Seeking entropy: Complex behavior from intrinsic motivation to occupy action-state path space

Jorge Ramírez-Ruiz[1], Dmytro Grytskyy[1], and Rubén Moreno-Bote[1,2]

[1]Center for Brain and Cognition, and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
[2]Serra Húnter Fellow Programme, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

Intrinsic motivation generates behaviors that do not necessarily lead to immediate reward, but help exploration and learning. Here we show that agents having the sole goal of maximizing occupancy of future actions and states, that is, moving and exploring on the long term, are capable of complex behavior without any reference to external rewards. We find that action-state path entropy is the only measure consistent with additivity and other intuitive properties of expected future action-state path occupancy. We provide analytical expressions that relate the optimal policy with the optimal state-value function, from where we prove uniqueness of the solution of the associated Bellman equation and convergence of our algorithm to the optimal state-value function. Using discrete and continuous state tasks, we show that 'dancing', hide-and-seek and a basic form of altruistic behavior naturally result from entropy seeking without external rewards. Intrinsically motivated agents can objectively determine what states constitute rewards, exploiting them to ultimately maximize action-state path entropy.

Introduction

Agents are endowed with a natural tendency to move, explore and interact with their environment with curiosity [1, 2]. For instance, human newborns unintentionally move their body parts [3], and 7 to 12-months infants spontaneously babble vocally [4] and with their hands [5]. Exploration and curiosity are major drives for learning and discovery through information-seeking [6–8]. These behaviors seem to elude a simple explanation in terms of external reward maximization. However, intrinsic motivation pushes agents to visit new states by performing novel courses of action, which helps learning and the discovery of even larger rewards in the long run [9, 10]. Therefore, it has been argued that exploration and curiosity could have arisen as a consequence of seeking external reward maximization by endowing agents with the necessary inductive biases to learn in complex and ever-changing natural environments [11, 12].

While most theories of rational behavior posit that agents are reward or utility maximizers [13–15], very few would agree that the sole goal of living agents is maximizing money gains or food intake. Indeed, expressing excessive emphasis on those goals is a sign of psychological disorders [16, 17]. Further, setting a reward function by design as the goal of intelligent agents is more often than not arbitrary [14, 18, 19], leading to the recurrent problem faced by theories of reward maximization of defining what rewards are [20–24]. In some cases, like in artificial games, rewards can be unambiguously defined, such as number of collected points or wins [25]. However, in most situations defining rewards is task-dependent, non-trivial and problematic. For instance, a vacuum cleaner robot could be designed to either maximize the weight or volume of dust collected, its energy efficiency, or a weighted combination of any of them [26]. In more complex cases, companies might aim at maximizing profit, but without a suitable innovation policy profit maximization can be self-defeating [27].

Here, we abandon the idea that the goal is maximizing external reward and that exploration of space is a means to achieve this goal. Instead, we adopt the opposite view, inspired by the nature of our intrinsic drives: we propose that the objective *is* to occupy action-state path space, understood in a broad sense, in the long term. According to this view, external rewards are the means to generate the work necessary to accomplish this goal, not the goals per se. The usual exploration–exploitation tradeoff [28] therefore disappears: agents that seek to occupy space “solve" this issue naturally because they care about rewards only as means to an end. Furthermore, in this sense, surviving is only preferred because it is needed to keep visiting space. We propose that the intrinsic motivation of agents to visit space –physically by moving, or mentally by generating novel neural activity states– is what ultimately defines intelligence. Indeed, we surmise that agents are intelligent when they are found in large regions of action-state space, by either visiting new spatial locations or by generating richer or unexpected behaviors, not when they maximize external reward. Our theory provides a rational account of exploratory and curiosity-driven behavior where the problem of defining an external reward goal vanishes, and captures the variability of perception and behavior [29–34] by taking it as principle.

We model an agent interacting with the environment as a Markov decision process where the intrinsic, immediate reward is the occupancy of the next action-state visited, which is largest when performing an uncommon action and visiting a rare state –there are no external rewards that drive the agent. We assume that the agent maximizes the occupancy of future action-state paths. We show that action-state path entropy is the only measure of occupancy consistent with additivity per time step, positivity and derivability. Then, we show that the Bellman equation for the state-value function of an agent maximizing the future time-discounted action-state path entropy has a unique solution that can be found with a straight-forward iterative map. In four simulated experiments we show that the sole goal of maximizing future action-state path entropy generates complex behaviors that, to the human eye, look genuinely goal-directed and playful, such as hide-and-seek in a prey-predator problem, dancing of a cartpole and a basic form of altruism in an agent-and-pet example.

Our work builds over an extensive literature on entropy-regularized reinforcement learning [35–42]. While these approaches emphasize the regularization aspects of entropy, external rewards still serve as the major drive of behavior. In our approach, in contrast, we take maximizing action-state path entropy as the exclusive agent's goal. This enables agents to generate rich behaviors constrained only by their body dynamics and the environment where they are situated, while avoiding absorbing states, where action-state entropy is zero. We also extend the notion of action entropy to action-state entropy, which objectively emphasizes visiting state space, not only generating all available actions. Our work also relates to the literature on intrinsically motivated agents and empowerment [19, 43–45], but our agents' goal is to maximize a non-negative linear combination of action and state path space, rather than the predictability of future states given the performed actions.

Results

Entropy measure of space occupancy and its state-value function

We model an agent as a finite action-state Markov decision process in discrete time. Here, the policy $\pi = \{\pi(a|s)\}$ describes the probability $\pi(a|s)$ of performing action a given that the agent is at state s at some time step, and $p(s'|s,a)$ is the transition probability from s to a successor state s' in the next time step given that action a is performed. Starting at $t=0$ in state s_0, an agent performing a sequence of actions and experiencing state transitions $\tau \equiv (a_0, s_1, ..., a_t, s_{t+1}, ...)$ gets a return defined as

$$R(\tau) = -\sum_{t=0}^{\infty} \gamma^t \ln\left(\pi^\alpha(a_t|s_t) p^\beta(s_{t+1}|s_t,a_t)\right) \tag{1}$$

with action and state weights $\alpha > 0$ and $\beta \geq 0$, respectively, and discount factor $0 < \gamma < 1$. Note that a larger return is obtained when, starting in s_t, a low-probability action a_t is performed and followed by a low-probability transition to a state s_{t+1}. Therefore, maximizing the return in Eq. (1) favors ‘visiting’ action-states (a_t, s_{t+1}) with a low transition probability. From s_{t+1}, another low-probability action-state transition is preferred and so on, such that low-probability trajectories τ are more rewarding than high-probability ones. Thus, the agent is pushed to visit action-states that are rare or ‘unoccupied’, implementing our intuitive notion of maximizing ‘path space occupancy’. Due to the freedom to choose action a_t given state s_t and the uncertainty of the resulting next state s_{t+1}, apparent in Eq. (1), the term ‘action-states’ used here is more natural than ‘state-actions’.

The agent is assumed to optimize the policy π to maximize the state-value $V_\pi(s)$, defined as the expected return

$$V_\pi(s) \equiv \mathbb{E}_\pi[R(\tau)|s_0 = s] = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left(\alpha\mathcal{H}(A|s_t) + \beta\mathcal{H}(S'|s_t, a_t)\right)\Big|s_0 = s\right] \tag{2}$$

given the initial condition $s_0 = s$ and following policy π, that is, the expectation is over the $a_t \sim \pi(a_t|s_t)$ and $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$, $t \geq 0$. In the last identity, we have rewritten the expectations of the terms in Eq. (1) as a discounted sum of action and successor state conditional entropies, defined as $\mathcal{H}(A|s) = -\sum_a \pi(a|s)\ln\pi(a|s)$ and $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a)\ln p(s'|s, a)$, respectively. We stress that this expected return is purely intrinsic, namely, there is no external reward (policy-independent reinforcer) that the agent seeks to maximize. This represents a major departure from most reinforcement learning approaches [14] by refraining from defining a goal based on external rewards. We allow for the existence of death (absorbing) states, where only one action-state is available forever, and thus they are naturally avoided by an entropy seeking agent, as they promise zero future action and state entropy. Therefore, our framework implicitly incorporates a survival instinct. Moreover, “rewarding" states that increase the energy reservoir of the agent will be more frequently visited than others so that further action-states can be reached.

We find that the discounted action-state path entropy in Eq. (2) is the only measure of action-state path occupancy in Markov chains consistent with the following intuitive conditions: if a path τ has probability p, visiting it results in an occupancy gain $C(p)$ that (i) decreases with p and (ii) is first-order differentiable (see Supplemental Sec. A.1 for details). Condition (i) implies that visiting a low probability path increases occupancy more than visiting a high probability path, and our agents should tend to occupy ‘unoccupied’ path space; condition (ii) requires that the measure should be smooth. We also ask that (iii) the occupancy of paths, defined as the expectation of occupancy gains over paths given a policy, should be the sum of the expected occupancies of their subpaths (additivity condition). This last condition implies that agents can accumulate occupancy over time by keeping visiting low-probability action-states, but the accumulation should be consistent with the Markov property of the decision process. Finally, we note that the condition (iii) in a time-homogeneous Markov chain is stronger than the additivity of information from two independent experiments assumed in information theory [46], as two independent experiments can be realized within a Markov chain.

Optimal policy and optimal state-value function

The state-value $V_\pi(s)$ in Eq. (2) can be recursively written using the values of successor states through the standard Bellman equation

$$\begin{aligned} V_\pi(s) &= \alpha\mathcal{H}(A|s) + \beta\sum_a \pi(a|s)\mathcal{H}(S'|s, a) + \gamma\sum_{a,s'}\pi(a|s)p(s'|s, a)V_\pi(s') \\ &= \sum_{a,s'}\pi(a|s)p(s'|s, a)\left(-\alpha\ln\pi(a|s) - \beta\ln p(s'|s, a) + \gamma V_\pi(s')\right), \end{aligned} \tag{3}$$

where the sum is over the available actions a from state s and over the successor states s' given the performed action at state s. The optimal policy π^* that maximizes the state-value is defined as $\pi^* = \arg\max_\pi V_\pi$ and the optimal state-value is

$$V^*(s) = \max_\pi V_\pi(s), \tag{4}$$

where the maximization is with respect to the $\{\pi(\cdot|\cdot)\}$ for all actions and states. To obtain the optimal policy, we first determine the critical points of the expected return $V_\pi(s)$ in Eq. (3) by taking partial derivatives respect to the probabilities $\{\pi(\cdot|\cdot)\}$ equal to zero using Lagrange multipliers (see Supplemental Sec. A.2 for details).

The optimal state-value $V^*(s)$ is found to obey the non-linear self-consistency set of equations

$$V^*(s) = \alpha \ln Z(s) = \alpha \ln \left[\sum_a \exp \left(\alpha^{-1}\beta \mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a) V^*(s') \right) \right], \tag{5}$$

where $Z(s)$ is the partition function, defined by substitution, and the critical policy satisfies

$$\pi^*(a|s) = \frac{1}{Z(s)} \exp \left(\alpha^{-1}\beta \mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a) V^*(s') \right). \tag{6}$$

We find that the solution to the non-linear system of Eqs. (5) is unique and, moreover, the unique solution is the absolute maximum of the state-values over all policies (Supplemental Sec. A.3).

