PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Rajendra D, Gokhale CS. 2026 Optimizing play for learning risky behaviour. *Proc. R. Soc. B* **293**: 20253111.
https://doi.org/10.1098/rspb.2025.3111

Received: 11 December 2025
Accepted: 7 January 2026

Subject Category:
Behaviour

Subject Areas:
behaviour, computational biology, theoretical biology

Keywords:
reinforcement learning, development, protected environments, dangerous environments

Author for correspondence:
Chaitanya S. Gokhale
e-mail: chaitanya.gokhale@uni-wuerzburg.de

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8250011.

THE ROYAL SOCIETY PUBLISHING

Optimizing play for learning risky behaviour

Dharanish Rajendra and Chaitanya S. Gokhale

Chair for Computational and Theoretical Biology, Julius-Maximilians-Universität Würzburg, Würzburg, Bavaria 97074, Germany

DR, 0000-0002-1399-154X; CSG, 0000-0002-5749-3665

Animals adapt their behaviour to current environmental conditions to enhance survival and reproductive success. While long-term adaptation occurs through evolutionary processes acting on heritable variation, individuals can also adapt within their lifetime via learning. Learning is particularly advantageous in environments that are uncertain or fluctuate across a lifespan or a few generations. However, reliance on individual learning entails a critical risk: juveniles may begin life poorly adapted, requiring costly and hazardous exploration, especially for species hunting dangerous prey. We explore how early-life learning in protected environments, such as those buffered by parental care, can facilitate behavioural adaptation in riskier adult contexts. Using reinforcement learning, grounded in dopaminergic reward circuits, we model decision-making in a predator hunting both safe and dangerous prey. Our results show that juvenile experiences can generalize to distinct adult environments when sufficient structural similarity exists between them. This framework helps explain phenomena such as meerkats provisioning disabled prey for pups and the benefits of extended human childhood. Our findings demonstrate that structured play or safe exploration in early life can significantly enhance learning-based adaptation to dangerous environments.

1. Introduction

Learning and evolution represent fundamental yet distinct mechanisms of biological adaptation [1,2]. While evolution shapes traits across generations through genetic selection [3], learning enables behavioural flexibility within an individual's lifetime [4]. This dual adaptive system is especially advantageous in variable environments, where learning can compensate for the relatively slow pace of evolutionary change [5,6]. Learning also confers benefits in complex environments by allowing individuals to tailor behaviour to local contingencies [7]. However, both learning and the capacity for learning are costly [8–10]. Organisms that rely heavily on learning produce individuals with the potential to acquire adaptive behaviours but who are initially unadapted or unspecialized to their environment [11,12], possibly even maladapted. This creates an apparent paradox: individuals must survive a vulnerable, uninformed phase before they can learn how to behave adaptively.

This paradox is particularly acute in the context of risky behaviours, where mistakes can be lethal [8,13]. Predators that hunt dangerous prey exemplify such scenarios, a strategy observed across the animal kingdom [14]. Some examples are African lion hunting elephants [15], wolves hunting elk and bison [16,17] and python hunting porcupine [18]. Effective predation on dangerous prey typically requires precise and often complex behavioural repertoires, which are not innate but must be acquired through individual experience or social learning [19]. The predators must also assess the situation—the danger posed by the prey, the potential food benefit and their own capture probability and energy levels—before making the decision to hunt. Jumping spiders

exemplify this, as they hunt a variety of dangerous prey, learn prey-specific strategies through trial and error [20] and consider their own energy level before choosing prey to hunt [21,22]. Theory shows that, paradoxically, hunting dangerous prey can emerge as an evolutionarily stable strategy even when safer alternatives exist [23,24], provided that the benefits justify the risks; for example, when energy from safe prey alone is not sufficient to survive, or when background mortality is high and a high rate of energy gain is needed to reproduce before death. Literature on risk-taking also shows that risky behaviour can be adaptive depending on resource availability and the presence of unpredictable interruptions [25]. Yet, the mechanisms through which naive individuals acquire these high-stakes behaviours remain poorly understood.

Extended parental care, particularly prevalent in species with advanced learning capabilities [26,27], may resolve this paradox. Developmental stages, such as infancy and the juvenile phase, may provide critical windows for safe learning [28,29]. These can be supported by parental behaviours that buffer young from environmental hazards [30]. Many species provide parental care, creating a relatively protected environment during early development. In such contexts, juveniles can engage in learning and behavioural exploration with reduced risk. Meerkats provide a safe environment where juveniles can interact with otherwise dangerous scorpion prey, which are presented disabled or dead [19]. African lions live in groups, caring for their young and allowing them to learn critical skills, such as hunting [31]. In humans, long and highly protected childhoods can allow for periods of learning through active exploration without costs [32,33]. During the juvenile phase, learning adapts to the protected conditions. However, once parental protection ends, the juvenile transitions into a potentially hostile adult environment, where previously learnt behaviours may no longer suffice. In this scenario, the individual must adapt once more, unlearning and/or re-learning through experience in a dangerous, unprotected environment. However, this raises a key question: to what extent can behaviours learnt in protected environments generalize to real-world, high-risk contexts?

In this study, we investigate how early-life learning in protected settings influences the development of adaptive prey choice strategies in predators. Using a combination of dynamic programming (DP) [34,35] and reinforcement learning (RL) frameworks [36], we model the acquisition of hunting strategies targeting both safe and dangerous prey. Animals make decisions and perform actions while considering both long-term impacts and short-term benefits. Both methods support such decision processes. DP is used in determining optimal behaviour for various scenarios where complete knowledge is present. It has been successfully applied not only to prey choice and foraging [23] but also to a range of other ecological and evolutionary contexts [37,38], including the evolution of mimicry and signal learning [39]. While DP provides us with the optimal behaviour, it does not encapsulate a mechanism of how such optimality is attained. An additional framework is, therefore, needed to bridge this gap. RL models how animals learn and make decisions by accounting for how current actions influence future states and rewards, a factor often overlooked in theoretical models of learning or its evolution (but see [40–42] for exceptions). The RL framework employed in our model is grounded in well-established neurobiology. Midbrain dopamine neurons encode reward prediction errors, the difference between received and expected rewards, providing a signal that updates value estimates and guides behavioural adaptation [43,44], and is observed across rodents, primates and humans [45]. The temporal difference (TD) RL algorithm in our model closely parallels this neural mechanism, as agents update action values based on the discrepancy between actual and predicted outcomes [46,47]. RL methods have been applied previously to model animal behaviour and learning, notably by McNamara & Leimar [48,49]. For a broad comparison between RL and the more classical DP methods, see [50].

Specifically, we explore: (i) the conditions under which behavioural skills learnt in protected environments can transfer to more dangerous ones; (ii) how environmental danger levels shape learning efficiency and performance in adulthood; and (iii) the relationship between the properties of the external and protected environments that result in a good transfer of skills and performance.

2. Methods

(a) Environment structure

We investigate prey choice dynamics through a stochastic energy budget model of a predator encountering heterogeneous prey types [51].

(i) Predator

In the model, we track the energy level of the predator as it changes over time because of its actions. It varies between zero and a fixed maximum value, denoted as E_c. The goal of the predator is to reach E_c, which corresponds to a fitness benefit, and to avoid the lower energy limit (zero). The energy level zero and E_c denote the endpoints of a single 'episode' of the process, following which, a new episode starts with a randomly chosen energy level. The predator alters its energy level by interacting with the environment, specifically through prey choice decisions.

