism for delusions in schizophrenia. Brain. 2019; 142: 1797–1812. https://doi.org/10.1093/brain/awz051 PMID: 30895299

79. Voon V, Reynolds B, Brezing C, Gallea C, Skaljic M, Ekanayake V, et al. Impulsive choice and response in dopamine agonist-related impulse control behaviors. Psychopharmacology (Berl). 2010; 207: 645–665. https://doi.org/10.1007/s00213-009-1697-y PMID: 19838863

80. Kashdan TB, Hofmann SG. The high-novelty–seeking, impulsive subtype of generalized social anxiety disorder. Depress Anxiety. 2008; 25: 535–541. https://doi.org/10.1002/da.20382 PMID: 17935217

81. Black DW, Shaw M, McCormick B, Bayless JD, Allen J. Neuropsychological performance, impulsivity, ADHD symptoms, and novelty seeking in compulsive buying disorder. Psychiatry Res. 2012; 200: 581–587. https://doi.org/10.1016/j.psychres.2012.06.003 PMID: 22766012

82. Wood AC, Rijsdijk F, Asherson P, Kuntsi J. Inferring causation from cross-sectional data: Examination of the causal relationship between hyperactivity-impulsivity and novelty seeking. Front Genet. 2011; 2: 6. https://doi.org/10.3389/fgene.2011.00006 PMID: 22303305

83. Noël X, Brevers D, Bechara A, Hanak C, Kornreich C, Verbanck P, et al. Neurocognitive Determinants of Novelty and Sensation-Seeking in Individuals with Alcoholism. Alcohol and Alcoholism. 2011; 46: 407–415. https://doi.org/10.1093/alcalc/agr048 PMID: 21596760

84. Costa VD, Averbeck BB. Primate orbitofrontal cortex codes information relevant for managing explore-exploit tradeoffs. Journal of Neuroscience. 2020; 40: 2553–2561. https://doi.org/10.1523/JNEUROSCI.2355-19.2020 PMID: 32060169

85. Costa VD, Mitz AR, Averbeck BB. Subcortical Substrates of Explore-Exploit Decisions in Primates. Neuron. 2019; 103: 533–545.e5. https://doi.org/10.1016/j.neuron.2019.05.017 PMID: 31196672

86. Wilson RC, Bonawitz E, Costa VD, Ebitz RB. Balancing exploration and exploitation with information and randomization. Current Opinion in Behavioral Sciences. Elsevier Ltd; 2021. pp. 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001 PMID: 33184605

87. Aloi J, Crum KI, Blair KS, Zhang R, Bashford-Largo J, Bajaj S, et al. Individual associations of adolescent alcohol use disorder versus cannabis use disorder symptoms in neural prediction error signaling and the response to novelty. Dev Cogn Neurosci. 2021; 48: 100944. https://doi.org/10.1016/j.dcn.2021.100944 PMID: 33773241

88. Busemeyer JR, Stout JC. A contribution of cognitive decision models to clinical assessment: decomposing performance on the Bechara gambling task. Psychol Assess. 2002; 14: 253–262. https://doi.org/10.1037//1040-3590.14.3.253 PMID: 12214432

89. Kvam PD, Romeu RJ, Turner BM, Vassileva J, Busemeyer JR. Testing the factor structure underlying behavior using joint cognitive models: Impulsivity in delay discounting and Cambridge gambling tasks. Psychol Methods. 2021; 26: 18–37. https://doi.org/10.1037/met0000264 PMID: 32134313

90. Lopez-Guzman S, Konova AB, Louie K, Glimcher PW. Risk preferences impose a hidden distortion on measures of choice impulsivity. PLoS One. 2018; 13: e0191357. https://doi.org/10.1371/journal.pone.0191357 PMID: 29373590

91. Andreoni J, Sprenger C. Risk Preferences Are Not Time Preferences. American Economic Review. 2012; 102: 3357–76. https://doi.org/10.1257/AER.102.7.3357

92. Pine A, Shiner T, Seymour B, Dolan RJ. Dopamine, Time, and Impulsivity in Humans. Journal of Neuroscience. 2010; 30: 8888–8896. https://doi.org/10.1523/JNEUROSCI.6028-09.2010 PMID: 20592211

93. Kopetz CE, Woerner JI, Briskin JL, Correspondence CE, Kopetz W. Another look at impulsivity: Could impulsive behavior be strategic? Soc Personal Psychol Compass. 2018; 12: e12385. https://doi.org/10.1111/spc3.12385 PMID: 34079587

94. Groman SM. The Neurobiology of Impulsive Decision-Making and Reinforcement Learning in Nonhuman Animals. Curr Top Behav Neurosci. 2020; 47: 23–52. https://doi.org/10.1007/7854_2020_127 PMID: 32157666

95. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. https://doi.org/101176/appi.ajp201009091379. 2010; 167: 748–751. PMID: 20595427

96. Luhmann CC, Chun MM, Yi DJ, Lee D, Wang XJ. Neural Dissociation of Delay and Uncertainty in Intertemporal Choice. The Journal of Neuroscience. 2008; 28: 14459. https://doi.org/10.1523/JNEUROSCI.5058-08.2008 PMID: 19118180

97. Yoshida N, Uchibe E, Doya K. Reinforcement learning with state-dependent discount factor. 2013 IEEE 3rd Joint International Conference on Development and Learning and Epigenetic Robotics, ICDL 2013—Electronic Conference Proceedings. 2013. https://doi.org/10.1109/DevLrn.2013.6652533

98. Humphreys KL, Lee SS, Telzer EH, Gabard-Durnam LJ, Goff B, Flannery J, et al. Exploration-exploitation strategy is dependent on early experience. Dev Psychobiol. 2015; 57: 313–321. https://doi.org/10.1002/dev.21293 PMID: 25783033

99. Lloyd A, McKay RT, Furl N. Individuals with adverse childhood experiences explore less and underweight reward feedback. Proc Natl Acad Sci U S A. 2022;119. https://doi.org/10.1073/pnas.2109373119 PMID: 35046026

100. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, et al. Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). J Exp Psychol Appl. 2002; 8: 75–84. https://doi.org/10.1037//1076-898x.8.2.75 PMID: 12075692

101. Birn RM, Roeber BJ, Pollak SD, Reyna VF. Early childhood stress exposure, reward pathways, and adult decision making. Proc Natl Acad Sci U S A. 2017; 114: 13549–13554. https://doi.org/10.1073/pnas.1708791114 PMID: 29203671

102. Gerin MI, Puetz VB, Blair RJR, White S, Sethi A, Hoffmann F, et al. A neurocomputational investigation of reinforcement-based decision making as a candidate latent vulnerability mechanism in maltreated children. Dev Psychopathol. 2017; 29: 1689–1705. https://doi.org/10.1017/S095457941700133X PMID: 29162176

103. Blair KS, Aloi J, Bashford-Largo J, Zhang R, Elowsky J, Lukoff J, et al. Different forms of childhood maltreatment have different impacts on the neural systems involved in the representation of reinforcement value. Dev Cogn Neurosci. 2022; 53: 101051. https://doi.org/10.1016/j.dcn.2021.101051 PMID: 34953316

104. Raabid HA, Foordid C, Ligneulid R, Hartleyid CA. Developmental shifts in computations used to detect environmental controllability. Hauser TU, editor. PLoS Comput Biol. 2022; 18: e1010120. https://doi.org/10.1371/journal.pcbi.1010120 PMID: 35648788

105. Bellman R. Dynamic Programming. Rand Corporation, editor. Princeton: Princeton University Press; 1957. Available: https://www.science.org/doi/10.1126/science.127.3304.976.a

106. Ryan TP. Sample Size Determination and Power. Sample Size Determination and Power. John Wiley & Sons; 2013. https://doi.org/10.1002/9781118439241

107. Gregorios-Pippas L, Tobler PN, Schultz W. Short-Term Temporal Discounting of Reward Value in Human Ventral Striatum. J Neurophysiol. 2009; 101: 1507. https://doi.org/10.1152/jn.90730.2008 PMID: 19164109

108. Hariri AR, Brown SM, Williamson DE, Flory JD, de Wit H, Manuck SB. Preference for Immediate over Delayed Rewards Is Associated with Magnitude of Ventral Striatal Activity. Journal of Neuroscience. 2006; 26: 13213–13217. https://doi.org/10.1523/JNEUROSCI.3446-06.2006 PMID: 17182771

109. Laibson D. Golden Eggs and Hyperbolic Discounting. Q J Econ. 1997; 112: 443–478. https://doi.org/10.1162/003355397555253

110. Madden GJ, Bickel WK, Jacobs EA. Discounting of delayed rewards in opioid-dependent outpatients: Exponential or hyperbolic discounting functions? Exp Clin Psychopharmacol. 1999; 7: 284. https://doi.org/10.1037//1064-1297.7.3.284 PMID: 10472517

111. Kim BK, Zauberman G. Perception of Anticipatory Time in Temporal Discounting. J Neurosci Psychol Econ. 2009; 2: 91–101. https://doi.org/10.1037/A0017686

112. Green L, Myerson J. Exponential Versus Hyperbolic Discounting of Delayed Outcomes: Risk and Waiting Time. Integr Comp Biol. 1996; 36: 496–505. https://doi.org/10.1093/ICB/36.4.496