To determine the actual value function from such non-linear set of equations, we derive an iterative map, a form of value iteration that exactly incorporates the optimal policy at every step. Defining $z_i = \exp(\alpha^{-1}\gamma V^c(s_i))$, $p_{ijk} = p(s_j|s_i,a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i,a_k)$, Eq. (5) can be turned into the iterative map

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} \tag{7}$$

for $n \geq 0$ and with initial conditions $z_i^{(0)} > 0$. Here, the matrix with coefficients $w_{ik} \in \{0,1\}$ indicate whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$), and j extends over all states, with the understanding that if a state s_j is not a possible successor from state s_i after performing action a_k then $p_{ijk} = 0$. We find that the infinite series $z_i^{(n)}$ defined in Eq. (7) converges to a finite limit $z_i^{(n)} \to z_i^{\infty}$ regardless of the initial condition in the positive first orthant, and that $V^*(s_i) = \alpha\gamma^{-1} \ln z_i^{\infty}$ is the optimal state-value function, which solves Eq. (5) (Supplemental Sec. A.3). Iterative maps similar to Eq. (7) have been studied before [35, 47], subsequently shown to have uniqueness [48] and convergence guarantees [41, 49] in the absence of state entropy terms.

We note that in the definition of return in Eq. (2) we could replace the absolute action entropy terms $\mathcal{H}(A|s)$ by relative entropies of the form $-D_{\mathrm{KL}}(\pi(a|s)||\pi_0(a|s)) = \sum_a \pi(a|s)\ln(\pi_0(a|s)/\pi(a|s))$, as in KL-regularization [35, 38, 42, 47], but in the absence of any external rewards. In this case, one obtains an equation identical to (7) where the coefficients w_{ik} are simply replaced by $\pi_0(a_k|s_i)$, one to one. This apparently minor variation undercovers a major qualitative difference between absolute and relative action entropy objectives: as $\sum_k w_{ik} \geq 1$, absolute entropy seeking favors visiting states with a large action accessibility, that is, where the sum $\sum_k w_{ik}$ and thus the argument of Eq. (7) tends to be largest. In contrast, as $\sum_k \pi_0(a_k|s_i) = 1$, maximizing relative entropies provides no preference for states s with large number of accessible actions $|\mathcal{A}(s)|$. This happens even if the default policy is uniform in the actions, as then the immediate intrinsic return becomes $-D_{\mathrm{KL}}(\pi(a|s)||\pi_0(a|s)) = \mathcal{H}(A|s) - \ln|\mathcal{A}(s)|$, instead of $\mathcal{H}(A|s)$. The negative logarithm penalizes visiting states with large number of actions, which is the opposite goal to occupying action-state path space.

An entropy seeking agent occupies physical space more efficiently than a reward seeking agent or a random walker

In very simple cases, like an homogeneous arena with an identical set of available actions at every state, maximizing action path occupancy (Eq. (2) with $\alpha = 1$ and $\beta = 0$) can trivially be accomplished by a random walk that chooses an available action at every step. However, in realistic examples where space is not homogeneous or there are energetic limitations for moving, a random walk is no longer optimal. We tested how an action entropy seeking agent moving in a 4-room and 4-food-sources environment (Fig. 1a) compares in occupying physical space to a random walker and to a reward seeking agent that gets a reward of 1 every time steps it is alive plus a reward of 1E-5 when it gets food. We allow variable behavior in the R agent by introducing an ϵ-greedy action selection, with ϵ matched to the

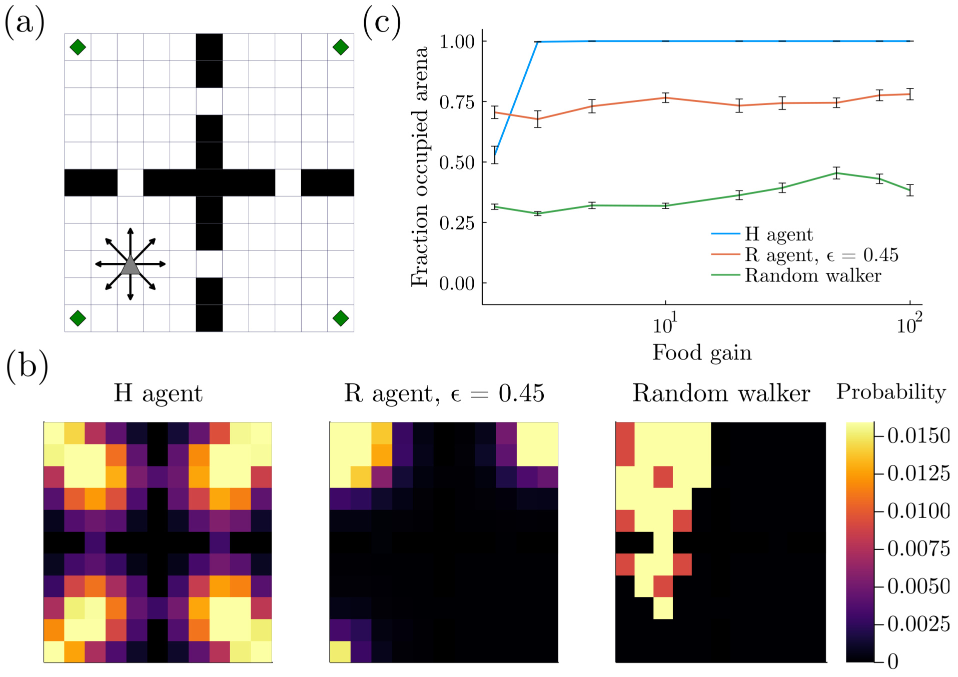

Figure 1: Maximizing future action path entropy leads to high occupancy of physical space. (a) Grid-world arena. The agent has 9 available actions (arrows, and `nothing`) when alive (that is, internal energy larger than zero) and far from walls. There are four rooms, each with a small food source in a corner (green diamonds). (b) Probability of visited spatial states for an entropy seeking (H) agent, an ϵ-greedy reward (R) agent that survives as long as the H agent, and a random walker, for a 5E4 time-step episode; food gain = 10 units; maximum reservoir energy = 100. Random walker survived 129 time steps in this episode. All agents are initialized in the middle of the lower left room, with maximum energy (100 units). (c) Percentage of locations visited at least once per episode as a function of food gain averaged over 50 episodes 5E4 time-steps long. Error bars correspond to s.e.m.

survival of the H agent (see Supplemental Sec. A.6 for more details). In the three cases, the agent can move in one of 8 possible directions if not at a boundary, or stay still, and has an internal state corresponding to the available energy, which reduces one unit at every time step. The agent can move as long as the energy is non-zero, and can increase its reservoir by a fixed amount (food gain) every time it visits a food source, up to a maximum. Note that the total state space is the Cartesian product between physical space and internal energy.

We find that the entropy seeking agent generates behaviors that can be dubbed goal-directed and curiosity-driven (Video 1, H agent). First, by storing enough energy in its reservoir, the agent reaches far, enters the four rooms on the long term (Fig. 1b, left panel), and visits 100% of the arena for moderately large enough food gains and long enough episodes (Fig. 1c, blue line). In contrast, the reward seeking agent lingers over one of the food sources for most of the time (Fig. 1b, middle panel; Video 1, R agent). Although its ϵ-greedy action selection allows for brief exploration of other rooms, the R agent still does not on average visit the whole arena (Fig. 1c, orange line). Finally, the random walker dies before it has time to visit a large fraction of the physical space (Fig. 1b, right panel). These differences hold for a large range of food gains (Fig. 1c — fraction of visited locations vs food gain for the 3 agents). Therefore, although none of the agents were designed to occupy physical space on the long term, the entropy seeking agent is more efficient in visiting physical space than the other agents by generating a guided random walking behavior.

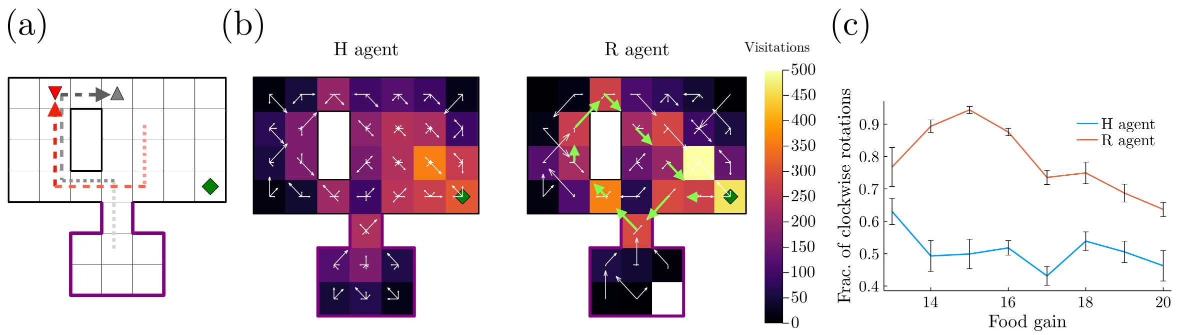

Figure 2: Complex hide-and-seek and escaping strategies in a prey-predator example. (a) Grid-world arena. The agent has 9 available actions when alive and far from walls. There is a small food source in a corner (green diamond). A predator (red, down triangle) is attracted to the agent (gray, up triangle), such that when they are at the same location, the agent dies. The predator cannot enter the locations surrounded by the purple border. Arrows show a clockwise trajectory. (b) Histogram of visited spatial states across episodes for the H and R agents. The vector field at each location indicates probability of transition at each location. Green arrows on R agent show major motion directions associated with the its dominant clockwise rotation. (c) Fraction of clockwise rotations (as in panel (a)) to total rotations as a function of food gain, averaged over epochs of 500 timesteps. Error bars are s.e.m.

Hide and seek in a prey-predator example

More interesting behaviors in entropy seeking agents arise in more complex environments. To show this, we next considered a prey and a predator in a grid world with a safe area (house) and a single food source (Fig. 2a). The prey (a "mouse", gray up triangle) is the agent whose behavior is optimized by maximizing future action path entropy, while the predator (a "cat", red down triangle) acts passively chasing the prey. The prey can move as in the previous 4-room grid world and has an energy reservoir as in the previous example. For simplicity, we only consider a food gain equal to the size of the energy reservoir, such that the agent fully replenishes its reservoir each time it visits the food source. The predator has the same available actions as the agent and is attracted to it stochastically: actions that move the predator towards the agent are more probable than those that move it away from it (see Supplemental Sec. A.6.4 for details).

The entropy seeking agent generates complex behaviors, not limited to visiting the food source to increase the energy buffer and hide at home. In particular, the agent very often first teases the cat and then performs a clockwise rotation around the obstacle, which forces the cat to chase it around, leaving the food source free for harvest (Fig. 2a, arrows show an example; Video 2, H agent). Importantly, this behavior is not restricted to clockwise rotations, as the agent performs an almost equal number of counterclockwise rotations to free the food area (Fig. 2c, H agent, blue line). The variability of these rotations in the entropy seeking agent are manifest in the lack of virtually any preferred directionality of movement in the arena at any single position: arrows pointing toward several directions indicate that on average the prey moves following different paths to get to the food source (Fig. 2b, H agent).

The behavior of the H agent was compared with a reward maximizer (R agent) that receives a reward of one each time it is alive and zero otherwise. To promote variable behavior in this agent as well, we implemented an ϵ-greedy action selection (See Supplemental Sec. A.6.4), where ϵ was chosen to match the expected lifetime of the H agent (Supplemental Fig. 6). The behavior of the R agent was strikingly less variable than that of the H agent, spending more time close to the food source (Fig. 2b, R agent). Most importantly, while the H agent performs an almost equal number of clock and counterclockwise rotations, the R agent strongly prefers the clockwise rotations, reaching 90% of all observed rotations (Video 3, R-agent; Fig. 2c, orange line). This shows that the R agent mostly exploits only one strategy to survive and displays a smaller behavioral repertoire than the H agent.

