Risking your Tail: Modeling Individual Differences in Risk-sensitive Exploration using Conditional Value at Risk and Bayes Adaptive Markov Decision Processes

Tingke Shen[1*], Peter Dayan[1, 2],

1 Max Planck Institute for Biological Cybernetics, Tübingen, Germany
2 University of Tübingen, Tübingen, Germany

* corresponding author: tingke.shen@tuebingen.mpg.de

Abstract

Novelty is a double-edged sword for agents and animals alike: they might benefit from untapped resources or face unexpected costs or dangers such as predation. The conventional exploration/exploitation tradeoff is thus coloured by risk-sensitivity. A wealth of experiments has shown how animals solve this dilemma, for example using intermittent approach. However, there are large individual differences in the nature of approach, and modeling has yet to elucidate how this might be based on animals' differing prior expectations about reward and threat and degrees of risk aversion. To capture these factors, we built a Bayes adaptive Markov decision process model with three key components: an adaptive hazard function capturing potential predation, an intrinsic reward function providing the urge to explore, and a conditional value at risk (CVaR) objective, which is a contemporary measure of trait risk-sensitivity. We fit this model to a coarse-grain abstraction of the behaviour of 26 animals who freely explored a novel object in an open-field arena (Akiti et al. *Neuron* 110, 2022). We show that the model captures both quantitative (frequency, duration of exploratory bouts) and qualitative (stereotyped tail-behind) features of behavior, including the substantial idiosyncrasies that were observed. We find that "brave" animals, though varied in their behavior, generally are more risk neutral, and enjoy a flexible hazard prior. They begin with cautious exploration, and quickly transition to confident approach to maximize exploration for reward. On the other hand, "timid" animals, characterized by risk aversion and high and inflexible hazard priors, display self-censoring that leads to the sort of asymptotic maladaptive behavior that is often associated with psychiatric illnesses such as anxiety and depression. Explaining risk-sensitive exploration using factorized parameters of reinforcement learning models could aid in the understanding, diagnosis, and treatment of psychiatric abnormalities in humans and other animals.

Author summary

Animals face a dilemma when they encounter novel objects in their environment. Approaching and investigating an object could lead to reward in the form of food, play, etc. but it also exposes the animal to dangers such as predation. Experiments have shown that animals solve this exploration dilemma by using intermittent strategies (alternately approaching the object and then retreating to a safe location) that gradually increase their level of risk. We built an abstract model of these exploration strategies and fit the model to the behavior of 26 mice freely exploring a novel object in

an arena. Our model accounts for the high-level physical and mental states of the mice, the actions the mice can take, and beliefs about the uncertain consequences of those actions. Our model provides a rational explanation for individual differences seen in experiments: individuals maximize their utility given different prior beliefs about the dangers and the rewards in the environment, and different tendencies to overestimate the probability of bad outcomes. Modeling individual differences in risk-sensitivity during exploration could aid in the understanding, diagnosis, and treatment of psychiatric diseases such as anxiety and depression in humans and animals.

1 Introduction

Novelty is a double-edged sword for agents and animals alike: they might benefit from untapped resources or face unexpected costs or dangers such as predation (Corey, 1978). The conventional exploration/exploitation trade-off (Mehlhorn et al., 2015) is thus coloured by risk (Kacelnik and Bateson, 1997); a factor to which different individuals may be differentially sensitive. Despite these duelling aspects, investigations of novelty in reinforcement learning (RL) have mostly focused on neophilia driven by optimism in the face of uncertainty, and so information-seeking (Dayan and Sejnowski, 1996; Duff, 2002b; Gottlieb et al., 2013; Wilson et al., 2014). Neophobia has attracted fewer computational studies, apart from some interesting evolutionary analyses (Greggor et al., 2015).

Both excessive novelty seeking and excessive novelty avoidance can be maladaptive – they are flip sides of a disturbed balance. Here, we seek to examine potential sources of such disturbances, for instance, in distorted priors about the magnitude or probabilities of rewards (which have been linked to mania; Bennett and Niv, 2020; Eldar et al., 2016; Radulescu and Niv, 2019) or threats (linked to anxiety and depression; Bishop and Gagne, 2018; Paulus and Angela, 2012), or in extreme risk attitudes (Gagne and Dayan, 2022).

To do this, we take advantage of a recent study by Akiti et al. (2022) on the behaviour of mice exploring a familiar open-field arena after the introduction of a novel object near to one corner. The mice could move freely and interact with the object at will. Akiti et al. (2022) performed detailed analyses of how individual animals' trajectories reflected the novel object, including using MOSEQ (Wiltschko et al., 2020) to extract behavioural 'syllables' whose prevalence was affected by it. The animals differed markedly in both how they approached the object, and in what pattern. For the former, Akiti et al. (2022) observed two characteristic positionings of the animals when near to the object: 'tail-behind' and 'tail-exposed', associated respectively with cautious risk-assessment and engagement. For the latter, there was substantial heterogeneity along a spectrum of timidity, with all animals initially performing tail-behind approach, but some taking much longer (or failing altogether) to transition to tail-exposed approach.

We model an abstract depiction of the behaviour of individual mice by combining two reinforcement learning (RL) frameworks: the Bayes-adaptive Markov Decision Process (BAMDPs) treatment of rational exploration (Dearden et al., 2013; Duff, 2002b; Guez et al., 2013), and the conditional value at risk (CVaR) treatment of risk sensitivity (Artzner et al., 1999; Bellemare et al., 2023; Chow et al., 2015; Gagne and Dayan, 2022). In a BAMDP, the agent maintains a belief about the possible rewards, costs and transitions in the environment, and decides upon optimal actions based on these beliefs. Since the agent can optionally reuse or abandon incompletely known actions based on what it discovers about them, these actions traditionally enjoy an exploration bonus or "value of information", which generalizes the famous Gittins indices (Gittins, 1979; Weber, 1992). These exploration bonuses are dependent on prior expectations about the

environment; and so are readily subject to individual differences.

However, in a conventional BAMDP, agents are assumed to optimize the long run expected value implied by their beliefs – implying risk neutrality. We consider optimizing the CVaR, in which agents concentrate on the average value within lower (risk-averse) or upper (risk-seeking) quantiles of the distribution of potential outcomes (Rigter et al., 2021). In the context of a BAMDP, this can force agents to pay particular attention to hazards. More extreme quantiles are associated with more extreme risk-sensitivity; and again are a potential locus of individual differences (as examined in regular Markov decision processes in the context of anxiety disorders in Gagne and Dayan, 2022).

Here, we present a behavioral model of risk sensitive exploration. Our agent computes optimal actions using the BAMDP framework under the CVaR objective. This acts on both aleatoric and epistemic uncertainty, with the latter coming from ignorance about how longer times spent at the object might lead to increases in the probability of predation. This model provides a normative explanation of individual variability – the agent makes decisions by trading off potential reward and threat in a principled way. Different priors and risk sensitivities lead to different exploratory schedules, from timid (indicative of neophilia) to brave. The model captures differences in duration, frequency, and type of approach (risk-assessment versus engagement) across animals, and through time. We report features of the different behavioural trajectories the model is able to capture, providing mechanistic insight into how the trade-off between potential reward and threat leads to rational exploratory schedules.

2 Results

2.1 Behavior Phases and Animal Groups

Our goal is to provide a computational account of the exploratory behavior of individual mice under the assumption that they have different prior expectations and risk sensitivities. We start from Akiti et al. (2022)'s observation that the animal approaches and remains within a threshold distance (determined by them to be 7cm) of the object in "bouts" which can be characterized as "cautious" or tail-behind (if the animal's nose lies between the object and tail) or otherwise "confident" or tail-exposed. We sought to capture both these qualitative differences (cautious versus confident) and aspects of the quantitative changes in bout durations and frequencies as the animal learns about their environment.

In order to focus narrowly on interaction with the object, we abstracted away from the details of the spatial interaction with the object, rather fitting *boxcar functions* to the percentage of its time $g^{\text{cau}}(t), g^{\text{con}}(t)$ that the animal spends in *cautious* and *confident* bouts around time t in the apparatus. We can then well encompass the behaviour of most animals via four coarse phases of behaviour that arise from two binary factors: whether the animal is mainly performing cautious or confident approaches, and whether bouts happen frequently, at a peak rate, or at a lower, steady-state rate. The time $g(t)$ an animal spends near the object in one of these phases reflects the product of how frequently it visits the object, and how long it stays per visit. We average these two factors across the phases.

