Evolutionary Reinforcement Learning Dynamics with Irreducible Environmental Uncertainty *

Wolfram Barfuss$^{1,2}$ and Richard P. Mann$^2$

Abstract

In this work we derive and present evolutionary reinforcement learning dynamics in which the agents are irreducibly uncertain about the current state of the environment. We evaluate the dynamics across different classes of partially observable agent-environment systems and find that irreducible environmental uncertainty can lead to better learning outcomes faster, stabilize the learning process and overcome social dilemmas. However, as expected, we do also find that partial observability may cause worse learning outcomes, for example, in the form of a catastrophic limit cycle. Compared to fully observant agents, learning with irreducible environmental uncertainty often requires more exploration and less weight on future rewards to obtain the best learning outcomes. Furthermore, we find a range of dynamical effects induced by partial observability, e.g., a critical slowing down of the learning processes between reward regimes and the separation of the learning dynamics into fast and slow directions. The presented dynamics are a practical tool for researchers in biology, social science and machine learning to systematically investigate the evolutionary effects of environmental uncertainty.

1 Introduction

Uncertainty is a fundamental feature of life. Even if we know how the world works, we might not know what will happen (stochastic uncertainty), what currently is (state uncertainty) and what others are doing (strategic uncertainty), among other forms of uncertainty (Kochenderfer, 2015, Halpern, 2017, Marchau et al., 2019).

Evolutionary game theory (Hofbauer and Sigmund, 1998) takes into account strategic uncertainty by not making assumptions about the rationality of other agents. Instead, the agents adapt to each other sequentially. Tools and methods from evolutionary game theory have also been used successfully to formally study the dynamics of multi-agent reinforcement learning (Bloembergen et al., 2015, Barfuss, 2020b). Bürgers and Sarin (1997) established the formal relationship between the learning behavior of one of the most basic reinforcement learning schemes, Cross learning (Cross, 1973), and the replicator dynamics of evolutionary game theory. Since then, this approach of evolutionary reinforcement learning dynamics has been extended to stateless Q-learning (Tuyts et al., 2003, Sato and Crutchfield, 2003), regret-minimization (Klos et al., 2010) and temporal-difference learning (Barfuss et al., 2019), as well as discrete-time dynamics (Galla and Farmer, 2013), continuous strategy spaces (Galla, 2013) and extensive-form games (Panozzo et al., 2014).

Apart from strategic uncertainty, representing stochastic uncertainty, i.e., uncertainty about what will happen in the form of probabilistic events within the environment, requires foremost the presence of an environment. Recent years have seen a growing interest to move evolutionary dynamics in stateless games to changing environments. Here, the term environment can mean external fluctuations (Assaf et al., 2013, Ashcroft et al., 2014), a varying population density (Hauert et al., 2006, Gokhale and Hauert, 2016), spatial network structure (Gracia-Lázaro et al., 2013, Szolnoki and Chen, 2018), or coupled systems out of evolutionary and environmental dynamics. Coupled systems may further be categorized into those with continuous environmental state spaces (Tavoni et al., 2012, Weitz et al., 2016, Chen and Szolnoki, 2018, Tilman et al., 2020) or discrete ones (Hilbe et al., 2018, Barfuss et al., 2019, Hauert et al., 2019, Su et al., 2019). We’ll be focusing on evolutionary and learning dynamics in stochastic games (Hilbe et al., 2018, Barfuss et al., 2019) which encode stochastic uncertainty via action-dependent transition probabilities between environmental states.

However, all dynamics discussed so far are either applicable only to stateless environments, assume that agents do not tailor their response to the current environmental state, or if they do, assume that agents observe the true states of the environment perfectly. Yet, often in real-world settings state observations are noisy and incomplete. Thus, they are lacking state uncertainty.

In this work, we relax the assumption of perfect observations and introduce evolutionary reinforcement learning dynamics for partially observable environments. Partial observability, state uncertainty or irreducible environmental uncertainty are synonymous terms to us. With the derived dynamics we are able to study the idealized reinforcement learning behaviour in a wide range of environmental classes, from partially observable Markov decision processes (POMDPs, Spaan, 2012), decentralized (Dec-)POMDPs (Oliehoek and Amato, 2016), and fully general partially observable stochastic games (Hansen et al., 2004).

Note, while a great deal of works on partially observable decision domains is of normative nature, this one is descriptive. Often, agents are enriched with, e.g., generative models and belief-state representations (Spaan, 2012, Oliehoek and Amato, 2016), abstractions (Sutton et al., 2006) or predictive state representations (Littman et al., 2001) in order to learn optimal strategies in partially observable decision domains.
Also the economic value of signal is often studied by asking how fully rational agents optimally deal with a specific from of state uncertainty (Bagh and Kusunose, 2020). However, such techniques can become computationally extremely expensive (Loch and Singh, 1998). It is unlikely that biological agents perform those elaborate calculations (Gigerenzer and Gaismaier, 2011) and the focus on unboundedly rational game equilibria lacks a dynamic perspective (Papadimitriou and Pilouras, 2019) making it unable to answer which equilibrium (of the many often) the agents select.

Instead, this work studies the evolutionary dynamics of individual learning agents employing the widely-occurring principle of temporal-difference reinforcement learning (Sutton, 1988) in which the agents simply treat their observations as if they were the true Markov states of the environment. Temporal-difference learning is not only a computational technique (Sutton and Barto, 2018), it also occurs in biological agents through the dopamine reward-prediction error signal (Schultz et al., 1997, Dayan and Niv, 2008). We focus on agents which employ either so called memoryless policies, at which they choose their actions based solely on their current observation (Singh et al., 1994), or they use a short and fixed history of current and past observations and actions to base the current action upon. This has the advantage of being simple to act upon (Williams and Singh, 1998) and they are easy to realize at no or little additional computational cost.

Put differently, we ask the question, how robust is the classic temporal-difference reinforcement learning process when agents are no longer capable to observe the true states of the environment? To be able to answer this question thoroughly we derive evolutionary reinforcement learning dynamics with irreducible environmental uncertainty (Sec. 3) after introducing the necessary background and notation (Sec. 2). Our dynamics show that learning under state uncertainty may cause worse learning outcomes, as expected and reported (Singh et al., 1994). Yet, we find that this worse learning outcome may manifest itself in the form of a catastrophic limit cycle. Interestingly, we also find that state uncertainty can lead to better learning outcomes faster, stabilize a chaotic learning process and overcome social dilemmas (Sec. 4). Thus, it depends on the environment and its representation whether temporal-difference learning alone is sufficient to utilize state-uncertainty to its advantage or to what extent sophisticated but expensive contemplations about the uncertainty are necessary (Sec. 5).

2 Background

2.1 Partially observable stochastic games

Definition. The game \( G = (N, S, \mathcal{A}, Q, T, R, O) \) is a stochastic game with \( N \in \mathbb{N} \) agents. The environment consists of \( Z \in \mathbb{N} \) states \( S = (S_1, \ldots, S_Z) \). In each state \( s \), each agent \( i \in \{1, \ldots, N\} \) has \( M_i \in \mathbb{N} \) available actions \( \mathcal{A}_i = (A_{i1}, \ldots, A_{iM}) \) to choose from. \( \mathcal{A} = \bigotimes_i \mathcal{A}_i \) is the joint-action set and agents choose their actions simultaneously. A joint action is denoted by \( \mathbf{a} = (a_1, \ldots, a_N) \in \mathcal{A} \). With \( a_{i1} = (a_{i1}, \ldots, a_{iM}) \in \mathcal{A}_i \) we denote the joint action except agent \( i \)'s. We chose an identical number of actions for all states and all agents out of notational convenience. Throughout this paper, we restrict ourselves to ergodic environments without absorbing states.