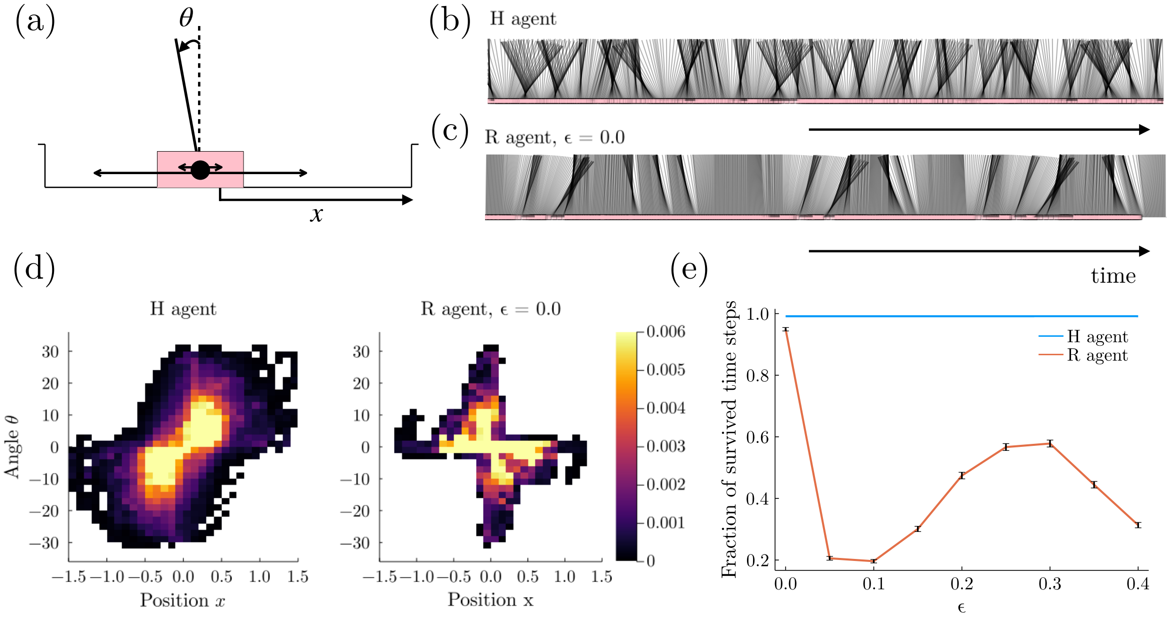

Figure 3: Dancing of an entropy seeking cartpole. (a) The cart (pink square) has a pole attached. The cartpole reaches an absorbing state if the magnitude of the angle θ exceeds 36 deg or its position x reaches the borders at $|x| = 2.4$. There are 5 available actions when alive: big and small acceleration to either side (arrows on cartpole), and doing nothing. (b)-(c) Time-shifted snapshots of the pole in the reference frame of the cart as a function of time for the H and R agents, respectively. (d) Position and angle occupation shows qualitatively different dynamics between H and R agents. (e) Introducing noise to the R agent in the form of ϵ-greedy action selection decreases the survivability of the agent significantly. The H agent survives all simulated 1E5-step episodes. Error bars correspond to s.e.m.

Dancing in an entropy seeking agent with detailed motion dynamics

In the previous examples, complex behaviors emerge as consequence of the presence of obstacles, predators and limited food sources, but the actual dynamics of the agents are very coarse-grained. Here, we consider a system with physically realistic dynamics, the balancing cartpole [50, 51], a system composed of a moving cart with an attached pole free to rotate (Fig. 3a). The cartpole is assumed to reach an absorbing state when either the angle of the pole surpasses 36 degrees in magnitude, or when the cart hits a border, similar to the specifications in a standard task [52] but with much wider allowed angle intervals and more available actions than those typically used. We discretize the state space to solve for the optimal value function in Eq. (4), and use value interpolation to implement the optimal policy in Eq. (6). The entropy seeking agent (H agent) produces a wide variety of angles for the pole (Fig. 3b), constantly swinging sideways, as if it were dancing (Video 4, H agent). It also displays a wide occupation of state space (Fig. 3d, left panel), but it is not homogenous: the bimodality of the histogram reflects the dancing of the cartpole.

We compared the behavior of the H agent with that of a reward maximizer (R agent) as before. In this case, a very small negative component (five orders of magnitude smaller than the survival reward) was added proportional to the angle, position and velocity magnitude (see Supplemental Sec. A.6.5 for details). This was done to break the degeneracy of the value function and to obtain large enough survival times comparable to the H agent. This small external reward was added to our H agent as well, without affecting its behavior. As expected, the R agent maintains the pole close to the balanced position throughout most of a long episode (Fig. 3c), producing very little behavioral variability (Fig. 3d, right panel) and no movement that could be dubbed 'dancing' (Video 4, R agent): instead, its behavior was better described as alternating between a bang-bang sort of control and letting the pole fall followed by a quick back and forth acceleration (high occupation in positive angles and negative positions and vice versa, Fig. 3d, right panel). Finally, by introducing variability in the form of an

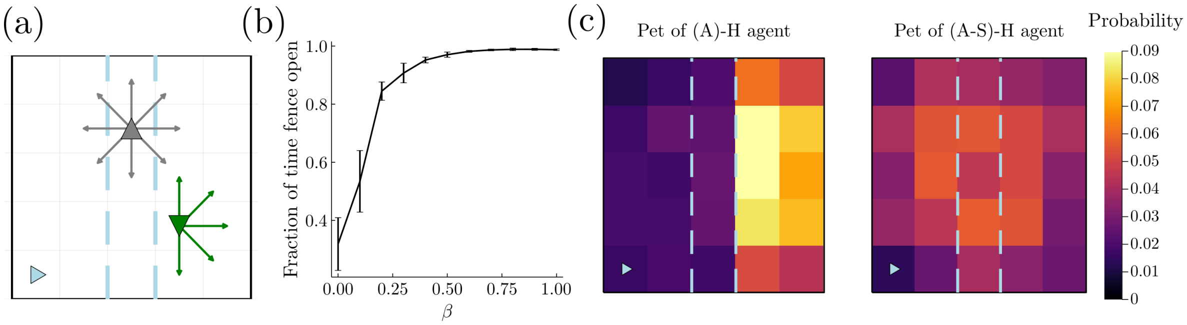

Figure 4: Altruism through an optimal tradeoff between own action entropy and other's state entropy. (a) An agent (gray up triangle) has access to 9 movement actions (gray arrows and doing nothing), and open or close a fence (dashed blue lines). This fence does not affect its movements. A pet (green, down triangle) has access to the same 9 movement actions, and chooses one randomly at each timestep, but is constrained by the fence when closed. Pet location is part of the state of the agent. (b) As β in Eq. (2) is increased, the agent tends to leave the fence open for a larger fraction of time. This helps its pet reach other parts of the arena. Error bars correspond to s.e.m. (c) Occupation heatmaps for 2000 timestep-episodes for $\beta = 0$ (left) and $\beta = 1$ (right). In all cases $\alpha = 1$.

ϵ-greedy action selection, the R agent did actually generate more variable behavior, but its survival time decreased significantly (Fig. 3e), showcasing how the H agent exhibits the most appropriate sort of variability while surviving 100% of the episodes.

Entropy seeking agents also seek entropy of other agents

In the previous examples, we have assumed that the agent maximizes action path entropy (Eq. (2) with $\alpha = 1$ and $\beta = 0$). Now, we consider an example where the goal is to maximize action-state path entropy ($\alpha = 1$ and $\beta = 1$). In the new arena (Fig. 4a), an agent can freely move (grey triangle). A fence can be opened and closed by the agent by pressing a lever in a corner (red triangle). The pet of the agent (green triangle) can freely move if the fence is open, but when the fence is closed the pet is confined to move in the region where it is currently located (see details in Sec. A.6.6). The pet moves randomly at each step, where its available actions are restricted by its available space.

The goal of the agent is to maximize the action-state entropy, which includes the state of its pet, not only its physical location. Therefore, the agent ought to trade off the state entropy resulting from letting the pet free with the action entropy resulting from using the open-close fence action when visiting the lever location. The optimal tradeoff depends on the relative strength of action and state entropies. In fact, when state entropy weighs as much as action entropy ($\alpha = \beta = 1$) the fraction of time that the agent leaves the fence open is close to 1 (rightmost point in Fig. 4b) so that the pet is free to move (Fig. 4c, right panel; (A-S)-H agent). However, when the state entropy weighs nothing ($\alpha = 1, \beta = 0$), the fraction of time that the fence remains open is close to 0.5 (leftmost point in Fig. 4b) and the pet remains confined on the right side for most of the time (Fig. 4c, left panel; (A)-H agent), the region where it was initially placed. As a function of β, the fraction of time the fence is open monotonically increases. Therefore, the agent is altruistic if β is sizeable: it increases the freedom of its pet (measured by the pet's state entropy) by curtailing its own action freedom (measured by its action entropy).

Discussion

Often, the success of agents in nature is not measured by the amount of reward or profit obtained, but by their ability to expand in state space and perform complex behaviors. Here we have proposed that the ultimate goal of intelligence is to 'occupy path space': by moving, agents can reach large portions of action-state space to spread energy beyond what their physical components could do in isolation.

In a Markov decision process setting, we have shown that an intuitive notion of path occupancy is well-formalized by the future action-state path entropy, and we have proposed that behavior is driven by the maximization of this sole intrinsic goal. External rewards are thus the means to move and reach a bigger action-state space, not the goal of behavior. We have solved the associated Bellman equation and provided a convergent iterative map to determine the optimal policy.

In four examples we have shown that, indeed, this single principle along with the embodiment of the agent into a physical system and an environment leads to complex behaviors that are not observed in other simple reward maximizing agents. Efficient occupancy of physical space by a moving agent, hide-and-seek behavior and variable escaping routes in a mouse-cat example, dancing in a realistic cartpole dynamical system and altruistic behavior in an agent-and-pet duet are all behaviors that strike us as being playful, curiosity-driven and energetic. Further, to the human eye, it is remarkable that these behaviors look genuinely goal-directed, although the agent does not have any externally designed goal, but rather intrinsic motivation to maximizing discounted future path action-state entropy.

Another major set of algorithms, known as empowerment, have also proposed using intrinsic rewards as the sole goal of behavior [19, 43, 45]. In this approach, the mutual information between a sequence of actions and the final state is maximized. This approach differs from ours in several fundamental ways. First, action-state path occupancy is maximized when both action and state entropy increase, while empowerment increases when action entropy increases and state entropy decreases given the performed actions. This makes empowerment agents to prefer states where actions leads to large and predictable changes, such as unstable fixed points [43]. One drawback is that empowered agents tend to remain close to those states without producing diverse behavioral repertoires, as it also happens in causal entropy approaches [53]. For instance, in the cartpole setting, both empowered and a causal entropy agents balance the pole upwards and cease behavior when that state is reached [43, 53]. In contrast, entropy seeking agents generate rich behavioral repertoires while avoiding states with little action-state entropy. Another crucial difference is that empowerment cannot be formalized as a cumulative per-step objective (see Sec. A.5 for a proof; [43, 45, 49]), in contrast to action-state path entropy. The fact that mutual information and channel capacity are not additive over Markov chains makes it difficult to formalize them as a cumulative objective that could benefit from the tools of dynamic programming [49, 54]. We note, however, that an approximation to empowerment having the desired additive property could be readily obtained from our framework by putting $\beta < 0$ in the expected return in Eq. (2), such that predictable state transitions would be preferred over more stochastic ones.

Several steps remain to have a more complete theory of entropy seeking behavior. One is to study learning in environments where state transitions are not known. Previous related attempts have introduced Z-learning [35, 47] and G-learning [55] using off-policy methods, so our results could be extended to learning following similar lines. As entropy seeking behavior obviate external rewards, an advantage of this approach is that those rewards do not need to be learned and optimized. Simply observing transitions could allow the estimation of action-state entropies and values without the need of separating them into immediate rewards and future expected value. In addition, modeling and injecting prior information could be particularly simple in our setting in view that intrinsic entropy rewards can be easily bounded before the learning process if action space is known. Therefore, initializing the state-value function to the lower or upper bounds of the action-state path entropy could naturally model pessimism or optimism during learning, respectively.