Consider the behaviour of the animal in the (left panel) of Fig 1. Here, $g^{\text{cau}}(t)$ (top graph) makes a transition from an initial level g_i^{cau} (during the "cautious" phase) to a final steady-state level g_s^{cau} (which we simplify as being $g_s^{\text{cau}} = 0$) at a transition point $t = t_1$. At the same timepoint, $g^{\text{con}}(t)$ (second row) makes a transition from 0 to a peak level g_p^{con} of confident approach (defining the "peak confident" phase). Finally, there is another transition at time t_2 from peak to a steady-state confident approach time g_s^{con}

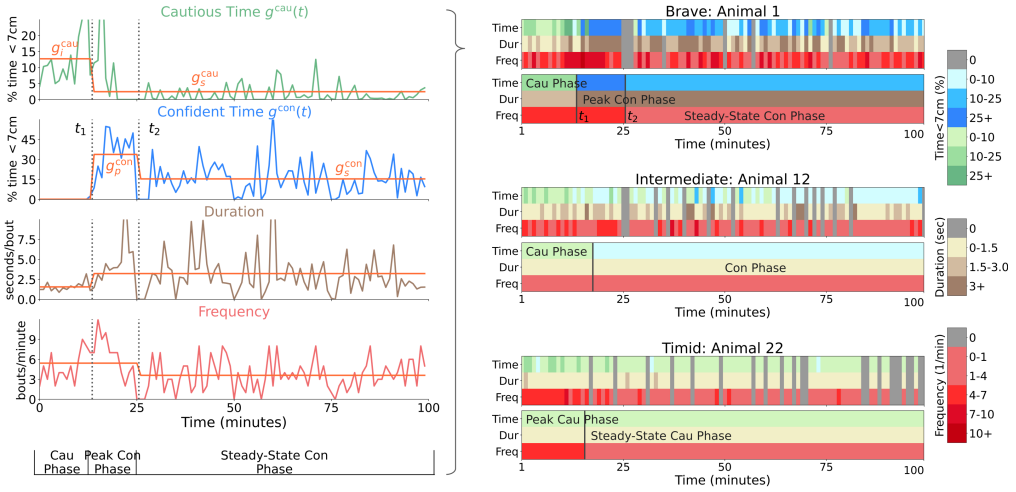

Fig 1. Left: detailed visualization of minute-to-minute statistics of animal 1 (in the sessions after the introduction of the novel object). Orange lines are the box-car functions fitted to segment phases and illustrate the change in time, duration, and frequency statistics across phases. The transition points t_1 and t_2 as well as the initial cautious g_i^{cau}, final cautious g_s^{cau}, peak confident g_p^{con} and steady-state confident g_s^{con} approach times are shown. Right: examples of minute-to-minute and phase-averaged approach time, duration, and frequency for brave (top), intermediate (middle), and timid (bottom) animals. Green indicates cautious and blue indicates confident approach. Darker colors indicate higher values. Averaging statistics over phases ignores idiosyncracies of behavior to provide a high-level summary of learning dynamics.

(in the "steady-state confident" phase). The lower two rows of figure 1 left panel show the duration of the bouts in the relevant phases, and the frequency per unit time of such bouts.

The top right panel of Fig 1 renders the actual and abstracted behaviour of this animal in an integrated form, showing how we generate "phase-level" statistics from minute-to-minute statistics. The colours in the top row indicate the type of approach (green is cautious; blue is confident). The second and third rows indicate the duration and frequency of approach. Darker colours represent higher values. Averaging statistics over phases ignores idiosyncrasies of behavior and allows us to fit the high-level statistics of behavior: phase-transiton times, phase-averaged durations and frequencies. We consider animal 1 to be a "brave" animal because of its transition to peak and then steady-state confident approach. There were 12 brave mice out of the 26 in total.

The middle panels on the right of Fig 1 show an example of another characteristic animal. They make a transition from cautious to confident approach (where both duration and frequency of visits can change), but the approach time during the confident phase g_s^{con} does not decrease. Hence, intermediate animals do not have a transition from peak to steady-state confident phase. There were 9 such "intermediate" mice.

The bottom panels in the right of Fig 1 show the last class of animals an example of another characteristic animal. This animal never makes a transition to confident approach. Hence, for it, $g^{\text{con}}(t) = 0$. However, the cautious approach time makes a transition to a non-zero steady state ($g_s^{\text{cau}} > 0$), often via a change in frequency, defining the fourth phase ("steady-state cautious"). There were 5 such "timid" mice.

Fig 2 summarizes our categorization of the animals into the three groups: brave, intermediate, and timid based on the phases identified in the animal's exploratory trajectories. Timid animals spend no time in confident approach. Brave animals differ

from intermediate animals in that their approach time during the first ten minutes of the confident phase is greater than the last ten minutes (steady-state phase). Fig 7 (top-left) shows that our categorization is different but correlated with the ranking of animals in Akiti et al. (2022) based on the total time spent near the object.

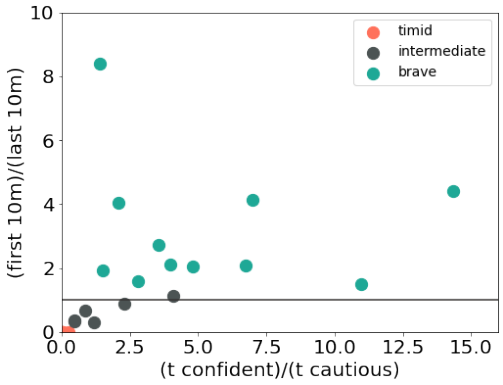

Fig 2. Separating the three animal groups. The x-axis is the ratio of total time spent in confident versus cautious bouts. The y-axis is the ratio of bout time in the first 10 minutes of confident approach and the last 10 minutes of confident approach (set to 0 for timid animals that do not have a confident phase). The horizontal line indicates $y = 1.0$. All timid animals are close to the origin. We separate brave and intermediate animals according to the $y = 1$ line.

2.2 A Bayes-adaptive Model-based Model for Exploration and Timidity

2.2.1 State description

We use a model-based Bayes-adaptive reinforcement learning model to provide a mechanistic account of the behavior of the mice under threat of predation. This extends the model-free description of threat in Akiti et al. (2022) by constructing various mechanisms to explain additional facets of the dynamics of the behavior.

Underlying the BAMDP is a standard multi-step decision-making problem of the sort that is the focus of a huge wealth of studies (Russell and Norvig, 2016). We cartoon the problem with the four real and four counterfactual states shown in Fig 3. The *nest* is a place of safety, (modelling all places in the environment away from the object, ignoring, for instance, the change to thigmotactic behaviour that the mice exhibit when the object is introduced. The animal can choose to stay at the nest (possibly for multiple steps) or choose to make a cautious or confident approach.

At an approach state, the modelled agent can either stay, or return to the nest via the retreat state; the latter happens anyhow after four steps. The animal also imagines the (in reality, counterfactual) possibility of being detected by a potential predator. It can then either manage to escape back to the nest, or alternatively expire. We parameterize costs associated with the various movements; and also the probability of unsuccessful escape starting from confident (p_1) or cautious ($p_2 < p_1$) approach.

We describe the dilemma between cautious and confident approach as a calculation of the risk and reward trade-off between the two types of approaches. Cautious approach (the "cautious object" state) has a lower (informational) reward (e.g. because in the cautious state the animal spends more cognitive effort monitoring for lurking predators rather than exploring the object). But cautious approach leads to a lower probability of expiring if detected than does confident approach (the "cautious object"

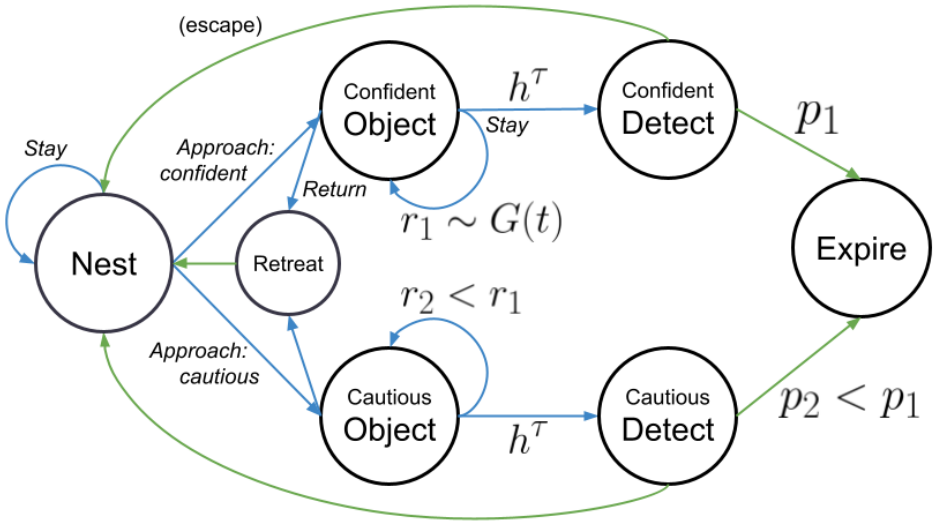

Fig 3. Markov decision process underlying the BAMDP model. Four real (nest, cautious object, confident object, retreat) and three imagined (cautious detect, confident detect, dead) states. Agent actions are italicized. Blue arrows indicate (possibly stochastic) transitions caused by agent actions. Green arrows indicate (possibly stochastic) forced transitions. Cautious approach provides less informational reward $r_2 < r_1$ but has a smaller chance of death $p_2 < p_1$ compared to confident approach. Travel and dying costs are not shown.

state) (e.g. because in the cautious state the animal is better poised to escape). Risk aversion modulates the agent's choice of approach type.

The next sections describe the characterization of the time-dependent risk of predation, the informational reward for exploration, the method of handling risk sensitivity, the way we fitted individual mice, and finally the full analysis of their behaviour. We report on recovery simulations in the supplement.

2.2.2 Modeling Threat with a Bayesian, Generalizing Hazard Function

Whilst exploring the novel object in the "object" state, we consider the animal as imagining that it might be detected, and then attacked, by a predator, whose appearance is governed by a temporal hazard function (see Fig 4).

Formally, the probability of detection given either cautious or confident approach is modelled using the *hazard* function h^τ, where τ is the number of steps the animal has so far spent at the object in the current bout. In a key simplification, this probability resets back to baseline upon a return to the nest. We treat the hazard function as being learned in a Bayesian manner, from the experience (in this case, of not being detected). We assume that the animal has the inductive bias that the hazard function is increasing over time, reflecting a potential predator's evidence accumulation process about the prey. Therefore, we derive it from a succession of independent Beta-distributed random variables $\theta^1 = 0; \theta^\tau \sim \text{Beta}(a^\tau, b^\tau), \tau > 1$ as:

$$h^\tau = 1 - \prod_{t=1}^{\tau}(1-\theta^t) \tag{1}$$

$$= h^{\tau-1} + \left(1 - h^{\tau-1}\right)\theta^\tau, \qquad \text{for } \tau > 1 \tag{2}$$

rather as in what is known as a stick-breaking process.