(ii) Prey types

There are two types of prey in the environment: safe and dangerous. The safe prey is a guaranteed catch and provides a small amount of net energy gain ΔE_{safe} after a handling time of Δt_{safe}. The capture of dangerous prey is not guaranteed upon attacking and occurs with probability ϕ. On a successful attack, the dangerous prey provides a large net energy gain ΔE_{dang}, after a corresponding handling time of Δt_{dang}. However, this energy benefit may have a high variance. Failure in capture (probability $1-\phi$)

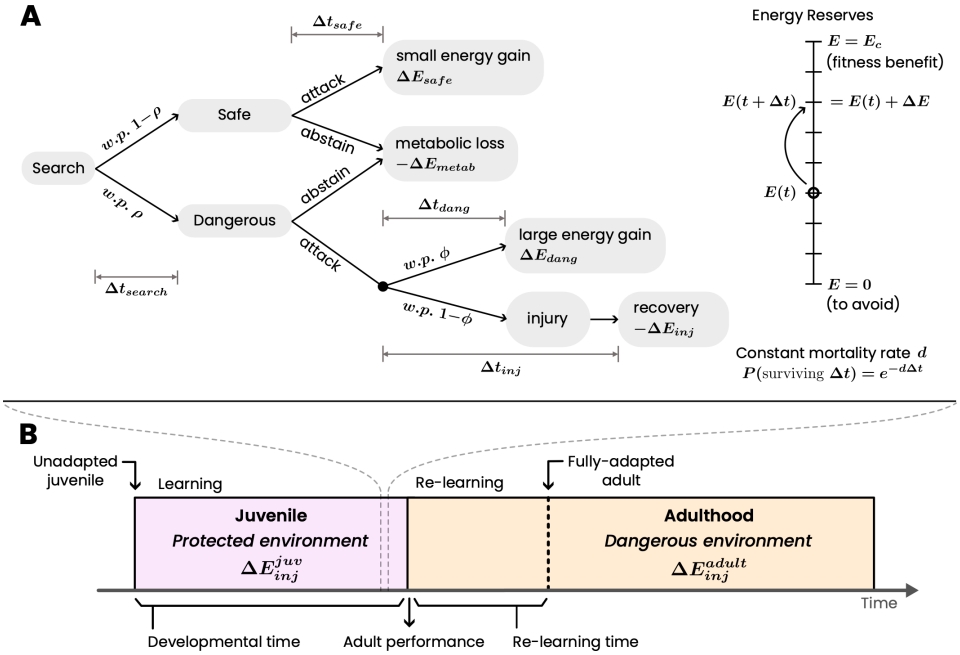

Figure 1. Depiction of the model processes. (A) Flow chart depicting the stochastic decision-making model of predator foraging behaviour and energy dynamics. At each step, the predator begins with the search. After a fixed search time Δt_{search}, it encounters dangerous prey with a probability ρ (availability of dangerous prey) or safe prey otherwise. The predator's decision of attacking the encountered prey or abstaining results in a corresponding outcome i with energy gains or losses ΔE_i and time investments Δt_i. Attacking a safe prey always results in success and a small energy gain. The predator succeeds probabilistically against a dangerous prey with a probability ϕ (capture probability). Upon failure, it gets injured and requires some time (Δt_{inj}) and some energy (ΔE_{inj}) from its reserves to recover. These outcomes change the predator's *energy reserves* ($E(t)$) stochastically, with a constant mortality rate (d) influencing survival probability. The goal is to reach $E = E_c$ to obtain a fitness benefit and avoid the lower limit $E = 0$. (B) Illustration of the two-phase learning model. A juvenile, initially unadapted, learns in a protected environment created by its parents for a duration of developmental time, which it adapts to. The juvenile environment is made protected by decreasing the energy cost of an injury from dangerous prey. As an adult, it is exposed to the truly dangerous elements of the environment. The performance in adulthood is dependent on the adult environment, the level of protection during the juvenile phase and the developmental time. The individual may need to undergo re-learning in order to become fully adapted to the (now changed) adult environment, a process that lasts for the duration of the re-learning time. The microscopic processes within each of the phases are the decision process described in (A).

occurs because of the retaliation of the prey against the predator, resulting in its injury. The predator expends some energy ΔE_{inj} from its reserves and some time Δt_{inj} to recover from the injury.

(iii) Decision-making process

We model prey choice dynamics as a semi-Markov decision process ([52]; figure 1A), where energy levels and prey type undergo discrete transitions separated by variable time intervals reflecting metabolic fluctuations, hunting outcomes and physiological variability. These transitions capture the differential risks of pursuing safe versus dangerous prey [53] and allow modelling of complex, nonlinear energy acquisition dynamics under uncertainty [54].

Each step consists of prey search, encounter, decision and outcome. After a fixed search time Δt_{search}, the predator encounters dangerous prey with probability ρ (dangerous prey availability) or safe prey with probability $1 - \rho$. The predator must be able to distinguish prey types, but need not assess the danger level directly. It then decides to attack or abstain. Abstaining incurs only the metabolic cost ΔE_{metab} from searching. Attacking yields stochastic energy and time changes, depending on the prey type and success probability. An episode ends when energy drops to zero, either by injury or metabolic costs, or, independently, through background mortality rate d. Survival probability over a time interval Δt is $s(\Delta t) = \exp(-d\Delta t)$. The transition times are fixed, while all the energies are random variables; their distributions and relative magnitudes between safe and dangerous prey determine optimal strategies. Parameter values and detailed descriptions are given in the electronic supplementary material, S1.

(iv) The policy

At each decision step, the predator can either attack the prey encountered or abstain. To simplify analysis, we will not directly consider the predator's individual actions of abstaining and attacking the two prey types, but rather consider the following three compound actions that the predator can choose between at each decision-making step:

(i) *indiscriminate attack*: attack the prey encountered regardless of its type;
(ii) *risk-prone attack*: attack the prey only if it is of the dangerous type; and
(iii) *risk-averse attack*: attack the prey encountered only if it is safe.

We must reiterate that these are not the identities of the agent, but rather behaviours that it can choose from at each step. Of critical importance is the fact that the decision taken by the predator can differ depending on its current energy level. Such an energy-dependent (in general, state-dependent) strategy is known as a *policy*. A policy π, formally, is a function that specifies the action

$A = \pi(E)$ that the predator takes when at energy level E, and describes its complete behaviour. This is equivalent to a behavioural strategy that adapts to the current energetic state and can vary across the predator's lifespan. For example, a predator can adopt a strategy that is indiscriminate at low energy levels and risk-prone at higher energy levels.

An example trajectory of this decision process is as follows. The predator is initially at energy level $E(t)$ at time t. It enacts its compound action determined by its policy $\pi(E(t)) = A_1$, which in this case is a risk-prone attack. After searching, it encounters a dangerous prey, attacks and succeeds. It obtains a net energy gain of ΔE_{dang} after a time of $\Delta t_{search} + \Delta t_{dang}$ and transitions to the energy level $E(t + \Delta t_{search} + \Delta t_{dang}) = E(t) + \Delta E_{dang} = E(t_2)$. At the next step, it enacts the policy determined at its current energy level $\pi(E(t_2)) = A_2$, which, in general, is different from A_1. In the example, A_2 = indiscriminate attack. After searching, it encounters a safe prey, attacks and succeeds. It has now gained sufficient energy to reach E_c. This is the goal state of the predator, and the episode is considered complete, indicating that the predator was successful this time. A new episode starts with the predator's energy initialized randomly.

(b) Dynamic programming

For each combination of environmental parameters, there exists an optimal policy, denoted π^*, that provides the highest fitness benefit to the predator. This optimal policy is determined using DP, specifically the value iteration method [34,35], which is well-suited to solving such sequential decision problems.

When the predator reaches an energy level of E_c, we consider it to have obtained a reward of 1. This reward makes the highest energy level directly valuable. No other state provides a direct reward. Therefore, all other energy levels acquire an indirect value through their probability of reaching E_c before death. Such a value is quantified by the *value function* $V(E)$, which is defined as the expected reproductive pay-off when starting from energy level E. An optimal policy, π^*, maximizes the value function $V(E)$, and hence the expected reproductive pay-off, for each energy level (see the electronic supplementary material, S2, for a detailed description). This approach is used to calculate the optimal policy and value functions, which serve as a baseline, and determine when the learning process has reached completion.