Acknowledgments

This work is supported by the Howard Hughes Medical Institute (HHMI, ref 55008742) and MINECO (Spain; BFU2017-85936-P) to R.M.-B, and MINECO/ESF (Spain; PRE2018-084757) to J.R.-R. We thank Jose Apesteguia, Luca Bonatti, Ignasi Cos, Benjamin Hayden and Chiara Mastrogiuseppe for very useful comments.

References

[1] Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[2] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

[3] Karen E Adolph and Sarah E Berger. Motor development. *Handbook of child psychology*, 2, 2007.

[4] Peter F MacNeilage and Barbara L Davis. On the origin of internal structure of word forms. *Science*, 288(5465):527–531, 2000.

[5] Laura Ann Petitto and Paula F Marentette. Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, 251(5000):1493–1496, 1991.

[6] Arne Dietrich. The cognitive neuroscience of creativity. *Psychonomic bulletin & review*, 11(6): 1011–1026, 2004.

[7] Celeste Kidd and Benjamin Y Hayden. The psychology and neuroscience of curiosity. *Neuron*, 88 (3):449–460, 2015.

[8] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17 (11):585–593, 2013.

[9] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[10] Bruno B Averbeck. Theory of choice in bandit, information sampling and foraging tasks. *PLoS computational biology*, 11(3):e1004164, 2015.

[11] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, 2012.

[12] Maya Zhe Wang and Benjamin Y Hayden. Latent learning, cognitive maps, and curiosity. *Current Opinion in Behavioral Sciences*, 38:1–7, 2021.

[13] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 2007.

[14] Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning. 1998.

[15] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[16] Carla J Rash, Jeremiah Weinstock, and Ryan Van Patten. A review of gambling disorder and substance use disorders. *Substance abuse and rehabilitation*, 7:3, 2016.

[17] Tamás Ágh, Gábor Kovács, Dylan Supina, Manjiri Pawaskar, Barry K Herman, Zoltán Vokó, and David V Sheehan. A systematic review of the health-related quality of life and economic burdens of anorexia nervosa, bulimia nervosa, and binge eating disorder. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 21(3):353–364, 2016.

[18] John M McNamara and Alasdair I Houston. The common currency for behavioral decisions. *The American Naturalist*, 127(3):358–378, 1986.

[19] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[20] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, pages 2601–2606. Cognitive Science Society, 2009.

[21] Tony Zhang, Matthew Rosenberg, Pietro Perona, and Markus Meister. Endotaxis: A universal algorithm for mapping, goal-learning, and navigation. *bioRxiv*, 2021.

[22] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[23] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. *Advances in neural information processing systems*, 30, 2017.

[24] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[25] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[26] TB Asafa, TM Afonja, EA Olaniyan, and HO Alade. Development of a vacuum cleaner robot. *Alexandria engineering journal*, 57(4):2911–2920, 2018.

[27] Stephen J Kline and Nathan Rosenberg. An overview of innovation. *Studies on science and the innovation process: Selected works of Nathan Rosenberg*, pages 173–203, 2010.

[28] Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current opinion in behavioral sciences*, 38: 49–56, 2021.

[29] Rubén Moreno-Bote, David C Knill, and Alexandre Pouget. Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30):12491–12496, 2011.

[30] Stefano Recanatesi, Ulises Pereira-Obilinovic, Masayoshi Murakami, Zachary Mainen, and Luca Mazzucato. Metastable attractors explain the variable timing of stable behavioral action sequences. *Neuron*, 110(1):139–153, 2022.

[31] Abel Corver, Nicholas Wilkerson, Jeremiah Miller, and Andrew Gordus. Distinct movement patterns generate stages of spider web building. *Current Biology*, 31(22):4983–4997, 2021.

[32] Paule Dagenais, Sean Hensman, Valérie Haechler, and Michel C Milinkovitch. Elephants evolved strategies reducing the biomechanical complexity of their trunk. *Current Biology*, 31(21):4727–4737, 2021.

[33] Gabriela Mochol, Roozbeh Kiani, and Rubén Moreno-Bote. Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior. *Current Biology*, 31(6):1234–1244, 2021.

[34] Fanny Cazettes, Masayoshi Murakami, Alfonso Renart, and Zachary F Mainen. Reservoir of decision strategies in the mouse brain. *bioRxiv*, 2021.

[35] Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

[36] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy.* Carnegie Mellon University, 2010.

[37] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[38] John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

[39] Karol Hausman, Jost Tobias Springenberg, Ziyu Wang, Nicolas Heess, and Martin Riedmiller. Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*, 2018.

[40] Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011.

[41] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[42] Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in kl-regularized rl. *arXiv preprint arXiv:1905.01240*, 2019.

[43] Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent—environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.

[44] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

[45] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

[46] János Aczél, Bruno Forte, and Che Tat Ng. Why the shannon and hartley entropies are 'natural'. *Advances in applied probability*, 6(1):131–146, 1974.

[47] Emanuel Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

[48] Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in mdps. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer, 2012.

[49] Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

[51] Razvan V Florian. Correct equations for the dynamics of the cart-pole system. *Center for Cognitive and Neural Studies (Coneural), Romania*, 2007.

[52] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[53] Alexander D Wissner-Gross and Cameron E Freer. Causal entropic forces. *Physical review letters*, 110(16):168702, 2013.

[54] Nicola Catenacci Volpi and Daniel Polani. Goal-directed empowerment: combining intrinsic motivation and task-oriented behaviour. *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[55] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

A Appendix

A.1 Entropy measures the occupancy of action-state paths

We consider a time-homogeneous Markov decision process with finite state set $\mathcal{S}$ and finite action set $\mathcal{A}(s)$ for every state $s \in \mathcal{S}$. Henceforth, the action-state $x_j = (a_j, s_j)$ is any joint pair of one available action a_j and one possible successor state s_j that results from making that action under policy $\pi \equiv \{\pi(a|s)\}$ from the action-state $x_i = (a_i, s_i)$. By assumption, the availability of action a_j depends on the previous state s_i alone, not on a_i. Thus, the transition probability from x_i to x_j in one time step is $p_{ij} = \pi(a_j|s_i)p(s_j|s_i, a_j)$, where $p(s_j|s_i, a_i)$ is the conditional probability of transitioning from state s_i to s_j given that action a_j is performed. Although there is no dependence of the previous action a_i on this transition probability, it is notationally convenient to define transitions between action-states. We conceive of rational agents as maximizing future action-state path occupancy. Any measure of occupancy should obey the intuitive Conditions *1-4* listed below.

Intuitive Conditions for a measure of action-state occupancy:

1. *Occupancy gain of action-state x_j from x_i is a function of the transition probability p_{ij}, $C(p_{ij})$*
2. *Performing a low probability transition leads to a higher occupancy gain than performing a high probability transition, that is, $C(p_{ij})$ decreases with p_{ij}*
3. *The first order derivative $C'(p_{ij})$ is continuous for $p_{ij} \in (0, 1)$*
4. *(Definition: the action-state occupancy of a one-step path from action-state x_i is the expectation over occupancy gains of the immediate successor action-states, $C_i^{(1)} \equiv \sum_j p_{ij}C(p_{ij})$)*

   The action-state occupancy of a two-steps path is additive,

   $C_i^{(2)} \equiv \sum_{jk} p_{ij}p_{jk}C(p_{ij}p_{jk}) = C_i^{(1)} + \sum_j p_{ij}C_j^{(1)}$

   for any choice of the p_{ij} and initial x_i

Condition *1* simply states that occupancy gain from an initial action-state is defined over the transition probabilities to successor action-states in a sample space. Condition *2* implies that performing a low probability transition leads to a higher occupancy of the successor states than performing performing a high probability transition. This is because performing a rare transitions allows the agent to occupy an space that was left initially unoccupied. Condition *3* imposes smoothness of the measure.

In Condition *4* we have defined the occupancy of the successor action-states (one-step paths) in the Markov chain as the expected occupancy gain. Condition *4* is the central property, and it imposes that the occupancy of action-states paths with two steps can be broken down into a sum of the occupancies of action-states at each time step. Note that the action-state path occupancy can be written as

$$C_i^{(2)} \equiv \sum_{jk} p_{ij}p_{jk}C(p_{ij}p_{jk}) = \sum_j p_{ij}C(p_{ij}) + \sum_{jk} p_{ij}p_{jk}C(p_{jk}) = \sum_{jk} p_{ij}p_{jk}\left(C(p_{ij}) + C(p_{jk})\right),$$

which imposes a strong condition on the function $C(p)$. Note also that the sum $\sum_{jk} p_{ij}p_{jk}C(p_{ij}p_{jk})$ extends the notion of action-state to a path of two consecutive action-states, each path having probability $p_{ij}p_{jk}$ due to the (time-homogeneous) Markov property. The last equality is an identity. While here we consider paths of length equal to 2, further below we show that there is no difference in imposing additivity to paths of any fixed or random length (Corollary 2).

Theorem 1. *$C(p) = -k \ln p$ with $k > 0$ is the only function that satisfies Conditions* 1-4

Corollary 1. *The entropy $C_i^{(1)} = -k \sum_j p_{ij} \ln p_{ij}$ is the only measure of action-state occupancy of successor action-states x_j from x_i with transition probabilities p_{ij} consistent with Conditions* 1-4.

Proof. Put $p_{1,1} = 1$ and $p_{1,j} = 0$ for $j \neq 1$. Then, Condition *4* reads $C(1) = C(1) + C(1)$ when the initial action-state is x_1, which implies $C(1) = 0$.

Now, take a Markov chain with $p_{0,0} = 1$, $p_{1,0} = 1 - t > 0$, $p_{1,2} = t > 0$, $p_{2,0} = p_{2,1} = 0$, $p_{2,j} = 1/n$ for $j = 3, ..., n+2$ and $n > 0$, and $p_{k,0} = 1$ for $k = 3, ..., n+2$. In this chain, the state 0 is absorbing and all others are transient (here action-states are simply referred to as states). Starting from state 1, it transitions to the transient state 2 with probability t and to the absorbing state 0 with probability $1 - t$. From state 2 a transition to states $j = 3, ..., n+2$ happens with equal probability. From any of those states, a deterministic transition to 0 ensues. (These last transitions can only happen in the third time step, and although it will be relevant later on, it is no used in the current proof, which focuses on paths of length two.) Then, Condition *4* with initial state 1 reads $tC(t/n) + (1-t)C(1-t) = tC(t) + (1-t)C(1-t) + tC(1/n) + (1-t)C(1)$, and hence $C(t/n) = C(t) + C(1/n)$ for any $0 < t < 1$ and integer $n > 0$. By Condition *3* and taking derivative with respect to t in both sides, we obtain $C'(t/n) = nC'(t)$, and multiplying in both sides by t we obtain $\frac{t}{n}C'(\frac{t}{n}) = tC'(t)$. By replacing t with nt, we get $tC'(t) = ntC'(nt)$, provided that $nt < 1$.

We will now show that $tC'(t)$ is constant. In the last equation replace t by t/m by integer $m > 0$ to get the last equivalence in $tC'(t) = \frac{t}{m}C'(\frac{t}{m}) = \frac{n}{m}tC'(\frac{n}{m}t)$ (the first equivalence is obvious). These equivalences are valid for positive $t < 1$ and $\frac{n}{m}t < 1$. Let $0 < s < 1$ and $n = \lfloor ms/t \rfloor$ be the largest integer smaller than ms/t. Therefore, as m increases $\frac{n}{m}t < 1$ and approaches s as close as desired. By Condition *3* the function $xC'(x)$ is continuous, and therefore $\lim_{m\to\infty} \frac{n}{m}tC'(\frac{n}{m}t) = sC'(s)$. The basic idea is that we can first compress t as much as needed by the integer factor m and then expand it by the integer factor n so that nt/m is as close as desired to s. This shows that $sC'(s) = tC'(t)$ for $s, t \in (0,1)$, and therefore $tC'(t)$ is constant.