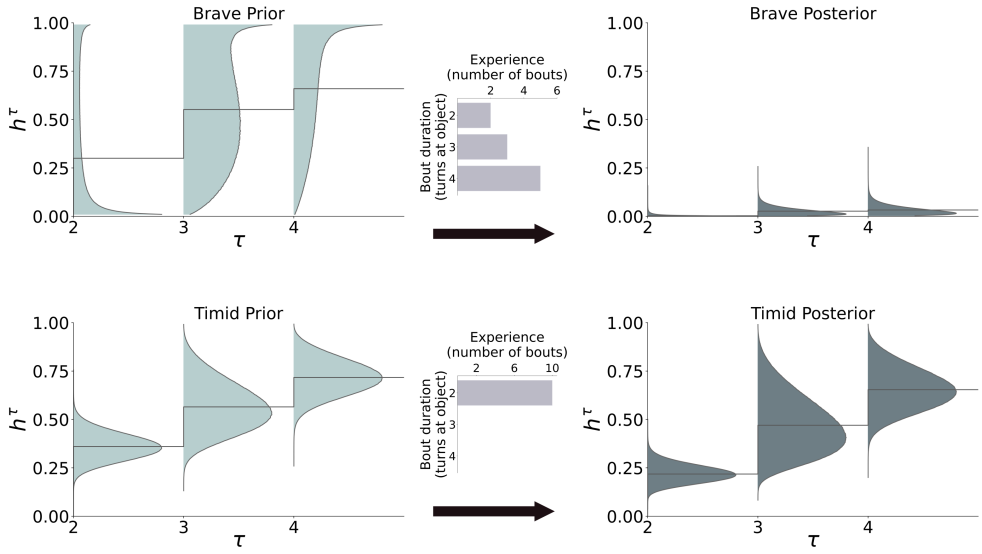

Fig 4. Hazard function learning for brave (top) and timid (bottom) animals. Brave animals start with a flexible hazard prior with a low mean for θ^1. This leads to longer bouts (first length 2, then 3 and 4), which imply that the hazard posterior quickly approaches zero (here, after 10 bouts). Timid animals start with an inflexible hazard prior with a higher mean θ^1, and are limited to length 2 bouts. The hazard posterior only changes slightly after 10 bouts.

Eq 2 shows that the hazard function is always increasing. As we will see, the duration of bouts at the object depend on the *slope* of the hazard function, with steep hazard functions leading to short bouts. In our model, the agent can stay at the object 2, 3 or 4 turns (we take $\theta^1 = 0$ as a way of coding actual approach).[1] Hence the collection of random variables, h^τ, is derived from six parameters (the mean μ^τ and the standard deviation σ^τ of the Beta distribution for the turn). These start at prior values (which we fit to the individual mice), and are subject to an update from experience, which, given the exclusively negative experience from the lack of actual appearance of a predator, has a closed form (see Methods). The animals' initial ignorance, which is mitigated by learning, makes the problem a BAMDP, whose solution is a risk-averse itinerant policy.

A particular characteristic of the noisy-or hazard function of Eq 1 is that the derived bout duration increases progressively. This is because not being detected at $\tau = 2$, say, provides information that θ^2 is small, and so reduces the hazard function for longer bouts $\tau > 2$.

Fig 4 shows the fitted priors of a brave (top) and timid (bottom) animal, as well as the posteriors at the end of model simulations. The brave animal starts with a high variance prior. This flexibility allows it to transition from short, cautious bouts (duration $\tau = 2$) to longer confident bouts (duration $\tau = 3, 4$), reducing the hazard function to near zero. The timid animals has a low variance prior, and does not stay long enough at the object to build sufficient confidence (only performing duration $\tau = 2$ bouts). As a result, its posterior hazard function remains similar to its prior.

[1] We therefore sometimes refer to cautious$-k$ or confident$-k$ bouts in which the model animal spends $k = \{2, 3, 4\}$ steps at the object.

2.2.3 Modeling the Motivation to Approach

We model the mouse's drive to approach the object using an exploration bonus $G(t)$ as an approximation to the value of information that would be derived from a fully Bayesian treatment. In the absence of actual reward, the model mouse will move from the "nest" state to the "object" state when these informational rewards exceed the costs implied by the risk of being attacked. The exploration bonus is implemented using an heuristic model with an initial bonus pool G_0 that becomes depleted, but is also replenished (through forgetting or potential change) at a steady rate f. We consider the animal to harvest the exploration bonus pool faster under confident than cautious approaches, for instance since it can pay more attention to the object. This underpins the transition between the two types of approach for non-timid animals. In simulations, when $G(t)$ is high, the agent has a high motivation to explore the object. In other words, the depletion from G_0 substantially influences the time point at which approach makes a transition from peak to steady-state; the steady-state time then depends on the dynamics of depletion (during time spent at the object) and replenishment (during time spent at the nest).

Finally, the animal is also motivated to approach by informational reward from the hazard function (which can be used exploited to collect more future reward) – according to a standard Bayes-adaptive bonus mechanism (Duff, 2002b).

2.2.4 Conditional Value at Risk Sensitivity

Along with varying degrees of pessimism in their prior over the hazard function, the mice could have different degrees of risk sensitivity in the aspect of the return that they seek to optimize. There are various ways in which the mice might be risk sensitive. Following Gagne and Dayan (2022), we consider a form called nested conditional value at risk (nCVaR). In general, CVaR_α, for risk sensitivity $0 \leq \alpha \leq 1$, measures the expected value in the lower α quantile of returns – thus over-weighting the worse outcomes. The lower α, the more extreme the risk-aversion; with $\alpha = 1$ being associated with the conventional, risk-neutral, expected value of the return. Section 5.2 details the optimization procedure concerned – it operates by upweighting the probabilities of outcomes with low returns – here, from detection and expiration. Thus, when α is low, confident and longer bouts are costly, inducing shorter, cautious ones. nCVaR_α affects behavior in a similar manner to pessimistic hazard priors, except that nCVaR_α acts on both the aleatoric uncertainty of expiring and epistemic uncertainty of detection, while priors only affect the latter. As we will see, despite this difference, we were not able to differentiate pessimistic priors from risk sensitivity using the data in (Akiti et al., 2022).

2.2.5 Model Fitting

The output of each simulation is a sequence of states from which we derive the statistics to fit the abstract mice data. The transition point from cautious to confident approach happens when the agent first ventures a confident approach; this switch is rarely reversed. Peak to steady-state transition points occur when the model mouse decreases its frequency of bouts, which tends to happen abruptly in the model. We fit the transition points in mouse data by mapping the length of a step in the model to wall-clock time. As in the abstraction of the experimental data, we average the duration (number of turns at the object) and frequency statistics in each phase. We characterize the relative frequencies of the bouts across phase transitions. Frequency mainly governs the total time at or away from the object and is formally defined as the inverse of the number of steps the model spends at the object and the nest.

We use a form of Approximate Bayesian computation Sequential Monte Carlo (ABCSMC; Toni et al. (2009)) to fit the elements of our abstraction of the approach

behaviour of the mice (section 2.1), namely change points, peak and steady-state durations as well as relative frequencies of bouts. See the Methods section 5.5 for details on the fitted statistics. At the core of ABCSMC is the ability to simulate the behaviour of model mice for given parameters. We do this by solving the underlying BAMDP problem approximately using receding horizon tree search with a maximum depth of 5 steps (which covers the longest allowable bout, defined as a subsequence of states where the model mouse goes from the nest to the object and back to the nest).

The full set of parameters includes 6 for the prior over the hazard function (given that we limit to four the number of time steps the model mouse can stay at the object), the risk sensitivity parameter α for CVaR_α, the initial reward pool G_0 and the forgetting rate f.

2.2.6 A Spectrum of Risk-Sensitive Exploration Trajectories

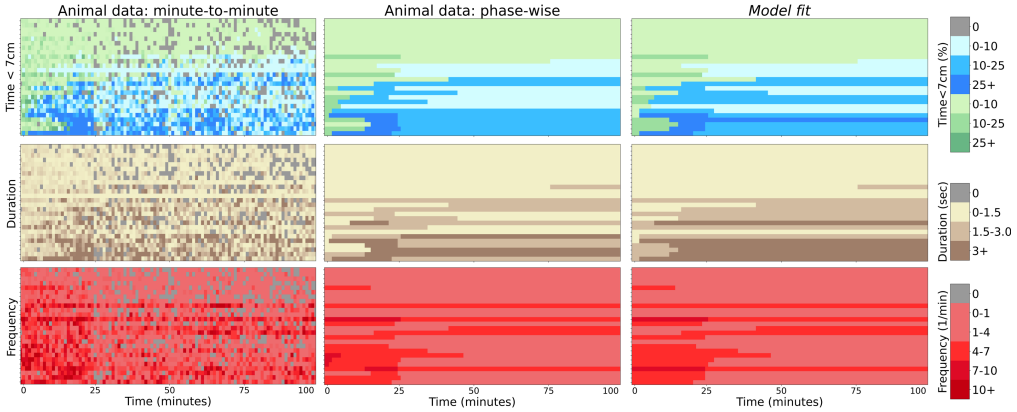

Fig 5. Summary of model fit. Left panels: minute-to-minute time the animals spend within 7cm of the novel object (top), duration (middle), and frequency (bottom). There are 26 animals (one per row) sorted by the animal ranking (see main text Section 2.2.6). Central panels: the same values averaged over behavioral phases. Right panels: time, duration and frequency of bouts generated as sample trajectories from the individual fits of the BAMDP model. Legend: green/blue distinguishes cautious and confident bouts. The intensity of colors indicates higher values, and gray indicates zeros.

Fig 5 shows model fits on the 26 mice from Akiti et al. (2022). The animal ranking is sorted firstly on animal group and secondly on total time spent near the object. We call this ranking the group-timidity animal index - it differs from the timidity index used in Akiti et al. (2022) which is only based on total time spent near the object. The model captures many details of the data across the entire spectrum of courage to timidity. The model explains the behavior of animals mechanistically. Differing schedules of exploration emerge because of the battle between learning about threat and reward.

