PLOS Computational Biology

RESEARCH ARTICLE

Deterministic dynamics of distributional multi-agent reinforcement learning

Clémence Bergerot[1,2*], **Pawel Romanczuk**[1,3], **Wolfram Barfuss**[4,5,6*]

1 Department of Biology, Humboldt Universität zu Berlin, Berlin, Germany, **2** Charité-Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany, **3** Science of Intelligence, Research Cluster of Excellence, Berlin, Germany, **4** Transdisciplinary Research Area Sustainable Futures, University of Bonn, Bonn, Germany, **5** Center for Development Research, University of Bonn, Bonn, Germany, **6** Institute for Food and Resource Economics, University of Bonn, Bonn, Germany

* clembergerot@gmail.com (CB); wbarfuss@uni-bonn.de (WB)

Check for updates

OPEN ACCESS

Citation: Bergerot C, Romanczuk P, Barfuss W (2026) Deterministic dynamics of distributional multi-agent reinforcement learning. PLoS Comput Biol 22(9): e1014723. https://doi.org/10.1371/journal.pcbi.1014723

Editor: Ulrik R. Beierholm, Durham University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: December 26, 2025

Accepted: August 17, 2026

Published: September 23, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014723

Copyright: © 2026 Bergerot et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution,

Abstract

Understanding how cognition shapes behavior across contexts remains a fundamental challenge for many disciplines. In particular, for the optimism heuristic–i.e., the tendency to overweight positive (relative to negative) information–knowledge remains fragmented, with models developed in specific domains in isolation. Here, we present a unifying computational framework by deriving the deterministic dynamics of distributional multi-agent reinforcement learning. Our approach discretizes return distributions through a finite set of neurons, consistent with recent empirical findings on distributional coding in the brain. We validate our framework by reproducing established results across three iconic domains spanning individual bandit choice under resource variability, social coordination, and risky choice. Beyond validation, we uncover novel interactions among optimism, return discretization, and temporal discounting. Specifically, we identify conditions under which return discretization generates choice hysteresis and, in extreme parameter regimes, inescapable perseveration. We further reveal "individual dilemmas": circumstances where agents gravitate toward suboptimal yet stable strategies, offering a mechanistic explanation for incoherent choice patterns. Our framework bridges neuroscience, psychology, and collective behavior, enabling empirically testable hypotheses about how cognitive biases propagate from individual cognition to social outcomes in complex environments.

Author summary

How do cognitive heuristics shape individual decisions and the collective outcomes that emerge when many agents interact? This question connects psychology, neuroscience, and artificial intelligence, yet computational models have developed separately across these fields. Here we present DDRL, a framework that combines distributional reinforcement learning, in which agents learn the full

UNCORRECTED PROOF

and reproduction in any medium, provided the original author and source are credited.

Data availability statement: All relevant data and code can be found under this Github repository: https://github.com/clembergerot/pyDDRL The code can also be found in the Supporting Information.

Funding: This work was supported by funding under the form of a PhD scholarship by the Einstein Center for Neurosciences Berlin, Charité – Universitätsmedizin Berlin to CB and funding by the German Research Foundation under Germany's Excellence Strategy EXC 2002/2 "Science of Intelligence" project 390523135 to PR. WB acknowledges support from the Cooperative AI Foundation. This publication was supported by the Open Access Publication Fund of the University of Bonn. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

spread of possible outcomes rather than their average, with deterministic learning dynamics that yield mathematically tractable trajectories in strategy space. DDRL is grounded in neuroscientific evidence: dopaminergic neurons encode outcome distributions rather than expected values alone. We illustrate DDRL with an optimism heuristic across three domains: bandit choice under resource variability, social coordination, and intertemporal risky choice. In each case, DDRL reproduces established results. Beyond replication, DDRL makes two novel predictions. Return discretization generates bistable regimes, in which agents settle into qualitatively different behaviors depending on their history. Under certain combinations of optimism and temporal discounting, agents can become locked into an individual dilemma, in which the better strategy is no longer a stable attractor. Together, these results provide a mechanistic route from neural reward encoding to emergent collective behavior, connecting individual cognitive heuristics to social outcomes in a biologically grounded and analytically tractable framework.

1. Introduction

Developing biologically grounded models of learning and collective decision-making is a shared priority across biology, economics, artificial intelligence, and sustainability science [1–7]. Reinforcement learning (RL) is a promising framework for this effort [8–11]: agents improve their decision strategies through trial and error, and dopaminergic neurons in the brain signal reward-prediction errors in a manner that closely matches RL learning rules [12,13]. In multi-agent settings, individual learning dynamics give rise to collective behavior — emergent outcomes arising from individual learning in a shared environment, without requiring explicit communication between agents — opening a route to mechanistic models of social decision-making [11,14,15]. Standard RL approaches, however, derive behavior from a single quantity: the expected return. Human and animal decision-making systematically deviates from this expectation-based framework — exhibiting biases and heuristics that are difficult to capture while preserving analytical tractability in a multi-agent setting with changing environmental contexts.

Behavioral heuristics—cognitive shortcuts that enable effective decisions in uncertain environments—offer one productive lens for this challenge [16,17]. Among the best-studied is the optimism heuristic: the tendency to weight positive over negative information. At the individual level, asymmetric RL documents it in humans, mice, and monkeys [18–21], using a higher learning rate for positive prediction errors than for negative ones. At the collective level, optimistic coordination algorithms have enhanced performance across a range of multi-agent tasks [22–26]. These efforts, however, have been developed in separate disciplinary traditions. Optimism and pessimism are treated as phenomenological hyperparameters: effective, but not grounded in a mechanistic account of how reward outcomes are represented.

Distributional RL (DistRL) provides the mechanistic substrate this picture is missing [27]. Rather than estimating the expected return, distributional agents learn the

UNCORRECTED PROOF

full distribution of possible returns. Neural signatures of this computation have been found in the dopaminergic system [28], connecting DistRL to the biological machinery of reward processing. Within DistRL, optimism and pessimism arise naturally: an agent selects an upper quantile of the return distribution [29,30] or distorts its shape [31,32], rather than adjusting learning rates. DistRL has been applied primarily in single-agent machine-learning settings; a deterministic formulation for multi-agent collective dynamics does not yet exist.

Deterministic RL (DetRL) dynamics provide analytical tractability in collective settings [33,34]. In DetRL, stochastic fluctuations in the learning process are integrated out: strategies are updated as deterministic functions of expected rewards and state transitions. With a formalism analogous to evolutionary game theory [35,36], this approach enables exact characterization of fixed points, their stability, and bifurcations of collective learning. Prior DetRL work has revealed how caring for the future shapes cooperation in social-ecological dilemmas [37] and how stochastic noise drives coordination [38]. DetRL, however, operates on scalar Q-values—point estimates of expected return. The distributional structure of reward outcomes is absent.

Here, we derive the deterministic dynamics of distributional multi-agent RL, introducing a framework we call DDRL (Section 3.1). DDRL combines the distributional representation of DistRL with the deterministic update dynamics of DetRL. Agents discretize the return spectrum using a finite population of reward neurons, consistent with empirical findings on distributional coding [28]. We apply our framework to the optimism heuristic. Optimism is implemented as a distortion of the learned return distribution. Our deterministic dynamics describe the expected trajectory of a learning agent, averaging out the intrinsic noise of the algorithmic learning update due to stochastic strategies, transitions, and rewards. DDRL is therefore a novel method of analysis for distributional multi-agent RL [39]. We validate it by testing whether its predictions (the fixed points, their stability, and bifurcations) reproduce established behavioral findings. We do so across three canonical task domains: a bandit choice task under resource variability (Section 3.2), a social coordination game (Section 3.3), and an intertemporal risky choice problem (Section 3.4). Our results replicate established behavioral findings in each domain. They also reveal novel interactions among optimism, return discretization, and temporal discounting—specifically, bistable strategy regimes and a form of incoherent individual choice we term the individual dilemma. This constitutes a step toward a unified, biologically grounded framework for studying behavior emerging from cognition across collective decision-making contexts.

2. Methods

2.1. Stochastic games

Stochastic games, also known as Markov games, are defined by the elements $\langle N, \mathcal{S}, \boldsymbol{\mathcal{A}}, T, \boldsymbol{R}\rangle$. $N \in \mathbb{N}$ agents reside in an environment of $Z \in \mathbb{N}$ states $\mathcal{S} = (S_1, ..., S_Z)$. In each state s, each agent $i \in \mathcal{N} = \{1, ..., N\}$ has $M \in \mathbb{N}$ available actions $\mathcal{A}^i = (A^i_1, ..., A^i_M)$ to choose from. $\boldsymbol{\mathcal{A}} = \bigotimes_i \mathcal{A}^i$ is the joint-action set where $\bigotimes_i$ denotes the Cartesian product over the sets indexed by i. Time advances in discrete steps, and agents choose their actions simultaneously. A joint action is denoted by $\boldsymbol{a} = (a^1, ..., a^N) \in \boldsymbol{\mathcal{A}}$. With $\boldsymbol{a}^{-i} = (a^1, ..., a^{i-1}, a^{i+1}, ..., a^N)$ we denote the joint action except agent i's, and we write the joint action in which agent i chooses a^i and all other agents choose $\boldsymbol{a}^{-i}$ as $a^i\boldsymbol{a}^{-i}$. We choose an equal number of actions for all states and agents out of notational convenience.

The **transition** tensor $\boldsymbol{T} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ determines the probabilistic state change. $T^{s,\boldsymbol{a},\acute{s}}$ is the transition probability from current state s to next state $\acute{s}$ under joint action $\boldsymbol{a}$.

The **reward** tensor $\boldsymbol{R} : \mathcal{N} \times \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to \mathbb{R}$ maps the triple of current state s, joint action $\boldsymbol{a}$ and next state $\acute{s}$ to an immediate reward scalar for each agent. $R^{i,s,\boldsymbol{a},\acute{s}}$ is the reward agent i receives.

At each time step, actions are chosen from the joint **strategy** tensor $\boldsymbol{X} : \mathcal{N} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$. In state s, agent i chooses action a^i with probability X^{i,s,a^i}.

Each agent i aims for a strategy X^{i,s,a^i} to maximize their **gain**,

UNCORRECTED PROOF

$$G_t^i = (1-\gamma^i)\sum_{k=0}^{\infty}(\gamma^i)^k r_{t+k}^i \tag{1}$$

where r_{t+k}^i denotes the reward agent i obtained at time step $t+k$. The discount factor, $\gamma^i \in [0, 1)$, denotes how much agent i cares for future rewards.

2.2. Reinforcement learning

Reinforcement learning is a trial-and-error method of mapping situations to actions to maximize a numerical reward signal [9]. When rewards are a delayed consequence of current actions, so-called temporal-difference or reward-prediction learning has been particularly influential. This type of learning summarizes the difference between value estimates from past and present experiences into a reward-prediction error, which is then used to adapt the current behavior to gain more rewards over time. There also exist remarkable similarities between computational reinforcement learning and the results of neuroscientific experiments [13]. Dopamine conveys reward-prediction errors to brain structures where learning and decision-making occur [12].

At time step t, in state s, agent i evaluates action a^i to be of quality Q_t^{i,s,a^i}. Learning then means updating these state-action value estimates, Q_t^{i,s,a^i}, after selecting action a_t^i and having observed state s_t according to

$$Q_{t+1}^{i,s_t,a_t^i} = Q_t^{i,s_t,a_t^i} + \alpha^i E_t^{i,s_t,a_t^i}, \tag{2}$$

where $\alpha^i \in [0, 1]$ denotes agent i's learning rate, which regulates how much new information the agent uses for the update, and E_t^{i,s_t,a_t^i} the current temporal-difference or reward-prediction error. The expression of the temporal-difference error depends on the chosen RL algorithm.

In this work, we focus on two different RL variants: SARSA learning, and Actor-Critic learning. SARSA learning takes into account the five elements of the current State, current Action, Reward, next State, and next Action for one learning update step. In Actor-Critic learning, an approximation of the State-Action value gets "criticized" by (i.e., compared against) an approximation of the state value. Another well-known variant is Q-learning. A comparison of these three variants can be found in [33].

In SARSA learning, the temporal-difference error is expressed as:

$$E_t^{i,s_t,a_t^i} = (1-\gamma^i)r_t^i + \gamma^i Q_t^{i,s_{t+1},a_{t+1}^i} - Q_t^{i,s_t,a_t^i}. \tag{3}$$

In Actor-Critic learning, by contrast, the temporal-difference error retains the same one-step bootstrap target $(1-\gamma^i)r_t^i + \gamma^i Q_t^{i,s_{t+1},a_{t+1}^i}$ as in SARSA, but replaces the baseline Q_t^{i,s_t,a_t^i} with the state value V_t^{i,s_t},

$$E_t^{i,s_t,a_t^i} = (1-\gamma^i)r_t^i + \gamma^i Q_t^{i,s_{t+1},a_{t+1}^i} - V_t^{i,s_t}. \tag{4}$$

The state value at time t, V_t^{i,s_t}, quantifies the overall quality of state s at time t, regardless of the chosen action a_t. In Actor-Critic learning, it is separately updated according to:

$$V_{t+1}^{i,s_t} = V_t^{i,s_t} + \alpha^i E_t^{i,s_t,a_t^i}. \tag{5}$$

Agents select actions based on the current state-action value beliefs Q_t^{i,s_t,a_t^i}, balancing exploitation (i.e., choosing the action of maximum quality) and exploration (i.e., selecting lower-quality actions to learn more about the environment). We employ the widely used Boltzmann policy function. The probability of choosing action a^i under observing state s is

UNCORRECTED PROOF

$$X_t^{i,s,a^i} = \frac{\exp\left(\beta^i Q_t^{i,s,a^i}\right)}{\sum_{b\in\mathcal{A}^i}\exp\left(\beta^i Q_t^{i,s,b}\right)}, \tag{6}$$

where the *intensity-of-choice* parameters, $\beta^i \in \mathbb{R}^+$, regulate the exploration-exploitation trade-off. For high β^i, agents exploit their learned knowledge about the environment, leaning toward actions with high estimated state-action values. For low β^i, agents are more likely to deviate from these high-value actions to explore the environment further with the chance of finding actions that eventually lead to even higher values. Our choice of the Boltzmann policy is motivated by the maximum-entropy principle [40]. It is the distribution over actions that maximizes entropy subject to a constraint on the expected Q-value $\sum_{a^i} X_t^{i,s,a^i} Q_t^{i,s,a^i}$ or equivalently, the distribution that maximizes the expected Q-value subject to a constraint on entropy, with β^i the Lagrange multiplier coupling the two objectives [41]. It also resembles the idea of reward-matching behavior in animal and human choice [42,43]. From a modeling standpoint, the Boltzmann policy is particularly convenient for the deterministic dynamics framework of Section 2.4: because the strategy update depends only on differences in Q-values across actions, action-constant quantities cancel, substantially simplifying the dynamics. Alternative exploration policies such as ε-greedy can also be combined with strategy averaging [38], but at the cost of a larger dynamical state space; more elaborate schemes such as UCB, which track uncertainty explicitly, complicate the dynamics further still, to the point of being analytically intractable unless the exploration process itself is the primary object of study.