Assume that $tC'(t) = -k$. Then, by integrating we obtain $C(t) = -k \ln t + a$, but $a = 0$ due to $C(1) = 0$, and $k > 0$ due to Condition *2*. Together with the above, we can now proof the theorem by noticing that the solution satisfies Condition *4* for any choice of the p_{ij}. □

Corollary 2. Condition *4* can be replaced by an equivalent condition that requires additivity of paths of any finite length n with no change in the above proof. We first introduce some notation: the probability of path $i_0, i_1, ..., i_n$ is $p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n}$, where i_t refers to the state visited at step t and i_0 is the initial state. Then the new Condition *4* reads in terms of the action-state occupancy of paths of length n as

$$\begin{aligned} C_{i_0}^{(n)} &= \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} C\left(p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n}\right) \\ &= \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_1,i_2}) + ... + \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} C\left(p_{i_{n-1},i_n}\right) \\ &= \sum_{i_1,i_2,...,i_n} p_{i_0,i_1}p_{i_1,i_2}...p_{i_{n-1},i_n} \left(C(p_{i_0,i_1}) + C(p_{i_1,i_2})... + C(p_{i_{n-1},i_n})\right) \ , \end{aligned}$$

for any time-homogeneous Markov chain. By choosing the particular chains used in Theorem 1, we arrive again to the same unique solution $C(p) = -k \ln p$ after using $C(1) = 0$ repeated times, which obviously solves the above equation for any chain and length path. Indeed, note that for the second chain in Theorem 1, from initial state 1 the absorbing state is reached in three time steps with probability one, and thus the above sum contains all $C(1)$ starting from the third terms, which contribute zero to the sum.

The above entropy measure of action-state path occupancy can be extended to the case where there is a discount factor $0 < \gamma < 1$. To do so, we assume now that the paths can have a random length $n \geq 1$ that follows a geometric distribution, $p_n = \gamma^{n-1}(1-\gamma)$. In this case, the occupancy of the paths is

$$\begin{aligned} C_{\text{global}} &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1}p_{i_1,i_2} C(p_{i_0,i_1}p_{i_1,i_2}) \\ &\quad + \gamma^2(1-\gamma)\sum_{i_1,i_2,i_3} p_{i_0,i_1}p_{i_1,i_2}p_{i_2,i_3} C(p_{i_0,i_1}p_{i_1,i_2}p_{i_2,i_3}) + ... \end{aligned} \tag{8}$$

where the n-th term in the sum is the expected occupancy gain of paths of length n weighted by the probability of a having a path with exactly such a length.

Equivalently, a path in course can grow one step further with probability γ or be extinguished with probability $1-\gamma$. Therefore, the occupancy in Eq. (8) should also be equal to the sum of the expected occupancy gains of the local states along the paths, defined as

$$C_{\text{local}} = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma \sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_1,i_2}) + \gamma^2 \sum_{i_1,i_2,i_3} p_{i_0,i_1} p_{i_1,i_2} p_{i_2,i_3} C(p_{i_2,i_3}) + \ldots \tag{9}$$

where the first term is the expected occupancy gain given by the initial condition, the second term is the expected occupancy gain in the next step weighted by the probability of having a path length of at least two steps, and so on.

Eqs. (8-9), after using the Markov chain in Corollary 2, reduce to

$$\begin{aligned} C_{\text{global}} &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_0,i_1} p_{i_1,i_2}) \\ &\quad + \gamma^2(1-\gamma)\sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_0,i_1} p_{i_1,i_2}) + \ldots \\ &= (1-\gamma)\sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma \sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_0,i_1} p_{i_1,i_2}) \end{aligned}$$

and

$$C_{\text{local}} = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \gamma \sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_1,i_2}),$$

where we have used $p_{i_2,i_3} = 1$ because all transitions in the third step are deterministic.

Equality of these two quantities leads to Condition *4*, specifically, $\sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_0,i_1} p_{i_1,i_2}) = \sum_{i_1} p_{i_0,i_1} C(p_{i_0,i_1}) + \sum_{i_1,i_2} p_{i_0,i_1} p_{i_1,i_2} C(p_{i_1,i_2})$. Therefore, the only consistent measure of occupancy with temporal discount is the entropy. Obviously, the equality of global and local time-discounted occupancies measured by entropy holds for any time-homogeneous or inhomogeneous Markov chain.

A.2 Critical policies and critical state-value functions

Here, the expected return following policy π in Eq. (9), known as the state-value function, is written recursively using the Bellman equation. Then, we find a non-linear system of equations for the critical policy and critical state-value function by taking partial derivatives with respect to the policy probabilities (Theorem 2).

Using Eq. (9) and Theorem 1 with $k = 1$, we define the expected return from state s under policy π as

$$V_\pi(s) = -\sum_{i_1} p_{s,i_1} \ln p_{s,i_1} - \gamma \sum_{i_1,i_2} p_{s,i_1} p_{i_1,i_2} \ln p_{i_1,i_2} - \gamma^2 \sum_{i_1,i_2,i_3} p_{s,i_1} p_{i_1,i_2} p_{i_2,i_3} \ln p_{i_2,i_3} + \ldots \tag{10}$$

where p_{s,i_1} is the transition probability from state s to action-state $x_{i_1} = (a_{i_1}, s_{i_1})$. Note that in Eq. (9) we have replaced the initial action-state i_0 by the initial state s alone, as the previous action that led to it does no affect the transition probabilities in the Markov decision process setting. The expected returns satisfy the standard recurrence relationship [14]

$$\begin{aligned} V_\pi(s) &= \sum_{a,s'} p_{s,(a,s')} \left(-\ln p_{s,(a,s')} + \gamma V_\pi(s')\right) \\ &= \sum_{a,s'} \pi(a|s) p(s'|s,a) \left(-\ln \pi(a|s) p(s'|s,a) + \gamma V_\pi(s')\right). \end{aligned} \tag{11}$$

Here, we have unpacked the sum over the action-state i_1 into a sum over (a, s'), where a is the action made in state s and s' is its successor. The second equation shows, in a more standard notation, the explicit dependence of the expected return on the policy. It also highlights that the intrinsic immediate reward takes the form $R_{\text{intrinsic}}(s, a, s') = -\ln \pi(a|s)p(s'|s, a)$, which is unbounded.

From Eq. (10) it is easy to see that the expected return exists (is finite) for any policy π if the Markov decision process has a finite number of actions and states. Due to the properties of entropy, Eq. (10) is a sum of non-negative numbers bounded by $H_{max} = \ln(|A|_{max}|S|)$ ($|A|_{max}$ is the maximum number of available actions from any state) weighted by the geometric series, which guarantees convergence of the infinite sum for $-1 < \gamma < 1$. An obvious, but relevant, implication of the above is that the expected return is non-negative and bounded, $0 \leq V_\pi(s) \leq H_{max}/(1 - \gamma)$, for any state and policy.

While in Eq. (11) the immediate intrinsic reward is the sum of the action and state occupancies, $R_{\text{intrinsic}}(s, a, s') = -\ln \pi(a|s)p(s'|s, a) = -\ln \pi(a|s) - \ln p(s'|s, a)$, we can generalize this reward to consider any weighted mixture of entropies as $R_{\text{intrinsic}}(s, a, s') = -\alpha \ln \pi(a|s) - \beta \ln p(s'|s, a)$ for any two numbers $\alpha > 0$ and $\beta \geq 0$. In particular, for $(\alpha, \beta) = (1, 1)$ we recover the action-state occupancy of Eq. (11), and for $(\alpha, \beta) = (1, 0)$ and $(\alpha, \beta) = (0, 1)$ we only consider action or state occupancy, respectively. The case $(\alpha, \beta) = (0, 1)$ is understood as the limit case where α becomes infinitely small. We note that the case $(\alpha, \beta) = (1, 0)$ has often been used along with an external reward with the aim of regularizing the external reward objective [35–39]. We also note that the case $(\alpha, \beta) = (1, -1)$, with negative β, constitutes an approximation to empowerment [19, 43]: the agent tries to maximize action entropy while minimizing state entropy conditioned to the previous action-state, which favors paths where there is more control on the resulting states. However, we do not consider this case in this paper.

Under the more general intrinsic reward, the expected return obeys

$$V_\pi(s) = \sum_{a,s'} \pi(a|s)p(s'|s, a) \left(-\ln \pi^\alpha(a|s)p^\beta(s'|s, a) + \gamma V_\pi(s')\right). \tag{12}$$

Our goal is to maximize the expected return over the policy probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$ to obtain the optimal policy. Note that for $\alpha > 0$ and $\beta \geq 0$ the expected return is non-negative, $V_\pi(s) \geq 0$.

Theorem 2. *The critical values $V^c(s)$ of the expected returns $V_\pi(s)$ in equation (12) with respect to the policy probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$ obey*

$$V^c(s) = \alpha \ln Z(s) = \alpha \ln \left[\sum_{a \in \mathcal{A}(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s, a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s, a)V^c(s')\right)\right] \tag{13}$$

where $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a)$ is the entropy of the successors of s after performing action a, and $Z(s)$ is the partition function.

The critical points (critical policies) are

$$\pi^c(a|s) = \frac{1}{Z(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s, a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s, a)V^c(s')\right), \tag{14}$$

one per critical value, where the partition function $Z(s)$ is the normalization constant.

Defining $z_i = \exp(\alpha^{-1}\gamma V^c(s_i))$, $p_{ijk} = p(s_j|s_i, a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i, a_k)$, Eq. (13) can be compactly rewritten as

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \tag{15}$$

where the matrix with coefficients $w_{ik} \in \{0, 1\}$ indicates whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$), and j extends over all states, with the understanding that if a state s_j is not a possible successor from state s_i and action a_k then $p_{ijk} = 0$.

Note that the we simultaneously optimize $|S|$ expected returns, one per state s, each with respect to the set of probabilities $\pi = \{\pi(a|s) : a \in A(s), s \in S\}$.

Proof. We first note that the expected return in Eq. (2) is continuous and differentiable in the policy except at the boundaries (i.e., $\pi(a|s) = 0$ for some action-state (a, s)). Choosing a state s, we first take partial derivatives with respect to $\pi(a|s)$ for each $a \in \mathcal{A}(s)$ in both sides of (12), and then evaluate them at a critical point π^c to obtain the condition

$$\begin{aligned}
\lambda(s,s) &= \sum_{s'} p(s'|s,a)\left(-\ln(\pi^c(a|s))^\alpha p^\beta(s'|s,a) + \gamma V^c(s')\right) - \alpha + \gamma \sum_{a,s'} \pi^c(a|s)p(s'|s,a)\lambda(s',s) \\
&= -\alpha \ln \pi^c(a|s) - \beta \sum_{s'} p(s'|s,a) \ln p(s'|s,a) - \alpha + \\
&\quad \gamma \sum_{s'} p(s'|s,a)V^c(s') + \gamma \sum_{a,s'} \pi^c(a|s)p(s'|s,a)\lambda(s',s), \qquad (16)
\end{aligned}$$

where we have defined the partial derivative at the critical point $\frac{\partial V_\pi(s')}{\partial \pi(a|s)}|_{\pi^c} \equiv \lambda(s', s)$ and used the fact that this partial derivative should be action-independent. This is because, for fix s, the $\{\pi(a|s) : a \in A(s), s \in S\}$ lie on a simplex, and the $\lambda(s', s)$ are the Lagrange multipliers corresponding to the state-value function at s', $V_\pi(s')$, associated to the constraint $\sum_a \pi(a|s) = 1$ with $0 \le \pi(a|s) \le 1$ defining that simplex. Noticing that the last term does not depend on the action, we can solve for the critical policy $\pi^c(a|s)$ to obtain equation (14). Eq. (14) implicitly relates the critical policy with the critical value of the expected returns from each state s. Inserting the critical policy (14) into Eq. (12), we get (13), which is an implicit non-linear system of equations exclusively depending on the critical values.