All animals initially assess risk with cautious approach, since potential predation significantly outweighs potential rewards. Brave animals assess risk either with short (length 2 bouts) or medium (length 3 bouts) depending on the hazard priors (Fig 6, left panel). If $E[h^3]$ is high, then the animal performs cautious length 2 bouts, otherwise, it performs cautious length 3 bouts. With more bout experience, the posterior hazard function becomes more optimistic (since there is no actual predator to observe; Fig 4), empowering it to take on more risk by staying even longer at the object and performing confident approach. How long brave animals spend assessing risk depends on hazard priors and the risk sensitivity nCVaR_α.

Fig 7 shows that the fitted hazard priors and nCVaR_α relate to the timidity of

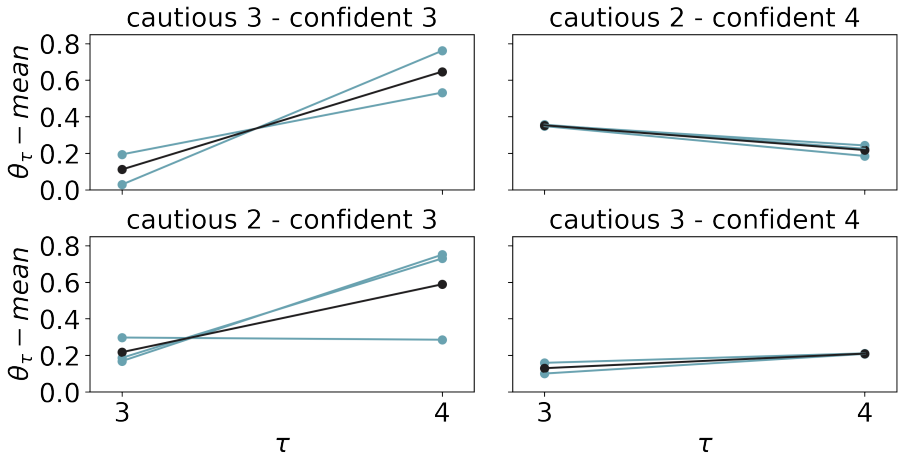

Fig 6. The bout durations of brave animals depend on the hazard prior. Top-left: brave animals that initially perform cautious-3 bouts, then confident-3 bouts. These animals are fitted with a low t=3 prior and high t=4 prior because they never perform duration-4 bouts. Blue indicates individual animals and black indicates the mean. Bottom-left: cautious-2 then confident-3 animals. The prior for t=3 is higher because there is some hazard to overcome before the animal does a duration-3 bout. Top-right: cautious-2 then confident-4 animals. Both t=3 and t=4 priors are low. Once the animal overcomes the t=2 hazard, it quickly transitions from duration 2 to 4. Bottom-right: cautious-3 then confident-4 animals. Because the t=3 prior is low, the animals begins with duration-3 bouts.

animals (as defined in Akiti et al. (2022) by time spent at the object). Brave animals are fitted by higher nCVaR_α and low slope and high variance (flexibility) hazard prior. In other words, the model brave mouse believes that the hazard probability for long bouts is low in its environment. Timid animals are fitted by lower nCVaR_α and high slope, inflexible hazard prior. Intermediate animals' parameters are between brave and timid animals'.

For brave animals, G_0 determines how much time brave animals spend in the peak-confident exploration phase, or the peak to steady-state change point. Animals with larger G_0 tend to have high bout frequency for a longer period (see Fig 8). Finally, how often brave animals revisit the object, which is related to the relative steady-state frequency, is determined by the forgetting rate.

Timid animals have short bouts and continue to assess risk with cautious approach in the steady-state. According to Fig 7, the reasons are the hazard prior is inflexible (low variance) and has a high slope and low nCVaR_α. The priors are slow to update and risk sensitivity causes timid agents to overestimate the probability of bad outcomes, leading to prolonged cautious behavior. Hence, the reward exploration pool is depleted (i.e. the agent transitions to the steady-state phase) before the agent overcomes its priors. This particular dynamic of approach-drive and hazard function updating leads to self-censoring and neophobia. In the steady-state phase, the agent stays long periods at the nest (how long depends again on the forgetting rate). As a result, the animal (during the course of the experiment) never accumulates sufficient evidence to learn the safety of the object or if the object yields rewards. However, Akiti et al's experiment did not last long enough to answer the question of whether all animals, even the timidest ones eventually perform confident approach. Our model predicts that they will since the agent only accumulates negative evidence for the hazard function. However, with sufficient low CVaR or pessimistic priors, this may take a very long time.

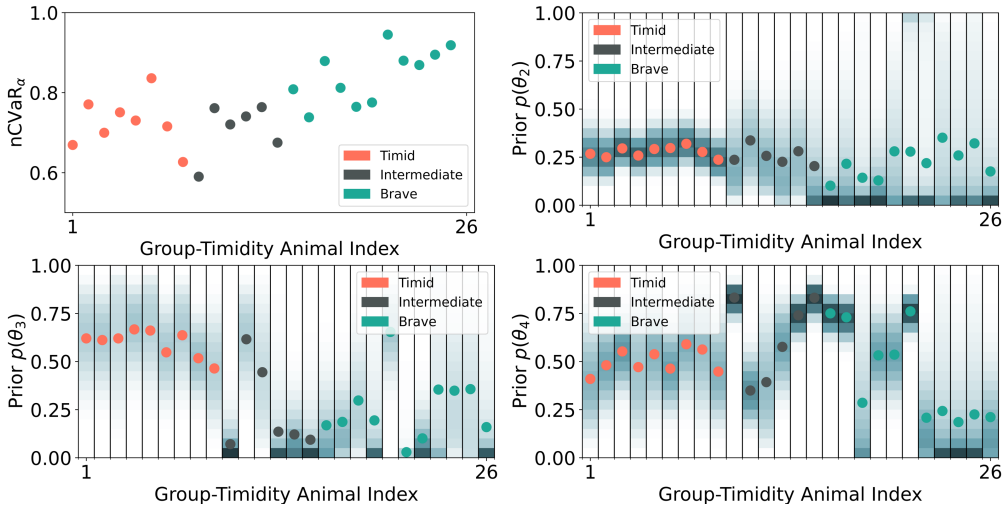

Fig 7. (Top-left) nCVaR versus the animal ranking defined in Section 2.2.6. Color indicates the animal group. More timid animals are generally fitted by a lower nCVaR$_\alpha$. (Top-right, bottom-left, bottom-right) Prior hazard parameter for t=2, 3, 4 respectively versus timidity ranking. Dots indicate the mean; the probability density is represented by color where darker means higher density regions. The t=2 prior mean is similar across all animals explaining the short, cautious bouts all animals initially use to assess risk. However, timid animals are best fit with lower variance (inflexible) and higher t=3 and t=4 prior means. This leads to shorter, cautious bouts in the long run. Brave animals are fitted by a low slope (indicated by lower mean for t=3 and t=4) and high variance (flexible) hazard prior. This allows them to perform longer bouts over time. t=4 mean is low (panel d) for brave animals that perform length 4 bouts. Like brave animals, most intermediate animals have flexible, gradual hazards up to t=3.

Intermediate animals, like brave animals, eventually switch to confident approach to maximize information gained about potential rewards. Similar to brave animals, the cautious to confident transition tends to be later with lower nCVaR$_\alpha$ and steeper, less flexible priors. Intermediate animals perform both cautious and confident bouts with medium duration. This is captured by a hazard prior with smaller $E[h^3]$ and larger $E[h^4]$. The percentage of time spent at the object is relatively constant throughout the experiment for intermediate animals. This can be explained by either large G_0 or a high forgetting rate. In other words, the animal is either slow to update its belief about the potential reward at the object, or it expects the reward probability to change quickly.

Fig 5 also illustrates several limitations of the model. In particular, the duration of bouts can only increase, whereas a few animals exhibit decreasing bout duration between confident-peak and confident-steady-state phases. Furthermore, the model has trouble capturing abrupt changes in duration (from 2 turns to 4) coinciding with an animal's transition from cautious to confident approach.

2.2.7 Risk Sensitivity versus Prior Belief Pessimism

We found that risk sensitivity and prior pessimism could not be teased apart in our model fits. This is illustrated in Fig 9. In the ABCSMC posterior distributions, nCVaR$_\alpha$ is correlated with θ_2-mean for timid and intermediate animals, θ_3-mean for cautious-2/confident-4 and cautious-2/confident-3 animals, and θ_4-mean for cautious-2/confident-4 and cautious-3/confident-4 animals. In other words, lower nCVaR$_\alpha$ (higher risk-sensitivity) can be traded off against lower (more optimistic)

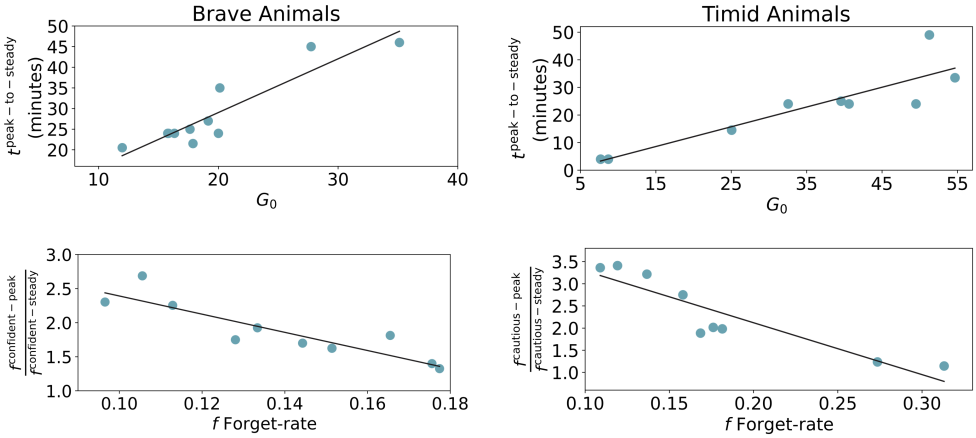

Fig 8. Top-left: the relationship between G_0 and the peak to steady-state change point for brave animals. The best fit line is shown in black. Higher G_0 means the agent explores longer, hence postponing the change point. Bottom-left: forgetting rate versus steady-state turns at the nest state for brave animals. A higher forgetting rate leads to quicker replenishment of the exploration pool and hence fewer turns at the nest before approaching the object. Top-right: G_0 versus peak to steady-state change point for timid animals. Bottom-right: forgetting rate versus turns at nest timid animal. All correlations are significant with $p < 0.002$.

priors to explain the observed risk-aversion in animals.