2.3. Distributional reinforcement learning

Distributional approaches to RL (DistRL) in artificial intelligence and machine learning research were pioneered by [44,45] and recently gained wider prominence [27]. DistRL diverges from the classical approach insofar as learning agents do not compute expected values, but rather the full distribution of outcomes. Therefore, the goal is not to compute an expected value or quality, but rather a **return distribution**, which is described by a distributional Bellman equation [27]:

$$\mathcal{Q}^{s,a} \overset{D}{=} (1-\gamma)R^{s,a} + \gamma\mathcal{Q}^{\acute{s},\acute{a}} \tag{7}$$

where $R^{s,a}$ is the reward for action a in state s, $(\acute{s},\acute{a})$ is the next-state action, and $\mathcal{Q}$ is its random return. The $(1-\gamma)$ factor reflects the normalized-gain convention used throughout this paper, $G_t = (1-\gamma)\sum_{k\geq 0}\gamma^k r_{t+k}$.

Over the past years, DistRL researchers have used various supports for discretizing the distributions. In categorical DistRL, the discretization support is a set $\{z_i = V_{\min} + i\Delta z : 0 \leq i < N\}$ of N categorical "atoms," which are fixed and evenly spaced (with $\Delta z = \frac{V_{\max}-V_{\min}}{N-1}$) [27]. Later versions of DistRL algorithms rely on the distributions' statistics as discretization support: the atoms are not fixed anymore, but are updated so as to converge towards quantiles or expectiles of the target distribution [46].

In this work, we will use a categorical discretization support because it maps naturally onto a fixed population of neurons, each tuned to a specific return value. In standard categorical DistRL, atoms span the range of possible returns $[V_{\min}, V_{\max}]$. Under the normalized gain convention (Eq. 1) used throughout this paper, returns are convex combinations of rewards and therefore lie in $[r_1, r_n]$. Our atoms, defined at reward values $r_1, \ldots, r_n$, cover this range exactly. Quantile- and expectile-based representations use adaptive atoms that shift during learning [46,47], which improves approximation quality but would render the deterministic dynamics derivation of this paper incompatible, as the representational state itself would become time-varying.

2.4. Deterministic learning dynamics

Deterministic dynamics of RL remove noise from the learning process. Hence, the agents' strategies over time can be computed and visualized as deterministic trajectories in phase space. To this end, we use temporal-difference RL in the deterministic limit (DetRL), which was developed by Barfuss *et al.*[33],. This method uses strategy-averages as a way to

UNCORRECTED PROOF

remove the noise from the learning dynamics: the prediction error is averaged over all strategies and state transitions. With a formalism analogous to evolutionary game theory [35,36], deterministic RL dynamics yield analytically tractable trajectories in strategy space and enable the exact characterization of fixed points, their stability, and bifurcations — properties we exploit to identify bistable choice regimes in Section 3.4. More precisely, SARSA dynamics with a Boltzmann strategy function are analogous to the replicator-mutator equation [33] — a mutation term prevents strategies from collapsing to a pure strategy — whereas Actor-Critic dynamics are analogous to the pure replicator equation [35], which drives strategies toward deterministic fixed points. Thus, deterministic RL dynamics can be viewed as distinct level of analysis of multi-agent learning [39]. Stochastic RL algorithms form their noisy counterparts: the stationary distribution of the stochastic system concentrates around the stable fixed points of the deterministic part and spends negligible time near unstable ones, provided the noise is sufficiently weak and the deterministic dynamics have isolated stable fixed points rather than more complex attractors [38,39]. Under these conditions, the deterministic approximation captures average learning behavior while abstracting away intrinsic fluctuations. A related but conceptually distinct line of work is the ODE method for stochastic approximation [48,49], which links stochastic learning algorithms to a limiting ordinary differential equation in the vanishing step-size regime. There, the ODE serves as an asymptotic tool to establish convergence of the stochastic recursion to its invariant sets. These are precisely the conditions under which the link between deterministic dynamics and stochastic algorithms holds. In contrast, however, we do not take a vanishing learning-rate limit, but derive the induced deterministic update dynamics directly at finite step sizes (Section 3.1), treating their discrete-time behavior, including fixed points, stability, and bifurcations, as objects of intrinsic interest rather than as auxiliary constructs for convergence analysis.

First, we compute the different quantities we need to calculate the average temporal-difference error. As explained in Section 2.1, X^{i,s,a^i} denotes the probability that agent i chooses action a^i in state s; $\boldsymbol{X} : \mathcal{N} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the joint strategy tensor.

We call $R_{\mathbf{X}}^{i,s,a^i}$ the average reward given the joint policy $\mathbf{X}$ that agent i obtains by taking action a^i in state s. More precisely, the average is taken over state transitions from s to $\acute{s}$, and over other agents' actions $\boldsymbol{a}^{-i}$:

$$R_{\mathbf{X}}^{i,s,a^i} = \sum_{\acute{s}} \sum_{\boldsymbol{a}^{-i}} \prod_{j \neq i} X^{j,s,a^j} T^{s,\boldsymbol{a},\acute{s}} R^{i,s,\boldsymbol{a},\acute{s}}. \tag{8}$$

Similarly, we call ${}^{next}Q_{\mathbf{X}}^{i,s,a^i}$ the average next-state Q-value given the joint policy $\mathbf{X}$ for agent i taking action a^i in state s:

$${}^{next}Q_{\mathbf{X}}^{i,s,a^i} = \sum_{\acute{s}} \sum_{\boldsymbol{a}^{-i}} \prod_{j \neq i} X^{j,s,a^j} T^{s,\boldsymbol{a},\acute{s}} \sum_{b^i} X^{i,\acute{s},b^i} Q_{\mathbf{X}}^{i,\acute{s},b^i}, \tag{9}$$

where $Q_{\mathbf{X}}^{i,\acute{s},b^i}$ denotes the average Q-value of action b^i taken by agent i in state $\acute{s}$, described by the recursive Bellman equation:

$$Q_{\mathbf{X}}^{i,\acute{s},b^i} = (1 - \gamma^i) R_{\mathbf{X}}^{i,\acute{s},b^i} + \gamma^i \cdot {}^{next}Q_{\mathbf{X}}^{i,\acute{s},b^i} \tag{10}$$

Throughout this section, a subscript $\mathbf{X}_t$ denotes the strategy-averaged quantity evaluated at the joint strategy at time t. In the deterministic approximation, these are the true on-strategy expectations under $\mathbf{X}_t$, not estimates, so $Q_{\mathbf{X}_t}^{i,s,a^i}$ satisfies the above Bellman recursion exactly. The time index t is therefore inherited entirely through $\mathbf{X}_t$, and substitutions that combine $(1 - \gamma^i) R_{\mathbf{X}_t} + \gamma^i {}^{next}Q_{\mathbf{X}_t}$ into $Q_{\mathbf{X}_t}$ are identities by construction. Unlike Eqs. 3 and 4, where t indexes the single state s_t and action a_t^i sampled at that step and only Q_t^{i,s_t,a_t^i} is updated, here s, a^i range over every state-action pair. All get updated at each step t together as a consequence of the joint strategy update (Eq. 13). The joint strategy $\mathbf{X}_t$ is the sole dynamic variable. Only it carries a time index t.

UNCORRECTED PROOF

Therefore, for a SARSA agent, the average prediction error for agent i taking action a^i in state s at time t is

$$\begin{aligned} E^{i,s,a^i}_{\mathbf{x}_t} &= (1-\gamma^i)R^{i,s,a^i}_{\mathbf{x}_t} + \gamma^i \cdot {}^{next}Q^{i,s,a^i}_{\mathbf{x}_t} - \frac{1}{\beta^i}\log X^{i,s,a^i}_t \\ &= Q^{i,s,a^i}_{\mathbf{x}_t} - \frac{1}{\beta^i}\log X^{i,s,a^i}_t . \end{aligned} \tag{11}$$

where β^i is the choice intensity, or inverse temperature, of agent i.

For an Actor-Critic agent, by contrast, the average Q-value $Q^{i,s,a^i}_{\mathbf{x}}$ is compared with the average state value $V^{i,s}_{\mathbf{x}}$ (the "critic"). Since this term is constant in action, we can omit it from the formula [33]. Thus, for an Actor-Critic agent, the average prediction error for agent i taking action a^i in state s at time t is:

$$\begin{aligned} E^{i,s,a^i}_{\mathbf{x}_t} &= (1-\gamma^i)R^{i,s,a^i}_{\mathbf{x}_t} + \gamma^i \cdot {}^{next}Q^{i,s,a^i}_{\mathbf{x}_t} \\ &= Q^{i,s,a^i}_{\mathbf{x}_t} . \end{aligned} \tag{12}$$

The combination of the Q-value update (Eq. 2) and the Boltzmann formula (Eq. 6) yields the deterministic update of strategy X^{i,s,a^i}, applied to every (s, a^i) pair at once:

$$X^{i,s,a^i}_{t+1} = \frac{X^{i,s,a^i}_t \exp[\alpha^i\beta^i E^{i,s,a^i}_t]}{\sum_{b^i} X^{i,s,b^i}_t \exp[\alpha^i\beta^i E^{i,s,b^i}_t]} \tag{13}$$

3. Results

3.1. Theoretical framework

In this section, we derive deterministic distributional reinforcement learning dynamics. In order to model the optimism-pessimism heuristic, we apply deterministic learning dynamics to distributional RL. By doing so, we combine the best of both worlds: on the one hand, DetRL provides a transparent, reproducible, and computationally fast approach to RL; on the other hand, DistRL allows us to bias the return distribution of the learning dynamics and devise a simple model at the neuronal level.

Our approach is presented in Fig 1. The agent-environment interface is formalized as a Markov game $\langle N, \mathcal{S}, \mathcal{A}, T, \boldsymbol{R}\rangle$ comprised of N agents, a set $\mathcal{S}$ of different states, a set $\mathcal{A}$ of available actions, a transition tensor T, and a reward tensor $\boldsymbol{R}$. In DDRL, rewards and qualities are not computed as expected values, but as full distributions. These distributions are discretized according to a categorical support and a smoothing procedure. Then, the Q-return distribution is distorted according to a weighting algorithm, implementing the optimism heuristic.

In the following subsections, we explain how we discretize and compute the relevant distributions.

3.1.1. Discretization and reward distribution. In contrast to standard distributional RL, we already discretize the rewards and not just the returns. We do so to be able to compute the return distribution in closed form (Sect. 3.1.4), which is necessary to formulate the deterministic dynamics. We assume an agent has $n \in \mathbb{N}$ reward "neurons," each neuron being sensitive to a given range of amount of reward. Throughout this work we use $n = 29$, balancing accuracy and computational cost across environments (see S1 Text, Part IV). We define $r_1 \in \mathbb{R}$ (*resp.* $r_n \in \mathbb{R}$) the minimal (*resp.* maximal) amount of reward that can be detected. Similarly to categorical DistRL [46], we assume a regular grid $(r_k)_{k\in[1,n]}$ such that $r_k = r_1 + (k-1)\Delta r$, where $\Delta r = \frac{r_n - r_1}{n-1}$ denotes the width separating two consecutive grid points.