It is easy to verify that the partial derivatives of $V_\pi(s)$ in Eq. (12) with respect to $\pi(a'|s')$ for $s \neq s'$ are

$$\lambda(s, s') = \gamma \sum_{s''} p(s''|s)\lambda(s'', s'),$$

and thus they provide no additional constraint on the critical policy.

□

We finally show that the optimal expected returns, as defined from the Bellman optimality equation

$$V^*(s) = \max_{\pi(\cdot|s)} \sum_{a,s'} \pi(a|s)p(s'|s,a)\left(-\ln \pi^\alpha(a|s)p^\beta(s'|s,a) + \gamma V^*(s')\right) \qquad (17)$$

obey the same Eq. (13) as the critical values of Eq. (12) do. To see this, note that after taking partial derivatives with respect to $\pi(a|s)$ for each $a \in \mathcal{A}(s)$ on the right-hand side of Eq. (17) we get

$$0 = -\alpha \ln \pi(a|s) - \beta \sum_{s'} p(s'|s,a) \ln p(s'|s,a) + \gamma \sum_{s'} p(s'|s,a)V^*(s') - \alpha + \lambda(s), \qquad (18)$$

which, except for the Lagrange multipliers $\lambda(s)$, is identical to Eq. (16). Eq. (13) follows from inserting the resulting optimal policy into the Bellman equation.

A.3 Unicity of the optimal value and policy, and convergence of the algorithm

We now prove that the critical value $V^c(s)$ is unique, in other words, equation (13) admits a single solution (Theorem 3). We later prove that the solution is the optimal expected return (Theorem 4).

Theorem 3. *With the definitions in Theorem 2, the system of equations*

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \qquad (19)$$

with $0 < \gamma < 1$, $\alpha > 0$ and $\beta \ge 0$ has a unique solution in the positive first orthant $z_i > 0$, provided that for all i there exists at least one k such that $w_{ik} = 1$. The solution satisfies $z_i \ge 1$.

Moreover, given any initial condition $z_i^{(0)} > 0$ for all i, the infinite series $z_i^{(n)}$ defined through the iterative map

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} \tag{20}$$

for $n \geq 0$ converges to a finite limit $z_i^{\infty} \geq 1$, and this limit is the unique solution of equation (19)

Note that the condition that for all i there exists at least one k such that $w_{ik} = 1$ imposes virtually no restriction, as it only asks for the presence of at least one available action in each state. For instance, in absorbing states, the action leads to the same state.

Importantly, proving that the map (20) has a single limit regardless of the initial condition in the positive first orthant $z_i^{(0)} > 0$ suffices to prove that equation (19) has a unique solution in that region, as then no other fix point of the map can exist. Additionally, since the solution is unique and satisfies $z_i^{\infty} \geq 1$, the critical state-value function that solves equation (13) is unique, and $V^c(s_i) = \alpha\gamma^{-1} \ln z_i^{\infty} \geq 0$, consistent with its properties.

The map (20) provides a useful value-iteration algorithm used in examples shown in the Results section, and empirically is found to rapidly converge to the solution.

Proof. We call the series $z_i^{(n)}$ with initial condition $z_i^{(0)} = 1$ for all i the *main* series. We first show that the main series is monotonic non-decreasing.

For $n = 1$, we get

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^{\gamma} \geq 1 = z_i^{(0)} \tag{21}$$

for all i, using that there exists k for which, $w_{ik} = 1$, w_{ik} is non-negative for all i and k, $\mathcal{H}_{ik} \geq 0$ and the power function x^{γ} is increasing with its argument.

Assume that for some $n > 0$, $z_i^{(n)} \geq z_i^{(n-1)}$ for all i. Then

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^{\gamma} \geq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n-1)} \right)^{p_{ijk}} \right)^{\gamma} = z_i^{(n)} \tag{22}$$

using the same properties as before, which proves the assertion for all n by induction.

Now let us show that the main series is bounded. Define $\mathcal{H}_{\max} = \max_{ik} \mathcal{H}_{ik}$, and obviously $\mathcal{H}_{\max} \geq 0$.

For $n = 1$ we have

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^{\gamma} \leq \left(|A|_{\max} e^{\mathcal{H}_{\max}} \right)^{\gamma} \equiv c^{\gamma} \tag{23}$$

(remember that $|A|_{\max}$ is the maximum number of available actions from any state).

For $n = 2$,

$$\begin{aligned} z_i^{(2)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(1)} \right)^{p_{ijk}} \right)^{\gamma} \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j c^{\gamma p_{ijk}} \right)^{\gamma} \\ &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} c^{\gamma} \right)^{\gamma} = c^{\gamma^2} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \right)^{\gamma} \leq c^{\gamma+\gamma^2} \end{aligned}$$

using the standard properties, $\sum_j p_{ijk} = 1$ and Eq. (23).

Assume that for some $n > 1$ we have $z_i^{(n)} \leq c^{\gamma+\gamma^2+\ldots+\gamma^n}$. We have just showed that this is true for $n = 2$. Then

$$\begin{aligned} z_i^{(n+1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)}\right)^{p_{ijk}}\right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} c^{\gamma+\ldots+\gamma^n}\right)^\gamma \\ &= c^{\gamma^2+\ldots+\gamma^{n+1}} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}}\right)^\gamma \leq c^{\gamma+\ldots+\gamma^{n+1}} \end{aligned}$$

and therefore it is true for all $n \geq 0$ by induction.

Therefore the series $z_i^{(n)}$ is bounded by $c^{1/(1-\gamma)}$. Together with the monotonicity of the series, we have now proved that the limit z_i^∞ of the series exists. Moreover, $z_i^\infty \geq z_i^0 = 1$.

The above results can be intuitively understood: the 'all ones' initial condition of the main series corresponds to an initial guess of the state-value function equal to zero everywhere. The iterative map corresponds to state-value iteration to a more optimistic value: as intrinsic reward based on entropy is always non-negative, the z-values monotonically increase after every iteration. Finally, the z-values reach a limit because the state-value function is bounded.

We now show the central result that the series obtained by using the iterative map starting from any initial condition in the positive first orthant can be bounded bellow and above by two series that converge to the main series. Therefore, by building 'sandwich' series we will confirm that any other series has the same limit as the main series.

Let the $y_i^{(0)} = u_i > 0$ be the initial condition of the series $y_i^{(n)}$ obeying the iterative map (20), and define $u_{\text{min}} = \min_i u_i$ and $u_{\text{max}} = \max_i u_i$. Obviously, $u_{\text{min}} > 0$ and $u_{\text{max}} > 0$. Applying the iterative map once, we get

$$\begin{aligned} y_i^{(1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(0)}\right)^{p_{ijk}}\right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j (u_{\text{max}})^{p_{ijk}}\right)^\gamma \\ &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} u_{\text{max}}\right)^\gamma = u_{\text{max}}^\gamma \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}}\right)^\gamma = u_{\text{max}}^\gamma z_i^{(1)} \end{aligned}$$

where in the last step we have used the values of the main series in the first iteration. We can similarly lower-bound $y_i^{(1)}$ to finally show that it is both lower- and upper-bounded by $z_i^{(1)}$ with different multiplicative constants,

$$u_{\text{min}}^\gamma z_i^{(1)} \leq y_i^{(1)} \leq u_{\text{max}}^\gamma z_i^{(1)} \tag{24}$$

Now, assume that

$$u_{\text{min}}^{\gamma^n} z_i^{(n)} \leq y_i^{(n)} \leq u_{\text{max}}^{\gamma^n} z_i^{(n)} \tag{25}$$

is true for some $n > 0$. Then, for $n + 1$ we get

$$\begin{aligned} y_i^{(n+1)} &= \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(n)}\right)^{p_{ijk}}\right)^\gamma \leq \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(u_{\text{max}}^{\gamma^n} z_i^{(n)}\right)^{p_{ijk}}\right)^\gamma \\ &= u_{\text{max}}^{\gamma^{n+1}} \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_i^{(n)}\right)^{p_{ijk}}\right)^\gamma = u_{\text{max}}^{\gamma^{n+1}} z_i^{(n+1)} \end{aligned}$$

by simply extracting the common factor in the fourth expression, remembering that $\sum_j p_{ijk} = 1$, and using the definition of the main series in the last one. By repeating the same with the lower bound, we finally find that (25) holds also for $n + 1$, and then, by induction, for every $n > 0$.

The proof concludes by noticing that the limit of both $u_{\max}^{\gamma^n}$ and $u_{\min}^{\gamma^n}$ is 1, and therefore using (25) the limit y_i^∞ of the series $y_i^{(n)}$ equals the limit of the main series, $y_i^\infty = z_i^\infty$.

Note that the iterative map (20) is not necessarily contractive in the Euclidian metric, as it is possible that, depending on the values of $u_{\min}$ and $u_{\max}$ and the changes in the main series, the bounds in Eq. (25) initially diverge to finally converge in the limit.

□

Theorem 4. *The (unique) critical value $V^c(s)$ is the optimal expected return, that is, the one that attains the maximum expected return at every state for any policy, and we write $V^c(s) = V^*(s)$*

Proof. To show that $V^c(s)$ is the optimal expected return, we note that the maximum of the functions $V_\pi(s)$ with respect to policy π should be at the critical policy or at the boundaries of the simplices defined by $\sum_a \pi(a|s) = 1$ with $0 \leq \pi(a|s) \leq 1$ for every a and s, as the expected return $V_\pi(s)$ is continuous and differentiable with respect to the policy except at the boundaries. At the policy boundary, there exists a non-empty subset of states s_i and a non-empty set of actions a_k for which $\pi(a_k|s_i) = 0$. Computing the critical value of the expected return along that policy boundary is identical to moving from the original to a new problem where we replace the graph connectivity matrix w_{ik} in Eq. (19) by a new one v_{ik} such that $v_{ik} \leq w_{ik}$ (remember that at the boundary there should be an action a_k that were initially available from state s_i, $w_{ik} = 1$, that at the policy boundary are forbidden, $v_{ik} = 0$). We now define the convergent series $z_i^{(n)}$ and $y_i^{(n)}$ for the original and new problems respectively by using the iterative map (20) with initial conditions equal to all ones. We prove now that $z_i^{(n)} \geq y_i^{(n)}$ for all i for $n = 1, 2, ...$, and thus their limits obey $z_i^\infty \geq y_i^\infty$.

For $n = 1$, we get

$$z_i^{(1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^\gamma \geq \left(\sum_k v_{ik} e^{\mathcal{H}_{ik}} \prod_j (1)^{p_{ijk}} \right)^\gamma = y_i^{(1)} \tag{26}$$

for all i, using that $w_{ik} \geq v_{ik}$ and that the power function x^γ is increasing with its argument.

Assuming that $z_i^{(n)} \geq y_i^{(n)}$ for all i for some $n > 0$, then

$$z_i^{(n+1)} = \left(\sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(z_j^{(n)} \right)^{p_{ijk}} \right)^\gamma \geq \left(\sum_k v_{ik} e^{\mathcal{H}_{ik}} \prod_j \left(y_j^{(n)} \right)^{p_{ijk}} \right)^\gamma = y_i^{(n+1)} \tag{27}$$

using the same properties as before, which proves the assertion for all n by induction.