In ablation studies (not shown), we found that hazard priors with a risk-neutral $\text{nCVaR}_{1.0}$ is capable of fitting the full range of animals equally well. The only advantage of fitting both nCVaR_α and hazard priors to each animal is greater diversity in the particles discovered by ABCSMC. While the model with $\text{nCVaR}_{1.0}$ is simpler, one might suspect, on general grounds, that both risk sensitivity and belief pessimism affect mice behavior.

We also found that nCVaR_α alone, with the same hazard prior for all animals, is incapable of fitting the full range of animal behavior (results not shown). This can be explained by the fact that nCVaR_α cannot model the different slopes in the hazard function. For example, a cautious-2/confident-3 animal must be modeled using a high θ_4-mean. Starting with the parameters for a cautious-2/confident-4 animal and decreasing nCVaR_α will not create a cautious-2/confident-3 animal. Instead, decreasing nCVaR_α will delay the cautious-to-confident transition of the cautious-2/confident-4 animal and eventually create a cautious-2 timid animal. This is unsurprising since nCVaR_α is a single free parameter while six hazard prior parameters are required to produce the different hazard functions that capture the full range of animal behavior. This illustrates that in general, structured prior beliefs are required in addition to nCVaR_α to model detailed behavior in complex environments.

3 Discussion

We combined a Bayes adaptive Markov decision process framework with a conditional value at risk objective to capture many facets of an abstraction of the substantially different risk-sensitive exploration of individual animals reported by Akiti et al. (2022). In the model, behaviour reflects a battle between learning about potential threat and potential reward. Individual variability in the schedules of exploratory approach was explained by different risk sensitivities, forgetting rates, exploration bonuses and prior

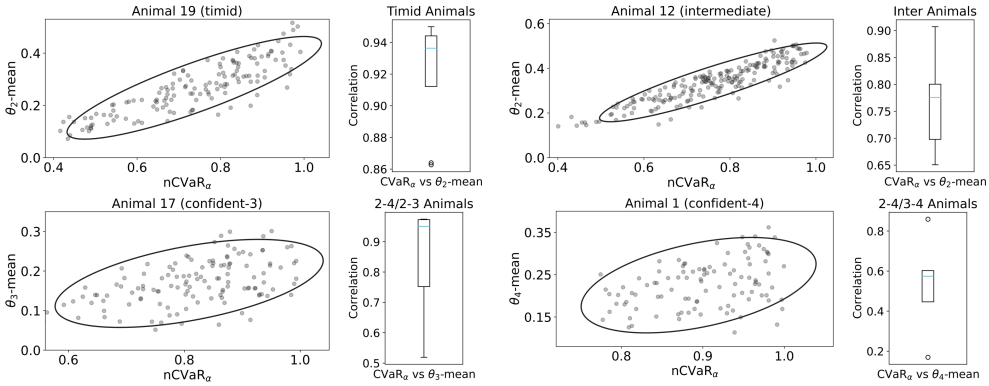

Fig 9. Non-identifiability of nCVaR$_\alpha$ against the hazard prior. Top-left: the scatter plot shows the t=2 prior mean (θ_2-mean) versus nCVaR$_\alpha$ for ABCSMC particles of timid animal 19. The ellipse indicates one standard deviation in a Gaussian density model. Animal 19 (and timid animals generally) can be either fit with a higher nCVaR$_\alpha$ and a higher θ_2-mean, or a lower nCVaR$_\alpha$ and a lower θ_2-mean. The box-and-whisker plot illustrates the correlation between θ_2-mean and nCVaR$_\alpha$ across all timid animals. Top-right: the scatter plot shows an example intermediate animal 12; the box-and-whisker plot shows θ_2-mean versus nCVaR$_\alpha$ for the intermediate population. Bottom-left: the scatter plot shows an example cautious-2/confident-4 and cautious-2/confident-3 animal 17. This group of animals starts with duration= 2 bouts and hence must overcome the prior θ_3-mean. The box-and-whisker plot shows θ_3-mean versus nCVaR$_\alpha$ for the population. Bottom-right: the scatter plot shows an example cautious-2/confident-4 and cautious-3/confident-4 animal 1. This group of animals eventually performs duration= 3 bouts and hence must overcome the prior θ_4-mean. The box-and-whisker plot shows θ_4-mean versus nCVaR$_\alpha$ for the population. nCVaR$_\alpha$ and θ-mean are correlated in the ABCSMC posterior for all animals and hence non-identifiable. $p < 0.05$ for all correlations.

beliefs about an assumed hazard associated with a novel object. Neophilia arises from a form of optimism in the face of uncertainty, and neophobia from the hazard. Critically, the hazard function is generalizing (reducing the $t = 2$ hazard reduces the $t = 4$ hazard) and monotonic. The former property induces an increasing approach duration over time (Arsenian, 1943). Furthermore, the exploration bonus associated with the object regenerates, as if the subjects consider its affordance to be non-stationary (Dayan et al., 2000). This encourages even the most timid animals to continue revisiting it.

A main source of persistent timidity is a sort of path-dependent self-censoring (Dayan et al., 2020). That is, the agents could be so pessimistic about the object that they never visit it for long enough to overturn their negative beliefs. This can in principle arise from either excessive risk-sensitivity or overly pessimistic priors. We found that it was not possible to use the model to disentangle the extent to which these two were responsible for the behavior of the mice, since they turn out to have very similar behavioural phenotypes. One key difference is that risk aversion continues to affect behaviour at the asymptote of learning; something that might be revealed by due choice of a series of environments. Certainly, according to the model, forced exposure Huys et al. (2022) would hasten convergence to the true hazard function and the transition to confident approach.

Due to the complexity of the dataset, we made several rather substantial simplifying assumptions. First, we adopted a particular set of state abstractions, for instance representing thigmotaxis as a notional "nest" (Simon et al., 1994). Second, we only allowed the agent to stay for a maximum of four turns at the object. In reality, the mice

could stay arbitrarily long near the object. Third, instead of maintaining appropriately Bayesian beliefs about potential rewards and deriving an exploration bonus accordingly, we used an heuristic regenerating exploration pool. Fourth, the model only allows the frequency of approach, and not its duration, to decrease during the steady-state phase. However, for some animals, duration or both decrease. Fifth, the probability of being detected was the same between cautious and confident approaches, which may not be true in general. Note that the agent decides the type of approach before the bout, and is incapable of switching from cautious to confident mid-bout or vice versa. This is consistent with behavior reported in Akiti et al. (2022). Sixth, we restricted ourselves to a monotonic hazard function for the predator. It would be interesting to experiment with a non-monotonic hazard function instead, as would arise, for instance, if the agent believed that if the predator has not shown up after a long time, then there actually is no predator. Of course, a sophisticated predator would exploit the agent's inductive bias about the hazard function – by waiting until the agent's posterior distribution has settled. In more general terms, the hazard function is a first-order approximation to a complex game-theoretic battle between prey and predator, which could be modeled, for instance using an interactive IPOMDP (Gmytrasiewicz and Doshi, 2005). How the predator's belief about the whereabouts of the prey diminishes can also be modeled game-theoretically, leading to partial hazard resetting rather than the simplified complete resetting in our model.

Our account is model-based, with the mice assumed to be learning the statistics of the environment and engaging in prospective planning (Mobbs et al., 2020). By contrast, Akiti et al. (2022) provide a model-free account of the same data. They suggest that the mice learn the values of threat using an analogue of temporal difference learning (Sutton, 1988), and explain individual variability as differences in value initialization (Akiti et al., 2022). The initial values are generalizations from previous experiences with similar objects, and are implemented by activity of dopamine in the tail of the striatum (TS) responding to stimuli salience (Akiti et al., 2022). By contrast, our model encompasses extra features of behavior such as bout duration, frequency, and type of approach – ultimately arriving at a different mechanistic explanation of neophobia. In the context of our model, TS dopamine could still respond to the physical salience of the novel object but might then affect choices by determining the potential cost of the encountered threat (a parameter we did not explore here) or perhaps the prior on the hazard function. An analogous mechanism may set the exploration pool or the prior belief about reward - perhaps involving projections from other dopamine neurons, which have been implicated in novelty in the context of exploration bonuses (Kakade and Dayan, 2002) and information-seeking for reward (Bromberg-Martin and Hikosaka, 2009; Ogasawara et al., 2022).

Of course, agents do not need to be fully model-free or model-based. They can truncate model-based planning using model-free values at leaf nodes (Keramati et al., 2016). Furthermore, prioritized model-based updates can update a model-free policy when environmental contingencies change (Antonov and Dayan, 2023). Finally, while online BAMDP planning can be computationally expensive, a model-based agent may simply amortize planning into a model-free policy which it can reuse in similar environments or even precompile model-based strategies into an efficient model-free policy using meta-learning (Wang et al., 2017). Agents may have faced many different exploration environments with differing reward and threat trade-offs through their lifetimes and across evolutionary scales that they have used to create fast, instinctive model-free policies that resemble prospective, model-based behavior (Mattar and Daw, 2018; Rusu et al., 2016). In turn, TS dopamine might reflect aspects of MF values or prediction errors that had been trained by a MB system following the precepts we outlined.