The transition function \( T : S \times \mathcal{A} \times S \to [0, 1] \) determines the probabilistic state changes. \( T(s, \mathbf{a}, s') \) is the transition probability from current state \( s \) to next state \( s' \) under joint action \( \mathbf{a} \).

The reward function \( R : S \times \mathcal{A} \times S \to \mathbb{R} \) maps the triple of current state \( s \), joint action \( \mathbf{a} \) and next state \( s' \) to an immediate reward scalar for each agent. \( R_i(s, \mathbf{a}, s') \) is the reward agent \( i \) receives.

Instead of observing the states \( s \in S \) directly, each agent \( i \) observes one of \( Q \in \mathbb{N} \) observations \( O^i = (O^i_1, \ldots, O^i_Q) \) according to the observation functions \( O^i : S \times \mathcal{A} \to [0, 1] \). \( O^i(s, a) \) is the probability that agent \( i \) observes observation \( o \in O^i \) given that the environment is in state \( s \in S \). \( \mathcal{O} = \bigotimes_i O^i : S \times \mathcal{A} \to [0, 1]^N \) is the joint observation set. We chose an identical number of observations for all agents out of notational convenience. By construction, this observation function can model both noisy state observations \( Q = Z \) and hidden states \( Q < Z \).

Policies. We consider agents that choose their actions probabilistically according to their memoryless policy \( X^i : O^i \times \mathcal{A}^i \to [0, 1] \). \( X^i(o^i, a^i) \) is the probability that agent \( i \) chooses action \( a^i \) given that it observed observation \( o^i \). We denote the joint policy by \( X = X(q, \mathbf{a}) = \bigotimes_i X^i(o^i, a^i) : \mathcal{O} \times \mathcal{A} \to [0, 1]^N \).

Histories. Besides memoryless policies we also consider policies with fixed histories \( H_h \) of type \( h \). The type \( h \) is composed of \( h = h_a \times h_o \) with \( h_a \in \mathbb{N} \) and \( h_o \in \mathbb{N}^h \). \( h_a \) represents how many of current and past observations are to be used to encode the histories. Likewise, \( h_o \) represents how many past actions of each agent are to be encoded in the histories. For example, the default memoryless policy is of type \( h = (1, 0) \).

Practically, histories induce an embedding of the game into a larger state space at which the histories \( H_h \) correspond to the larger state set and transitions, rewards and observations are adjusted accordingly.

2.2 Temporal-difference reinforcement learning

Temporal-difference Q-learning is one of the most widely studied reinforcement learning processes (Watkins and Dayan, 1992, Schultz et al., 1997, Sutton and Barto, 2018). Agents successively improve their evaluations of the quality of the available actions. Originally developed under the assumption that agents can observe the true Markov state of the environment, we here present the basic temporal-difference Q-learning algorithm in the more general formulation, where agents use observations instead of states. When observations exactly map onto the states, the original algorithm is recovered.

At time step \( t \) agent \( i \) evaluates action \( a^i \) at observation \( o^i \) to be of quality \( Q^i_t(o^i, a^i) \). Those state-action values \( Q^i_t(o^i, a^i) \) are then updated after selecting action \( a^i_{t+1} \) after observing observation \( o^i_t \) according to

\[
Q^i_{t+1}(o^i_t, a^i_{t+1}) = Q^i_t(o^i_t, a^i_{t+1}) + \alpha \cdot \delta^i_t(o^i_t, a^i_{t}),
\]

with the temporal-difference error

\[
\delta^i_t(o^i_t, a^i_{t}) := (1-\gamma)r^i_t + \gamma \max_b Q^i_{t}(o^i_{t+1}, b) - Q^i_{t}(o^i_t, a^i_{t}).
\]

The discount factor parameter \( \gamma \in [0, 1] \) regulates how much the agent cares for future rewards. The learning rate
parameter $\alpha \in (0,1)$ regulates how much new information is used for an observation-action-value update. For the sake of simplicity, we assume identical parameters across agents throughout this paper and therefore do not equip parameters with agent indices. The variable $r^i_t$ refers to the immediate reward at time step $t$. Note that the $(1-\gamma)$ prefactor in front of the reward occurs when we assume that agents aim to maximize a return defined as $G_t^i = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k r^i_{t+k}$ (Barfuss et al., 2019). This leads the values to be on the same scale as the rewards.

Agents select actions based on the current observation-action values $Q^i_t(o^i, a^i)$ balancing exploitation (i.e., selecting the action of maximum quality) and exploration (i.e., selecting lower quality actions in order to learn more about the environment). We here use the widely used Boltzmann policy. The probability of choosing action $a^i$ under observation $o^i$ is

$$X^i_t(o^i,a^i) = \frac{e^{\beta Q^i_t(o^i,a^i)}}{\sum_{b \in A^i} e^{\beta Q^i_t(o^i,b)}}$$

(3)

where the intensity of choice parameter $\beta$ controls the exploration-exploitation trade-off. Throughout this paper, we are interested in the idealized learning process with fixed parameters $\alpha$, $\beta$ and $\gamma$ throughout learning and evaluating a policy.

3 Derivation

In this section we derive the evolutionary reinforcement learning dynamics under partial observability in discrete time. As classic evolutionary dynamics operate in the theoretical limit of an infinite population, the learning dynamics are derived by considering an infinite memory batch (Barfuss, 2020a, 2021). A learning dynamic update of the current policy uses policy-averages instead of individual samples. Thus, we need to construct the policy-average temporal-difference error $\delta^i$ to be inserted in the update for the joint policy,

$$X^i_{t+1}(o^i,a^i) = \frac{X^i_t(o^i,a^i) \cdot \exp[\alpha \beta \delta^i_t(o^i,a^i)]}{\sum_{b \in A^i} X^i_t(o^i,b) \cdot \exp[\alpha \beta \delta^i_t(o^i,b)]}.$$  

(4)

Eq. 4 can be derived by combining Eqs. 1 and 3. The bar on top of $\delta^i$ indicates implicitly that $\delta^i$ depends fully on the current joint policy $X_t$. Computing $\delta^i(o^i,a^i)$ involves averaging over policies, environmental transitions and observations for the first two terms of the temporal-difference error (Eq. 2), the immediate rewards and the qualities of the next observation. The quality of the current observation, $Q^i_t(o^i,a^i)$ becomes $\beta^{-1} \ln X^i_t(o^i,a^i)$ in the average temporal-difference error and serves as regularization term. This can be derived by inverting Eq. 3 and realizing that the dynamics induced by Eq. 4 are invariant under additive transformations which are constant in actions.

The challenge is that the rewards $R^i_t(s,a^i)$ in the stochastic game model depend on the true states, not on the observations of the agents. Thus, in order to obtain the average observation-action rewards $R^i(o^i,a^i)$, we need a mapping from observations to states. The observation function is a mapping from states to observations. With Bayes rule,

$$\bar{B}^i(o^i,s) = \frac{O^i(o^i,s)P(s)}{\sum_{s'} O^i(o^i,s')P(s)}$$

(5)

we can transform the observation function into a belief function, following the rules of probability. $\bar{B}^i(o^i,s)$ is the belief of agent $i$ (or simply the probability) that the environment is in state $s$ when it observed observation $o^i$.