Remembering that the expected return $V(s_i)$ is increasing with z_i, we conclude that the expected return obtained from policies restricted on the boundaries of the simplices is no better than the original critical value of the expected return.

□

A.4 Particular examples

Here we summarize the main results and specialize them to specific cases. We assume $0 < \gamma < 1$, $\alpha > 0$ and $\beta \geq 0$ and use the notation $z_i = \exp(\alpha^{-1}\gamma V^*(s_i))$, where $V^*(s)$ is the optimal expected return, $p_{ijk} = p(s_j|s_i, a_k)$ and $\mathcal{H}_{ik} = \alpha^{-1}\beta\mathcal{H}(S'|s_i, a_k)$, where $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a)$.

A.4.1 Action-state entropy maximizers

Agents that seek to maximize the discounted action-state path entropy follow the optimal policy

$$\pi^*(a_k|s_i) = \frac{1}{Z_i} \left(w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \right) \tag{28}$$

with

$$Z_i = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk'}} \tag{29}$$

The matrix with coefficients $w_{ik} \in \{0, 1\}$ indicate whether action a_k is available at state s_i ($w_{ik} = 1$) or not ($w_{ik} = 0$)

The expected return (state-value function) in terms of the z variables obeys

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} e^{\mathcal{H}_{ik}} \prod_j z_j^{p_{ijk}} \tag{30}$$

A.4.2 Action-only entropy maximizers

Agents that ought to maximize the time-discounted action path entropy correspond to the above case with $\beta = 0$, and therefore the optimal policy reads as

$$\pi^*(a_k|s_i) = \frac{1}{Z_i} \left(w_{ik} \prod_j z_j^{p_{ijk}} \right) \tag{31}$$

with

$$Z_i = \sum_k w_{ik} \prod_j z_j^{p_{ijk}} \tag{32}$$

The state-value function in terms of the z variables obeys

$$z_i^{\gamma^{-1}} = \sum_k w_{ik} \prod_j z_j^{p_{ijk}} \tag{33}$$

A.4.3 Entropy maximizers in deterministic environments

In a deterministic environment $p_{i,j(i,k),k} = 1$ for successor state $j = j(i, k)$, and zero otherwise. In this case, at every state i we can identify an action k with its successor state j. Therefore, the optimal policy is

$$\pi^*(a_k|s_i) = \frac{w_{ij} z_j}{Z_i} \tag{34}$$

with

$$Z_i = \sum_j w_{ij} z_j \tag{35}$$

The state-value function in terms of the z variables reads

$$z_i^{\gamma^{-1}} = \sum_j w_{ij} z_j \tag{36}$$

A.5 Non-additivity of mutual information and channel capacity

Here we show that mutual information over Markov chains does not obey the additive property. As the result applies to any policy, channel capacity is not additive either. It suffices to prove our statement for paths of length two. Thus, we ask whether the mutual information between actions (a_0, a_1) and states (s_1, s_2) given initial state s_0

$$\mathrm{MI}_{\mathrm{global}} = \sum_{a_0, a_1, s_1, s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0, s_1, a_1, s_2|s_0)}{p(a_0, a_1|s_0) p(s_1, s_2|s_0)}$$

equals the sum of the per-step mutual information

$$\mathrm{MI}_{\mathrm{local}} = \sum_{a_0,s_1} p(a_0, a_1|s_0) \ln \frac{p(a_0, s_1|s_0)}{p(a_0|s_0)p(s_1|s_0)} + \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1, s_2|s_1)}{p(a_1|s_1)p(s_2|s_1)}$$

where $p(a_0, s_1, a_1, s_2|s_0) = \pi(a_0|s_0)p(s_1|s_0, a_0)\pi(a_1|s_1)p(s_2|s_1, a_1)$ and $p(a_0, a_1|s_0) = \pi(a_0|s_0)p(s_1|s_0, a_0)$. Using Bayes' rule and the Markov property, the above quantities can be rewritten as

$$\begin{aligned}
\mathrm{MI}_{\mathrm{global}} &= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0, a_1|s_0, s_1, s_2)}{p(a_0, a_1|s_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{p(a_0, a_1|s_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{\pi(a_0|s_0)p(a_1|s_0, a_0)} \\
&= \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_0|s_0, s_1)p(a_1|s_1, s_2)}{\pi(a_0|s_0)\sum_s \pi(a_1|s)p(s|s_0, a_0)} \\
&= \sum_{a_0,s_1} p(a_0, a_1|s_0) \ln \frac{p(a_0|s_0, s_1)}{\pi(a_0|s_0)} + \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1|s_1, s_2)}{\sum_s \pi(a_1|s)p(s|s_0, a_0)}
\end{aligned}$$

and

$$\mathrm{MI}_{\mathrm{local}} = \sum_{a_0,s_1} p(a_0, a_1|s_0) \ln \frac{p(a_0|s_0, s_1)}{\pi(a_0|s_0)} + \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \frac{p(a_1|s_1, s_2)}{\pi(a_1|s_1)}$$

The quantities $\mathrm{MI}_{\mathrm{global}}$ and $\mathrm{MI}_{\mathrm{local}}$ are remarkable similar except for the denominator in the ln of the last term in each expression. Therefore, equality between $\mathrm{MI}_{\mathrm{global}}$ and $\mathrm{MI}_{\mathrm{local}}$ holds iff

$$\sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \sum_s \pi(a_1|s)p(s|s_0, a_0) = \sum_{a_0,a_1,s_1,s_2} p(a_0, s_1, a_1, s_2|s_0) \ln \pi(a_1|s_1)$$

which is not true for all choices of policy and transitions probabilities. To see this, take a Markov chain where the action $a_0 = 0$ from $s_0 = 0$ is deterministic, but results in two possible successor states $s_1 = 1$ or $s_1 = 2$ with equal probability 1/2. From $s_1 = 1$ the policy takes actions $a_1 = 1$ and $a_1 = 2$ with probability 1/2. From $s_1 = 2$ the policy is deterministic, that is, $a_1 = 3$ with probability 1. A simple calculation shows that the left side equals $-\frac{3}{2}\ln 2$, while the right side equals a different quantity, $-\frac{1}{2}\ln 2$.

A.6 Experiments

In this subsection, we present the details for the numerical simulations performed for the different experiments in the manuscript. First, we discuss the construction of the H and R agents, and afterwards we present the details of each particular experiment.

A.6.1 H agent

In all the experiments presented, we introduce the H agent, whose name comes from the usual notation for using H to denote entropy. Therefore, the objective function that this agent maximizes in general is Eq. (2). As described in section A.4, the α and β parameters control the weights of action and next-state entropies to the objective function, respectively. Unless indicated otherwise, we always use $\alpha = 1, \beta = 0$ for the experiments. It is important to note, as we have done before, that if the environment is deterministic, then the next-state entropy $\mathcal{H}(S'|s, a) = -\sum_{s'} p(s'|s, a) \ln p(s'|s, a) = 0$, and therefore β does not change the optimal policy, Eq. (6).

We have implemented the iterative map, Eq. (7), to solve for the optimal value, using $z_i^{(0)} = 1$ for all i as initial condition. Theorem (3) ensures that this iterative map finds a unique optimal value regardless of the initial condition in the first orthant. To determine a degree of convergence, we compute the supremum norm between iterations,

$$\delta = \max_i |V_i^{(n+1)} - V_i^{(n)}|,$$

where $V_i = \frac{\alpha}{\gamma} \log(z_i)$, and the iterative map stops when $\delta < 10^{-3}$.

A.6.2 R agent

We also introduce a reward-maximizing agent in the usual RL sense. In this case, the reward is $r = 1$ for living and $r = 0$ when dying. In other words, this agent maximizes life expectancy. Additionally, to emphasize the typical reward-seeking behavior and avoid degenerate cases induced by the tasks, we introduced a small reward for the Four-room grid world, and for the Cartpole experiments (see below). In all other aspects, the modelling of the R agent is identical to the H agent. To allow for reward-maximizing agents to display some stochasticity, we used an ϵ-greedy policy, the best in the family of ϵ-soft policies [14]. At any given state, a random admissible action is chosen with probability ϵ, and the action that maximizes the value is chosen with probability $1 - \epsilon$. Given that the world models $p(s'|s, a)$ are known and the environments are static, this ϵ-greedy policy does not serve the purpose of exploration (in the sense of learning), but only to inject behavioral variability. Therefore, we construct an agent with state-independent variability, whose value function satisfies the optimality Bellman equation for this ϵ-greedy policy,

$$V_\epsilon(s) = (1 - \epsilon) \max_a \sum_{s'} p(s'|s, a) \left(r + \gamma V_\epsilon(s')\right) + \frac{\epsilon}{|\mathcal{A}(s)|} \sum_{a,s'} p(s'|s, a) \left(r + \gamma V_\epsilon(s')\right), \qquad (37)$$

where $|\mathcal{A}(s)|$ is the number of admissible actions at state s. To solve for the optimal value in this Bellman equation, we perform value iteration [14]. The ϵ-greedy policy for the R agent is therefore given by

$$\pi(a|s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(s)|}, & \text{if } a = \arg\max_{a'} \sum_{s'} p(s'|s, a') \left(r + \gamma V_\epsilon(s')\right) \\ \frac{\epsilon}{|\mathcal{A}(s)|}, & \text{otherwise} \end{cases}$$

where ties in $\arg\max$ are broken randomly. Note that if $\epsilon = 0$, we obtain the usual greedy optimal policy that maximizes reward.

A.6.3 Four-room grid world

Environment The arena is composed of four rooms, each having size 5×5 locations where the agent can be in. From each room, the agent can go to two adjacent rooms through small openings, each located in the middle of the wall that separates the rooms. At each of these rooms, there is a food source located in the corner furthest from the openings. See Fig. 1 for a graphic description. The discount factor is set to $\gamma = 0.99$.

States The states are the Cartesian product between (x, y) location and internal state u, which is simply a scalar value between a minimum of 0 and a maximum capacity of 100. All states such that $(x, y, u = 0)$ are absorbing states, independently of the location (x, y). The particular internal state $u = 100$ is the maximum capacity for energy, such that even when at a food source, this internal state does not change. Therefore, the number of states in this experiment is $|\mathcal{S}| = 104\,\text{external states} \times 101\,\text{internal states} = 10504$.

Actions The agent has a maximum of 9 actions: `up`, `down`, `left`, `right`, `up left`, `up right`, `down left`, `down right`, and `nothing`. Whenever the agent is close to a wall, the number of available actions decreases such that the agent cannot choose to go into walls. Finally, whenever the agent is in an absorbing state, only `nothing` is available.

Transitions At any transition, there is a cost of 1 unit of energy for being alive. On the other hand, whenever the agent is located at a food source, there is an increase in energy that we vary parametrically that we call food gain g. For example, if the agent is in location $(2,1)$ at time t and moves towards $(1,1)$ (where food is located), the change in energy would be $\Delta u_t = -1$, given that the change in internal energy depends only on the current state and action. If the agent decides to stay in $(1,1)$ at time $t+1$, then $\Delta u_{t+1} = -1 + g$.

R agent As stated above, in this experiment we introduced an extra reward for the R agent when it reaches the food source. The magnitude is small compared to the survival reward ($1E-5$ smaller) and it mainly serves to break the degeneracy of the value function. The variability of the R agent is thus coming purely from the ϵ-greedy action selection.

Survivability To allow for the maximum uniform variability for the R agent, we tested various values for ϵ and observed the survivability of the agents as a function of ϵ, across all the food gains tested (see Results section). The value of ϵ for which the R agent still survives as much as the H agent is $\epsilon = 0.45$ (see Figure 5).