In Akiti et al. (2022), ablating TS-projecting dopamine neurons made mice "braver". They spent more time near the object, performed more tail-exposed approach and transitioned faster to tail-exposed approach compared to control. In Menegas et al. (2018) TS ablation affected the learning dynamics for actual, rather than predicted threat. Both ablated and control animals initially demonstrated retreat responses towards airpuffs but only control mice maintained this response (Menegas et al., 2018). After airpuff punishment, ablated individuals surprisingly did not decrease their choices of water ports associated with airpuffs (while controls did). One possibility is that this additional exposure could have caused acclimatization to the airpuffs in the same way that brave animals in our study acclimatize to the novel object by approaching more, and timid animals fail to acclimatize because of self-censoring. Indeed, future experiments might investigate why punishment-avoidance does not occur in ablated animals and whether the same holds in risk-sensitive exploration settings (Menegas et al., 2018). In other words, would mice decrease approach after reaching the "detected" state, as expected by our model, or would they maladaptively continue the same rate of approach? Finally, while our study has focused on threat, Menegas et al. (2017) showed that TS also responds to novelty and salience in the context of rewards and neutral stimuli. That TS ablated animals spend more, rather than less time near the novel object suggests that the link from novelty to neophilia and exploration bonuses might not be mediated by this structure.

The behaviour of the mice in Akiti et al. (2022) somewhat resembles attachment behaviour in toddlers (Ainsworth, 1964; Bowlby, 1955), albeit with the care-giver's trusty leg (a secure base from which to explore) replaced by thigmotaxis (or, in our case, the notional 'nest'). Characteristic to this behaviour is an intermittent exploration strategy, with babies venturing away from the leg for a period before retreating back to its safety. Through the time course of the experiment, the toddler progressively ventures out longer and farther away, spending more time actively playing with the toys rather than passively observing them in hesitation (Arsenian, 1943). This is another example of a dynamic exploratory strategy, putatively arising again from differential updates to beliefs about threats and the rewards in the environment (Ainsworth, 1964; Arsenian, 1943).

Variability in timidity during exploration has been reported in other animal species and can be caused by differences in both prior experience and genotype. Fish from predator-dense environments tend to make more inspection approaches but stay further away, avoid dangerous areas (attack-cone avoidance) and approach in larger shoals compared to fish from predator-sparse environments (Dugatkin, 1988; Magurran, 1986; Magurran and Seghers, 1990). Dugatkin (1988) and Magurran (1986) report significant within-population differences in the inspection behavior of guppies and minnows respectively. Brown and Dreier (2002) directly manipulates the predator experience of glowlight tetras, leading to changes to inspection behavior. Similar inter- and intra-population differences in timidity have been reported in mammals. In Coss and Biardi (1997), the squirrel population sympatric with the tested predators stayed further away and spent less time facing the predator compared to the allopatric population. Furthermore, the number of inspection bouts differed between litters, between individuals within the same litter, and even between the same individuals at different times during development (Coss and Biardi, 1997). In Kemp and Kaplan (2011), marmosets differed in risk-aversion when inspecting a potential (taxidermic) predator but risk-aversion was not stable across contexts for some individuals. FitzGibbon (1994) reports age differences in inspection behavior - adolescent gazelles inspected cheetahs more than adults or half-growns. Finally, Eccard et al. (2020); Mazza et al. (2019) report substantial individual differences in the foraging behavior of voles in risky environments.

4 Conclusion

In conclusion, our model shows that risk-sensitive, normative, reinforcement learning can account for individual variability in exploratory schedules of animals, providing a crisp account of the competition between neophilia and neophobia that characterizes many interactions with an incompletely known world.

5 Materials and methods

5.1 BAMDP Hyperstate

A Bayes-Adaptive Markov Decision Process (BAMDP; Duff, 2002a; Guez et al., 2013) is an extension of model-based MDP and a special case of a Partially Observable Markov Decision Process (POMDP; Kaelbling et al., 1998) in which the agent models its uncertainty about the (unchanging) transition dynamics. In a BAMDP, the agent extends its state representation into a hyperstate consisting of the original MDP state s, and the belief over the transition dynamics $b(T)$.

In our model s is the conjunction of the "physical state" (the location of the agent, as shown in Fig 3) and the number of turns the agent has spent at the object so far τ. $b(T)$ is the agent's posterior belief over the hazard function. In this simple case, $b(T)$ is parameterized as a vector of beta distributions, with parameters $\vec{\eta}_1$ and $\vec{\eta}_0$.

$$b(T) = p(T; \vec{\eta}^1, \vec{\eta}^0) \tag{3}$$

Our hyperstate additionally contains the nCVaR static risk preference $\bar{\alpha}$, and the parameters of the heuristic exploration bonus G, n_1, n_0 (see Section 5.4).

5.2 Bellman Updates for BAMDP nCVaR

As for a conventional MDP, the nCVaR objective for a BAMDP can be solved using Bellman updates. We use Eq 4 which assumes a deterministic, state-dependent, reward.

$$V^*(b(T), s, \bar{\alpha}) = \max_a \left[r(s) + \gamma \min_{\xi \in \mathcal{U}(\bar{\alpha})} \sum_{s'} \xi(b'(T), s') \bar{T}(s, a, s') V^*(b'(T), s', \bar{\alpha}) \right] \tag{4}$$

s' is the next state and $b'(T)$ is the posterior belief over transition dynamics after observing the transition (s, a, s'). $\bar{T}(s, a, s')$ is the expected transition probability.

$$\bar{T}(s, a, s') = \int T(s, a, s') b(T) dT \tag{5}$$

Proof of Eq 4.

$$V^*(b(T), s, \bar{\alpha}) = \max_a \{ r(x) + \gamma \min_{\xi \in \mathcal{U}(\bar{\alpha})} \int_{\hat{b}(T), s'} \xi(\hat{b}(T), s') \cdot p([b(T), s], a, [\hat{b}(T), s']) \cdot V^*(\hat{b}(T), s', \bar{\alpha}) d[\hat{b}(T), s'] \}$$

where $\mathcal{U}(\bar{\alpha}) = \{ \xi : \xi(\hat{b}(T), s') \in [0, \frac{1}{\bar{\alpha}}], \int_{\hat{b}(T), s'} \xi(\hat{b}(T), s') p([b(T), s], a, [\hat{b}(T), s']) = 1 \}$ is the risk envelope for CVaR (Chow et al., 2015). But $p([b(T), s], a, [\hat{b}(T), s'])$ is only

non-zero when $\hat{b}(T) = b'(T)$.

$$\begin{aligned} p([b(T), s], a, [\hat{b}(T), s']) &= \int_T p(\hat{b}(T), s' | T, s, a) b(T) dT \\ &= \int_T \delta(\hat{b}(T) - b'(T)) T(s'|s, a) b(T) dT \\ &= \begin{cases} \bar{T}(s, a, s') = \int_T T(s'|s, a) b(T) dT, & \text{for } \hat{b}(T) = b'(T) \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Hence we can drop the independent integration over $\hat{b}(T)$, and only integrate over s'.

$$\begin{aligned} V^*(b(T), s, \bar{\alpha}) &= \max_a \{r(s) + \gamma \min_{\xi \in \mathcal{U}(\bar{\alpha})} \int_{s'} \xi(b'(T), s') \cdot \bar{T}(s, a, s') \cdot V^*(b'(T), s', \bar{\alpha}) ds'\} \\ &= \max_a \{r(s) + \gamma \min_{\xi \in \mathcal{U}(\bar{\alpha})} \sum_{s'} \xi(b'(T), s') \cdot \bar{T}(s, a, s') \cdot V^*(b'(T), s', \bar{\alpha})\} \end{aligned}$$

□

Epistemic uncertainty about the transitions only generates risk in as much as it affects the probabilities of realizable transitions in the environment.

5.3 Noisy-Or Hazard Function

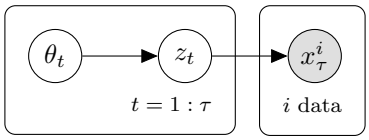

Fig 10. Bayes-net showing the relationship between the random variables in the noisy-or model. Only x_τ is shown. $x_{\tau+1}$ depends on $z_{t=1:\tau+1}$, and so on.

$$X_\tau = Z_1 \cup Z_2 \cup \ldots Z_\tau \tag{6}$$

In our model, the hazard function defines a binary detection event X_τ for each number of turns the agent spends at the object $\tau = 2, 3, 4$. The predator detects the agent when $X_\tau = 1$. We use a noisy-or hazard function which defines X_τ as the union of Bernoulli random variables $Z_t \sim \text{Bernoulli}(\theta_t)$ (Eq 6) with priors $\theta_t \sim \text{Beta}(\eta_t^1, \eta_t^0)$ for $t = 2, 3, 4$. Fig 10 shows the relationships between the random variables in plate notation.