The only problem is to obtain the policy-average stationary state distribution $\bar{P}(s)$. $\bar{P}(s)$ is the left-eigenvector of the average transition matrix $\bar{T}(s,s')$ which the entry $\bar{T}(s,s')$ denotes the probability of transitioning from state $s$ to state $s'$. This matrix could be obtained as $\bar{T}(s,s') = \prod_{j} \sum_{o^j} \bar{Y}^j(s,a^j)T(s,g^j,s')$ if we had the probability for each agent $j$ to choose action $a^j$ in state $s$, $\bar{Y}^j(s,a^j)$. However, we assumed that agents condition their actions only on observations, $X^j(o^j,a^j)$. Yet, whenever the environment is in state $s$, agent $j$ observes observation $o^i$ with probability $O^j(s,o^j)$ and than chooses action $a^j$ with probability $X^j(o^j,a^j)$. Thus, with

$$\bar{Y}^j(s,a^j) := \sum_{o^j \in O^j} O^j(s,o^j)X^j(o^j,a^j),$$

(6)

we can average out the observation and obtain the policy-average state-policies $\bar{Y}^j(s,a^j)$. Note that $\bar{Y}^j(s,a^j)$ are proper conditional probabilities, which can be seen by applying $\sum_{o^j}$ to both sides of Eq. 6. With $\bar{Y}^j(s,a^j)$ we can then compute the policy-average transition matrix $\bar{T}(s,s')$, its left-eigenvector, the stationary state distribution $\bar{P}(s)$, and thus, the policy-average belief of agent $i$ that the environment is in state $s$ when it observed observation $o^i$, $\bar{B}^i(o^i,s)$.

Rewards. Whenever agent $i$ observes observation $o^i$, with probability $\bar{B}^i(o^i,s)$ the environment is in state $s$ where all other agents $j \neq i$ behave according to $\bar{Y}^j(s,a^j)$, the environment transitions to a next state $s'$ with probability $T(s,g,s')$, and agent $i$ receives the reward $R^i(s,g,s')$. Mathematically, the policy-average reward for action $a^i$ under observation $o^i$ reads

$$\bar{R}^i(o^i,a^i) := \sum_{s} \sum_{s'} \sum_{a^j \neq i} \bar{B}^i(o^i,s)\bar{Y}^j(s,a^j)T(s,g,s')\bar{R}^i(s,g,s').$$

(7)

Qualities. Second, the policy-average of the quality of the next observation (max$_{a^j} Q^j_t(o^j_{t+1},b)$ in Eq. 2) is computed by averaging over all states, all actions of the other agents, next states and next observations. Whenever agent $i$ observers observation $o^i$, the environment is in state $s$ with probability $\bar{B}^i(o^i,s)$. There, all other agents $j \neq i$ choose their action $a^j$ with probability $\bar{Y}^j(s,a^j)$. Consequently, the environment transitions to the next state $s'$ with probability $T(s,g,s')$. At $s'$, the agent observes observation $o^i$ with probability $O^i(s',o^i)$ and estimates the quality to be of value max$_{a^j} Q^j_t(o^j,b)$. Mathematically, we write

$$\max Q^i_t(o^i,a^i) := \sum_s \sum_{s'} \sum_{a^j \neq i} \sum_{b} \bar{B}^i(o^i,s)\bar{Y}^j(s,a^j)T(s,g,s')O^i(s',o^i) \max Q^j_t(o^j,b).$$

(8)

Here, we replace the quality estimates $Q^j_t(o^j,a^i)$, which evolve in time $t$ (Eq. 1), with the policy-average observation-action quality $\bar{Q}^i_t(o^i,a^i)$, which is the expected discounted sum of future rewards from executing action $a^i$ at observation $o^i$ and then following along the joint policy $X$. It is obtained
by a discount factor weighted average of the current policy-average reward $R^i(o', a')$ and the policy-average observation quality of the next observation $V^i(o')$,

$$Q^i(o', a') = (1 - \gamma)T^i(o', a') + \gamma \sum_{o' \in O} T^i(o', a', o')V^i(o'). \quad (9)$$

Here, $T^i(o', a', o')$ is agent $i$’s policy-average transition probability of observing observation $o'$ at the next time step given it observed observation $o'$ at the current time step and chose action $a'$. It is computed by averaging over all states, next states and all actions of the other agents. Whenever agent $i$ observes observation $o'$ and selects action $a'$, the environment is in state $s$ with probability $B^i(o', s)$, where all other agents $j \neq i$ select action $a_j$ with probability $Y^j(s, a_j)$. Consequently, the environment will transition to the next state $s'$ with probability $T(s, a_i, s')$ which is observed with probability $O^i(s', o')$ as $o'$ by agent $i$. Mathematically, we write

$$T^i(o', a', o') = \sum_s \sum_{a_i} \sum_{s'} \prod_j B^i(o', s)Y^j(s, a^j)T(s, a_i, s')O^i(s', o'). \quad (10)$$

Further at Eq. 9, $V^i(o)$ is the policy-average observation quality, i.e., the expected discounted sum of future rewards from observation $o'$ and then following along the joint policy $X$. They are computed via matrix inversion according to

$$V^i(o) = (1 - \gamma)[\mathbb{I}_Q - \gamma T^i(o, o)]^{-1}R(o). \quad (11)$$

This equation is a direct conversion of the Bellman equation $V^i(o') = (1 - \gamma)R(o') + \gamma \sum_{o'} T^i(o', o')V^i(o')$, which expresses that the value of the current observation is the discount factor weighted average of the current reward and the value of the next observation. Underlined observation variables indicate that the corresponding object is a vector or matrix and $\mathbb{I}_Q$ is a $Q$-by-$Q$ identity matrix.

$T^i(o, o)$ denotes the policy-averaged transition matrix for agent $i$. The entry $T^i(o, o')$ indicates the probability that agent $i$ will observe observation $o'$ after observing observation $o'$ at the previous time step, given all agents follow the joint policy $X$. We compute them by averaging over all states, all actions from all agents and all next states,

$$T^i(o, o') = \sum_s \sum_{a_i} \sum_{s'} \prod_j B^i(o, s)Y^j(s, a^j)T(s, a_i, s')O^i(s', o'). \quad (12)$$

For any observation $o'$, $B^i(o, s)$ is the probability to be in state $s$, where all agents $j$ act according to $Y^j(s, a^j)$. Therefore, the environment transitions with probability $T(s, a_i, s')$ from state $s$ to the next state $s'$, which is observed by agent $i$ as observation $o'$ with probability $O^i(s', o')$. Note that $T^i(o, o)$ is a proper probabilistic matrix. This can be seen by applying $\sum_v$ to both sides of Eq. 12.

Further in Eq. 11, $R^i(o')$ denotes the policy-average reward agent $i$ obtains from observation $o'$. We compute them by averaging over all states, all actions from all agents and all next states. Whenever agent $i$ observes observation $o'$, the environment is in state $s$ with probability $B^i(o', s)$. Here, all agents $j$ choose action $a_j$ with probability $Y^j(s, a_j)$. Hence, the environment transitions to the next state $s'$ with probability $T(s, a_i, s')$ and agent $i$ receives the reward $R^i(s, a_i, s')$.

$$R^i(o') := \sum_s \sum_{a_i} \sum_{s'} \prod_j B^i(o', s)Y^j(s, a^j)T(s, a_i, s')R^i(s, a_i, s'). \quad (13)$$

Note that the quality $\max_{a_i}Q^i(o', a')$ depends on $o'$ and $a'$ although it is the policy-averaged maximum observation-action value of the next observation.

**TD error.** All together, the policy-average temporal-difference error, to be inserted into Eq. 4, reads

$$\delta^i(o', a') = (1 - \gamma)R^i(o', a') + \gamma \sum_{o'} T(s, a_i, s')R^i(s, a_i, s') - \frac{\ln X^i(o', a')}{\beta}. \quad (14)$$

4 **Evaluation**

We evaluate the derived dynamics across five test environments: three single-agent decision problems and two multi-agent games. Three environments will cover noisy observations, the other two focus on a reduced observation space, where a given observation is consistent with multiple true states of the world. As evaluation metric we use the average reward, $\sum_j P(s)R^j(s)$, where $P(s)$ is the stationary state-distribution and $R^j(s) = \sum s' \sum_j Y^j(s, a^j)T(s, a_i, s')R^j(s, a_i, s')$ is the average reward for each state given the current policy $X$ (see Sec. 3). We defined a learning trajectory as having converged if the norm between old and updated policy (according to Eq. 4) is below $10^{-5}$. Since we defined the return with the $(1 - \gamma)$ prefactor we also consider a scaled version of the intensity of choice parameter $\beta = \beta/(1 - \gamma)$ for some of the evaluation examples. Doing so preserves the ratio of exploration and exploitation in the temporal-difference error (Eq. 14) under changes in the discount factor $\gamma$.