(c) Reinforcement learning

We use TD RL to model how a predator learns about its environment and develops an optimal foraging policy [36]. TD learning is a class of algorithms that updates estimates of future reward by comparing successive predictions and capturing how the difference in value between states informs learning.

An important note is that in such a model, the learning agent learns how to behave in the environment *while* interacting with the environment. In fact, it directly uses its experience from these interactions to inform its future decisions. The learning agent estimates the expected value (total expected reward) of each action at each energy level. These estimates are updated based on actual experiences during interaction with the environment.

Specifically, updates occur when outcomes deviate from expectations: if a chosen action yields a higher reward than previously estimated, the agent increases the corresponding action-value proportionally to this discrepancy. The magnitude of this change is governed by the learning rate α. This can also be thought of as the rate at which the individual absorbs new information. The agent's behaviour is guided by an ϵ-greedy policy: most of the time, it selects the action with the highest estimated value at its current energy level, but with a small probability ϵ, it explores other actions at random. Such a policy allows for exploring different (previously unknown) possibilities. As learning proceeds, the policy stochastically approaches the optimal one. Provided the learning rate is sufficiently low or gradually decays, convergence to the optimal policy is guaranteed in the long run. The exact details of the learning algorithm are outlined in the electronic supplementary material, S3.

At the start of an *experiment*, the agent is initialized at a random energy level with a random value function (and therefore policy). That is, the value of taking each action at each energy level is given a random number between 0 and 1. The corresponding (ϵ-greedy) policy has an equal chance of being any of the three actions at every energy level. The learning process continues until it reaches one of the boundaries ($E = 0$ or $E = E_c$) and an *episode* is completed. Then, a new episode starts with a random initial energy, while retaining the learning from the previous episodes. The process continues until a fixed time or number of episodes is completed, or it reaches a learning target. At this point, one experimental run is considered complete. Multiple (10 000) independent experimental runs are performed for each parameter value to obtain the mean and variance for the metrics of interest.

(d) Two-phase model

Our model captures the developmental transition of an individual by dividing the lifespan into two distinct stages: a juvenile phase and an adult phase (figure 1B). We term the duration of the juvenile phase as the developmental time. In both the juvenile and adult phases, the individual faces the prey-choice problem, aiming to reach the maximum energy level. In the adult phase, reaching the maximum energy level corresponds to a fitness benefit. However, in the juvenile phase, it learns the policy to reach the maximum energy level, which it can apply in its adult phase. The (parameters of the) environment experienced by the animal is constant within each phase but can be potentially different between phases. The key feature of this set-up is that the juvenile phase provides an opportunity for learning in a more protected context, while the adult phase presents the true conditions under which survival and reproduction depend. While there can be different protection strategies during the juvenile stage, we quantify this by comparing the cost of injury upon failure to capture a dangerous prey between the adult phase and the juvenile phase.

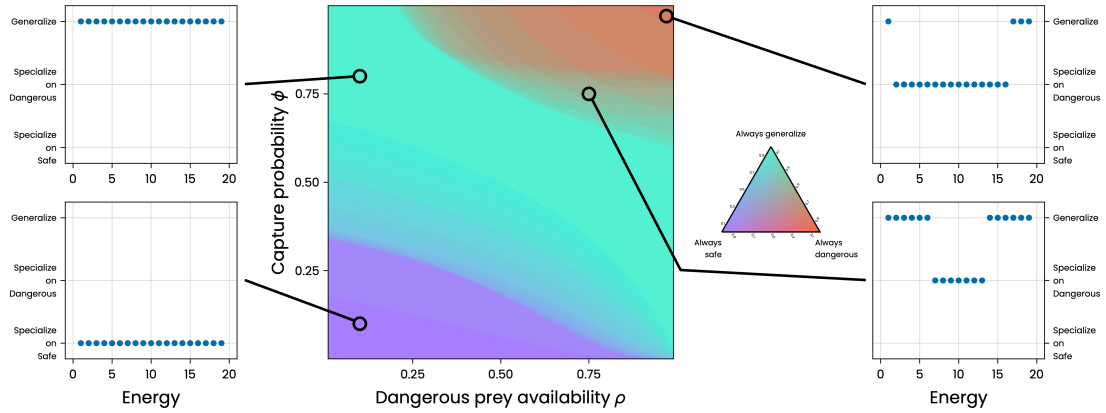

Figure 2. Optimal policy. A representation of the optimal policy for a range of parameter values and example policies for different regions. The colour of the points in the heatmap depicts the composition of the three compound actions of the policy.

That is, the protection level $\psi = \Delta E_{inj}^{adult} - \Delta E_{inj}^{juv}$. The cost of injury in the adult environment ΔE_{inj}^{adult} is kept constant, while that in the juvenile phase ΔE_{inj}^{juv} is reduced to increase the level of protection. The other two parameters, dangerous prey availability ρ and capture probability ϕ, are kept constant across the two phases. However, the parameters of ρ, ϕ, ψ and developmental time vary across experiments. These parameters enable us to investigate how various qualities of the early and adult environment influence subsequent adaptation.

An individual begins its life in the juvenile phase, with a randomly chosen value function and, hence, a corresponding random policy. While experiencing the protected juvenile environment, it learns to adapt to it, i.e. learn the optimal policy of the juvenile environment. The maximum duration of the juvenile phase is fixed at the developmental time. If the juvenile learns the optimal policy and value of the juvenile environment (determined using DP; §2b) before the entire developmental time has elapsed, then the juvenile phase ends early. Following the juvenile phase, the adult phase begins. Here, the individual experiences the unprotected adult environment. The policy it has learnt in the juvenile phase may not be optimal in the adult phase. Thus, it may need to re-learn in the adult phase. We do not fix the duration of the adult phase, but let the (re-)learning process continue until it has learnt the optimal policy and value of the adult environment (determined using DP). With that, a single experiment of the two-phase model is completed. Several (10 000) experiments are performed to obtain average values of the metrics described below.

(e) Metrics of learning

To assess performance after learning in a protected environment, we use two primary metrics: relative adult performance and relative learning speed.

(i) Learning time

In a constant environment, the learning agent converges to the optimal policy and value (electronic supplementary material, figure S1A; compare with figure 2). Learning time is the time required to reach optimality (within a margin of error) from a randomly initialized value and policy. It varies stochastically across realizations and depends on environmental parameters (electronic supplementary material, figure S1B) and learning hyperparameters. See the electronic supplementary material, S3.2, for more details.

(ii) Scaled developmental time

To compare developmental time across parameter values, we scale it by dividing by the learning time required in the corresponding adult environment, yielding scaled developmental time T_{dev}. Thus, $T_{dev} = 0$ indicates no juvenile learning, while $T_{dev} = 1$ means the juvenile phase duration equals the time needed to learn the optimal policy in an unprotected environment.

(iii) Adult performance and relative adult performance

Following the developmental time, we assess the expected reward in a single episode of the adult environment, calculated as the average reward over 10 000 episodes. This expected reward is then normalized to lie between two benchmarks—the reward of a random policy and the reward of the optimal policy—to obtain the adult performance $P(T_{dev}, \psi)$, which depends on the developmental time and protection level.

To compare protected with unprotected learning, we define relative adult performance as:

$$\mathcal{P}(T_{dev}, \psi) = P(T_{dev}, \psi) - P(T_{dev}, \psi = 0). \tag{2.1}$$

We measure this at two developmental time points: after a short juvenile phase ($T_{dev} = 0.1$), capturing early learning benefits, and after an extended phase ($T_{dev} = 2$) when performance typically saturates.