Posterior inference for the noisy-or model is intractable in the general case (Jaakkola and Jordan, 1999). However, there is a closed-form solution for the posterior when the agent only makes *negative* observations, meaning $x_\tau^i = 0 \; \forall i$ (in our case, since there is no actual predator). For example, given a single observation $x_\tau = 0$,

$$\begin{aligned} p(\theta_{t=1:\tau} | x_\tau = 0) &= \frac{p(x_\tau = 0 | \theta_{t=1:\tau}) p(\theta_{t=1:\tau})}{p(x_\tau = 0)} \\ &= \frac{\prod_{t=1:\tau} p(z_t = 0 | \theta_t) p(\theta_t)}{p(x_\tau = 0)} \\ &= \frac{\prod_{t=1:\tau} (1 - \theta_t) \text{Beta}(\theta_t; \eta_t^1, \eta_t^0)}{p(x_\tau = 0)} \end{aligned}$$

By conjugacy,

$$p(\theta_{t=1:\tau}|x_\tau = 0) \sim \prod_{t=1:\tau} \text{Beta}(\theta_t; \eta_t^1, \eta_t^0 + 1) \tag{7}$$

Hence the posterior update simply increments the Beta pseudocounts for the '0' outcomes. The hazard probability is the posterior predictive distribution $h(\tau) = p(x_\tau = 1|D)$ where D are a set of observations of $X_1, X_2, \ldots X_\tau$.

$$p(x_\tau = 1|D) = 1 - \prod_{t=1}^{\tau}(1 - \mu_t) \tag{8}$$

Where $\mu_t = \mathbb{E}[\theta_t]$ is the expected value of the posterior on θ_t.

Proof of Eq 8.

$$\begin{aligned}
p(x_\tau = 1|D) &= 1 - p(x_\tau = 0|D) \\
&= 1 - \int p(x_\tau = 0|\theta_{t=1:\tau}) P(\theta_{t=1:\tau}|D) d\theta_{t=1:\tau} \\
&= 1 - \int \prod_{t=1}^{\tau} p(z_t = 0|\theta_t) P(\theta_t|D) d\theta_t \\
&= 1 - \int \prod_{t=1}^{\tau} (1 - \theta_t) \text{Beta}(\theta_t; \eta_t^1, \tilde{\eta}_t^0) d\theta_t \\
&= 1 - \prod_{t=1}^{\tau} \int (1 - \theta_t) \text{Beta}(\theta_t; \eta_t^1, \tilde{\eta}_t^0) d\theta_t \\
&= 1 - \prod_{t=1}^{\tau} (1 - \mu_t)
\end{aligned}$$

□

where $\tilde{\eta}_t^0$ are the pseudocounts of negative observations after updating the Beta prior with D using Eq 7. It can be shown that $h(\tau)$ is recursive.

$$h(\tau) = h(\tau - 1) + [1 - h(\tau - 1)]\mu_\tau \tag{9}$$

This recursion has two implications. First, the hazard function is monotonic since $(1 - h(\tau - 1)) > 0$ and $\mu_\tau > 0$. Second, the hazard function generalizes. From Eq 9 it is clear if $h(\tau - 1)$ increases, then $h(\tau)$ increases. It is this generalization that allows the agent to progressively spend more turns at the object.

5.4 Heuristic Exploration Bonus Pool

The heuristic reward function approximates the sort of exploration bonus (Gittins, 1979) that would arise from uncertainty about potential exploitable benefits of the object. It incentivizes approach and engagement. In the experiment, there is no actual reward so the motivation is purely intrinsic (Oudeyer and Kaplan, 2007). The exploration bonus depletes as the agent learns about the object; but regenerates if the agent believes that the object can change over time (or, equivalently, if the agent forgets what it has learnt). Since we imagine the agent as finding more out about the object through confident than cautious approach, the former generates a greater bonus per step, but also depletes it more quickly.

We model the exploration-based reward as an exponentially decreasing resource. $G(t)$ is the "exploration bonus pool" and can be interpreted as the agent's remaining

motivation to explore in the future. We fit the size of the initial exploration pool $G(0) = G_0$ to the behavior of each animal. During planning, the agent imagines receiving rewards at the cautious and confident object states proportional to $G(t)$.

$$\hat{r}_{\text{cautious}} = \omega_{\text{cautious}} \cdot G(t) \tag{10}$$

$$\hat{r}_{\text{confident}} = \omega_{\text{confident}} \cdot G(t) \tag{11}$$

$$\hat{r}_{\text{cautious}} < \hat{r}_{\text{confident}} \tag{12}$$

On every turn at the cautious or confident object states, the agent *extracts* reward $\hat{r}_{\text{cautious}}$ or $\hat{r}_{\text{confident}}$ from its budget G, depleting G at rates ω_{cautious} or $\omega_{\text{confident}}$. This leads to an exponential decrease in $G(t)$ with turns spent at the object which is clear from Eq 13. For example, at the cautious object state the update to $G(t)$ is,

$$G(t+1) = G(t) - \hat{r}_{\text{cautious}} = (1 - \omega_{\text{cautious}})G(t) \tag{13}$$

However, a secondary factor affects the update to $G(t)$. G linearly regenerates back to G_0 at the forgetting rate f which we also fit for each animal. The full update to the reward pool for spending one turn at the cautious object state is,

$$G(t+1) = \min\{(1 - \omega_{\text{cautious}})G(t) + f, G_0\} \tag{14}$$

Note that $G(t)$ regenerates by f in all states, not only at the object states. We use linear forgetting for its simplicity although other mechanisms such as exponential forgetting are possible.

Finally, for completeness in other environments, the reward the agent imagines receiving also depends on the actual reward it has received in the past. Let n^1 and n^0 be the number of times the agent has received one or zero reward at the object state, analogous to the pseudocounts of a Beta posterior in a fully Bayesian treatment of reward. Furthermore, let n_0^1 and n_0^0 be the (fitted) values at $t = 0$. We use $n_0^1 = 1$ and $n_0^0 = 1$. The agent imagines receiving reward

$$r_{\text{cautious}} = \hat{r}_{\text{cautious}} + \frac{n^1}{n^1 + n^0} \tag{15}$$

after spending one turn in the cautious object state. A similar equation applies to the confident object state.

We define the depletion rates as $\omega_{\text{confident}} = \frac{R}{G_0}$ and $\omega_{\text{cautious}} = K \cdot \omega_{\text{confident}}$ with constants $R = 1.1$ and $K = 0.89 < 1.0$. These values were fitted to capture the full range of behavior of the 26 animals.

5.5 Data Fitting

Data fitting aims to elucidate individual differences and population patterns in behavior by searching for the model parameters that best describe the behavior of each animal. We map the behavior of model and animals to a shared abstract space using a common set of statistics and then fit the model to data using ABCSMC.

5.5.1 Animal Statistics

To extract animal statistics, we first coarse-grain behavior into phases and subsequently classify the animals into three groups: brave, intermediate, and timid (as described in the main text). This allows us to maintain the temporal dynamics of the behavior while reducing the dimension of the data. We average the approach type, duration, and frequency over each phase and fit a subset of statistics that capture the high-level temporal dynamics of behavior of animals in each group.

The behavior of brave animals comes in three phases: cautious, confident-peak and confident-steady-state. We fit five statistics: the transition time from cautious to confident-peak phase $t^{\text{cautious-to-confident}}$, the transition time from confident-peak to confident-steady-state phase $t^{\text{peak-to-steady}}$, the average durations during the cautious and confident-peak phases $d^{\text{cautious}}, d^{\text{peak-confident}}$, and the ratio of confident-peak and confident-steady-state phases' frequencies $\frac{f^{\text{confident-peak}}}{f^{\text{confident-steady}}}$.

Intermediate animals only exhibit two phases: cautious and confident. We fit four statistics: the transition time from cautious to confident phase $t^{\text{cautious-to-confident}}$, the durations of the two phases $d^{\text{cautious}}, d^{\text{confident}}$, and the ratio of the cautious and confident phases frequencies $\frac{f^{\text{cautious}}}{f^{\text{confident}}}$. However, one limitation of the model is that frequency can only decrease, not increase, because of the dynamics of depletion and replenishment of the exploration bonus pool. Hence we instead fit $\max\{\frac{f^{\text{cautious}}}{f^{\text{confident}}}, 1.0\}$.

Timid animals also only exhibit two phases, albeit different ones from the intermediate animals: cautious-peak and cautious-steady-state. We fit four statistics: the transition time from cautious-peak to cautious-steady-state phase $t^{\text{peak-to-steady}}$, the durations of the two phases $d^{\text{cautious-peak}}, d^{\text{cautious-steady}}$, and the ratio of the frequencies of the two phases $\frac{f^{\text{cautious-peak}}}{f^{\text{cautious-steady}}}$.

5.5.2 Model Statistics

By design, our BAMDP agent also enjoys a notion of bouts and behavioral phases. We map the behavior of the agent to the same abstract space of duration, frequency, and transition time statistics as the animals to allow the fitting.

We consider the agent as performing a bout when it leaves the nest, stays at the object state for some turns, and finally returns to the nest. We parse bouts and behavioral phases from the overall state trajectory of the agent which, like the animals, has what we can describe as contiguous periods of cautious or confident approach and low or high approach frequency.

The transition from cautious to confident phase (measured in the number of turns) is when the model begins visiting the confident-object state rather than the cautious-object state (this transition never happens for low $\bar{\alpha}$). The transition from peak to steady-state phase is when the model starts spending > 1 consecutive turns at the nest (to regenerate G), which happens when G reaches its steady-state value determined by the forgetting rate. We linearly map the agent's transition times (in units of turns) to the space of animals' transition times (units of minutes) using the relationship: 2 turns to 1 minute. Therefore, agent is simulated for 200 turns corresponding to 100 minutes in the experiment.

Bout duration is naturally defined as the number of consecutive turns the agent spends at the object. Because the agent lives in discrete time, we map its duration (units of turns) to the space of animal duration (units of seconds) using the formula,

$$d_{\text{animal}} = 0.75 + 1.5(d_{\text{agent}} - 2) \tag{16}$$

Hence the agent is capable of having durations from 0.75 to 3.75 seconds. This captures a large range of the animals' phase-averaged durations.

We define the momentary frequency with which the agent visits the object as the inverse of the period, which is the number of turns between two consecutive bouts (sum of turns at nest and object states). Frequency ratios are computed by dividing the average periods of two phases (in units of turns) and are unitless. Hence, no mapping between agent and animal frequency ratios is necessary.