4.1 **Simple coordination task**

**Environment description.** The first environment is a simple coordination task in which the agent must move between the left and right environmental state in order to obtain a maximum reward of 1. Coordinating which of the two available actions (Left, Right) to choose from is complicated by observational noise $\nu$, letting the agent perceive the correct state only with probability $1 - \nu$ (Fig. 1A). This environment is adapted from Singh et al. (1994).

**Results.** Fig. 1 shows how irreducible environmental uncertainty can cause the evolutionary learning dynamics to enter low-rewarding limit cycles under a high intensity of choice. Often, learning a policy involves a trade-off between the amount of reward from that policy and the amount of time required learning it. In the simple coordination task with perfect observation, a high intensity of choice can speed up the learning process by a factor of 6. The trajectories with $\beta' = 40$ require about 18 time steps to arrive at the optimal policy with average reward 1 (green lines, top row), the trajectories with $\beta' = 400$ require only 3 time steps (green lines, bottom row).
**Figure 1. Evolutionary learning dynamics in a simple coordination task.** The agent has to anti-coordinate its actions on the environmental state, which it observes through a noisy channel (Panel A). Learning trajectories are computed for three observational noise levels (0.0, 0.49, 0.5) and two intensities of choice, a low one ($\beta' = \beta / (1 - \gamma) = 40$), shown in the top row, and a high one ($\beta' = 400$), shown in the bottom row. Panel B shows the corresponding policy spaces in which the agent’s probability of choosing Left, given the agent perceived the environment to be in the left state, is plotted on the x-axes; and the agent’s probability of choosing Left, given the agent perceived the environment to be in the right state, is plotted on the y-axes. 5 individual trajectories, whose initial policies were centered around the center of the policy space, are plotted in color. Arrows in gray indicate the flow of the learning dynamical system. Panel C shows the corresponding reward trajectories. Remaining hyper-parameters were set as $\alpha = 0.01$, $\gamma = 0.9$. Irreducible environmental uncertainty can cause the learning to enter low-rewarding limit cycles under high intensity of choice.

**Figure 2. Evolutionary learning dynamics with history in a simple coordination task.** Same situation as in Fig. 1 with high intensity of choice $\beta' = 400$ and observational noise level 0.5 but here, the agent remembers and conditions its action not only on the last observation but also on its last action. Thus, the left (right) panel shows a projection of the learning dynamics, given the last action was Left (Right). X-axes show the probabilities of choosing Left, given the last environmental observation was left. Y-axes show the probabilities of choosing Left, given the last environmental observation was right. The agent learns to alternate between Left and Right, which yield the highest reward possible, in - at most - only two time steps.

Thus, a high intensity of choice is clearly preferable under perfect observation.

With fully uninformative observations (observational noise level $\nu = 0.5$, Fig. 1 B, third column) a more explorative agent (i.e. lower intensity of choice, top row) has an advantage. From all initial policies, it takes the agent about 580 time steps to learn to fully randomize its actions. This yields an average reward of zero and is also the optimal memoryless policy (Singh et al., 1994). The more exploitative agent (bottom row) on the other hand enters a limit cycle between choosing Left and Right almost deterministically, irrespective of its current observations. Thus, while choosing Left the agent is trapped in the LEFT state, obtaining an average reward of $-1$. While choosing Right, the agent is trapped in the RIGHT state also obtaining an average reward of $-1$. The positive reward obtained through the move between states is neglected, since the derived dynamics consider the theoretical limit of an infinite memory batch (Barfuss, 2020a). This can also be interpreted as a complete separation of the interaction time scale and the adaptation time scale (Barfuss et al., 2019, Barfuss, 2021). The agent experiences an infinite amount of negative reward during interaction and only one single positive reward after the policy adaptation. It will be interesting to reexamine this scenario under relaxed conditions when interaction and adaption time scales are not completely separated in future work (cf., Galla, 2009).

When observations are almost completely noisy, yet still contain some information about the true environmental state, the more exploitative agent learns a slightly more rewarding policy faster (Fig. 1 B second column). An observational noise level of $\nu = 0.49$ means that out of 100 times being in the LEFT environmental state, the agent will observe on average left 51 times and right 49 times. Here, from all initial policies, the more explorative agent (top row) converges to a fixed point
in the upper left part of the policy space in about 600 time steps, i.e., slower than under completely noisy observations. This policy yields an average reward of about 0.013. The more exploitative agent (bottom row) learns on an interesting transient resembling the limit cycle of the fully uninformative case, yet manages to converge to the deterministic policy in the upper left of the policy space. This yields an average reward of about 0.02 and takes at most 250 times, depending on the initial policy. This is still distinctly faster than the more explorative agent.

Overall, it is interesting to observe how the environmental uncertainty separated the learning dynamics into a fast eigendirection along the diagonal from the bottom left to the top right of the policy space and a slow eigendirection perpendicular to that (Strogatz, 2018). The slow eigendirection corresponds to a coordinated policy where the agent’s observation is decisive for its actions. Along the fast eigendirection the agent’s policy is independent of its observations. The more explorative agent moves along these axes whereas the more exploitative agent overrides. Yet, as long as there is some information in the observations about the environmental state, the more exploitative agent learns better policies faster.

So far we examined only memoryless policies, i.e., policies that condition their choice of action only on the current observation. If the more exploitative agent ($\beta' = 400$) is able to condition its choice of action not only on the current observations but also on its last action, it learns the optimal policy with an average reward of 1 in at most only 2 time steps - even under fully uninformative state observations (Fig. 2). The agent learns to alternate between Left and Right. This learned policy and even the whole learning dynamics do not depend on the state-observation, as shown by the straight line trajectories and corresponding dynamical flow arrows in Fig. 2.

4.2 Navigation task

Environment description. The next environment is the single-agent navigation task adapted from Parr and Russell’s Grid World (1995). It consists of 11 states, 6 observations, 4 actions and 1 agent (Fig. 3 A). The agent can move north, south, east and west. If the agent would move into a wall it stays on its current patch. The agent wants to reach the patch in the upper right, which is rewarded by a reward of 1. However, entering the patch below is punished by a reward of −1. In both cases, the episode ends and the agent begins a new episode on one randomly chosen patch out of the nine other patches. All other state-action combinations yield zero reward. We use this environment to compare the effect of various hyper-parameter combinations on the learning behavior of an agent with partial observability to an agent with full observability. Under partial observability the agent can only observe whether or not there is a wall east and west of its current patch. Imagine, for example, a robot equipped only with haptic sensors on its sides or an insect with corresponding antennae. With full observability, the agent can distinguish each grid patch separately.

Results. In contrast to a fully observant agent, under environmental uncertainty neither a high weight on future rewards (large $\gamma$) nor a high intensity of choice (large $\beta'$) leads to the highest reward. Instead, the highest reward depends on the mutual combination of the two hyper-parameters.
points in policy space (light blue dots) are increasingly farther apart from the optimal policy (light red dots) the more steps the grid cell is away from the goal. However, when we turn to the agent with partial observability, it is the other way around. Here, less weight on future rewards and more exploration lead to a better average reward at convergence (dark colored dashed lines). This result can be explained as follows. In a fully observable Markov decision process there is always an optimal deterministic policy (Puterman, 2014). See how the light red straight lines converges to edge of most grid cells, indicating a deterministic action in that direction (Fig. 3 A). Yet, policies in a partially observable Markov decision process require stochasticity (Singh et al., 1994). More exploration directly ensures that, although not in a reward-targeted way. Less weight on future rewards might be advantageous under partial observability since too much weight on too distant rewards in the future cannot pay off when there is a fundamental uncertainty about which state the agent occupies or even about what the real states of the environment are. When the environment is irreducibly uncertain, anticipating too distant rewards does not have to be beneficial.