(iv) Re-learning time and relative learning speed

Following a protected juvenile phase, individuals may require re-adaptation in the adult environment. Re-learning time $\mathcal{T}_{re}(T_{dev}, \psi)$ measures the time required to learn the optimal adult policy and value after experiencing protection level ψ for duration T_{dev} (figure 1B). This decays approximately exponentially: $\mathcal{T}_{re}(T_{dev}, \psi) \sim \exp(-v_\psi T_{dev})$. The exponential factor v_ψ is the learning speed for protection level ψ, and the relative learning speed (comparing protected with unprotected) is:

$$\mathcal{V}_\psi = \log_{10}\left(\frac{v_\psi}{v_{\psi=0}}\right). \tag{2.2}$$

(v) Benefit of a protected juvenile environment

A protected environment benefits learning through: (i) high initial relative adult performance ($\mathcal{P}(T_{dev} = 0.1, \psi) > 0.02$); (ii) high maximum relative adult performance ($\mathcal{P}(T_{dev} = 2, \psi) \geq -0.02$); or (iii) positive relative learning speed ($\mathcal{V}_\psi > 0$). These thresholds exclude noise while capturing biologically meaningful differences in performance.

The model was implemented in Julia [55], with code and data available on Zenodo (https://doi.org/10.5281/zenodo.18345461).

3. Results

We find that learning in protected environments can significantly enhance an individual's ability to adapt to later, riskier environments, especially when there is structural similarity between the two. For certain environmental conditions, simply providing a safe and protected environment during the juvenile phase is sufficient to promote learning. For other environmental conditions, a more involved strategy of providing a highly protected environment for the initial part of the juvenile phase and then transitioning to a less protected environment increases the performance and adaptation of the individual in adulthood. There exists an optimal duration of juvenile learning that maximizes adult reproductive success, balancing the trade-off between early safety and delayed experience.

(a) Optimal foraging policies across risk profiles

We compute the optimal foraging policy over a range of environmental and individual parameters, specifically dangerous prey availability (ρ) and predator capture probability (ϕ) (figure 2). The optimal policy is a vector of decisions across energy levels from 1 to E_c (we use $E_c = 100$), where each entry specifies the best action at that energy state. For visualization, we transform this policy vector into a tuple: (in, rp, ra), denoting the number of energy levels where the predator should attack indiscriminately, be risk-prone or be risk-averse, respectively. This tuple is then mapped to a three-component colour and displayed as a heatmap in figure 2.

The resulting phase diagram reveals three broad behavioural regimes. In the bottom-left region of the heatmap, with low ϕ and intermediate ρ, the optimal policy is to exclusively specialize on safe prey across all energy levels. This reflects the high risk and low success rate of attacking dangerous prey under these conditions. By contrast, the top-right region, with high ϕ and high ρ, favours a strategy of specializing on dangerous prey for most energy levels. The predator's high skill and the availability of high-energy prey make it optimal to fully rely on the risky but rewarding option. This is true, however, only for very high ϕ. For slightly lower ϕ and ρ, it is a mixture of specializing on dangerous prey and being indiscriminate. In the top-left and bottom-right corners, where one of the parameters is low and the other is high, it is optimal to be indiscriminate regardless of energy level. In general, between these extremes lies a region of mixed policies. Here, the policy is not homogeneous: some energy levels favour being indiscriminate, while others favour specialization.

This pattern can be understood by considering the trade-off between the risk of starvation and reproductive gain. When both ϕ and ρ are both low, the costs of attacking dangerous prey are greater than the potential benefits, and it is optimal to ignore them entirely. As ϕ increases from here, incorporating dangerous prey at intermediate energy levels becomes beneficial, as it provides a chance to rapidly increase energy without a significant risk of starvation. However, attacking a dangerous prey at low energy levels is still not optimal, as it can result in death from injury. As ϕ increases further, including dangerous prey in the diet at all energy levels is beneficial, as the risk of fatal injuries at low energy levels has reduced. At high ϕ, beyond a certain threshold of ρ, it becomes beneficial to ignore safe prey altogether. In this region, abstaining from safe prey and waiting until a dangerous prey is encountered provides a higher rate of energy gain. This then transitions to complete dependence on dangerous prey, characterized by a very high ϕ and ρ. When ρ is high and ϕ is low, it is beneficial to attack any prey encountered, as waiting for safe prey may result in the predator not obtaining any food and starving to death.

These are the optimal behaviours that a predator must perform in different environmental conditions to maximize its reproductive pay-off. We assume that these prey choice behaviours are not genetically determined but learnt through trial and error throughout an individual's life. In the next section, we describe how juvenile play can benefit learning in this scenario.

(b) Adult performance

Once the individual transitions from the juvenile to the adult phase, their performance in the adult environment becomes critical for reproductive success. Generally, individuals who spend more time learning during the juvenile phase tend to achieve

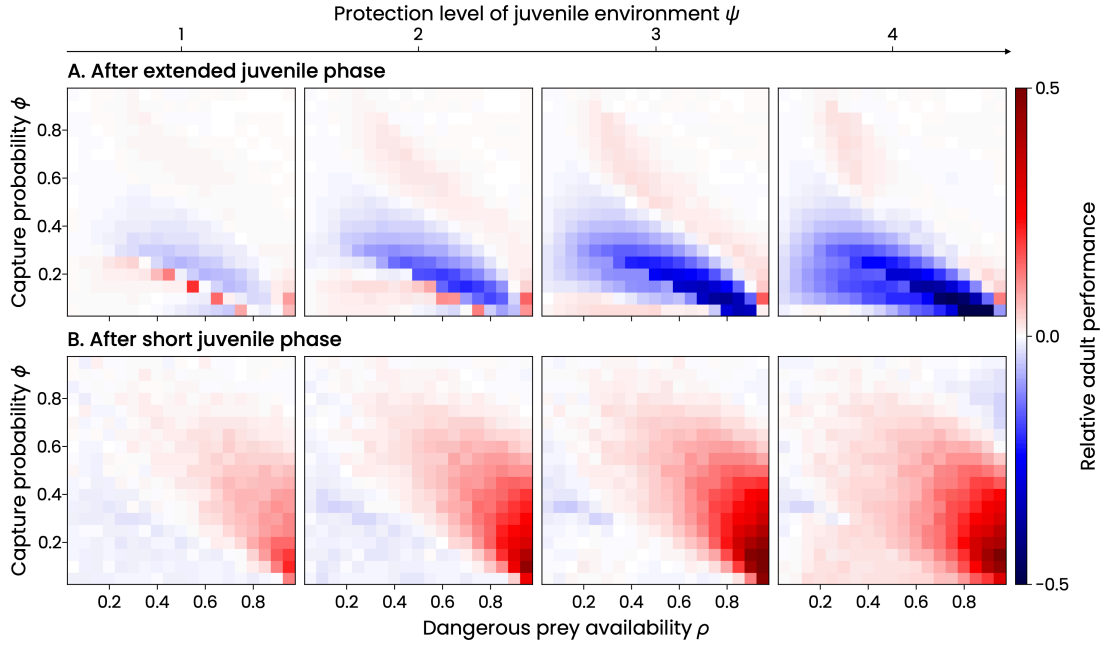

Figure 3. Adult performance after different developmental times. Adult performance is shown as a function of dangerous prey availability (ρ), capture probability (ϕ) and juvenile environment protection level (ψ), measured by the reduction in injury cost from the dangerous to the protected environment $\Delta E_{inj}^{juv} - \Delta E_{inj}^{adult}$. The corresponding adult environment for each panel has the danger level of $\Delta E_{inj}^{adult} = 4$. (A) Adult performance is measured after an extended juvenile phase, i.e. a long developmental time ($T_{dev} = 2$). In the upper right triangle, adult performance approaches optimal performance (white/clear regions) for all ψ. In most other areas, adult performance falls short of optimal and worsens progressively as the juvenile environment becomes increasingly protected. (B) Relative adult performance after a short juvenile phase with a developmental time of $T_{dev} = 0.1$. Positive values (red) indicate that a short juvenile phase in a protected environment leads to higher adult performance compared to a similar juvenile phase in a dangerous environment. Negative values (blue) indicate the opposite. White regions indicate similar adult performance between protected and unprotected juvenile environments. There is a sizeable region of parameter space where a protected environment leads to a higher adult performance for short juvenile durations. This region expands and becomes more prominent as the protection level of the juvenile environment increases.

higher adult performance. The adult performance first increases rapidly with developmental time and then saturates. However, the exact outcome strongly depends on the match or mismatch between the juvenile and adult environments (see the electronic supplementary material, figure S2). When the juvenile environment is identical to the adult environment (i.e. both are dangerous), performance improves steadily with developmental time and can approach optimality. By contrast, when the juvenile environment is a more protected proxy for the adult environment, the developmental trajectory of performance can differ substantially.