Fig 1 describes the process of discretization by reward neurons. In particular, it shows how an arbitrary reward amount $\tilde{r}$ gets represented as a discretized distribution by defining a Gaussian kernel centered around $\tilde{r}$. To smooth out the effects

UNCORRECTED PROOF

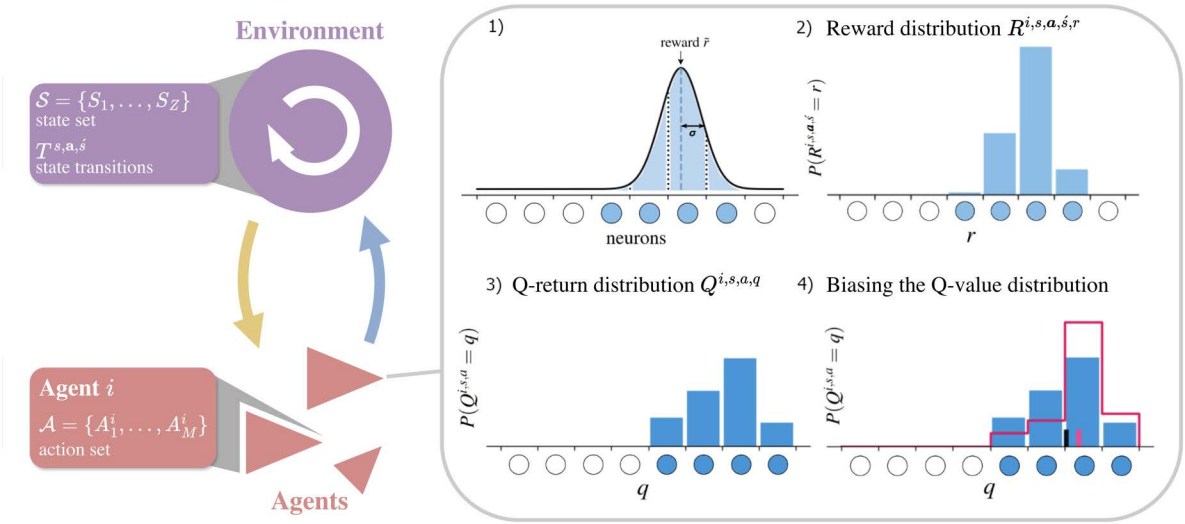

Fig 1. The agent-environment interface. N agents play in an environment formalized as a set of Z states $\mathcal{S} = \{S_1, ..., S_Z\}$. In each state, agent i can choose between M actions from the action set $\mathcal{A} = \{A_1^i, ..., A_M^i\}$. The probability that a transition from state s to state $\acute{s}$ occurs when agents play the joint action $\boldsymbol{a}$ is $T^{s,\boldsymbol{a},\acute{s}}$. The right panel represents reward discretization and Q-return distribution estimation for one agent i. 1) An environmental reward quantity, $\tilde{r}$, is detected with smoothing width σ by several neurons (in light blue). Each neuron k represents the probability that the environmental reward equals a quantity $r_k \pm \Delta r$; this probability is the area, taken between $r_k - \Delta r$ and $r_k + \Delta r$, under the Gaussian curve centered around $\tilde{r}$ with standard deviation σ. 2) This yields the discretized reward distribution $R^{i,s,\boldsymbol{a},\acute{s},r}$. In DDRL, all quantities are no longer calculated as average estimates, but rather as discretized distributions. 3) Example Q-return distribution, which is calculated according to our DDRL dynamics (Section 3.1). 4) The optimistic heuristic is applied to the Q-return distribution (in blue). It transforms the original distribution into a distorted one (outlined in magenta). The mean of this new distribution (magenta bar) is an optimistic estimate of the average Q-value (black bar).

https://doi.org/10.1371/journal.pcbi.1014723.g001

of discretization, we assume that each neuron k is maximally sensitive to the reward amount r_k, quantified by a smoothing width, σ. Each reward neuron r_k "fires" with intensity $P_\sigma(r_k|\tilde{r})$, which quantifies the probability that $r_k - \frac{\Delta r}{2} < \tilde{r} < r_k + \frac{\Delta r}{2}$. More precisely, $P_\sigma(r_k|\tilde{r})$ is calculated as the area between $r_k - \frac{\Delta r}{2}$ and $r_k + \frac{\Delta r}{2}$ under the Gaussian bell curve centered around $\tilde{r}$ and with standard deviation σ,

$$P_\sigma(r_k|\tilde{r}) = \int_{r_k - \frac{\Delta r}{2}}^{r_k + \frac{\Delta r}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\rho - \tilde{r})^2}{2\sigma^2}\right] d\rho. \tag{14}$$

Therefore, this smoothing procedure captures some uncertainty about the categorization into different bins. This uncertainty is modulated by the smoothing width, σ. This smoothing is a soft projection: each atom or neuron r_k receives probability mass $P_\sigma(r_k|\tilde{r})$ rather than all mass going to the nearest neuron. It is analogous to label smoothing in supervised learning [50] and has been applied in distributional RL by Farebrother et al. [51]. Biologically, it reflects the overlapping tuning curves of reward-sensitive neurons, which respond to nearby stimuli [52,53]. Ultimately, this discretization method allows us to transform any stochastic game's reward tensor into a reward distribution. For all $r \in \{r_1, ..., r_n\}$, we define the reward distribution $R^{i,s,\boldsymbol{a},\acute{s},r}$ for agent i, current state s, joint action $\boldsymbol{a}$, and next state $\acute{s}$, as:

$$R^{i,s,\boldsymbol{a},\acute{s},r} = P_\sigma(r|R^{i,s,\boldsymbol{a},\acute{s}}). \tag{15}$$

Note that each $R^{i,s,\boldsymbol{a},\acute{s}}$ is an entry of the original reward tensor $\boldsymbol{R}$. In other terms, $R^{i,s,\boldsymbol{a},\acute{s},r}$ is the area between $r - \frac{\Delta r}{2}$ and $r + \frac{\Delta r}{2}$ under the bell curve centered around $R^{i,s,\boldsymbol{a},\acute{s}}$.

UNCORRECTED PROOF

3.1.2. Reward distributions $R_{\mathbf{X}}^{i,s,a^i,r}$ and $R_{\mathbf{X}}^{i,s,r}$. The previous reward distribution can be averaged over joint actions $\boldsymbol{a}$ and next states $\acute{s}$. We call $R_{\mathbf{X}}^{i,s,a^i,r}$ the reward distribution for agent i taking action a^i in state s, averaged over transitions to the next state $\acute{s}$, and over other agents' actions $\boldsymbol{a}^{-i}$:

$$R_{\mathbf{X}}^{i,s,a^i,r} = \sum_{\acute{s}} \sum_{\boldsymbol{a}^{-i}} \prod_{j \neq i} X^{j,s,a^j} T^{s,\boldsymbol{a},\acute{s}} R^{i,s,\boldsymbol{a},\acute{s},r}. \tag{16}$$

The reward distribution $R_{\mathbf{X}}^{i,s,r}$ denotes the probability that agent i receives reward r in state s. It is averaged over joint actions $\boldsymbol{a}$ (including agent i's) and next states $\acute{s}$:

$$R_{\mathbf{X}}^{i,s,r} = \sum_{\acute{s}} \sum_{\boldsymbol{a}} \prod_{j} X^{j,s,a^j} \cdot T^{s,\boldsymbol{a},\acute{s}} \cdot R^{i,s,\boldsymbol{a},\acute{s},r}. \tag{17}$$

3.1.3. Bellman equation and smoothed Bellman Kronecker $\delta^{i,v,r,\acute{v}}$. In traditional, expected-value RL, the value of a state s for agent i is computed via the Bellman equation:

$$V_{\mathbf{X}}^{i,s} = (1-\gamma^i) R_{\mathbf{X}}^{i,s} + \gamma^i \sum_{\acute{s}} T_{\mathbf{X}}^{s,\acute{s}} V_{\mathbf{X}}^{i,\acute{s}}$$

where $V_{\mathbf{X}}^{i,\acute{s}}$ is the value of the next state, $\acute{s}$, and $T_{\mathbf{X}}^{s,\acute{s}}$ is the effective state transition matrix given joint policy $\mathbf{X}$.

In distributional form, we want to compute the probability that (discretized) values r, v and $\acute{v} \in \{r_1, ..., r_n\}$ satisfy the Bellman equation $v = (1-\gamma^i)r + \gamma^i \acute{v}$. To this end, we introduce a smoothed Bellman Kronecker delta,

$$\delta^{i,v,r,\acute{v}} = P_\sigma(v | (1-\gamma^i)r + \gamma^i \acute{v}), \tag{18}$$

using the same discretization method as above. The idea is that $\delta^{i,v,r,\acute{v}}$ peaks when the Bellman condition $v = (1-\gamma^i)r + \gamma^i \acute{v}$ is fulfilled for agent i, and its reward bins v, r, and $\acute{v}$, allowing for some uncertainty, modulated by the smoothing width, σ, around the discretization into different bins. We use this smoothed Bellman Kronecker delta to compute the V-return distribution $V_{\mathbf{X}}^{i,s,v}$ in closed form.

3.1.4. V-return distribution $V_{\mathbf{X}}^{i,s,v}$. The V-return distribution $V_{\mathbf{X}}^{i,s,v}$ represents the probability that agent i in state s receives a long-term return v. It results from four instances that must occur together. For each short-term reward bin $r \in \mathcal{R}$, for all possible next states $\acute{s} \in \mathcal{S}$, and each long-term return bin $\acute{v} \in \mathcal{R}$, i) v, r and $\acute{v}$ must fulfill the Bellman equation according to probability $\delta^{i,v,r,\acute{v}}$, ii) agent i must receive short-term reward r in state s, according to probability $R_{\mathbf{X}}^{i,s,r}$, iii) state s must transition to state $\acute{s}$, according to probability $T_{\mathbf{X}}^{s,\acute{s}}$, and finally iv) in state $\acute{s}$, agent i must receive a long-term return $\acute{v}$. More concisely,

$$V_{\mathbf{X}}^{i,s,v} = \sum_{r,\acute{v},\acute{s}} \delta^{i,v,r,\acute{v}} R_{\mathbf{X}}^{i,s,r} T_{\mathbf{X}}^{s,\acute{s}} V_{\mathbf{X}}^{i,\acute{s},\acute{v}}. \tag{19}$$

We now summarize the smoothed Bellman-Kronecker delta, reward, and transition probabilities into a tensor $M_{\mathbf{X}}^{i,s,v,\acute{s},\acute{v}} = \sum_r \delta^{i,v,r,\acute{v}} R_{\mathbf{X}}^{i,s,r} T_{\mathbf{X}}^{s,\acute{s}}$. Doing so, we can rewrite Eq. 19 as $V_{\mathbf{X}}^{i,s,v} = \sum_{\acute{v},\acute{s}} M_{\mathbf{X}}^{i,s,v,\acute{s},\acute{v}} V_{\mathbf{X}}^{i,\acute{s},\acute{v}}$. We now combine the indices $s \times v = w$ and $\acute{s} \times \acute{v} = \acute{w}$, which transforms Eq. 19 into,

$$V_{\mathbf{X}}^{i,w} = \sum_{\acute{w}} M_{\mathbf{X}}^{i,w,\acute{w}} V_{\mathbf{X}}^{i,\acute{w}} \tag{20}$$

UNCORRECTED PROOF

In other terms, for each agent i and joint policy X, the V-return distribution $\mathbf{V}^i_{\mathbf{X}}$ fulfills $\mathbf{V}^i_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}\mathbf{V}^i_{\mathbf{X}}$. Thus, $\mathbf{V}^i_{\mathbf{X}}$ is the eigenvector of the matrix $\mathbf{M}_{\mathbf{X}}$ corresponding to eigenvalue 1, normalized in such a way that for each state s, the probabilities to receive values $v \in \mathcal{R}$ sum up to 1, $\sum_v V^{i,s,v}_{\mathbf{X}} = 1$.

3.1.5. Q-return distribution $Q^{i,s,a^i,v}_{\mathbf{X}}$. The probability for agent i to receive long-term return v in state s under action a^i is the joint probability that from all possible short-term reward values r, all possible next states $\acute{s}$, and all possible next long-term returns $\acute{v}$, i) v, r and $\acute{v}$ fulfill the Bellman equation according to probability $\delta^{i,v,r,\acute{v}}$, ii) agent i receives short-term reward r in state s under action a^i, according to probability $R^{i,s,a^i,r}_{\mathbf{X}}$, iii) state s transitions to state $\acute{s}$ under agent i's action a^i, according to probability $T^{i,s,a^i,\acute{s}}_{\mathbf{X}}$ (the effective transition given joint strategy $\mathbf{X}$, marginalized over the other agents' actions), and iv) in state $\acute{s}$, agent i receives a long-term return $\acute{v}$, according to probability $V^{i,\acute{s},\acute{v}}_{\mathbf{X}}$. More concisely,

$$Q^{i,s,a^i,v}_{\mathbf{X}} = \sum_{\acute{s},r,\acute{v}} \delta^{i,v,r,\acute{v}} \cdot R^{i,s,a^i,r}_{\mathbf{X}} \cdot T^{i,s,a^i,\acute{s}}_{\mathbf{X}} \cdot V^{i,\acute{s},\acute{v}}_{\mathbf{X}} \tag{21}$$

3.1.6. Temporal-difference error. The SARSA and Actor-Critic temporal-difference errors differ slightly (see Section 2.4). Denoting $\langle . \rangle_d$ the mean of a distribution on dimension d, the deterministic approximation of the temporal-difference error at time t for a SARSA agent i taking action a^i in state s, $E^{i,s,a^i}_{\mathbf{X}_t}$, is:

$$E^{i,s,a^i}_{\mathbf{X}_t} = \langle Q^{i,s,a^i,v}_{\mathbf{X}_t} \rangle_v - \frac{1}{\beta^i} \cdot \log X^{i,s,a^i}_t. \tag{22}$$

For an Actor-Critic agent, the last term is not needed, as the "critic" (*i.e.*, $\langle V^{i,s,v}_{\mathbf{X}_t} \rangle_v$) is constant in action (see Section 2.4). The remaining $\langle Q^{i,s,a^i,v}_{\mathbf{X}_t} \rangle_v$ in Eq. 23 is therefore not the full TD error in the classical sense; it is the action-dependent component, which yields the same deterministic strategy update as the full TD error because the omitted critic is constant across actions and cancels in the Boltzmann update. Therefore, the deterministic approximation of the temporal-difference error at time t for an Actor-Critic agent i taking action a^i in state s, $E^{i,s,a^i}_{\mathbf{X}_t}$, is:

$$E^{i,s,a^i}_{\mathbf{X}_t} = \langle Q^{i,s,a^i,v}_{\mathbf{X}_t} \rangle_v. \tag{23}$$

3.1.7. Biasing the Q-return distribution. Optimism (*resp.* pessimism) causes agents to overestimate (*resp.* underestimate) Q-values. There are different ways to obtain an optimistic (*resp.* pessimistic) estimate of a distribution. Here, we adopt a weights method: return distribution mass is rescaled by w_U above the threshold $V^{i,s}_{\mathbf{X}}$ and by unity below (reversed for pessimism). The primary motivation is a formal equivalence with asymmetric RL at a two-armed bandit choice task: a weights-DDRL agent with upper weight w_U reaches the same steady-state Q-values as an asymmetric-RL agent with bias ratio $b = \alpha^+/\alpha^- = w_U$, where α^+ and α^- are the learning rates applied to positive and negative temporal-difference errors, respectively, (proven in S1 Text, Section 12.2), directly connecting this probability distortion to the established cognitive science literature [54]. In S1 Text, Section 12.1 we compare this against quantile- and expectile-based distortions of the return distribution [29]. While the qualitative outcomes are similar, these other distortions exhibit a less smooth parameter dependence as a result of our categorical discretization. In contrast, our weights method yields smooth, monotonic dependence on w_U.