A systematic analysis of the hyper-parameter grid (Fig. 3 C) confirms that partial observability requires the right combination of the two hyper-parameters in order to learn the highest reward. Under full observability, simply setting a sufficiently high weight on future rewards $\gamma$ and a sufficiently strong intensity of choice $\beta$ (i.e., little exploration) leads to the average reward of the optimal policy. In contrast, for partial observability too much far-sightedness and too intense exploitation can hurt the performance of the agent. Instead, the optimal reward at convergence is obtained by a more randomly explorative and myopic agent.

With memoryless policies, the average reward of the partially observant agent is smaller by an order of magnitude compared to the fully observant agent. Fig. 4 compares the results of the two simplest types of history, i.e., where the agent uses one more piece of information. Thus, the agent conditions its actions either on the current observation and the last action $h = (1, 1)$ or on the current and last observations $h = (2, 0)$. Both types of history are able to obtain a similar maximum average reward, with a slight advantage for history $h = (1, 1)$. Although both are of the simplest type of history conceivable the difference in maximum reward between the partially observant agent and fully observant agent is already halved, compared to the partially observant agent without history.

Also, the set of hyper-parameter combinations that obtain a high average reward is shifted to the lower right corner of the parameter space where also the fully observant agent obtains its maximum. Interestingly though, the set of high rewarding hyper-parameter combinations is not identical across the two types of history. The action-depended history ($h = (1, 1)$) performs best with a high weight on future rewards $\gamma$ and more exploration, whereas the two-observation history ($h = (2, 0)$) obtains the highest rewards by more exploitation across a wider range of weights on future rewards $\gamma$.

Furthermore, the learner experiences a critical slowing down of its learning dynamics Strogatz (2018) before a hyper-parameter bifurcation into the high rewarding regime (Fig. 4, bottom row). In the area around the hyper-parameter regions which obtain high average reward (yellow area in the plots in the top row), the number of time steps it takes the agent to converge is distinctly higher compared to other hyper-parameter regions. Interestingly, this effect is absent in the memoryless learner of Fig. 3 (not shown). Utilizing such dynamical phenomena have the potential to improve the efficiency of hyper-parameter search.

4.3 Renewable Resource Harvesting

Environment description. Harvesting a renewable resource is a foundational challenge in environmental economics, the earth and sustainability science (Perrman et al., 2003, Lindkvist and Norberg, 2014, Barfuss et al., 2017, Geier et al., 2019).

Here, we use a standard logistic growth model, in which the resource stock $t+1 = (1 + r/C)$. The agent has three possible actions: harvest nothing, harvest a small amount, or harvest a large amount. What is small and large depends on the maximum amount, $\Delta s_{\text{max}}$, the resource regrows from environmental states $s \in \{0, ..., C - 1\}$. The agent has three possible actions: harvest nothing, harvest a small amount, or harvest a large amount. What is small and large depends on the maximum amount, $\Delta s_{\text{max}}$, the resource regrows from environmental states $S$. The small harvest amounts to $(1 - DE)\Delta s_{\text{max}}$, the large harvest amounts to $(1 + DE)\Delta s_{\text{max}}$, with $DE$ representing the deviation in the agent’s harvesting effort.

State transitions work as follows: The harvest amount is
Figure 5. **Evolutionary learning dynamics in renewable resource environments.** Panel A sketches the functioning of the renewable resource harvesting environment. 1) The agent decides on a harvest, which is subtracted from the current environmental stock. 2) The stock regrows according to a logistic function. 3) The stock is discretized by a normal distribution to have the number of states equaling the capacity $C$ of the logistic growth function. We set the growth rate $r = 0.8$, the effort deviation $\Delta E = 0.2$, the stock base level $\tilde{s}_{base} = 0.1$, and the environmental stochasticity $\sigma = 0.5$. Panel B shows the possible observation spaces—how the environment is represented by the agent—ordered by decreasing complexity, for a world in which there are five possible true environmental states. In the most complex (at the top) the agent perceives all real states of the world as distinct; in the least complex (at the bottom), the agent makes the same observation regardless of the true state. We investigate all representations where the agent perceives several adjacent states as a single coherent observation. Panels C shows the reward out-performance $R = r/r_{ac} - 1$ (red), the speed out-performance $S = 1 - 1/l_{ac}$ (blue), and the combined reward-speed out-performance $R \cdot S$ (if $R > 0 \land S > 0$) (purple) for all possible representations, for the four renewable resource environments with capacities $C$ and likewise number of states, 6 – 9. Out-performance is measured with respect to the agent which used the accurate representation of the environment and obtained a reward $r_{ac}$ in $l_{ac}$ time steps. For each environment, each dot represents the average out of 100 Monte Carlo simulations from random initial policies of a single representation, ordered from the most complex, i.e., the accurate one, on the left to the simplest, i.e., perceiving all states as one, on the right. Violin plots show the distribution of rewards and speed, relative to the agent with the accurate representation. The three top performing representations are shown schematically by the dashed lines. Additionally, the average rewards of the optimal discounted policy $R^*_{\gamma}$ and the optimal average-reward policy $R^*_{avg}$ are shown. The agent’s discount factor $\gamma = 0.9$, intensity of choice $\beta' = 25$ and learning rate $\alpha = 0.02$. There exist imperfect representations of the environment that lead to a better learning outcome faster than the perfect representation.
subtracted from the current stock state $s_t$. The stock regrows according to the logistic growth equation, yielding a new hypothetical stock $\tilde{s}_{t+1}$. In order to avoid the complete depletion of the resource, the minimum hypothetical stock yields a value proportional to a base level $\delta_{base}$. Since the agent should have an influence on the regrowth of the resource, $\delta_{base}$ is multiplied by $(1 + \Delta E)$ if the agent chose to harvest nothing, by $(1 - \Delta E)$ if the agent chose to harvest a lot. The resource stock is then discretized by a normal distribution around $\tilde{s}_{t+1}$ with variance $\sigma^2$. The probability mass that lies between stock $s_{t+1} - 0.5$ and $s_{t+1} + 0.5$ gives the probability to transition to the new state $s_{t+1}$. (For $s_{t+1} = 0$ the lower bound is $-\infty$, for $s_{t+1} = C$ the upper bound is $+\infty$.) Thus, $\sigma$ represents the level of stochasticity within the environmental dynamics.

The rewards are identical to the harvest amount. Harvesting a lot yields a higher immediate reward than harvesting a little. Except when the resource is degraded, i.e., either the current state $s_t$ or the next state $s_{t+1}$ equals zero, than the rewards are only 10% of the harvest amount. Thus, the agent has always an immediate incentive to harvest more over a little. The optimal policy depends on the weight the agent puts on future rewards ($\gamma$).

We use this environment to investigate the effect of different (imperfect) representations of the environment. We focus on representations under which the agent perceives several adjacent states as a single coherent observations. Fig. 5 A & B illustrates the renewable resource harvesting environment and the investigated observation representations for capacity $C = 5$.

**Results.** We find that inaccurate (reduced complexity) representations of the environment can lead to a better learning outcome faster, when compared to an agent which perceives the environment accurately (at full complexity) (Fig 5).