The adult performance for long developmental times may or may not be close to the optimal performance for the adult environment. This depends on several factors, including the availability of dangerous prey (ρ), capture probability (ϕ) and the protection level (ψ) of the juvenile environment (difference in injury cost ΔE_{inj}). As shown in figure 3A, there is a region in the parameter space in which the maximum performance from the protected juvenile environment is significantly worse than the maximum performance from the unprotected juvenile environment. This discrepancy increases as the juvenile environment becomes more protected. Nevertheless, this outcome is not universal. A substantial region of parameter space exists in which protected juvenile environments can support learning trajectories that eventually reach performance levels comparable to those developed in dangerous environments. This observation highlights that not all mismatches lead to maladaptive outcomes, particularly when adult challenges are not too dissimilar from juvenile experiences.

Nevertheless, the difference between adult performance in a protected juvenile environment and in a dangerous juvenile environment is not the same over developmental time (see the electronic supplementary material, S4). For adults coming from a protected environment, the performance can increase more rapidly than those from a dangerous juvenile environment for short developmental periods, but it saturates at a lower level (see the electronic supplementary material, figure S4, showing the relative adult performance against a range of developmental times and environmental parameter values). For low developmental times (figure 3B), there is a large region of parameter space for all ψ, for which the adult performance is better than that from the dangerous juvenile environment. The size of this region is larger for more protected juvenile environments and decreases as the developmental time increases. As developmental time increases, a different region becomes more prominent, where the adult performance following a protected juvenile phase is worse than that following an unprotected juvenile phase. Even for long developmental periods, this region persists. This means that for these parameter values, the adult performance saturates significantly below the optimal performance (see the electronic supplementary material, figure S2C).

(c) Relative learning speed

As the time spent in the juvenile phase increases, the time required to learn in the adult phase decreases (see the electronic supplementary material, figure S3). If the environment in the juvenile phase is the same as the adult, then the decrease in re-learning

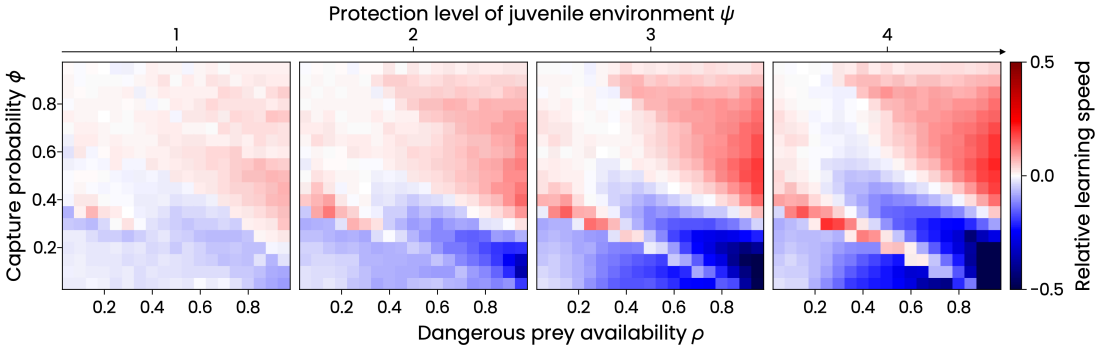

Figure 4. Relative learning speed across juvenile environment protection levels. Re-learning time as a function of developmental time can be approximated and fitted with an exponential decay for developmental times of less than 1. The ratio of the decay exponents of a protected and dangerous juvenile environment gives the relative learning speed in the protected environment. Relative learning speed is shown as a function of dangerous prey availability (ρ), capture probability (ϕ) and juvenile environment protection level (ψ). The adult phase has an injury cost of $\Delta E^{inj}_{adult} = 4$. Positive values (red) indicate that a juvenile phase in a protected environment needs a shorter re-learning time compared to a juvenile phase in a dangerous environment of the same developmental length. Negative values (blue) indicate that a protected juvenile phase results in a longer re-learning time than an unprotected juvenile phase of a similar duration. The red and blue regions are roughly the same size and remain so even as the juvenile environment becomes more protected. However, they do become more prominent as ψ increases. That is, for some parameters, the learning is highly accelerated, while for others, it becomes highly decelerated.

time with an increase in developmental time is nearly linear, with a deviation at high developmental times owing to stochasticity. If, instead, the juvenile environment is different from the adult environment, then this curve is, in general, different. For some parameter values, the protected juvenile environment may accelerate re-learning and make this curve fall more rapidly than before (electronic supplementary material, figure S3B). For other parameter values, the re-learning time from a protected juvenile environment may still be very high even for long developmental times (electronic supplementary material, figure S3C).

The second metric we use to quantify learning is relative learning speed. It is the speed of re-learning following an unprotected juvenile phase, relative to that following a protected juvenile phase. We plot the relative learning speed value for a range of parameter values and protection level (ψ) of the juvenile environment in figure 4. For high capture probability (ϕ) and high availability of dangerous prey (ρ), the learning speed in a protected juvenile environment is higher than that in a dangerous environment. For low ϕ, the learning speed in a protected juvenile environment is, in general, lower than that in a dangerous juvenile environment. Even for low ϕ, there is a small region in parameter space with a high relative learning speed. These differences are only exaggerated as the juvenile environment becomes more protected.

The relative learning speed primarily provides information about what happens during short developmental times. However, the curve of learning time may deviate from exponential decay at higher developmental times. Thus, we plot the learning time for a range of developmental times in the electronic supplementary material, figure S5. Here, a positive value indicates that the protected juvenile environment has a longer learning time than the dangerous juvenile environment, i.e. learning is slower in the protected juvenile environment compared to the dangerous juvenile environment. We observe that the difference between protected and dangerous juvenile environments increases as developmental time progresses and the juvenile environment becomes more protected. There is a large region in the parameter space at all ψ, where the learning time is still very high, even for long developmental times.

(d) Benefit of a protected juvenile environment

In this section, we examine the overall benefits of a protected juvenile environment on learning. Learning can be benefited from in the form of high relative adult performance or high relative learning speed. The benefit to adult performance can come after a short juvenile phase or after an extended juvenile phase. Depending on the environmental parameters and the protection level (ψ) during the juvenile phase, the type and level of benefit differ. This is shown in figure 5. We claim that a protected juvenile environment offers one of these benefits, and the value of that metric for a protected juvenile environment is higher than that for a dangerous juvenile environment. For the metric of adult performance after an extended juvenile phase, we consider a benefit even if the protected environment matches the metric of the dangerous environment.

For a single environmental condition, there can be one or more of these benefits. As the parameters of the environment and protection change, the kind and extent of the benefit change in highly nonlinear ways. This is because these benefits depend on the optimal policies and values of the dangerous and protected environments, their properties and their extent of match or mismatch. Since these quantities vary in highly nonlinear ways, so do the metrics of learning benefits.