The weights method involves applying a weights tensor, $\boldsymbol{w}^{i,s,a^i,v}$, to the discretized distribution $Q^{i,s,a^i,v}_{\mathbf{X}}$ so as to distort it. Although there are different ways to distort a distribution, we use a stepwise function for simplicity, and because it is equivalent with asymmetric RL on a two-armed bandit task (see S1 Text, Section 12.2). For agent i in state s, the step is located at the scalar $V^{i,s}_{\mathbf{X}}$ defined below: the mean of the action-averaged return distribution. When the agent is optimistic (*resp.* pessimistic), a higher weight is put on the part of the distribution which is above (*resp.* below) that step.

UNCORRECTED PROOF

We define the threshold $V_{\mathbf{X}}^{i,s}$ as the expected value of the action-averaged return distribution at state s under joint strategy $\mathbf{X}$. Marginalizing $Q_{\mathbf{X}}^{i,s,a^i,v}$ over agent i's strategy yields the state-value distribution $V_{\mathbf{X}}^{i,s,v} = \sum_{a^i} X^{i,s,a^i} Q_{\mathbf{X}}^{i,s,a^i,v}$ (consistent with Eq. 19); taking its mean over v gives the scalar threshold,

$$V_{\mathbf{X}}^{i,s} = \langle V_{\mathbf{X}}^{i,s,v} \rangle_v = \sum_{v \in \mathcal{R}} v \sum_{a^i \in \mathcal{A}^i} X^{i,s,a^i} Q_{\mathbf{X}}^{i,s,a^i,v}. \tag{24}$$

We define the weights tensor $\boldsymbol{w}^{i,s,a^i,r_k}$, $r_k \in \mathcal{R}$, as follows. For all agent i, state s, action a^i, and neuron r_k,

- $\boldsymbol{w}^{i,s,a^i,r_k} = 1$ if $r_k < V_{\mathbf{X}}^{i,s}$;
- $\boldsymbol{w}^{i,s,a^i,r_k} = w_U$ if $r_k \geq V_{\mathbf{X}}^{i,s}$,

where $w_U < 1$ if the agent is pessimistic, and $w_U > 1$ if it is optimistic.

We obtain the weighted estimate $\langle Q^{i,s,a^i,v} \rangle_{\mathbf{w}}$ of distribution $Q_{\mathbf{X}}^{i,s,a^i,v}$ by computing the product of the distribution with the weights vector, normalizing this product by a normalization factor N^{i,s,a^i}, and computing the mean $\langle Q_{\mathbf{X}}^{i,s,a^i,v} \rangle_{\mathbf{w}}$ of the resulting distribution:

$$\langle Q_{\mathbf{X}}^{i,s,a^i,v} \rangle_{\mathbf{w}} = \frac{1}{N^{i,s,a^i}} \sum_{r_k \in \mathcal{R}} r_k \cdot \boldsymbol{w}^{i,s,a^i,r_k} \cdot Q_{\mathbf{X}}^{i,s,a^i,r_k} \tag{25}$$

The normalization factor, $N^{i,s,a^i} = \sum_{r_k \in \mathcal{R}} \boldsymbol{w}^{i,s,a^i,r_k} \cdot Q_{\mathbf{X}}^{i,s,a^i,r_k}$, helps us make sure the integral of the weighted distribution over $\mathcal{R}$ is 1.

To validate our framework and highlight its usefulness for uncovering new insights, we test our model across three different choice environments: an individual bandit choice task under resource variability, a social coordination challenge, and an intertemporal risky choice problem.

3.2. Example application: Bandit choice under resource variability

Here, we use our model to confirm previous findings showing that the influence of optimism on performance in a two-armed bandit task is modulated by resource scarcity. The two-armed bandit is a standard task in cognitive psychology to study asymptotic strategy formation between options with different reward probabilities. An agent must choose between two options, a_1 and a_2. a_1 gives a reward +1 with probability p_1, and a punishment –1 with probability $1 - p_1$; a_2 gives a reward +1 with probability p_2, and a punishment –1 with probability $1 - p_2$. By repeatedly choosing between the two options, the agent must learn which is most rewarding on average (Fig 2A). We use SARSA rather than Actor-Critic dynamics here because Actor-Critic dynamics (analogous to the pure replicator equation), drive agents toward deterministic strategies, whereas SARSA (analogous to the replicator-mutator equation) admits mixed asymptotic strategies, as observed in previous research on this task [54].

In previous research using two-armed bandit tasks, it has been established via computational modeling that optimism (implemented as an asymmetric pair of learning rates) increases or decreases an agent's performance depending on resource scarcity. Optimism is beneficial for performance in scarce environments, and detrimental in abundant environments [54]. Empirical research on human decision-making also revealed an asymmetric sensitivity to positive and negative prediction errors. Individual differences in the ratio of positive to negative learning rates predict risk-seeking behavior in bandit tasks [55] and reflect variation in dopaminergic signaling [56,57]. Humans tend toward a stable positivity bias across environments rather than adaptively modulating this asymmetry with resource scarcity [58,59], consistent with the non-adaptive optimism weights in DDRL. Therefore, we tested our framework on a "poor" two-armed bandit, in which the

UNCORRECTED PROOF

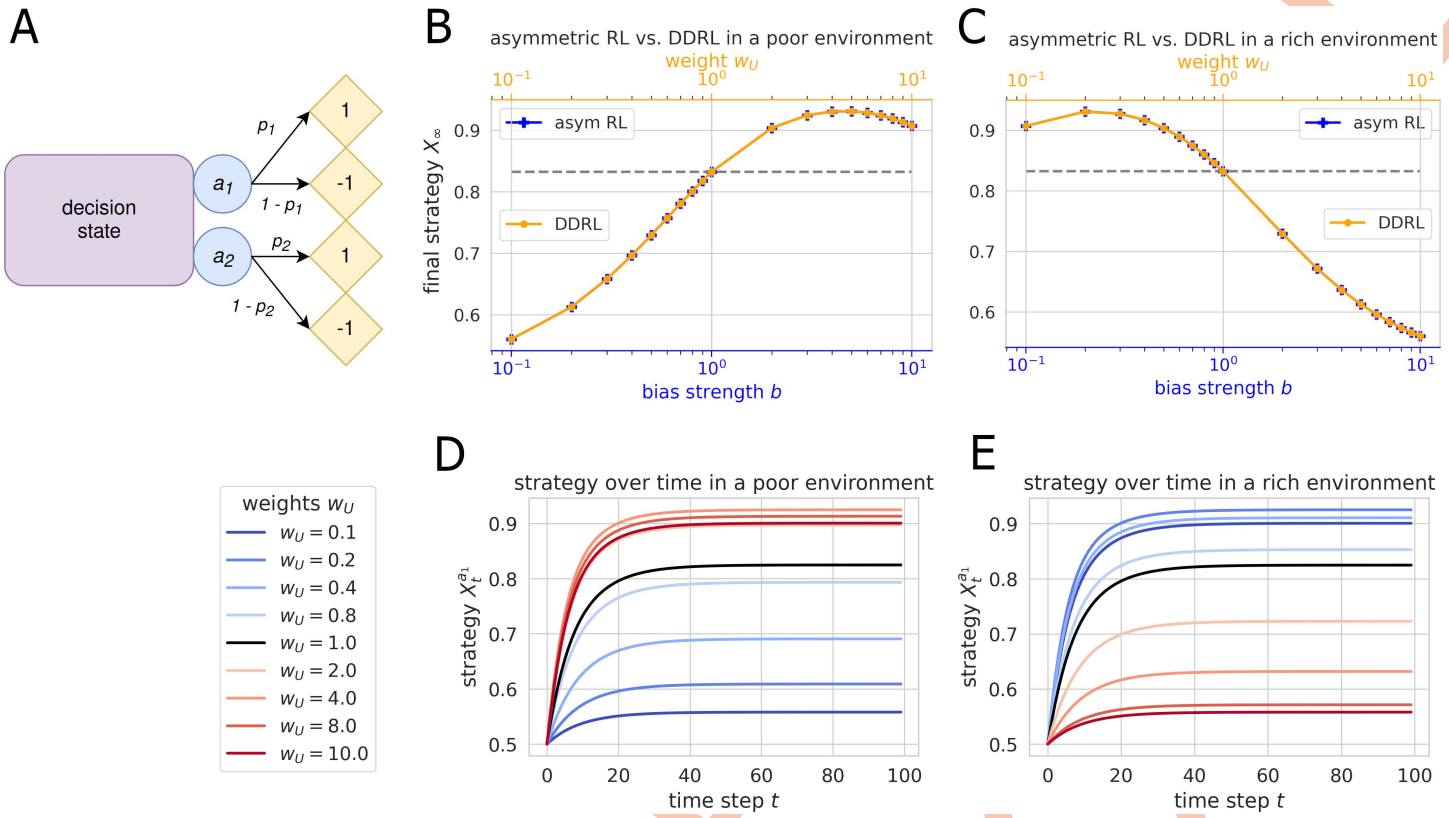

Fig 2. Influence of optimism/pessimism on a two-armed bandit environment with varying resource availability. A The two-armed bandit environment. **B**–**C** Comparison between final strategies X_∞ obtained *via* DDRL (yellow curve) vs. asymmetric RL (blue curve), **B** on a "poor" two-armed bandit task ($p_1 = 0.3$, $p_2 = 0.1$); **C** on a "rich" two-armed bandit task ($p_1 = 0.9$, $p_2 = 0.7$). These final strategies vary as a function of bias strength $b = \frac{\alpha^+}{\alpha^-}$. For DDRL agents, the final strategies were obtained from a trajectory of 150 time steps, after checking that this number of time steps was sufficient for convergence (see Fig 2D and 2E). For asymmetric agents, the final strategies were obtained analytically, using the formula in [54]. **D**–**E** Strategy plots of a DDRL agent performing a two-armed bandit task with various levels of optimism/pessimism, **D** in a poor environment ($p_1 = 0.3$, $p_2 = 0.1$); **E** in a rich environment ($p_1 = 0.9$, $p_2 = 0.7$). These plots represent the evolution of an agent's probability to choose the best arm over time. Blue lines indicate pessimistic agents ($w_U < 1$), while red lines indicate optimistic agents ($w_U > 1$). Neutral agents ($w_U = 1$) are represented by a black line. In all panels, all DDRL agents had $n = 29$ neurons, and a smoothing width $\sigma = 0.0001$ to minimize errors (see S1 Text, Part **IV**).

https://doi.org/10.1371/journal.pcbi.1014723.g002

arms' reward probabilities are $p_1 = 0.3$ and $p_2 = 0.1$, and on a "rich" version of the same task, in which the arms' reward probabilities are $p_1 = 0.9$ and $p_2 = 0.7$.

First, we find that DDRL closely approximates the behavior of an asymmetric-RL agent on a two-armed bandit task (Fig 2B and 2C). To do so, we checked that both algorithms converged towards the same final strategy X_∞–i.e., an agent's probability to choose the best arm after convergence. We computed final strategies for different bias strengths. In asymmetric RL, bias strength is defined as $b = \frac{\alpha^+}{\alpha^-}$ (where α^+ updates positive temporal-difference errors and α^-, negative temporal-difference errors). An analytical calculation of X_∞ as a function of b is given in [54]. In DDRL, bias strength is modulated by the upper weight, w_U. Here, we simulated the deterministic trajectory of an agent's strategy X over time until convergence for different values of w_U.

Second, we replicate the aforementioned findings on the influence of optimism/ pessimism in poor/rich environments (Fig 2B and 2C). In both scarcity conditions, DDRL closely approximates asymmetric RL. More precisely, the final strategy X_∞ of a DDRL agent with weight w_U matches the final strategy of an asymmetric-RL agent with bias strength $b = w_U$. We give a formal proof of this equivalence in S1 Text, Section 12.2. Fig 2B and 2C shows how the influence of optimism/

UNCORRECTED PROOF

pessimism on performance is modulated by resource scarcity. In a poor environment (Fig 2B), optimistic agents ($w_U > 1$) converge to a higher probability of choosing the best arm than neutral ($w_U = 1$) and pessimistic ($w_U < 1$) agents. In a rich environment (Fig 2C), by contrast, pessimistic agents converge to a higher probability of choosing the best arm than neutral ($w_U = 1$) and optimistic ($w_U > 1$) agents. Furthermore, both curves indicate the presence of an optimal bias strength (or upper-weight value), for which the final strategy X_∞ is maximal. In our poor environment, this optimal upper-weight value lies close to $w_U = 5$; whereas, in our rich environment, it is close to $w_U = 0.2$.

This interaction between optimism/pessimism and resource scarcity, previously established in Ref [54]., is further visualized in Fig 2D and 2E. These strategy plots depict the evolution of a DDRL agent's strategy–i.e., its probability of choosing the best arm–over time for different levels of optimism/pessimism. In a poor environment (Fig 2D), starting from a uniform random strategy ($X = 0.5$), optimistic DDRL agents ($w_U > 1$, red lines) converge towards a higher probability of choosing the best arm than the neutral agent ($w_U = 1$, black line); whereas pessimistic DDRL agents ($w_U < 1$, blue lines) converge towards lower performance. In line with previous findings [54] and the results mentioned above, this pattern is reversed in a rich environment (Fig 2E).