In the majority of cases an inaccurate representation of the environment leads to a speed out-performance in the order of 10%, i.e., a smaller number of time steps it takes the learner to converge to a fixed point. Only four representations of the 9-state environment take distinctly longer to converge. Overall, there is a slight tendency that simpler representations lead to faster convergence. Representations (dots) are ordered from the most complex, i.e., the accurate one, on the left to the simplest, i.e., perceiving all states as one, on the right (per environment). All top speed representations (dashed bars) cluster the resource stock 0 and 1 together but separate between stock 1 and 2. In the environments with capacity 8 and 9, a resource stock of 2 is represented completely separate by all top speed representations.

In contrast, the majority of inaccurate representations lead to a worse reward at convergence. Clearly visible by the red dots on the right for each environment, the simpler the representation the worse the performance. Nevertheless, a few representations of intermediate complexity lead to a reward out-performance in the order of 1%. This is remarkable, since Blackwell (1953)’s theorem showed that a rational decision maker cannot improve by an inaccurate representation. Of course, our result does not contradict Blackwell, since we investigate a learning process.

To better understand the relationship between the learning process and the rational optimal policies Tab. 1 shows the average reward of the optimal policy $R^{*}\gamma$ and the average-reward optimal policy $R^{*,avg}$ relative to the reward obtained by the fully observant agent (shown in Fig 5 by diamonds and downward triangles). The optimal policy maximizes the state values for each state and depends on the discount factor $\gamma$. The average-reward optimal policy maximizes the average reward. Since in this environment $R^{*}\gamma$ approaches $R^{*,avg}$ under $\gamma \to 1$, the rewards between $R^{*,avg}$ and $R^{*}\gamma|_{\gamma=0.9}$ represent the rewards more patient or future caring agents could obtain. Thus, the out-performing representations cause the learner to behave as if it were more patient or future-oriented than it actually is (defined by its discount factor $\gamma$). However, it is not obvious to identify regularities across the environments between the top rewarding representations. Moreover, Tab. 1 shows that the learning process under full observability yields decent results. For the environment with 8 states the learner obtains the exact same reward as the optimal policy. In the environment with 6 states the learner obtains an average reward which is even above the one of the optimal policy. Taken together, a speed out-performance in the order of 10% multiplied by a reward out-performance in the order of 1% leads to combined speed-reward out-performance in the order of 0.1%. Along the four environments investigated, the magnitude in out-performance is increasing with the number of environmental states. Future work is needed to investigate this effect in larger, more complex resource harvesting environments and also how to obtain those representations which lead to better outcomes faster.

Notably, this result resembles the one by Mark et al. (2010) who show also that simpler views on the world can be of advantage. However, in their model perceiving the truths comes with a cost which is subtracted from the rewards of the environment. If this cost parameter is sufficiently large, perceiving the truths cannot pay off by design. We do not model such a cognitive cost of being close to the truths and still find that some inaccurate representations lead to better outcomes faster.

### 4.4 Uncertain Social Dilemma

**Environment description.** The emergence of cooperation in social dilemmas is another key research challenge for evolutionary biology, the social and sustainability sciences (Nowak, 2006, Kollock, 1998, Barfuss et al., 2020). We’ll focus on the situation where two agents can either cooperate (C) or defect (D) and either face a Prisoner’s Dilemma or a Stag Hunt game with equal probability (Fig. 6 A, cf., Levine and Ponsard, 1977, LiCalzi and Mühlenbernd, 2019). In the pure Prisoner’s Dilemma defection is the Nash equilibrium, which leads to a sub-optimal reward for both agents, also known

| Env. states | 6 | 7 | 8 | 9 |
|-------------|---|---|---|---|
| Reward $R^{*,avg}$ | 0.25 | 0.013 | 0.019 | 0.024 |
| Reward $R^{*}\gamma$ | -0.015 | 0.013 | 0 | 0.003 |

Table 1. Average reward of the optimal average-reward policy $R^{*,avg}$ and the optimal policy of the discounted reward setting $R^{*,\gamma}$ for the same four renewable resource environments as in Fig. 5. Rewards are also transformed in the same way ($R = r/r_{ac} - 1$, with $r_{ac}$ being the reward the fully observant agent obtained at convergence).
Figure 6. Evolutionary learning dynamics in an uncertain social dilemma. Panel A illustrates the environment. Panels B and C show the average rewards at convergence for agent 1 in red and agent 2 in blue (top row) and the time steps it takes the learners to convergence (bottom row) for various observational noise levels from 0 to 0.5. For each noise level, the plots show a histogram via the color scale. Each histogram results from a Monte Carlo simulation from 100 random initial policies. Panel B shows the case of homogeneous uncertainty where both agents’ observations are corrupted equally by noise. In Panel C only agent 2 is increasingly unable to observe the environment correctly (Heterogeneous Uncertainty). The discount factor was set to $\gamma = 0.5$ since future states are independent of the agents’ actions, which makes the discount factor irrelevant for the learning in this case. Remaining hyper-parameters were set to $\alpha = 0.01$ and $\beta' = 50$.

Homogeneous uncertainty can overcome the social dilemma through the emergence of a stable, mutually high rewarding fixed point above a critical level of observational noise. Heterogeneous uncertainty, however, leads to reward inequality. In both cases, the transition is accompanied by a critical slowing down of the convergence speed.

Results. Homogeneous uncertainty can overcome the social dilemma through the emergence of a stable, mutually high rewarding fixed point above a critical level of observational noise. Under perfect observation both agents converge to full defection when observing the Prisoner’s Dilemma. When observing the Stag Hunt game it depends on the initial joint policy whether the agents converge to mutual defection or mutual cooperation. Reward values are as such that the defective basin of attraction is comparable small (see the light line at an average reward of 0 in Fig. 6 B). Increasing the observational noise level from zero under homogeneous uncertainty will first decrease the average reward at convergence. The agents still converge to the perfect observation policy which leads them to defect when they observe the Prisoners’ Dilemma but the situation is actually the Stag Hunt. However, increasing observational noise further eventually leads to a bifurcation (Fig. 6 B). Mutual cooperation under both observations becomes a stable fixed point. As a consequence both agents obtain an average reward of 5 at convergence. Interestingly, there seems to be a small range of observational noise at which all three rewards 0, 2 and 5 are supported by equilibria. For large noise levels only the rewards at 0 and 5 are stable.

Thus, we find that the evolutionary learning dynamics under homogeneous partial observability are able to converge to mutually more rewarding policies compared to the perfect observation case. The existence of those equilibria is long known in traditional static game theory (Levine and Ponssard, 1977). Here we show the dynamic counterpart. These static equilibria not only correspond to fixed points of the derived learning dynamics, the transition to the more rewarding equilibrium is again accompanied by a critical slowing down of the convergence speed (Fig. 6 B, bottom).

However, the mutual benefit of uncertainty vanishes when not all agents’ observations are uncertain (Fig. 6 C). Under slight uncertainty only the reward of the ill-informed agent (Agent 2 in Fig. 6) decreases. After the bifurcation point under large uncertainty, the ill-informed agent converges to full cooperation under both observations, whereas the well-informed agent still defects in the Prisoner’s Dilemma which earns it an average reward of even more than 5. The knowledgable agent exploits the ill-informed and heterogeneous uncertainty leads to reward-inequality between the agents.

Interestingly, Fig. 6 suggests a difference in the type of phase transition between the policy of mediocre reward at low observational noise levels and the policies at high noise levels. The phase transition under homogeneous uncertainty seems to
be discontinuous whereas the transition under heterogeneous uncertainty seems to be continuous. Investigating the relationship between the learning dynamics, free energy equivalents (Barfuss, 2021) and phase transitions is a promising direction of future work.