However, there are some general patterns. The region with no learning benefit first increases in size and then decreases as ψ is increased. For a highly protected environment, this region is very small. That is, there is some benefit to learning in all environmental conditions. There is a sizeable region where only the adult performance for an extended juvenile phase is beneficial. This region shrinks in size as ψ increases. The regions of no benefit and only maximum performance benefit are replaced by the region with only an early performance benefit. For higher ψ, a small region with a benefit in both early and maximum adult performance becomes prominent for a slightly higher capture probability than the previously mentioned regions. For high capture probability,

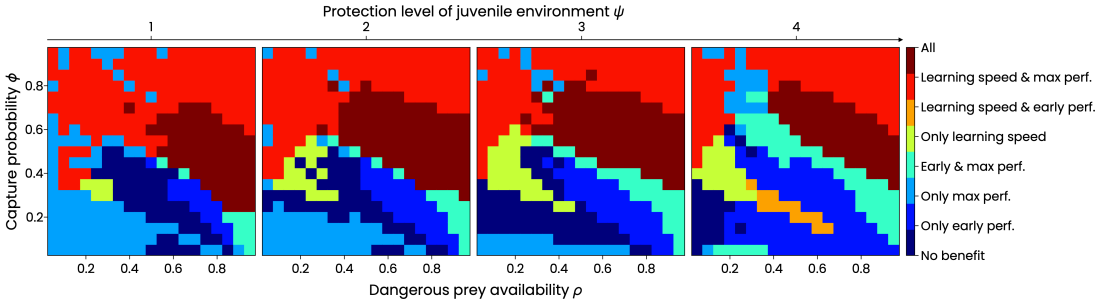

Figure 5. Benefit of protected juvenile environment. The benefit of a protected juvenile environment can be either in terms of a high (i) adult performance after a short juvenile phase (early perf.), (ii) adult performance after an extended juvenile phase (max perf.), or (iii) learning speed, or a combination of these. The type and extent of benefit changes across environmental conditions, including the availability of dangerous prey (ρ), capture probability (ϕ) and the protection level of the juvenile environment (ψ).

there are two regions, one in which learning is benefited in the form of a high learning speed and high maximum performance, and the other in which learning is benefited in all manners. These regions are roughly the same size for all ψ.

4. Discussion and conclusion

Our results demonstrate that learning risky behaviours in protected environments can significantly enhance adaptation through learning, depending on the structural similarity between developmental and adult contexts. Protected environments confer a benefit towards adaptation through learning in three distinct (but sometimes overlapping) ways. For a large fraction of parameter values (high ρ and ϕ), provision of a sufficiently long, highly protected juvenile environment enables individuals to attain near-optimal adult performance (figure 3A). For another region of parameter space (high ρ and low ϕ), protection allows rapid adaptation and increase in performance, even faster than in a juvenile environment that exactly matches the adult environment (figure 3B). In a third region of parameter space (two disjoint regions with high ρ and ϕ, and intermediate ρ and ϕ), protected learning may not provide full adaptation but accelerates the process of re-learning in adulthood (figure 4). The metrics of learning and the benefits of protection depend on the optimal value, policy and transition functions of the two phases and their relationship to each other [56].

Thus, structural similarity between developmental and adult contexts is essential for such benefits to arise. This environmental matching principle helps explain puzzling aspects of parental care across taxa. A strong match in environments confers near-optimal adaptation to the adult environment post-development. In such a scenario, the parental care strategy of providing a highly protected environment for the entire duration of a (long) juvenile phase is sufficient to produce well-adapted adults. When the match is not strong, complete protection yields incomplete adaptation. A more nuanced parental care strategy is required. Provided there is a benefit of rapid performance increase or accelerated re-learning, a gradual exposure to danger would allow for rapid performance increase until optimality.

Meerkat teaching behaviour exemplifies this: helpers initially provision dead scorpions to young pups, then progressively provide mobile prey as pups develop, ultimately presenting fully functional, deadly scorpions to older juveniles [19,57]. Our model reveals that such graduated exposure is not merely cautious parenting, but an adaptive strategy that balances safety and prepares individuals for the challenges of adulthood. Play behaviour in mammals has long been hypothesized to serve as preparation for challenges in adulthood, allowing juveniles to develop flexible behavioural responses in low-risk contexts [58,59]. Lions provide communal crèches for play and observational learning [31,60]. Humans maintain extraordinarily extended childhoods [27], and children engage in extensive, diverse play [32,61]. Protected environments are also, in general, widespread among species that face significant risks during foraging or hunting [62].

Our framework generates testable comparative predictions. The environment, its danger and complexity, along with the constraints of the developmental process, drives selection on optimal parenting and learning strategies. This may involve long developmental periods, gradual parental provisioning or both, in order to adapt to a dangerous environment. Comparative analyses within clades showing variation in dangerous prey specialization would be particularly valuable. For example, within Salticidae, do dangerous prey specialists (e.g. *Portia* hunting web-building spiders [20]) show extended learning periods compared to safe prey specialists? Within social carnivores, do species hunting dangerous megafauna (wolves hunting bison [17]) show different developmental strategies than those hunting smaller, safer prey? Testing these requires measuring dangerous prey availability from dietary analyses, age-specific capture probability from hunting success rates [16,63] and injury costs from wounding rates and post-attack mortality. Human ancestral environments probably included hunting dangerous megafauna [26], processing dangerous resources and navigating complex social hierarchies with severe costs of error. Under such conditions, our model predicts that selection should favour extended protected periods, allowing for exploration, graduated exposure to realistic risks and cognitive mechanisms that support cross-context transfer. Teaching, active facilitation of learning at cost to the teacher [57], may evolve to address environmental mismatch, with teaching content compensating for the specific mismatch.

Our RL framework is grounded in the neurobiology of midbrain dopamine neurons encoding reward prediction errors across taxa [44–47]. These rewards are not obtained from an external source, but rather through an internal mechanism that provides reward signals based on external stimuli [64–66]. Evolution generates an appropriate reward mechanism that signals rewards for stimuli leading to a higher expected fitness [67,68]. If protected environments accelerate learning through enhanced exploration,

juveniles should show higher dopaminergic responsivity to novel encounters and more rapid policy updating than adults learning equivalent tasks. The evolution of sensitive periods [29,69] may represent developmental specialization for protected learning when juvenile environments reliably predict adult conditions.

The learning process can further evolve to account for the differences between juvenile and adult environments. There can be an efficient transfer of knowledge from the juvenile environment to the adult phase (transfer learning [70]). While in a protected juvenile environment, the individual could directly learn the optimal policy of the adult environment (off-policy RL [36]).

Several extensions would strengthen the model's biological realism and generalizability. Incorporating skill learning (dynamic ϕ) alongside prey choice would reveal whether protected practice on disabled prey accelerates skill acquisition or creates false confidence and also allows the analysis of more teaching strategies. Exploring diverse protection mechanisms—physical (den sites), social (group defence), temporal (restricting foraging) and dietary (provisioning)—may reveal differential transfer properties. Many species with extended parental care also show pronounced social learning [71–73]. Investigating how social learning interacts with protected development could clarify whether social transmission reduces the need for graduated exposure or creates challenges when juveniles learn outdated strategies [74,75].

Protected juvenile environments represent a widespread solution to a fundamental challenge: how naive individuals learn dangerous behaviours without lethal errors. This solution works through environmental matching; protected contexts benefit learning when they preserve the structure of adult decision problems despite reducing consequences. When matching is good, simple protection suffices; when it is poor, graduated exposure becomes essential. This framework unifies observations about extended parental care, teaching, developmental timing and play across taxa, while generating testable predictions about which species should show longer development and more complex parental strategies.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. All simulation code, datasets and plotting scripts are available at Zenodo [76].

Supplementary material is available online [77].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. D.R.: conceptualization, formal analysis, investigation, methodology, validation, visualization, writing—original draft, writing—review and editing; C.S.G.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, visualization, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. Funding support from Julius-Maximilians-University Würzburg is gratefully acknowledged.

Acknowledgements. We thank the reviewers for their constructive feedback on the manuscript. The authors thank members of the T-Eco-Evo group for inspiring discussions.