Overall, our results confirm that DDRL closely approximates asymmetric RL on a two-armed bandit task with varying resource scarcity. In particular, we replicate previous findings indicating that optimism is beneficial in poor environments, while pessimism enhances performance on the rich version of the task. This means DDRL is a suitable model for optimism/pessimism in this basic choice task. Next, we turn to an environment which is widely used in biology, economics, and sustainability science: the Stag-Hunt game.

3.3. Example application: Social coordination task

Here, we use our model to confirm previous findings that show optimism improves coordination in the Stag-Hunt game. The Stag-Hunt game is an iconic coordination problem in the space of social dilemmas [60]. A social dilemma is a situation in which agents individually have an incentive to deviate from outcomes that would be best for the group. Each agent must choose between two actions: cooperate (c) or defect (d). In the two-agent case, if both cooperate, they get a payoff R (called "reward"). If both defect, they get a payoff P (called "punishment"). If one cooperates and the other defects, the cooperator receives a payoff of S (called "sucker") and the defector gets a payoff of T (called "temptation") (Fig 3A). In the Stag-Hunt game, the payoff ordering is as follows: $R > T > P > S$. Therefore, both mutual defection and mutual cooperation are Nash equilibria. However, if both agents defect, they receive the punishment payoff P, which is lower than the maximum reward payoff R obtainable by mutual cooperation. This means individual interests align with collective interests, but the best outcome can only be achieved through coordination among agents.

Previous research in multi-agent reinforcement learning suggests that variants of optimism enhance coordination on various tasks. For instance, Lenient Learning, an algorithm that initially has agents dismiss negative outcomes, has led to performance improvements across multiple coordination games, such as the climb and penalty games [23]. Likewise, Hysteretic Q-learning (a multi-agent equivalent of asymmetric RL) has been shown to improve coordination on several tasks: the aforementioned climb and penalty games, a collaborative ball balancing task, and a pursuit game (where predators need to coordinate to capture prey) [22]. In this section, we investigate whether optimism enhances coordination in the Stag-Hunt game. The climb and penalty games are two three-action versions of the Stag-Hunt problem, which received its name from a story in which hunters need to coordinate to capture a stag, analogous to predators coordinating to capture prey. Thus, the Stag-Hunt game is well-suited to capture the essence of these coordination problems. Human experimental evidence corroborates this connection. When participants reason under time pressure, favoring intuitive, optimistic social expectations, they cooperate more in Stag-Hunt tasks than when reasoning deliberatively [61]. Shared optimistic beliefs about a partner's trustworthiness increase cooperation by up to 40 percentage points [62], and optimism about partner behavior predicts payoff-dominant choices in experimental coordination tasks [63].

UNCORRECTED PROOF

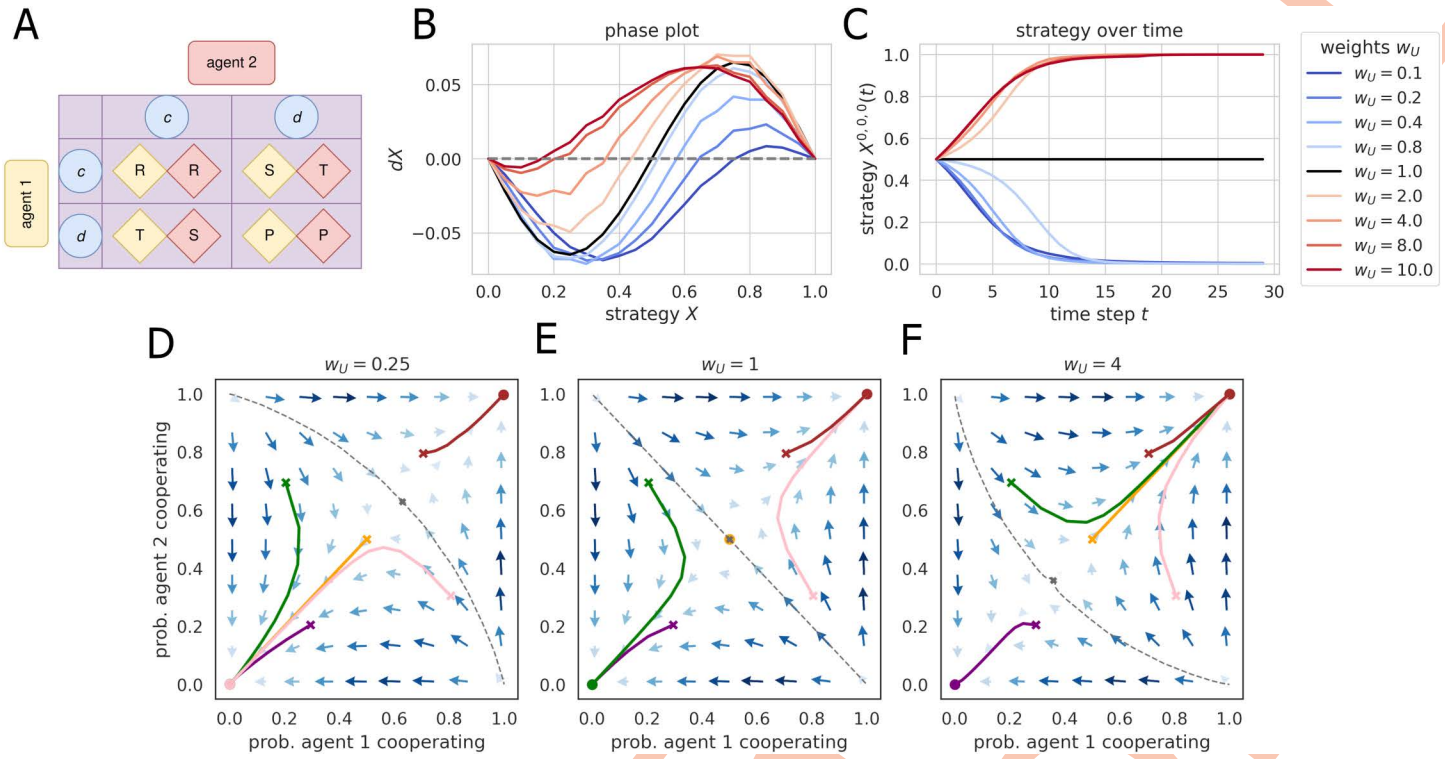

Fig 3. Influence of optimism/pessimism on a Stag-Hunt game. A Structure of a social dilemma. In the Stag-Hunt game, $R > T > P > S$. For our analyses, we set $R = 3$, $T = 1$, $P = 0$ and $S = -2$ as illustrative values satisfying this ordering; the qualitative structure of fixed points and their stability depends only on the ordering, not the specific magnitudes. **B** Phase plot and **C** strategy plot of an agent playing the Stag-Hunt game for various levels of optimism/pessimism. Since the Stag-Hunt game is symmetric among agents, we assumed the second agent had the same strategy as the first. On the phase plot **(B)**, $X = 1$ indicates maximal cooperation, and $X = 0$, maximal defection. Intersections of the curves with the x-axis indicate stable and unstable fixed points, depending on the sign of the derivative dX. The strategy plot (**C**) represents the evolution of an agent's probability to cooperate over time, given an initial cooperation probability of 0.5. Blue lines represent pessimistic agents ($w_U < 1$), while red lines represent optimistic agents ($w_U > 1$). Neutral agents ($w_U = 1$) are represented by a black line. **D–F** Flow plots of two agents playing the Stag-Hunt game **D** with a pessimistic heuristic ($w_U = 0.25$); **E** with no heuristic ("neutral," $w_U = 1$); **F** with an optimistic heuristic ($w_U = 4$). Arrows represent the direction and intensity of the strategies' derivative dX. Colored lines are sample trajectories starting from various initial conditions (+), while round dots indicate fixed points. Dashed gray lines delineate the flow plots' separatrices–i.e., the threshold of initial cooperation probability above which agents learn towards maximal cooperation, and below which they learn towards maximal defection. In all panels, all DDRL agents had $n = 29$ neurons, and a smoothing width $\sigma = 0.08$ to minimize errors (see S1 Text, Part **IV)**.

https://doi.org/10.1371/journal.pcbi.1014723.g003

In line with previous findings, we find that optimistic DDRL agents converge to mutual cooperation when starting from a uniformly random strategy. In contrast, pessimistic agents converge to the suboptimal equilibrium of mutual defection. A neutral agent remains at its mixed strategy for the payoff values of our Stag-Hunt game (Fig 3C). Fig 3B shows that the uniform random strategy (i.e., a cooperation and defection probability of 0.5) is an unstable fixed point of the learning dynamics for neutral agents. Therefore, if a neutral agent's initial strategy is higher than 0.5 (and provided the second agent's initial strategy is the same), it converges toward cooperation. If, by contrast, both agents' initial strategies are below 0.5, they converge toward defection. Optimism (red lines in Fig 3B) lowers this threshold–i.e., two optimistic DDRL agents may converge towards cooperation even if they started with a strategy $X < 0.5$–while pessimism heightens it. This is why, in Fig 3C, when all agents start with a uniformly random strategy, optimistic agents end up mutually cooperating, whereas pessimistic agents converge towards mutual defection. Thus, optimism facilitates cooperation, while pessimism hinders it.

UNCORRECTED

The flow plots in Fig 3D, 3E and 3F give an overview of the learning dynamics in the two-dimensional phase space spanned by each agent's individual strategy. Thus, the phase plots in Fig 3B are equivalent to the diagonal in the flow plots in Fig 3D–3F, for the relevant optimism levels. The strategy changes dX are represented by the flow arrows. The separatrix curves separating the two basins of attraction of the two equilibria, mutual cooperation and mutual defection, are shown with dashed gray lines. Optimism distorts these separatrices. When agents are pessimistic, it takes higher initial levels of cooperation to eventually learn to fully cooperate (Fig 3D). When they are optimistic, learning to cooperate fully is possible even from initially lower levels of cooperation (Fig 3F).

Additionally, we tested our model on a prisoner's dilemma with $T > R > P > S$. Here, agents have the incentive to defect, regardless of what the other agent does. Thus, mutual cooperation is no longer an equilibrium despite being the optimal solution for the collective. We find that optimism/pessimism does not affect the learning dynamics. In other terms, whatever the upper-weight value w_U, DDRL agents always converge towards mutual defection in this environment (see S1 Text, Section 13.3). In this dilemma, indeed, the reward R is lower than the temptation T. Therefore, if an agent is optimistic that the other will cooperate, it is still incentivized to defect, since it will then obtain T instead of R. The mutual-defection equilibrium thus cannot be changed through the optimism/pessimism heuristic. In a Stag-Hunt game, by contrast, if an agent is optimistic that the other agent will cooperate, it increases its incentive to cooperate as well (see S1 Text, Section 13.4 for an analytical explanation of this difference between the two games).

In short, our results confirm that optimism enhances coordination in DDRL agents, while pessimism hinders it. These results echo previous findings across several multi-agent reinforcement learning algorithms in coordination games. Additionally, our flow plot visualization provides further insights into the collective learning dynamics in coordination games and into how the optimistic/pessimistic heuristic distorts this landscape. Next, we consider a lesser-known intertemporal risky choice problem.

3.4. Example application: Intertemporal risky choice problem

Here, we use our model to confirm previous findings that show optimism increases risk-taking. We also uncover new findings in this lesser-known choice environment [39,64], as it incorporates a novel component of decision-making into our analysis: intertemporal choice. In the choice environment, an agent must trade off a low but safe reward with a high reward with potentially negative future consequences. This is implemented via two environmental states: a prosperous and a degraded one. The agent must choose between a risky action and a safe action. In the prosperous state, taking the risky action leads to a reward $r_r = 1$, whereas the safe action leads to a reward $r_s = 0.5$. However, choosing the risky action in the prosperous state results in a transition to the degraded state with probability $p_c = 0.2$. There, the agent gets no reward $r_d = 0$, and can only transition back to the prosperous state with probability $p_r = 0.1$ if it chooses the safe action (Fig 4A). To simplify the analysis, we use an Actor-Critic agent here, since it has no explicit exploration term (see Section 2.2). We presume that a SARSA agent would exhibit additional, exploration-related effects and interactions.

To our knowledge, no behavioral experiment has studied an environment combining these features: repeated stochastic choice with state-dependent consequences and temporal recovery dynamics. Risk-taking and intertemporal discounting are each extensively studied in isolation; their combination has been explored in simplified choice tasks [65], but not in a multi-state repeated structure of this form. Previous research on risk-sensitivity in RL agents uses asymmetric learning rates to implement different attitudes towards risk, where optimism corresponds to risk-seeking and pessimism to risk-aversion [66,67]. Thus, we hypothesize that in our intertemporal risky choice problem, an optimistic DDRL agent is more likely to choose the risky action than a neutral agent. We also hypothesize that this pattern is reversed for pessimistic DDRL agents.