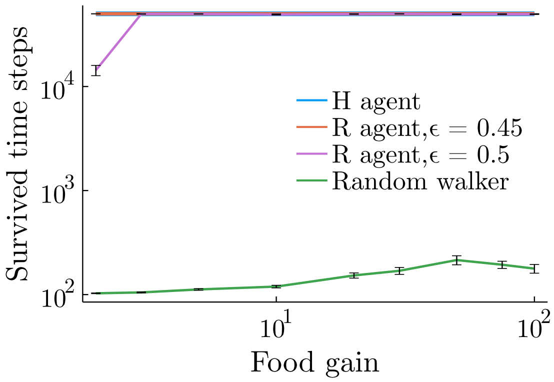

Figure 5: Survivability of the various agents tested in the four-room grid world. At each 5E4 timestep episode, we recorded the survived time and averaged across episodes. The ϵ-greedy R agent that survives as much as the H agent is the one for $\epsilon = 0.45$.

A.6.4 Predator-prey scenario

Here we provide all details of the simulated experiments. Results are shown in Fig. 2.

Environment The environment is similar to that one used for the 4-room grid world described in A.6.3. Apart from the agent (prey), there is also another moving subject (predator) with a simple predefined policy. The grid world consists of a "home" area, a rectangle 2x3 where the agent may enter, but the predator cannot. This home area has a small opening that leads to a bigger 4x7 rectangle arena available for both the agent and the predator. The only food source is located at the bottom-right corner of the common part of the arena, so that the agent needs to leave its home to boost its energy. Additionally, there is an obstacle which separates the arena in two parts with two openings, above and under the obstacle. This obstacle allows the agent to "hide" from the predator behind it.

States The location of the predator is part of the agent's state, such that a particular state consists of the position of the agent, the position of the predator and the amount of energy of the agent. For this case, we set the maximum amount of energy F equal to the food gain. Positions are 2-dimensional, and therefore the states are 5-dimensional. In the used arena there are 33 possible locations for the agent and 26 ones for the predator, so that the total number of states ranges from 11154 for $F = 10$ to 17160 for $F = 20$.

Actions The agent has the same actions as in the four-room grid world. The maximum number of available actions is therefore 9. Moving towards obstacles or walls is not allowed.

Transitions The agent loses one unit of energy every time step and increases the amount of energy up to a given maximum capacity level F only at the food source. If the position of both the agent and the predator are the same, then the agent is "eaten" and moves to the absorbing state of death as well as in the case of energy equal to 0. After entering the absorbing state the agent stays there forever.

The predator also moves as the agent (horizontally, vertically, diagonally on one step or to stay still). Steps of the agent and the predator happen synchronously. The predator is “attracted” to the agent: the probability of moving to some direction is an increasing function on the cosines $\cos\alpha_k$ of the angle α_k between this direction of motion k and the direction of the radius vector from the predator to the agent. In particular, this probability is

$$p_k^c = C^{-1}\exp(\kappa\cos\alpha_k) \tag{38}$$

where κ is the inverse temperature of the predator and $C = \sum_k \exp(\kappa\cos\alpha_k)$ is a normalization factor. These probabilities are computed only for motions available at the current location of the predator, so that e.g. for the location at the wall the motions along the wall are taken into account, but not the motion towards the wall.

Goal The goal of the H agent is to maximize discounted action entropy, and thus to find the optimal state-value function using the iterative map in Eq. (7) with $\mathcal{H}_{ik} = 0$ ($\beta = 0$). While using the iterative map, we take advantage of the fact that given an action the physical transition of the agent is deterministic, but the physical transition of the predator is stochastic. Therefore, the sum over successor states j in Eq. (7) is simply a sum over the predator successor states.

Parameters $\gamma = 0.98$, $F = 15$, $\kappa = 2$. Simulation time is 5000 steps.

Counting rotations We define a clockwise (counterclockwise) half-rotation as the event when the agent came from the left part of the arena to the right part over the field above (under) the wall and from the right part to the left one over the field under (above) the wall without crossing the vertical line of the wall in between. One full rotation consists of two half-rotations in the same directions performed one after another. We counted the number of full rotations in both directions in 70 episodes of 500 time steps each for both H and R agents for different values of the food gain F. Error bars where computed based on these 70 repetitions. The fraction of clockwise rotations to total rotations (sum of clockwise and anticlockwise rotations) for different values of F is shown at Fig. 2.

Survivability The ϵ-greedy R agents display some variability that depends on ϵ. To select this parameter, we matched expected lifetimes (measured in simulations of 5000 steps length) between the H and R agents, separately for every F. Lifetimes are plotted in Figure 6.

Videos We have generated one video for the H agent (Video 2) and another for the R agent (Video 3), both for $F = 15$, $\kappa = 2$, and $\epsilon = 0.06$ for the R agent so as to match their expected lifetimes as described above. In the videos, green vertical bar indicates the amount of energy by the agent at current time. When the agent makes at least one full rotation around the wall, it is indicated by the written phrase "clockwise rotation" or "anticlockwise rotation". Black vertical arrow indicates direction ('up' for clockwise and 'down' for anticlockwise directions) of the half-rotation in the part of arena left from the wall.

A.6.5 Cartpole

Environment A cart is placed in a one-dimensional track with boundaries at $|x| = 2.4$. It has a pole attached to it, that rotates like an inverted pendulum with its pivot point on the cart.

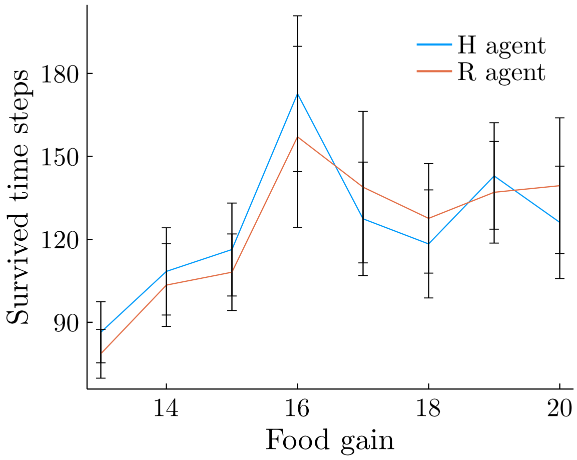

Figure 6: Survivability of the mouse for both H and R agents.

States The dynamical system can be described by a four-dimensional external state (x, v, θ, ω), where x is the position of the cart, v is its linear velocity, θ is the angle of the pole with respect to the vertical which grows counterclockwise, and ω is its angular velocity. In this case, we model the internal state u simply with the binary variable `alive, dead`, where the agent enters the absorbing state `dead` if its position exceeds the boundaries, or if its angle exceeds 36 degrees. This amplitude of angles is larger than that typically assumed (12 degrees in [52]), and therefore our system can be less stable. The state space is $[-2.4, 2.4] \times (-\infty, \infty) \times [-36, 36] \times (-\infty, \infty) \times \{0, 1\}$. To solve for the state value function in Eq. (7), we discretize the state space by setting a maximum value for the velocities. Given all the parameters (allowed x and θ, magnitude of the forces, masses of cart and pole, length of pole and gravity, below), we empirically set the maximum values for $|v| = 3$ and $|\omega| = 3$, which the cart actually never exceeds. Therefore, we computed the state value function in a $31 \times 31 \times 31 \times 31 \times 2$ grid (number of states $= 1.8 \times 10^6$).

Actions Any time the agent is `alive`, it has 5 possible actions: forces of $\{-50, -10, 0, 10, 50\}$, where zero force is understood as `nothing`. If the agent is `dead`, then only `nothing` is allowed.

Transitions This dynamical system is a standard task in reinforcement learning, namely the `cartpole-v0` system of the OpenAI gym [52]. The solution of this dynamical system is given in Ref. [51], where we use a frictionless cartpole. The equations for angular and linear accelerations are thus

$$\ddot{\theta} = \frac{-g\sin(\theta) + \frac{\cos(\theta)}{M+m}\left(-F + m\dot{\theta}^2 l\sin(\theta)\right)}{l\left(\frac{4}{3} - \frac{m\cos^2(\theta)}{M+m}\right)} \tag{39}$$

$$\ddot{x} = \frac{1}{\cos(\theta)}\left(\frac{4}{3}l\ddot{\theta} - g\sin(\theta)\right). \tag{40}$$

Given a force F, a deterministic transition can be computed from these dynamical rules, and a real-valued state transition is observed by the agents.

R agent As stated above, in this experiment we introduced an extra reward for the R agent, in this case a negative reward proportional to the state magnitude (position, angle, linear velocity and angular velocity), with a small strengh ($1E - 5$ smaller than the survival reward), and it mainly serves to break the degeneracy of the value function. The variability of the R agent is thus coming purely from the ϵ-greedy action selection. To check that this extra reward does not dictate the behavior in general, we introduced the same cost to the H agent as an extrinsic reward (additional to the entropy), and found no differences in behavior.

Parameters Mass of the cart $M = 1$, mass of the pole $m = 0.1$, length of the pole $l = 1$, acceleration due to gravity $g = 9.81$, time discretization $\Delta t = 0.02$. The discount factor was set to $\gamma = 0.96$.

Value interpolation Since the observed external state is a continuous four-dimensional variable, there are two options to compute the optimal policy, Eq. (6), which needs the value at a given state. One can either use the value of the nearest neighbor state and its successors, or interpolate the value in the grid. Given the dimensionality of the space, we opted for a linear interpolation of the optimal state value function, which we passed to the optimal policy. The interpolated state values to compute $Z(s)$ do not normalize the policy exactly, so we normalized accordingly to produce a proper policy.

A.6.6 Agent-pet scenario

An agent and a pet move in an arena with degrees of freedom that depend on the actions made by the agent, as explained next in detail.

Environment A 5 × 5 arena. The middle column of arena can be blocked by a fence, a vertical obstacle that the pet cannot cross. The agent can cross it freely regardless of whether it is open or closed. The agent can open or close the fence by performing the corresponding action when visiting the lever location, at the left bottom corner.

States The system's state consists of the Cartesian product of agent´s location, pet's location and binary state of the fence. So, the number of states is 1250. For the sake of simplicity there is no internal states for the energy, and thus there are not terminal states. The initial states of the agent and pet at the start of each episode are the middle of the second column and the right lower corner of the arena, respectively.

Actions As in Sec. A.6.4 the agent's actions are movements to one of the 8 neighbour locations as well as staying on the current one. Additionally, if the agent is on the "lever" location, an additional action is available, namely to open or close the fence, depending on its previous state.

Transitions The pet has same available movements as the agent when the fence is open. The pet performs a random transition to any of the neighbour locations, or stay still, with the same probability. If the agent closes the fence, then the pet can only move on the side where it lies when closed. For simplicity, if the fence is closed by the agent when the pet lies in the middle column, then the pet can only move to the right or left locations such that it will be at one side of the fence in the next time step.

Goal The goal of the H agent is to maximize discounted action-state entropy using the iterative map in Eq. (7) with $\alpha = 1$ and $\beta \in [0, 1]$, parameters that measure the weight of action and state entropies, respectively. As in the prey-predator example, we take advantage of the fact that given an action the physical transition of the agent is deterministic, while the physical transition of the pet is stochastic. Thus, the sum over successor states j in Eq. (7) is a sum over the pet successor states.

Simulation details We ran simulations for several values of β, from 0 to 1 in 0.1 steps, to interpolate between pure action entropy ($\beta = 0$) and action-state entropy ($\beta = 1$). We measured the fraction of time the gate was open using episodes of 2000 steps averaged over 70 simulations for each β, shown in Fig 4. Heat-maps in that figure correspond to the occupation probability by the pet for $\beta = 0$ (left panel) and $\beta = 1$ (right panel) using an episode of 5000 steps.