References

1. Dewitt TJ, Sih A, Wilson DS. 1998 Costs and limits of phenotypic plasticity. *Trends Ecol. Evol*. **13**, 77–81. (doi:10.1016/s0169-5347(97)01274-3)
2. Pigliucci M, Murren CJ, Schlichting CD. 2006 Phenotypic plasticity and evolution by genetic assimilation. *J. Exp. Biol.* **209**, 2362–2367. (doi:10.1242/jeb.02070)
3. Futuyma DJ, Kirkpatrick M. 2017 *Evolution*, 4th edn. Sunderland, MA: Sinauer Associates.
4. Shettleworth SJ. 2009 *Cognition, evolution, and behavior*, 2nd edn. Oxford, UK: Oxford University Press. (doi:10.1093/oso/9780195319842.001.0001)
5. Krebs JR, Davies NB. 1981 *An introduction to behavioural ecology*. London, UK: Blackwell Scientific Publications.
6. Dukas R. 2004 Evolutionary biology of animal cognition. *Annu. Rev. Ecol. Evol. Syst*. **35**, 347–374. (doi:10.1146/annurev.ecolsys.35.112202.130152)
7. Dridi S, Lehmann L. 2016 Environmental complexity favors the evolution of learning. *Behav. Ecol*. **27**, 842–850. (doi:10.1093/beheco/arv184)
8. Johnston TD. 1982 Selective costs and benefits in the evolution of learning. In *Advances in the study of behavior, vol. 12* (eds JS Rosenblatt, RA Hinde, C Beer, MC Busnel), pp. 65–106. New York, NY: Academic Press. (doi:10.1016/S0065-3454(08)60046-7)
9. Mery F, Kawecki TJ. 2003 A fitness cost of learning ability in *Drosophila melanogaster*. *Proc. R. Soc. Lond. B* **270**, 2465–2469. (doi:10.1098/rspb.2003.2548)
10. Mery F, Kawecki TJ. 2004 An operating cost of learning in *Drosophila melanogaster*. *Anim. Behav*. **68**, 589–598. (doi:10.1016/j.anbehav.2003.12.005)
11. Snell-Rood EC. 2013 An overview of the evolutionary causes and consequences of behavioural plasticity. *Anim. Behav*. **85**, 1004–1011. (doi:10.1016/j.anbehav.2012.12.031)
12. Frankenhuis WE, Panchanathan K. 2011 Balancing sampling and specialization: an adaptationist model of incremental development. *Proc. R. Soc. B* **278**, 3558–3565. (doi:10.1098/rspb.2011.0055)
13. Mayley G. 1996 Landscapes, learning costs, and genetic assimilation. *Evol. Comput*. **4**, 213–234. (doi:10.1162/evco.1996.4.3.213)
14. Mukherjee S, Heithaus MR. 2013 Dangerous prey and daring predators: a review. *Biol. Rev.* **88**, 550–563. (doi:10.1111/brv.12014)
15. Loveridge AJ, Hunt JE, Murindagomo F, Macdonald DW. 2006 Influence of drought on predation of elephant (*Loxodonta africana*) calves by lions (*Panthera leo*) in an African wooded savannah. *J. Zool*. **270**, 523–530. (doi:10.1111/j.1469-7998.2006.00181.x)
16. MacNulty DR, Smith DW, Mech LD, Eberly LE. 2009 Body size and predatory performance in wolves: is bigger better? *J. Anim. Ecol*. **78**, 532–539. (doi:10.1111/j.1365-2656.2008.01517.x)
17. MacNulty DR, Tallian A, Stahler DR, Smith DW. 2014 Influence of group size on the success of wolves hunting bison. *PLoS ONE* **9**, e112884. (doi:10.1371/journal.pone.0112884)
18. Shine R, Harlow PS, Keogh JS, Boeadi. 1998 The influence of sex and body size on food habits of a giant tropical snake, *Python reticulatus*. *Funct. Ecol.* **12**, 248–258. (doi:10.1046/j.1365-2435.1998.00179.x)
19. Thornton A, McAuliffe K. 2006 Teaching in wild meerkats. *Science* **313**, 227–229. (doi:10.1126/science.1128727)
20. Jackson RR, Pollard SD. 1996 Predatory behavior of jumping spiders. *Annu. Rev. Entomol.* **41**, 287–308. (doi:10.1146/annurev.en.41.010196.001443)
21. Li D, Jackson RR, Barrion A. 1997 Prey preferences of *Portia labiata, P. africana,* and *P. schultzi,* araneophagic jumping spiders (Araneae: Salticidae) from the Philippines, Sri Lanka, Kenya, and Uganda. *N. Z. J. Zool*. **24**, 333–349. (doi:10.1080/03014223.1997.9518129)
22. Li D. 2000 Prey preferences of *Phaeacius malayensis*, a spartaeine jumping spider (Araneae: Salticidae) from Singapore. *Can. J. Zool*. **78**, 2218–2226. (doi:10.1139/z00-176)