### 4.5 Uncertain Zero-sum Competition

**Environment description.** The last environment we use as a test bed is a two-agent, two-state, two-action zero-sum competition, also known as the two-state matching pennies game (Hennes et al., 2010). It roughly models the situation of a test bed is a two-agent, two-state, two-action zero-sum competition. The keeper agent scores one point if it catches the ball (when both agents have chosen the same action), otherwise the kicker agent receives one point. The two states of the environment encode which agent is the keeper and which one is the kicker. In state *Keep* agent 1 is the keeper and agent 2 is the kicker. In the state *Kick* it is the other way around. Agents change roles under state transitions, which depend only on agent 1’s actions. When agent 1 selects either *left* as keeper or *right* as kicker both agents will change roles. With symmetrical rewards but asymmetrical state transitions, this two-state zero-sum game presents the challenge of coordinating both agents on playing a mixed strategy with equiprobable actions. Similarly as in Sects. 4.1 and 4.4, the agents’ observations of the environmental states are obscured by a noise level $\nu$.

**Results.** Fig. 7 shows how environmental uncertainty can stabilize the learning process. When both agents observe the environment perfectly the learning dynamics are prone to be unstable, either unpredictably chaotic or on periodic orbits and limit cycles (Panel A, Barfuss et al., 2019). The rewards of agent 1 and 2 are circulating around zero. Under a medium observational noise level of $\nu = 0.25$ the learning dynamics are still unstable. Especially the transient dynamics in the policy space (Panel B, on the right) appear strange. The average reward trajectory looks damped compared to the fully observant agents. Increasing the observational noise further such that the agents perceive the two environmental states (*Keep* and *Kick*) as a single observation (*KK*), is able to stabilize the learning process. Interestingly, the flow of the learning dynamics is separated into two half circles directed at the upper half of the line at which agent 1 chooses both actions with equal probabilities. As shown by the gray arrows, the circled flow is on a fast time scale compared to the movement downward to the center of the policy space (which is not reached here within 1000 time steps). At this downward movement, both agents play the different roles of kicker and keeper in equal amounts, since only agent 1 is responsible for the state transitions. Any advantage agent 2 gains from deviating from the equiprobable policy as kicker is balanced by the same amount of disadvantage agent 2 looses as keeper. Thus, the reward for both agents quickly stabilizes at zero.

![Figure 7](image_url)

**Figure 7. Evolutionary learning dynamics in an uncertain zero-sum competition.** Policy spaces and reward trajectories are shown for three different observational noise levels: (A) $\nu = 0.0$, i.e., perfect observation, (B) $\nu = 0.25$, and (C) $\nu > 0.5$, i.e., both states are observed inseparably as one. The probability of choosing action *left*, conditioned on the current observation, is plotted on the x-axis for agent 1 and on the y-axis for agent 2. Learning trajectories are shown from 5 initial policies around the center of the policy spaces. For better visual inspection only one of those trajectories is portrayed in color. Arrows in gray indicate the flow of the learning dynamical system. Hyper-parameter were $\alpha = 0.005$, $\beta = 200$, and $\gamma = 0.9$. Here, full environmental uncertainty is able to stabilize the learning process.
ments, from partially observable Markov decision processes to partially observable stochastic games. This is favorable, especially since other ways of formal analysis become increasingly limited, the more general the agent-environment systems become (Bernstein et al., 2002, Hansen et al., 2004, Xie et al., 2020). The main limitation of the evolutionary learning dynamics approach is its lack of scalability to large environments. However, its main purpose is to facilitate a principled and deeper understanding (cf., Zdeborová, 2020), for which the focus on conceptual and stylized environments seems most appropriate.

Providing agents with only a partial view of the true state of the world might be expected to result in poorer learning decision-making outcomes. However, we have demonstrated that irreducible environmental uncertainty can instead lead to better learning outcomes, even in a single-agent environment, stabilize the learning process and overcome social dilemmas in multi-agent domains. These results depend on agents continuing to apply the specific process of learning that temporal difference reinforcement learning implies; a sufficiently sophisticated agent could always reconstruct the partial state space perspective from a more detailed view, and thus must perform at least as well as the agent with partial observability. However, given that temporal-difference reinforcement learning is a relatively simple and widely effective algorithm, and one which closely matches known features of neurological learning (Schultz et al., 1997, Dayan and Niv, 2008), this points to a potential evolutionary pressure for agents to develop internal models of the world that do not match the true state space of their environment (cf. Mark et al., 2010, Hoffman et al., 2015).

Whether state uncertainty in the classic principle of temporal-difference learning is advantageous depends on the specific nature of the environment and its representation, (c.f., ecological rationality, Todd and Gigerenzer, 2012, Hertwig et al., 2019). The proposed dynamics are therefore a suitable tool to advance theoretical research in cognitive ecology, which studies how animals acquire, retain, and use information within their ecology, evolution and behavior (Shettleworth, 2009, Dukas and Ratcliffe, 2009). Within this area, research has begun to ask how agents’ may evolve non-veridical or incomplete representations of the world (Mark et al., 2010, Prakash et al., 2020, Mann, 2021); the dynamic model presented here offers a tool to study the effect of non-veridical representations in greater depth.

We also showed that partial observability can lead to better collective outcomes in the case of social dilemmas. The question for the preconditions of cooperation and sustainable behavior presents an important area for deeper investigation (Barfuss et al., 2018, Strnad et al., 2019, Bak-Coleman et al., 2021). Temporal-difference learning is a widespread principle in neuroscience and psychology (Dayan and Niv, 2008) and there is indeed evidence that humans use a payoff-based learning rule in social dilemmas (Burton-Chellew et al., 2015). The topic of uncertainty is of special relevance in the mitigation of the climate crisis through global cooperation agreements (Kolstad, 2007, Milinski et al., 2008, Kolstad and Ulph, 2011, Barnett and Dannenberg, 2012, Domingos et al., 2020). Our results highlight the potential for a systematic investigation of mechanisms that incorporate useful uncertainty (Nax et al., 2018). In our examples, the mutual benefit of uncertainty in the social dilemma vanishes when not all agents are likewise ill-informed causing reward-inequality between the agents. This suggests that partial observability as a mechanism for solving social dilemmas may need to be regulated externally (e.g. by authorities that monitor information flow, or as a feature of the environment) rather than something that is likely to be generated as an evolutionary adaptation amongst individuals in competition with each other.

We also found a range of interesting dynamical effects induced by irreducible environmental uncertainty. Where, as expected, partial observability caused worse learning outcomes, we found that this can happen in the form of a catastrophic limit cycle, within which the agent obtains the worst possible reward. We also found instances where environmental uncertainty induces phase transitions between low and high rewarding regimes accompanied with a critical slowing down of the learning processes. Further, we saw a state uncertainty induced separation of the learning dynamics into fast and slow eigendirections, as well as multi-stability of the learning process. Since all agent behaviour is ultimately dynamic, these phenomena highlight what the study of static game-equilibria must miss (Smith, 2005, Papadimitriou and Piliouras, 2019).

Moreover, these results may be of use in technological applications of multi-agent reinforcement learning, with respect to training regimes, hyper-parameter tuning, and the development of novel algorithms. For example, if agents are able to detect that they entered a slow eigendirection, they can safely increase their learning rate for a faster convergence. Or training regimes and hyper-parameter search techniques might be on the lookout for a critical slowing down since this can indicate a phase transition towards high rewarding solutions. With respect to the hyper-parameter values required for a decent performance we found across environments that learning with irreducible environmental uncertainty demands more exploration and less weight on future rewards, compared to fully observant agents. Moreover, the learning outcome under state uncertainty might depend crucially on the precise combination of the two parameters, whereas without uncertainty both parameters can be tuned fairly independently. The fast computation speed and visualization capabilities of the evolutionary learning dynamics approach might be particular suited for the challenge to engineer interpretable and safety-critical learning systems.

A promising directions for future work is the integration of model uncertainty through an analytical treatment of noisy dynamics (cf., Galla, 2009). The stochastic noise models the finiteness of a reasonable learning algorithm compared to the theoretical limit of the infinite memory batch of the present dynamics. The challenge is that this problem is ill-defined and many reasonable learning algorithms exist. Furthermore exciting is the embedding of representation and generalization dynamics into the nonlinear dynamics of learning, acting and environment to study the principles of advantageous representations.