We confirm both hypotheses. Optimism increases risky behavior while pessimism favors safe behavior (Fig 4B and 4C). Fig 4B and 4C show the strategy change $dX^{\text{prosp, safe}}$ versus the strategy $X^{\text{prosp, safe}}$ to choose the safe action in the prosperous environmental state for different levels of optimism and pessimism and two discount factors. In the degraded

UNCORRECTED PROOF

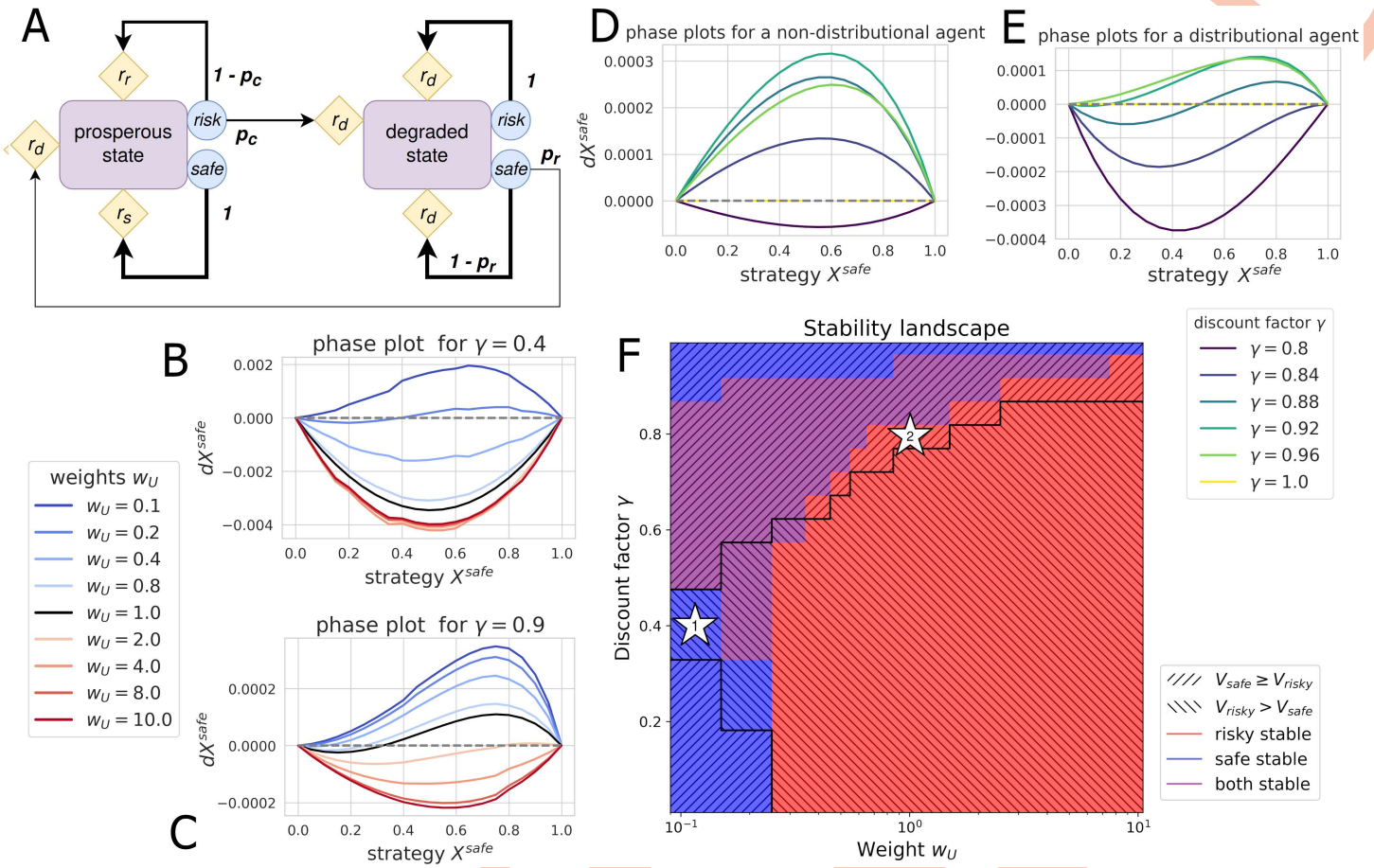

Fig 4. Influence of optimism/pessimism in a risk-reward dilemma. A Structure of a risk-reward dilemma. In our analyses, we set the collapse probability $p_c = 0.2$ and the recovery probability $p_r = 0.1$; we fixed the risky reward $r_r = 1$, the safe reward $r_s = 0.5$, and the degraded reward $r_d = 0$. **B–C** Phase plots of a DDRL agent performing a risk-reward dilemma with various levels of optimism/pessimism, **B** with a low discount factor ($\gamma = 0.4$); **C** with a high discount factor ($\gamma = 0.9$). Blue lines indicate pessimistic agents ($w_U < 1$), while red lines indicate optimistic agents ($w_U > 1$). Neutral agents ($w_U = 1$) are represented by a black line. **D–E** Phase plots of an agent performing a risk-reward dilemma with various discount factors, **D** using a regular deterministic Actor-Critic algorithm, **E** using a neutral ($w_U = 1$) DDRL Actor-Critic algorithm. On all phase plots **(B–E)**, $X = 1$ corresponds to maximal safety, and $X = 0$, to maximal risk. Intersections with the x-axis indicate stable and unstable fixed points (depending on the sign of the derivative dX). **F** Stability landscape of a DDRL on a risk-reward dilemma for various combinations of weights w_U and discount factors γ. Red patches correspond to regions in the parameter space where the risky policy ($X^{\text{prosp, risky}} = 1$) is stable, while blue patches indicate that the safe policy ($X^{\text{prosp, safe}} = 1$) is stable. Purple patches correspond to regions in the parameter space where both policies are stable. Stars mark regions of interest, where the safe or risky policy has highest value but is unstable. In all panels, the agent had $n = 29$ neurons, and a smoothing width $\sigma = 0.05$ to minimize errors (see S1 Text, Part **IV)**.

https://doi.org/10.1371/journal.pcbi.1014723.g004

state, only the safe action leads to a recovery to the rewarding prosperous state. Thus, we assumed $X^{\text{deg,safe}} = 1$ for Fig 4B and 4C. Flow plots of the entire strategy space for given discount factors and weight values can be found in S1 Text, Fig 14.6 and S1 Text, Fig 14.7.

When the agent cares little about future rewards (low discount factor $\gamma = 0.4$, Fig 4B), only intense pessimism can push an agent to act safely. Only the most pessimistic agents ($w_U = 0.1$ and $w_U = 0.2$) have a chance to converge towards complete safety in the prosperous state ($X^{\text{prosp,safe}} = 1$). All the other agents, including the moderately pessimistic and neutral ones, converge towards maximal risk-taking ($X^{\text{prosp,safe}} = 0$). This means that, in such situations, to act safely, intense pessimism can compensate for an agent's lack of concern for future rewards.

UNCORRECTED PROOF

When the agent cares more about future rewards (high discount factor $\gamma = 0.9$, Fig 4C), the more optimistic an agent is, the more likely it is to converge toward the risky strategy. In fact, for a moderately-to-strongly optimistic agent ($w_U \geq 4$), the risky strategy is the only stable one. A mildly optimistic agent ($w_U = 2$) could converge to a maximally safe approach, but only if its initial safety levels are very high. A reversed pattern is found for the different levels of pessimism. Because γ is high, risk-seeking is overall lower than with a low discount factor.

In both cases, we observe instances of bistability in the strategy space. Which strategy the agent converges to depends on the initial strategy. This is particularly interesting as non-distributional agents (using expected values from a continuous reward spectrum) do not exhibit any bistable regimes (Fig 4D). This observation aligns with the finding that an optimal strategy in a Markov Decision Process is always deterministic [68]. Fig 4E compares our distributional agent dynamics without any bias ($w_U = 1$) for different discount factors. We find certain discount factors γ (e.g., $\gamma = 0.88$, $\gamma = 0.92$) at which the learning dynamics are bistable. The neutral weight $w_U = 1$ rules out optimism or pessimism as a cause; the only remaining structural difference between Fig 4D and 4E is therefore the discretization of the return distribution. Thus, bistability does not solely emerge from optimism or pessimism, but from an intrinsic feature of our DDRL framework: the discretization of the return distribution. Fig 4F visualizes how the stability of the safe and risky strategies depends on the interaction of bias weight values w_U and discount factors γ. The red zone denotes the stability of the risky strategy ($X^{\text{prosp, risky}} = 1$); the blue zone, that of the safe one ($X^{\text{prosp, safe}} = 1$); whereas, in the purple zone, both strategies are stable. Thus, bistability emerges for a broad range of weights and discount factors, including for neutral agents ($w_U = 1$). Overall, we find that these bistable regimes depend on the interplay among the discount factor, the optimism weight, and return discretization, whose granularity is set by the smoothing width σ.

We also find a kind of incoherent choice pattern, which we coin "individual dilemma." Just as in a social dilemma, an agent would be better off playing the strategy that yields the highest value, but it cannot realize this, as it is pulled towards the other, suboptimal strategy, which is stable (Stars in Fig 4F). To see this, we highlighted the region in parameter space where the safe policy yields a higher value to the agent than the risky policy ($V_{\text{safe}} > V_{\text{risky}}$) and likewise, $V_{\text{risky}} > V_{\text{safe}}$. We define $V_{\text{safe}} = \langle V_{\mathbf{x}}^{\text{prosp}} \rangle_{w_U} |_{X^{\text{prosp, safe}}=1}$ and $V_{\text{risky}} = \langle V_{\mathbf{x}}^{\text{prosp}} \rangle_{w_U} |_{X^{\text{prosp, risky}}=1}$. For most parameter values, when a strategy yields a higher overall value, it is also stable. Nevertheless, Fig 4F exhibits some deviations from this pattern. For example, at $w_U = 0.1$ and $\gamma = 0.402$, the safe strategy is stable while the risky strategy yields a higher value (Star 1 in Fig 4F). At $w_U = 1$ and $\gamma = 0.794$, the risky strategy is stable, while the safe strategy yields a higher value (Star 2 in Fig 4F). Although the state value of the safe strategy is higher than that of the risky one, the state-action value of the risky action is higher than that of the safe action. In S1 Text, Fig 14.10, we show the corresponding agent's V-return and Q-return distributions.

Overall, our results confirm previous findings regarding the relationship between optimism, pessimism, and risky behavior. Optimism enhances risky behavior, whereas pessimism favors risk-averse behavior. Additionally, our framework generates two novel predictions: bistable regimes and individual dilemmas. Both have no counterpart in non-distributional dynamics and await empirical test.

4. Discussion

In this work, we derived the deterministic dynamics of distributional multi-agent reinforcement learning to obtain a unifying, biologically plausible, and tractable framework for modeling individual cognition and behavior across contexts. We applied our framework to model the widely studied optimism heuristic. We established the soundness of our framework by reproducing results across three iconic choice domains: a bandit choice task under resource variability, a social coordination challenge, and a risky choice problem. This demonstrates the potential of our model as a unifying framework to study the optimism-pessimism heuristic in collective learning dynamics.

Our framework rests on a simple, biologically plausible mechanism in which outcome distributions are discretized by a set of neurons. Empirical research has recently provided evidence for an analogous computation of distributions during learning and decision-making in the brain [28,69]. Thus, our return discretization should not be viewed simply as

UNCORRECTED PROOF

a technical trick, but as a meaningful biological feature. Distributional coding is empirically distinguishable from a population code for expected value. A code representing only the mean return would not generate the systematic diversity in prediction-error asymmetries observed in dopaminergic neurons: some neurons learn faster from better-than-expected outcomes, others from worse-than-expected outcomes, together spanning the full distribution of returns [28]. D1 and D2 striatal neurons encode opposite tails of the return distribution through an opponent circuit [70], and distributional coding has also been observed in prefrontal cortex [71]. These findings point to a specific neural substrate for distributional rather than mean-value coding.

In this regard, our framework could help enhance understanding of how reward information is encoded neurologically. How many neurons are involved in categorizing environmental rewards, and how does this number impact the accuracy of this distribution? In S1 Text, we show how the smoothing width, σ, affects the error between the expected values of the continuous and discretized reward spectra across various environments (see S1 Text, Part IV). In particular, we found that, for a fixed number of neurons n, the smoothing width σ that minimizes this error is environment-dependent. Thus, is there a context-dependent mechanism that modulates the smoothing width? Recent studies suggest such a context-dependent mechanism. It has been found that humans do not encode objective outcomes but rather evaluate options with respect to available alternatives [72]. This has been modeled with RL as range normalization [73], in which the value of an option is rescaled by the difference between its maximum and minimum values in the environment. Our framework could incorporate such a mechanism to explore how varying assumptions about neurological reward representations affect outcomes, leading to empirically testable hypotheses.

Furthermore, our framework could help uncover the empirical basis for optimistic reinforcement learning in the brain. In particular, it is an open question whether optimism in the brain manifests at the level of state-action values or temporal-difference errors. Indeed, asymmetric RL, which is widely used in psychological experiments, locates optimism/pessimism at the level of temporal-difference errors. These are updated differentially through asymmetric learning rates. By contrast, distributional reinforcement learning distorts the state-action return distribution by applying weights. According to prior work, these implementations should produce the same results in stateless environments, but they are expected to yield different outcomes in stateful environments [29]. Thus, comparing our framework with asymmetric RL algorithms in stateful choice environments could generate empirically testable hypotheses that would distinguish between signatures of state-action values versus temporal-difference-error optimism in the brain.

Applying our model to a single-agent intertemporal risky-choice task, we found that return discretization with Gaussian smoothing can lead to a bistable choice regime. In other words, strategies become path-dependent. This bistable structure is analytically tractable precisely because of the deterministic dynamics framework: fixed points and their stability can be identified exactly from the flow equations, whereas stochastic algorithms offer no such direct access to the stability landscape. By contrast, non-distributional agents that compute expected values of a continuous reward spectrum do not exhibit bistable regimes. This finding suggests a mechanistic account of choice hysteresis [74]. Choice hysteresis is the tendency to repeat past choices as a function of one's history, reflecting path-dependence in an agent's choice history [75]. Hysteresis in intertemporal choice has been documented from eye-tracking studies [76] to studies about addictive behaviors [77]. It has been shown that hysteresis in sequential decision-making increases in aging rats [78], and hysteresis has even been accounted for in terms of attractor models [79] — a description that aligns naturally with the bistable structure identified here.