23. Merad S, McNamara JM. 1994 Optimal foraging of a reproducing animal as a discounted reward problem. *J. Appl. Probab.* **31**, 287–300. (doi:10.1017/S0021900200044831)
24. Gokhale CS, Wignall AE. 2019 On the innovation and evolution of predatory tactics. *bioRxiv.* (doi:10.1101/530238)
25. Fenneman J, Frankenhuis WE. 2020 Is impulsive behavior adaptive in harsh and unpredictable environments? A formal model. *Evol. Hum. Behav*. **41**, 261–273. (doi:10.1016/j.evolhumbehav.2020.02.005)
26. Uomini N, Fairlie J, Gray RD, Griesser M. 2020 Extended parenting and the evolution of cognition. *Phil. Trans. R. Soc. B* **375**, 20190495. (doi:10.1098/rstb.2019.0495)
27. Joffe TH. 1997 Social pressures have selected for an extended juvenile period in primates. *J. Hum. Evol.* **32**, 593–605. (doi:10.1006/jhev.1997.0140)
28. Bateson P. 1981 Discontinuities in development and changes in the organization of play in cats. In *Behavioral development the Bielefeld interdisciplinary project* (eds K Immelmann, GW Barlow, L Petrinovich, MB Main), pp. 281–295. Cambridge, UK: Cambridge University Press.
29. Panchanathan K, Frankenhuis WE. 2016 The evolution of sensitive periods in a model of incremental development. *Proc. R. Soc. B* **283**, 20152439. (doi:10.1098/rspb.2015.2439)
30. Fagen R. 1981 *Animal play behavior*. Oxford, UK: Oxford University Press.
31. Packer C, Scheel D, Pusey AE. 1990 Why lions form groups: food is not enough. *Am. Nat*. **136**, 1–19. (doi:10.1086/285079)
32. Gopnik A. 2020 Childhood as a solution to explore–exploit tensions. *Phil. Trans. R. Soc. B* **375**, 20190502. (doi:10.1098/rstb.2019.0502)
33. Frankenhuis WE, Gopnik A. 2023 Early adversity and the development of explore-exploit tradeoffs. *Trends Cogn. Sci*. **27**, 616–630. (doi:10.1016/j.tics.2023.04.001)
34. Bellman R. 1957 A Markovian decision process. *J. Math. Mech.* **6**, 679–684. (doi:10.1512/iumj.1957.6.56038)
35. Puterman ML, Shin MC. 1978 Modified policy iteration algorithms for discounted Markov decision problems. *Manag. Sci*. **24**, 1127–1137. (doi:10.1287/mnsc.24.11.1127)
36. Sutton RS, Barto AG. 2018 *Reinforcement learning: an introduction*. Cambridge, MA: The MIT Press.
37. Houston AI, McNamara JM. 1999 *Models of adaptive behaviour: an approach based on state*. Cambridge, UK: Cambridge University Press.
38. Venkateswaran VR, Gokhale CS, Mangel M, Eliassen S. 2024 Effects of time spent in pregnancy or brooding on immunocompetence. *Ecol. Evol*. **14**, e10764. (doi:10.1002/ece3.10764)
39. Kikuchi DW, Sherratt TN. 2015 Costs of learning and the evolution of mimetic signals. *Am. Nat*. **186**, 321–332. (doi:10.1086/682371)
40. Teichmann J, Broom M, Alonso E. 2014 The application of temporal difference learning in optimal diet models. *J. Theor. Biol*. **340**, 11–16. (doi:10.1016/j.jtbi.2013.08.036)
41. Whalen A, Cownden D, Laland K. 2015 The learning of action sequences through social transmission. *Anim. Cogn*. **18**, 1093–1103. (doi:10.1007/s10071-015-0877-x)
42. Enquist M, Lind J, Ghirlanda S. 2016 The power of associative learning and the ontogeny of optimal behaviour. *R. Soc. Open Sci*. **3**, 160734. (doi:10.1098/rsos.160734)
43. Bayer HM, Glimcher PW. 2005 Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141. (doi:10.1016/j.neuron.2005.05.020)
44. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. 2013 A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci*. **16**, 966–973. (doi:10.1038/nn.3413)
45. Glimcher PW. 2011 Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl Acad. Sci. USA* **108**, 15647–15654. (doi:10.1073/pnas.1014269108)
46. Schultz W, Dayan P, Montague PR. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)
47. Stetsenko A, Koos T. 2023 Neuronal implementation of the temporal difference learning algorithm in the midbrain dopaminergic system. *Proc. Natl Acad. Sci. USA* **120**, e2309015120. (doi:10.1073/pnas.2309015120)
48. Leimar O, Dall SRX, Houston AI, McNamara JM. 2022 Behavioural specialization and learning in social networks. *Proc. R. Soc. B* **289**, 20220954. (doi:10.1098/rspb.2022.0954)
49. McNamara JM, Dall SRX, Houston AI, Leimar O. 2024 The evolutionary consequences of learning under competition. *Proc. R. Soc. B* **291**, 20241141. (doi:10.1098/rspb.2024.1141)
50. Frankenhuis WE, Panchanathan K, Barto AG. 2019 Enriching behavioral ecology with reinforcement learning methods. *Behav. Process*. **161**, 94–100. (doi:10.1016/j.beproc.2018.01.008)
51. Stephens DW, Krebs JR. 1986 *Foraging theory. Monographs in behavior and ecology*. Princeton, NJ: Princeton University Press.
52. Ibe OC. 2013 *Markov processes for stochastic modeling*, 2nd edn. Oxford, UK: Elsevier. (doi:10.1016/B978-0-12-407795-9.00015-3)
53. Lima SL. 1991 The influence of models on the interpretation of vigilance. In *Interpretation and explanation in the study of animal behavior, 1st edn* (eds M Bekoff, D Jamieson), pp. 246–267. New York, NY: Routledge. (doi:10.4324/9780429042799-14)
54. Mangel M, Clark CW. 1988 *Dynamic modeling in behavioral ecology*. Princeton, NJ: Princeton University Press. (doi:10.2307/j.ctvs32s5v)
55. Bezanson J, Edelman A, Karpinski S, Shah VB. 2017 Julia: a fresh approach to numerical computing. *SIAM Rev*. **59**, 65–98. (doi:10.1137/141000671)
56. Lazaric A. 2012 Transfer in reinforcement learning: a framework and a survey. In *Reinforcement learning: state-of-the-art* (eds M Wiering, M van Otterlo), pp. 143–173. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg. (doi:10.1007/978-3-642-27645-3_5)
57. Thornton A, Raihani NJ. 2008 The evolution of teaching. *Anim. Behav.* **75**, 1823–1836. (doi:10.1016/j.anbehav.2007.12.014)
58. Spinka M, Newberry RC, Bekoff M. 2001 Mammalian play: training for the unexpected. *Q. Rev. Biol*. **76**, 141–168. (doi:10.1086/393866)
59. Burghardt GM. 2005 *The genesis of animal play: testing the limits*. Cambridge, MA: The MIT Press. (doi:10.7551/mitpress/3229.001.0001)
60. Pusey AE, Packer C. 1994 Non-offspring nursing in social carnivores: minimizing the costs. *Behav. Ecol*. **5**, 362–374. (doi:10.1093/beheco/5.4.362)
61. Bjorklund DF. 2022 Children's evolved learning abilities and their implications for education. *Educ. Psychol. Rev*. **34**, 2243–2273. (doi:10.1007/s10648-022-09688-z)
62. Clutton-Brock T. 2002 Breeding together: kin selection and mutualism in cooperative vertebrates. *Science* **296**, 69–72. (doi:10.1126/science.296.5565.69)
63. Packer C. 2019 The African lion: a long history of interdisciplinary research. *Front. Ecol. Evol* **7**, 259. (doi:10.3389/fevo.2019.00259)
64. Oudeyer PY, Kaplan F. 2007 What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobotics* **1**, 6. (doi:10.3389/neuro.12.006.2007)
65. Lewis RL, Singh S, Barto AG. 2010 Where do rewards come from. In *Proc. of the int. symposium on AI-inspired biology*, De Montfort University, Leicester, UK, 29 March – 1 April 2010, pp. 2601–2606.
66. Juechems K, Summerfield C. 2019 Where does value come from? *Trends Cogn. Sci*. **23**, 836–850. (doi:10.1016/j.tics.2019.07.012)
67. Singh S, Lewis RL, Barto AG, Sorg J. 2010 Intrinsically motivated reinforcement learning: an evolutionary perspective. *IEEE Trans. Auton. Ment. Dev*. **2**, 70–82. (doi:10.1109/tamd.2010.2051031)
68. Kanagawa Y, Doya K. 2024 Evolution of rewards for food and motor action by simulating birth and death. In *ALIFE 2024: Proc. of the 2024 Artificial Life Conference*, Online, 22–26 July 2024, p. 35. Cambridge, MA: The MIT Press.
69. Frankenhuis WE, Walasek N. 2020 Modeling the evolution of sensitive periods. *Dev. Cogn. Neurosci*. **41**, 100715. (doi:10.1016/j.dcn.2019.100715)
70. Taylor ME, Stone P. 2009 Transfer learning for reinforcement learning domains: a survey. *J. Mach. Learn. Res* **10**, 1633–1685. https://dl.acm.org/doi/10.5555/1577069.1755839
71. Heyes C. 2012 What's social about social learning? *J. Comp. Psychol.* **126**, 193–202. (doi:10.1037/a0025180)
72. Gariépy JF, Watson KK, Du E, Xie DL, Erb J, Amasino D, Platt ML. 2014 Social learning in humans and other animals. *Front. Neurosci*. **8**, 58. (doi:10.3389/fnins.2014.00058)
73. van Schaik C, Graber S, Schuppli C, Burkart J. 2016 The ecology of social learning in animals and its link with intelligence. *Span. J. Psychol*. **19**, E99. (doi:10.1017/sjp.2016.100)

74. Hoppitt W, Laland KN. 2013 *Social learning: an introduction to mechanisms, methods, and models*. Princeton, NJ: Princeton University Press. See https://www.jstor.org/stable/j.ctt2jc8mh.
75. Whiten A, van de Waal E. 2018 The pervasive role of social learning in primate lifetime development. *Behav. Ecol. Sociobiol*. **72**, 80. (doi:10.1007/s00265-018-2489-3)
76. Theoretical Eco-evolutionary Dynamics. 2025 tecoevo/riskyplay: Publication release (v-publication). Zenodo. (doi:10.5281/zenodo.18345461)
77. Rajendra D, Gokhale CS. 2026 Supplementary material from: Optimising play for learning risky behaviour. Figshare. (doi:10.6084/m9.figshare.c.8250011)