5.5.3 Approximate Bayesian Computation

We fit each of the 26 animals from Akiti et al. (2022) separately using an Approximate Bayesian Computation Sequential Monte Carlo (ABCSMC) algorithm (Toni et al., 2009). We use an adaptive acceptance threshold schedule that sets ϵ_t to the lowest 30-percentile of distances $d(x, x_0)$ in the previous population. We use a Gaussian transition kernel $K_t(\theta|\theta^*) = \mathcal{N}(0, \Sigma)$, where the bandwidth of Σ is set using the Silverman heuristic. We ran ABC-SMC for $T = 30$ populations for each animal but most animals converged earlier. We used uniform priors. Table 1 contains a list of ABCSMC parameters.

Table 1. Table of ABCSMC Parameters

Parameter	Description	Value
T	Number of populations	30
B	Population size	100
ϵ_t	Set adaptively to lowest 30-percentile	
$\pi(\theta)$	Prior distributions for fitted parameters	Uniform
$K_t(\theta\|\theta^*)$	Transition kernel	$\mathcal{N}(0, \Sigma)$
$d(x, x_0)$	Distance function	L_1 distance

Given agent statistics $\mathbf{x}$ and animal statistics $\mathbf{x}_0$ in a joint space, we compute the ABC distance $d(\mathbf{x}, \mathbf{x}_0)$ using the a normalized L_1 distance function.

$$d(\mathbf{x}, \mathbf{x}_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{C^i(x^i)} \left| x^i - x_0^i \right| \tag{17}$$

where i indexes the statistics. $C^i(x^i)$ is a normalization constant that depends on the statistic and possibly the value x^i. Normalization is necessary because the statistics have different units and value ranges.

We normalize durations using a constant $C^i(x^i) = 4.0$ seconds. We normalize the transition times using a piece-wise linear function to prevent extremely small or large values from dominating the distance.

$$C^i(x^i) = \min(30, 10 + 0.8 \max(x^i - 5, 0)) \tag{18}$$

We also normalize the frequency ratio using a piece-wise linear function.

$$C^i(x^i) = \min(20, 2 + \frac{18}{19} \max(x^i - 1, 0)) \tag{19}$$

6 Acknowledgments

We are grateful to Chris Gagne, Mitsuko Watabe-Uchida, Vikki Neville, Mike Mendl, Elizabeth S. Paul, and Richard Gao for their helpful discussion and feedback. Funding was from the Max Planck Society and the Humboldt Foundation. PD is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764 and of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research. We thank the IT team from the Max Planck Institute for Biological Cybernetics for technical support.

References

Ainsworth, M. D. (1964). Patterns of attachment behavior shown by the infant in interaction with his mother. *Merrill-Palmer Quarterly of Behavior and Development*, 10(1):51–58.

Akiti, K., Tsutsui-Kimura, I., Xie, Y., Mathis, A., Markowitz, J. E., Anyoha, R., Datta, S. R., Mathis, M. W., Uchida, N., and Watabe-Uchida, M. (2022). Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron*, 110:3789–3804.e9.

Antonov, G. and Dayan, P. (2023). Exploring replay. *bioRxiv*, pages 2023–01.

Arsenian, J. M. (1943). Young children in an insecure situation. *The Journal of Abnormal and Social Psychology*, 38(2):225.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical finance*, 9(3):203–228.

Bellemare, M. G., Dabney, W., and Rowland, M. (2023). *Distributional Reinforcement Learning*. MIT Press. `http://www.distributional-rl.org`.

Bennett, D. and Niv, Y. (2020). Opening burton's clock: Psychiatric insights from computational cognitive models. *The Cognitive Neurosciences*, pages 439–450.

Bishop, S. J. and Gagne, C. (2018). Anxiety, depression, and decision making: A computational perspective. *Annual Review of Neuroscience*, 41(1):371–388. PMID: 29709209.

Bowlby, J. (1955). (b) the growth of independence in the young child. *Journal (Royal Society of Health)*, 76(9):587–591.

Bromberg-Martin, E. and Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, 63:119–26.

Brown, G. E. and Dreier, V. M. (2002). Predator inspection behaviour and attack cone avoidance in a characin fish: the effects of predator diet and prey experience. *Animal Behaviour*, 63(6):1175–1181.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. (2015). Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28.

Corey, D. T. (1978). The determinants of exploration and neophobia. *Neuroscience & Biobehavioral Reviews*, 2(4):235–253.

Coss, R. G. and Biardi, J. E. (1997). Individual variation in the antisnake behavior of california ground squirrels (spermophilus beecheyi). *Journal of Mammalogy*, 78(2):294–310.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature neuroscience*, 3(11):1218–1223.

Dayan, P., Roiser, J. P., and Viding, E. (2020). The first steps on long marches: The costs of active observation. In Savulescu, J., Roache, R., Davies, W., and Loebel, J. P., editors, *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine*. Oxford University Press.

Dayan, P. and Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, 25:5–22.

Dearden, R., Friedman, N., and Andre, D. (2013). Model-based bayesian exploration. *arXiv preprint arXiv:1301.6690*.

Duff, M. (2002a). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD thesis, University of Massachusetts Amherst.

Duff, M. O. (2002b). *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst.

Dugatkin, L. A. (1988). Do guppies play tit for tat during predator inspection visits? *Behavioral Ecology and Sociobiology*, 23:395–399.

Eccard, J. A., Liesenjohann, T., and Dammhahn, M. (2020). Among-individual differences in foraging modulate resource exploitation under perceived predation risk. *Oecologia*, 194:621–634.

Eldar, E., Rutledge, R. B., Dolan, R. J., and Niv, Y. (2016). Mood as representation of momentum. *Trends in cognitive sciences*, 20(1):15–24.

FitzGibbon, C. D. (1994). The costs and benefits of predator inspection behaviour in thomson's gazelles. *Behavioral Ecology and Sociobiology*, 34:139–148.

Gagne, C. and Dayan, P. (2022). Peril, prudence and planning as risk, avoidance and worry. *Journal of Mathematical Psychology*, 106:102617.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164.

Gmytrasiewicz, P. J. and Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., and Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593.

Greggor, A. L., Thornton, A., and Clayton, N. S. (2015). Neophobia is not only avoidance: improving neophobia tests by combining cognition and ecology. *Current Opinion in Behavioral Sciences*, 6:82–89. The integrative study of animal behavior.

Guez, A., Silver, D., and Dayan, P. (2013). Scalable and efficient bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883.

Huys, Q. J. M., Russek, E. M., Abitante, G., Kahnt, T., and Gollan, J. K. (2022). Components of behavioral activation therapy for depression engage specific reinforcement learning mechanisms in a pilot study. *Computational Psychiatry*.

Jaakkola, T. S. and Jordan, M. I. (1999). Variational probabilistic inference and the qmr-dt network. *Journal of artificial intelligence research*, 10:291–322.

Kacelnik, A. and Bateson, M. (1997). Risk-sensitivity: crossroads for theories of decision-making. *Trends in cognitive sciences*, 1(8):304–309.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

Kakade, S. and Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559.

Kemp, C. and Kaplan, G. (2011). Individual modulation of anti-predator responses in common marmosets. *International Journal of Comparative Psychology*, 24(1).

Keramati, M., Smittenaar, P., Dolan, R. J., and Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, 113(45):12868–12873.

Magurran, A. E. (1986). Predator inspection behaviour in minnow shoals: differences between populations and individuals. *Behavioral ecology and sociobiology*, 19:267–273.

Magurran, A. E. and Seghers, B. H. (1990). Population differences in predator recognition and attack cone avoidance in the guppy poecilia reticulata. *Animal Behaviour*, 40(3):443–452.

Mattar, M. and Daw, N. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21.

Mazza, V., Jacob, J., Dammhahn, M., Zaccaroni, M., and Eccard, J. A. (2019). Individual variation in cognitive style reflects foraging and anti-predator strategies in a small mammal. *Scientific Reports*, 9(1):10157.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., and Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191.

Menegas, W., Akiti, K., Amo, R., Uchida, N., and Watabe-Uchida, M. (2018). Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nature neuroscience*, 21(10):1421–1430.

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

Mobbs, D., Headley, D. B., Ding, W., and Dayan, P. (2020). Space, time, and fear: survival computations along defensive circuits. *Trends in cognitive sciences*, 24(3):228–241.

Ogasawara, T., Sogukpinar, F., Zhang, K., Feng, Y.-Y., Pai, J., Jezzini, A., and Monosov, I. E. (2022). A primate temporal cortex–zona incerta pathway for novelty seeking. *Nature Neuroscience*, 25(1):50–60.

Oudeyer, P.-Y. and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6.

Paulus, M. P. and Angela, J. Y. (2012). Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in cognitive sciences*, 16(9):476–483.

Radulescu, A. and Niv, Y. (2019). State representation in mental illness. *Current Opinion in Neurobiology*, 55:160–166. Machine Learning, Big Data, and Neuroscience.

Rigter, M., Lacerda, B., and Hawes, N. (2021). Risk-averse bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1142–1154.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Rusu, A. A., Colmenarejo, S. G., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., and Hadsell, R. (2016). Policy distillation.

Simon, P., Dupuis, R., and Costentin, J. (1994). Thigmotaxis as an index of anxiety in mice. influence of dopaminergic transmissions. *Behavioural Brain Research*, 61(1):59–64.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2017). Learning to reinforcement learn.

Weber, R. (1992). On the Gittins Index for Multiarmed Bandits. *The Annals of Applied Probability*, 2(4):1024 – 1033.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074.

Wiltschko, A. B., Tsukahara, T., Zeine, A., Anyoha, R., Gillis, W. F., Markowitz, J. E., Peterson, R. E., Katon, J., Johnson, M. J., and Datta, S. R. (2020). Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature Neuroscience*, 23(11):1433–1443.