In experimental psychology, there is a debate about whether optimism can be reduced to perseveration. While related, choice hysteresis and perseveration are distinct phenomena [80]: stuck-in-set perseveration [81] refers to the deployment of a previously appropriate strategy that is no longer goal-congruent. Several studies and meta-analyses have striven to disentangle optimism from perseveration in empirical datasets [75,82]. They concluded that both heuristics were present and irreducible to one another–but it was also shown that the optimism heuristic leads to higher performance in a wide range of two-armed bandit tasks, compared to perseveration [83]. These findings establish that the two phenomena are

UNCORRECTED PROOF

empirically distinguishable, but leave open the mechanistic question of what generates perseveration in the first place. DDRL provides a mechanistic answer. Bistability alone does not produce perseveration: it does so only when the agent is locked by its history into the lower-valued attractor–that is, when parameters lie on the suboptimal side of the value-ordering boundary (thick black line, Fig 4F). In that regime, the agent's strategy is goal-incongruent; yet, in a stochastic learning system, noise-induced transitions across the basin boundary could in principle restore goal-congruent behavior. DDRL further predicts a more extreme case: in the individual dilemma, the better strategy is no longer a stable fixed point, and no noise-induced correction is possible. This constitutes inescapable stuck-in-set perseveration, consistent with accounts that trace perseverative behavior to constraints on policy representation [80,84]. Return discretization thus generates a hierarchy of path-dependent phenomena–from choice hysteresis, to escapable perseveration, to fully inescapable stuck-in-set perseveration–none of which arise in non-distributional dynamics.

The individual dilemma–the inescapable extreme of the perseverative hierarchy described above–offers a novel mechanistic account of incoherent choice. In analogy with the well-known social dilemma, an agent would be better off choosing the strategy that provides the highest value, but cannot do so because it is drawn toward the other, suboptimal strategy that is stable. Our results show that this situation arises in a wide range of combinations between discount factors and optimism weights. While individual dilemmas might appear as flaws in our model, there are many cases where human choices are inconsistent. Perhaps one of the most compelling illustrations lies in compulsive behavior. Compulsion occurs when one keeps engaging in actions that feel immediately relieving but are harmful in the long run. It is a key component of, e.g., Tourette's syndrome and obsessive-compulsive disorder (OCD), and can occur even when patients are completely aware of the counterproductivity of their behavior. On the one hand, because of the ritualistic nature of compulsions, it has been suggested that habit plays a central role in compulsivity; on the other hand, there is a significant overlap between compulsivity and impulsivity–i.e., the reduced ability to delay rewards, which is quantified by the discount factor γ [85,86]. Our findings suggest there could be a complex interplay between reward discounting, optimism/pessimism, and discretization/habit in the emergence of compulsive behavior. Indeed, individual dilemmas appear for certain values of the discount factor and optimism weight. Thus, our framework could help disentangle the contributions of impulsivity, optimism and habit formation in giving rise to incoherent choice patterns. Moreover, our results show that an individual dilemma can be escaped by tuning up the optimism weight or the discount factor, depending on the region of the parameter space one finds oneself in. This opens the door to developing treatments which, by acting on impulsivity and/or optimism levels, could alleviate compulsive symptoms.

Empirically testing the behavioral predictions of DDRL — including the presence of individual dilemmas and their dependence on discount factor and optimism weight — would require fitting the framework to individual-subject trial-by-trial data. Adapting the deterministic dynamics to a stochastic likelihood model is non-trivial but constitutes an important direction for future empirical work.

Because of its distributional basis, our framework opens the door to representing decision-making heuristics as distortions on state-action return distributions. The optimism/pessimism distortion implemented here via the weights method is only one of several available approaches; cumulative prospect theory, CVaR, and the Wang distortion [87] provide alternative formal representations, and comparing outcomes across methods remains an open question. Among candidate heuristics, self-efficacy [88] and confirmation bias [59] could be interesting candidates to model in individual and collective scenarios. This would help address outstanding debates and questions in several disciplines. For instance, it has been suggested that optimism could be a confirmation bias in disguise, at the individual [59,89] and collective levels [90]. Thus, distinguishing between optimism and confirmation heuristics in individual and collective settings is an essential task for experimental researchers. In this regard, our framework could be used to generate hypotheses and predictions.

Besides empirical questions, our model could be applied to study the role of optimism/pessimism in complex collective-action problems, such as climate change. For tractability reasons, cognitive, social, and environmental factors are often studied in isolation, missing an integrated perspective [11]. Our model can contribute to such an

UNCORRECTED PROOF

integrated perspective. For example, it could enable insights into the influence of optimism/pessimism on cooperation in the face of climate tipping points [91]. Previous work found that the severity of the tipping impact, the certainty of the tipping threshold, and caring for future rewards are conducive to cooperation [37,92,93]. But what if these factors are not enough? Is it better to be optimistic, to induce coordination in the social dilemma, as we have seen in our social coordination task? Or is it better to be pessimistic and risk-averse, to avoid triggering the tipping element, as we have seen in our intertemporal risky-choice problem? DDRL is suited to address this tension: its deterministic dynamics yield an analytically tractable stability landscape — analogous to Fig 4F — in which the conditions for coordination, collapse, and bistability can be identified as functions of optimism weight, discount factor, and return discretization. The climate scenario combines both a coordination structure (as in our Stag-Hunt) and an intertemporal collapse risk (as in our risk-reward dilemma), and DDRL provides a natural framework for analyzing their interaction in a stateful multi-agent setting. More broadly, DDRL opens a route to studying how neurocognitive heuristics shape collective behavior across complex social-ecological systems [6,7,11].

We also highlight some limitations of our framework. First, although it builds upon stochastic games, which are the foundation of multi-agent reinforcement learning in machine learning [94], our framework does not scale to large environments (such as large grid worlds or 3D computer games). Such choice environments have large state and reward spaces, leading to a combinatorial explosion and the curse of dimensionality. The analysis of collective dynamics in such high-dimensional environments requires the use of more suitable methods. Second, our framework presupposes categorically discretized neurons. This limits the precision of approximating return distributions, compared to more sophisticated parameter-estimation methods used in the machine-learning literature [46,47]. Thus, our framework's application to machine learning might be limited, as it is not an optimization method. It is a biologically inspired method of analysis designed to improve understanding. Third, the Gaussian smoothing projection used throughout is a biologically motivated choice — its graded, overlapping allocation reflects the tuning-curve structure of reward-sensitive neurons — but it is not the only option. Mean-preserving projections such as the Cramér projection [27,95] deterministically allocate mass to at most two adjacent atoms and are standard in categorical distributional RL. These projections would be integrable into our framework in principle, but would deviate from our initial motivation. Using a mean-preserving projection would align the framework more closely with standard categorical distributional RL, potentially opening it to a broader range of questions and applications.

In conclusion, our deterministic distributional multi-agent reinforcement learning framework bridges the gap between the microscopic constraints of neural reward encoding and the macroscopic dynamics of collective behavior. It offers a unified perspective through which seemingly disparate phenomena–from optimism and choice hysteresis to compulsive behavior and social coordination–can be understood as emergent properties of a common underlying mechanism. While we acknowledge the trade-off between tractability and computational scalability, our approach provides a necessary middle ground for cross-disciplinary inquiry across the animal, human, and machine behavioral sciences. Moving forward, this framework offers a robust toolkit for generating testable hypotheses and investigating how individual-level cognitive biases shape collective outcomes in complex real-world challenges, ultimately advancing our understanding of the intricate interplay between neural computation, individual choice, and collective dynamics.

Supporting information

S1 Text. Supplementary Information for "Deterministic dynamics of distributional multi-agent reinforcement learning. " Contains full mathematical derivations, supplementary analyses, and all simulation code rendered as a self-contained document, ensuring that results remain reproducible independently of the availability of external code repositories.
(PDF)

UNCORRECTED PROOF

Author contributions

Conceptualization: Pawel Romanczuk, Wolfram Barfuss.

Data curation: Clémence Bergerot.

Formal analysis: Clémence Bergerot.

Investigation: Clémence Bergerot.

Methodology: Clémence Bergerot, Wolfram Barfuss.

Project administration: Pawel Romanczuk, Wolfram Barfuss.

Resources: Pawel Romanczuk.

Software: Clémence Bergerot, Wolfram Barfuss.

Supervision: Pawel Romanczuk, Wolfram Barfuss.

Validation: Clémence Bergerot, Wolfram Barfuss.

Visualization: Clémence Bergerot.

Writing – original draft: Clémence Bergerot.

Writing – review & editing: Pawel Romanczuk, Wolfram Barfuss.

References

1. Couzin ID. Collective cognition in animal groups. Trends Cogn Sci. 2009;13(1):36–43. https://doi.org/10.1016/j.tics.2008.10.002 PMID: 19058992
2. Sayin S, et al. The behavioral mechanisms governing collective motion in swarming locusts. Science. 2025;387(6737):995–1000. https://doi.org/10.1126/science.adq7832
3. McNamara JM. Towards a richer evolutionary game theory. J R Soc Interface. 2013;10(88):20130544. https://doi.org/10.1098/rsif.2013.0544 PMID: 23966616
4. Johnson SGB, Schotanus PR, Kelso JAS. Minds and markets as complex systems: an emerging approach to cognitive economics. Trends Cogn Sci. 2024;28(11):1037–50. https://doi.org/10.1016/j.tics.2024.07.003 PMID: 39127602
5. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, et al. Machine behaviour. Nature. 2019;568(7753):477–86. https://doi.org/10.1038/s41586-019-1138-y PMID: 31019318
6. Schill C, Anderies JM, Lindahl T, Folke C, Polasky S, Cárdenas JC, et al. A more dynamic understanding of human behaviour for the Anthropocene. Nat Sustain. 2019;2(12):1075–82. https://doi.org/10.1038/s41893-019-0419-7
7. Constantino SM, et al. Cognition and behavior in context: A framework and theories to explain natural resource use decisions in social-ecological systems. Sustainability Science. 2021;16(5):1651–71. https://doi.org/10.1007/s11625-021-00989-w
8. Niv Y. Reinforcement learning in the brain. Journal of Mathematical Psychology. 2009;53(3):139–54. https://doi.org/10.1016/j.jmp.2008.12.005
9. Sutton RS, Barto AG. Reinforcement learning: An introduction. 2018.
10. Botvinick M, et al. Deep reinforcement learning and its neuroscientific implications. Neuron. 2020;107(4):603–16. https://doi.org/10.1016/j.neuron.2020.06.014
11. Barfuss W, Flack J, Gokhale CS, Hammond L, Hilbe C, Hughes E, et al. Collective cooperative intelligence. Proc Natl Acad Sci U S A. 2025;122(25):e2319948121. https://doi.org/10.1073/pnas.2319948121 PMID: 40523168
12. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593–9. https://doi.org/10.1126/science.275.5306.1593 PMID: 9054347
13. Dayan P, Niv Y. Reinforcement learning: the good, the bad and the ugly. Curr Opin Neurobiol. 2008;18(2):185–96. https://doi.org/10.1016/j.conb.2008.08.003 PMID: 18708140
14. Barfuss W. Towards a Unified Treatment of the Dynamics of Collective Learning. In: Challenges and Opportunities for Multi-Agent Reinforcement Learning, AAAI Spring Symposium, 2020.
15. Bloembergen D, Tuyls K, Hennes D, Kaisers M. Evolutionary Dynamics of Multi-Agent Learning: A Survey. jair. 2015;53:659–97. https://doi.org/10.1613/jair.4818
16. Gigerenzer G, Goldstein DG. Reasoning the fast and frugal way: models of bounded rationality. Psychol Rev. 1996;103(4):650–69. https://doi.org/10.1037/0033-295x.103.4.650 PMID: 8888650

UNCORRECTED PROOF

17. Korteling JE, Brouwer A-M, Toet A. A Neural Network Framework for Cognitive Bias. Front Psychol. 2018;9:1561. https://doi.org/10.3389/fpsyg.2018.01561 PMID: 30233451
18. Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S. Behavioural and neural characterization of optimistic reinforcement learning. Nat Hum Behav. 2017;1(4). https://doi.org/10.1038/s41562-017-0067
19. Gagne C, Zika O, Dayan P, Bishop SJ. Impaired adaptation of learning to contingency volatility in internalizing psychopathology. Elife. 2020;9:e61387. https://doi.org/10.7554/eLife.61387 PMID: 33350387
20. Ohta H, Satori K, Takarada Y, Arake M, Ishizuka T, Morimoto Y, et al. The asymmetric learning rates of murine exploratory behavior in sparse reward environments. Neural Netw. 2021;143:218–29. https://doi.org/10.1016/j.neunet.2021.05.030 PMID: 34157646
21. Farashahi S, Donahue CH, Hayden BY, Lee D, Soltani A. Flexible combination of reward information across primates. Nat Hum Behav. 2019;3(11):1215–24. https://doi.org/10.1038/s41562-019-0714-3 PMID: 31501543
22. Matignon L, Laurent GJ, Le Fort-Piat N. Hysteretic Q-learning: an algorithm for Decentralized Reinforcement Learning in Cooperative Multi-Agent Teams. In: 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007. 64–9. https://doi.org/10.1109/iros.2007.4399095
23. Panait L, Sullivan K, Luke S. Lenient learners in cooperative multiagent systems. In: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems, 2006. 801–3. https://doi.org/10.1145/1160633.1160776
24. Bowling M, Veloso M. Multiagent learning using a variable learning rate. Artificial Intelligence. 2002;136(2):215–50. https://doi.org/10.1016/s0004-3702(02)00121-2
25. Kapetanakis S, Kudenko D. Improving on the reinforcement learning of coordination in cooperative multi-agent systems. In: Proceedings of the Second Symposium on Adaptive Agents and Multi-agent Systems (AISB02), 2002.
26. Matignon L, Laurent GJ, Le Fort-Piat N. A study of FMQ heuristic in cooperative multi-agent games. In: The 7th International Conference on Autonomous Agents and Multiagent Systems. Workshop 10: Multi-Agent Sequential Decision Making in Uncertain Multi-Agent Domains, aamas' 08, 2008. 77–91.
27. Bellemare MG, Dabney W, Munos R. A distributional perspective on reinforcement learning. In: International conference on machine learning, 2017. 449–58.
28. Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, et al. A distributional code for value in dopamine-based reinforcement learning. Nature. 2020;577(7792):671–5. https://doi.org/10.1038/s41586-019-1924-6 PMID: 31942076
29. Rowland M, et al. Temporal difference and return optimism in cooperative multi-agent reinforcement learning. In: 2021.
30. Oh J, et al. Toward risk-based optimistic exploration for cooperative multi-agent reinforcement learning. 2023. https://arxiv.org/abs/2303.01768
31. Keramati R, Dann C, Tamkin A, Brunskill E. Being Optimistic to Be Conservative: Quickly Learning a CVaR Policy. AAAI. 2020;34(04):4436–43. https://doi.org/10.1609/aaai.v34i04.5870
32. Cho T, Han S, Lee H, Lee K, Lee J. Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion. In: Advances in Neural Information Processing Systems 36, 2023. 56802–24. https://doi.org/10.52202/075280-2481
33. Barfuss W, Donges JF, Kurths J. Deterministic limit of temporal difference reinforcement learning for stochastic games. Phys Rev E. 2019;99(4–1):043305. https://doi.org/10.1103/PhysRevE.99.043305 PMID: 31108579
34. Barfuss W. Reinforcement Learning Dynamics in the Infinite Memory Limit. In: International Joint Conference on Autonomous Agents and Multiagent Systems, 2020. 1768–70. https://doi.org/10.65109/crff7553
35. Hofbauer J, Sigmund K. Evolutionary Games and Population Dynamics. 1998. https://doi.org/10.1017/CBO9781139173179
36. Börgers T, Sarin R. Learning Through Reinforcement and Replicator Dynamics. Journal of Economic Theory. 1997;77(1):1–14. https://doi.org/10.1006/jeth.1997.2319
37. Barfuss W, Donges JF, Vasconcelos VV, Kurths J, Levin SA. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. Proc Natl Acad Sci U S A. 2020;117(23):12915–22. https://doi.org/10.1073/pnas.1916545117 PMID: 32434908
38. Barfuss W, Meylahn JM. Intrinsic fluctuations of reinforcement learning promote cooperation. Sci Rep. 2023;13(1):1309. https://doi.org/10.1038/s41598-023-27672-7 PMID: 36693872
39. Barfuss W. Dynamical systems as a level of cognitive analysis of multi-agent learning. Neural Comput & Applic. 2021;34(3):1653–71. https://doi.org/10.1007/s00521-021-06117-0
40. Jaynes EDT, Bretthorst GL. Probability Theory: The Logic of Science. 2003.
41. Wolpert DH, Harré M, Olbrich E, Bertschinger N, Jost J. Hysteresis effects of changing the parameters of noncooperative games. Phys Rev E Stat Nonlin Soft Matter Phys. 2012;85(3 Pt 2):036102. https://doi.org/10.1103/PhysRevE.85.036102 PMID: 22587144
42. Herrnstein RJ. Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav. 1961;4(3):267–72. https://doi.org/10.1901/jeab.1961.4-267 PMID: 13713775
43. Soltani A, Wang X-J. A biophysically based neural model of matching law behavior: melioration by stochastic synapses. J Neurosci. 2006;26(14):3731–44. https://doi.org/10.1523/JNEUROSCI.5159-05.2006 PMID: 16597727
44. Morimura T, et al. Nonparametric return distribution approximation for reinforcement learning. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010. 799–806.

UNCORRECTED PROOF

45. Morimura T, et al. Parametric return density estimation for reinforcement learning. In: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, 2010. 368–75.
46. Rowland M, et al. Statistics and samples in distributional reinforcement learning. In: 2019. 5528–36.
47. Dabney W, Rowland M, Bellemare M, Munos R. Distributional Reinforcement Learning With Quantile Regression. AAAI. 2018;32(1). https://doi.org/10.1609/aaai.v32i1.11791
48. Borkar S. Stochastic approximation: a dynamical systems viewpoint. 2008. https://doi.org/10.1007/978-93-86279-38-5
49. Tsitsiklis J, Van Roy B. Analysis of Temporal-Difference Learning with Function Approximation. Advances in Neural Information Processing Systems. 1996.
50. Imani E, White M. Improving regression performance with distributional losses. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 2157–66.
51. Farebrother J, et al. Stop Regressing: Training Value Functions via Classification for Scalable Deep RL. In: Proceedings of the 41st International Conference on Machine Learning, 2024. 13049–71.
52. van Duuren E, Lankelma J, Pennartz CMA. Population coding of reward magnitude in the orbitofrontal cortex of the rat. J Neurosci. 2008;28(34):8590–603. https://doi.org/10.1523/JNEUROSCI.5549-07.2008 PMID: 18716218
53. Ferrari-Toniolo S i m o n e, Schultz W o l f r a m. Reliable population code for subjective economic value from heterogeneous neuronal signals in primate orbitofrontal cortex. Neuron. 2023;111(22):3683-3696.e7. https://doi.org/10.1016/j.neuron.2023.08.009
54. Cazé RD, van der Meer MAA. Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern. 2013;107(6):711–9. https://doi.org/10.1007/s00422-013-0571-5 PMID: 24085507
55. Niv Y, Edlund JA, Dayan P, O'Doherty JP. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci. 2012;32(2):551–62. https://doi.org/10.1523/JNEUROSCI.5498-10.2012 PMID: 22238090
56. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004;306(5703):1940–3. https://doi.org/10.1126/science.1102941 PMID: 15528409
57. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007;104(41):16311–6. https://doi.org/10.1073/pnas.0706111104 PMID: 17913879
58. Gershman SJ. Do learning rates adapt to the distribution of rewards?. Psychon Bull Rev. 2015;22(5):1320–7. https://doi.org/10.3758/s13423-014-0790-3 PMID: 25582684
59. Palminteri S, Lefebvre G, Kilford EJ, Blakemore S-J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. PLoS Comput Biol. 2017;13(8):e1005684. https://doi.org/10.1371/journal.pcbi.1005684 PMID: 28800597
60. Skyrms B. Proceedings and Addresses of the American Philosophical Association. 2001;75(2):31–41. https://doi.org/10.2307/3218711
61. Belloc M, Bilancini E, Boncinelli L, D'Alessandro S. Intuition and Deliberation in the Stag Hunt Game. Sci Rep. 2019;9(1):14833. https://doi.org/10.1038/s41598-019-50556-8 PMID: 31619691
62. Jansson F, Eriksson K. Cooperation and Shared Beliefs about Trust in the Assurance Game. PLoS One. 2015;10(12):e0144191. https://doi.org/10.1371/journal.pone.0144191 PMID: 26640892
63. Jagau S. To Catch a Stag: Identifying payoff- and risk-dominance effects in coordination games. Games and Economic Behavior. 2024;147:429–48. https://doi.org/10.1016/j.geb.2024.08.011
64. Barfuss W, Donges JF, Lade SJ, Kurths J. When optimization for governing human-environment tipping elements is neither sustainable nor safe. Nat Commun. 2018;9(1):2354. https://doi.org/10.1038/s41467-018-04738-z PMID: 29907743
65. Andreoni J, Sprenger C. Risk Preferences Are Not Time Preferences. American Economic Review. 2012;102(7):3357–76. https://doi.org/10.1257/aer.102.7.3357
66. Mihatsch O, Neuneier R. Risk-Sensitive Reinforcement Learning. Machine Learning. 2002;49(2–3):267–90. https://doi.org/10.1023/a:1017940631555
67. Shen Y, Tobia MJ, Sommer T, Obermayer K. Risk-sensitive reinforcement learning. Neural Comput. 2014;26(7):1298–328. https://doi.org/10.1162/NECO_a_00600 PMID: 24708369
68. Puterman L. Markov decision processes: discrete stochastic dynamic programming. 1994.
69. Lowet AS, Zheng Q, Matias S, Drugowitsch J, Uchida N. Distributional Reinforcement Learning in the Brain. Trends Neurosci. 2020;43(12):980–97. https://doi.org/10.1016/j.tins.2020.09.004 PMID: 33092893
70. Lowet AS, Zheng Q, Meng M, Matias S, Drugowitsch J, Uchida N. An opponent striatal circuit for distributional reinforcement learning. Nature. 2025;639(8055):717–26. https://doi.org/10.1038/s41586-024-08488-5 PMID: 39972123
71. Muller TH, Butler JL, Veselic S, Miranda B, Wallis JD, Dayan P, et al. Distributional reinforcement learning in prefrontal cortex. Nat Neurosci. 2024;27(3):403–8. https://doi.org/10.1038/s41593-023-01535-w PMID: 38200183
72. Palminteri S, Lebreton M. Context-dependent outcome encoding in human reinforcement learning. Current Opinion in Behavioral Sciences. 2021;41:144–51. https://doi.org/10.1016/j.cobeha.2021.06.006
73. Bavard S, Palminteri S. The functional form of value normalization in human reinforcement learning. Elife. 2023;12:e83891. https://doi.org/10.7554/eLife.83891 PMID: 37428155

UNCORRECTED PROOF

74. Katahira K. The statistical structures of reinforcement learning with asymmetric value updates. Journal of Mathematical Psychology. 2018;87:31–45. https://doi.org/10.1016/j.jmp.2018.09.002
75. Sugawara M, Katahira K. Dissociation between asymmetric value updating and perseverance in human reinforcement learning. Sci Rep. 2021;11(1):3574. https://doi.org/10.1038/s41598-020-80593-7 PMID: 33574424
76. Senftleben U, Schoemann M, Scherbaum S. Choice Repetition Bias in Intertemporal Choice: An Eye-Tracking Study. Behavioral Decision Making. 2024;37(3). https://doi.org/10.1002/bdm.2388
77. Massin S, Nguyen-Van P, Dubois D, Willinger M, Ventelou B. Hysteresis in addictive consumption depends on time preferences. Theory Decis. 2025;100(3):769–806. https://doi.org/10.1007/s11238-025-10072-w
78. Breton Y-A, Seeland KD, Redish AD. Aging impairs deliberation and behavioral flexibility in inter-temporal choice. Front Aging Neurosci. 2015;7:41. https://doi.org/10.3389/fnagi.2015.00041 PMID: 25870560
79. Senftleben U, Schoemann M, Rudolf M, Scherbaum S. To stay or not to stay: The stability of choice perseveration in value-based decision making. Q J Exp Psychol (Hove). 2021;74(1):199–217. https://doi.org/10.1177/1747021820964330 PMID: 32976065
80. Banavar NV, Bornstein A. Multi-plasticities: distinguishing context-specific habits from complex perseverations. Habits: their definition, neurobiology, and role in addiction. 2024. p. 87–106. https://doi.org/10.1007/978-3-031-55889-4_5
81. Sandson J, Albert ML. Varieties of perseveration. Neuropsychologia. 1984;22(6):715–32. https://doi.org/10.1016/0028-3932(84)90098-8
82. Palminteri S. Choice-confirmation bias and gradual perseveration in human reinforcement learning. Behav Neurosci. 2023;137(1):78–88. https://doi.org/10.1037/bne0000541 PMID: 36395020
83. Hoxha I, Sperber L, Palminteri S. Evolving choice hysteresis in reinforcement learning: Comparing the adaptive value of positivity bias and gradual perseveration. Proc Natl Acad Sci U S A. 2025;122(36):e2422144122. https://doi.org/10.1073/pnas.2422144122 PMID: 40892918
84. Gershman SJ. Origin of perseveration in the trade-off between reward and complexity. Cognition. 2020;204:104394. https://doi.org/10.1016/j.cognition.2020.104394 PMID: 32679270
85. Torregrossa MM, Quinn JJ, Taylor JR. Impulsivity, compulsivity, and habit: the role of orbitofrontal cortex revisited. Biological Psychiatry. 2008;63(3):253.
86. Gillan CM, Robbins TW, Sahakian BJ, van den Heuvel OA, van Wingen G. The role of habit in compulsivity. Eur Neuropsychopharmacol. 2016;26(5):828–40. https://doi.org/10.1016/j.euroneuro.2015.12.033 PMID: 26774661
87. Dabney W, et al. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 1096–105.
88. Bandura A. Self-efficacy mechanism in human agency. American Psychologist. 1982;37(2):122–47. https://doi.org/10.1037/0003-066x.37.2.122
89. Palminteri S, Lebreton M. The computational roots of positivity and confirmation biases in reinforcement learning. Trends Cogn Sci. 2022;26(7):607–21. https://doi.org/10.1016/j.tics.2022.04.005 PMID: 35662490
90. Bergerot C, Barfuss W, Romanczuk P. Moderate confirmation bias enhances decision-making in groups of reinforcement-learning agents. PLoS Comput Biol. 2024;20(9):e1012404. https://doi.org/10.1371/journal.pcbi.1012404 PMID: 39231162
91. Lenton TM, et al. Global tipping points report 2025. 2025.
92. Barrett S, Dannenberg A. Sensitivity of collective action to uncertainty about climate tipping points. Nature Clim Change. 2013;4(1):36–9. https://doi.org/10.1038/nclimate2059
93. Tavoni A, İriş D. Tipping and reference points in climate change games. Handbook on the Economics of Climate Change. Edward Elgar Publishing. 2020. https://doi.org/10.4337/9780857939067.00018
94. Littman ML. Markov games as a framework for multi-agent reinforcement learning. Machine Learning Proceedings 1994. Elsevier. 1994. p. 157–63. https://doi.org/10.1016/b978-1-55860-335-6.50027-1
95. Rowland M, et al. An analysis of categorical distributional reinforcement learning. In: Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, 2018. 29–37.

UNCORRECTED PROOF