Acknowledgements

This work was supported by UK Research and Innovation Future Leaders Fellowship MR/S032525/1 and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A.
References

Peter Ashcroft, Philipp M Altrock, and Tobias Galla. Fixation in finite populations evolving in fluctuating environments. *Journal of The Royal Society Interface*, 11(100):20140663, 2014.

Michael Assaf, Mauro Mobilia, and Elijah Roberts. Cooperation dilemma in finite populations under fluctuating environments. *Physical Review Letters*, 111(23):238101, 2013.

Adib Bagh and Yoko Kusunose. On the economic value of signals. *The BE Journal of Theoretical Economics*, 20(1), 2020.

Joseph B Bak-Coleman, Mark Alfano, Wolfram Barfuss, Miguel A Centeno, Iain D Couzin, Jonathan F Donges, Mirta Galesic, Andrew S Gersick, Jennifer Jacquet, et al. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118(27), 2021.

Wolfram Barfuss. Reinforcement learning dynamics in the infinite memory limit. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1768–1770, 2020a.

Wolfram Barfuss. Towards a unified treatment of the dynamics of collective learning. In *AAAI Spring Symposium: Challenges and Opportunities for Multi-Agent Reinforcement Learning*, 2020b.

Wolfram Barfuss. Dynamical systems as a level of cognitive analysis of multi-agent learning. *Neural Computing and Applications*, pages 1–19, 2021.

Wolfram Barfuss, Jonathan F Donges, Marc Wiedermann, and Wolfgang Lucht. Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution. *Earth System Dynamics*, 8(2):255–264, 2017.

Wolfram Barfuss, Jonathan F Donges, Steven J Lade, and Jürgen Kurths. When optimization for governing human-environment tipping elements is neither sustainable nor safe. *Nature communications*, 9(1):1–10, 2018.

Wolfram Barfuss, Jonathan F Donges, and Jürgen Kurths. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 99(4):043305, 2019.

Wolfram Barfuss, Jonathan F Donges, Vitor V Vasconcelos, Jürgen Kurths, and Simon A Levin. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proceedings of the National Academy of Sciences*, 117(23):12915–12922, 2020.

Scott Barrett and Astrid Dannenberg. Climate negotiations under scientific uncertainty. *Proceedings of the National Academy of Sciences*, 109(43):17372–17376, 2012.

Scott Barrett and Astrid Dannenberg. Tipping versus cooperating to supply a public good. *Journal of the European Economic Association*, 15(4):910–941, 2017.

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pages 265–272, 1953.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Doebeli. Evolutionary dynamics of multi-agent learning: a survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

Maxwell N. Burton-Chellew, Heinrich H. Nax, and Stuart A. Weber. Payoff-based learning explains the decline in cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences*, 282(1801):20142678, Feb 2015.

Xiaojie Chen and Attila Szolnoki. Punishment and inspection for governing the commons in a feedback-evolving game. *PLoS Computational Biology*, 14(7):e1006347, 2018.

John G. Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87(2):239, 1973.

Peter Dayan and Yael Niv. Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008.

Elias Fernández Domingos, Jelena Grujić, Juan C Burguillos, George Kirchsteiger, Francisco C Santos, and Tom Lenaerts. Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. *Isiscience*, 23(12):101752, 2020.

Reuven Dukas and John M Ratcliffe. *Cognitive ecology II*. University of Chicago Press, 2009.

Tobias Galla. Intrinsic noise in game dynamical learning. *Physical Review Letters*, 103(19):198702, 2009.

Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

Aram Galstyan. Continuous strategy replicator dynamics for multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems*, 26(1):37–53, 2013.

Fabian Geier, Wolfram Barfuss, Marc Wiedermann, Jürgen Kurths, and Jonathan F Donges. The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources. *The European Physical Journal Special Topics*, 228(11):2357–2369, 2019.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011.

Chaitanya S Gokhale and Christoph Hauert. Eco-evolutionary dynamics of social dilemmas. *Theoretical Population Biology*, 111:28–42, 2016.

Carlos Gracia-Lázaro, Luis M Floría, Jesús Gómez-Gardeñes, and Yamir Moreno. Cooperation in changing environments: Irreversibility in the transition to cooperation in complex networks. *Chaos, Solitons & Fractals*, 56:188–193, 2013.

Joseph Y Halpern. *Reasoning about uncertainty*. MIT Press, 2017.

Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *National Conference on Artificial Intelligence (AAAI)*, pages 709–715, 2004.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

Christoph Hauert, Miranda Holmes, and Michael Doebeli. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2565–2571, 2006.

Christoph Hauert, Camille Saade, and Alex McAvoy. Asymmetric evolutionary games with environmental feedback. *Journal of Theoretical Biology*, 462:347–360, 2019.

Daniel Hennes, Michael Kaisers, and Karl Tuyls. RESQ-learning in stochastic games. In *Adaptive and Learning Agents Workshop at AAMAS (ALA)*, 2010.

Ralph Hertwig, Timothy J Pleskac, and Thorsten Pachur. Taming uncertainty. *MIT Press*, 2019.

Christian Hilbe, Štěpán Šimsa, Krishnendu Chatterjee, and Martin A Nowak. Evolution of cooperation in stochastic games. *Nature*, 559(7713):246–249, 2018.

Josef Hofbauer and Karl Sigmund. *Evolutionary games and population dynamics*. Cambridge University Press, 1998.

Donald D Hoffman, Manish Singh, and Chetan Prakash. The interface theory of perception. *Psychonomic Bulletin & Review*, 22(6):1480–1506, 2015.

Tomas Klos, Gerrit Jan Van Ahee, and Karl Tuyls. Evolutionary dynamics of regret minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 82–96. Springer, 2010.

Mykel J Kochenderfer. Decision making under uncertainty: theory and application. *MIT Press*, 2015.

Peter Kollock. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24(1):183–214, 1998.
Charles D Kolstad. Systematic uncertainty in self-enforcing international environmental agreements. *Journal of Environmental Economics and Management*, 53(1):68–79, 2007.

Charles D Kolstad and Alistair Ulph. Uncertainty, learning and heterogeneity in international environmental agreements. *Environmental and Resource Economics*, 50(3):389–403, 2011.

Pierre Levine and Jean-Pierre Ponssard. The values of information in some nonzero sum games. *International Journal of Game Theory*, 6(4):221–229, 1977.

Marco LiCalzi and Roland Mühlenbernd. Categorization and cooperation across games. *Games*, 10(1):5, 2019.

Emilie Lindkvist and Jon Norberg. Modeling experiential learning: The challenges posed by threshold dynamics for sustainable renewable resource management. *Ecological Economics*, 104:107–118, 2014.

Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 14, page 30, 2001.

John Loch and Satinder P. Singh. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In *International Conference on Machine Learning (ICML)*, pages 323–331, 1998.

Richard P Mann. Optimal use of simplified social information in sequential decision-making. *bioRxiv*, 2021.

Vincent AWJ Marchau, Warren E Walker, Pieter JTM Bloemen, and Steven W Poppet. *Decision making under deep uncertainty: from theory to practice*. Springer Nature, 2019.

Justin T Mark, Brian B Marion, and Donald D Hoffman. Natural selection and veridical perceptions. *Journal of Theoretical Biology*, 266(4):504–515, 2010.

Manfred Milinski, Ralf D Sommerfeld, Hans-Jürgen Krambeck, Floyd A Reed, and Jochem Marotzke. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences*, 105(7):2291–2294, 2008.

Heinrich H Nax, Stefano Barietti, Ryan O Murphy, and Dirk Helbing. Adding noise to the institution: an experimental welfare investigation of the contribution-based grouping mechanism. *Social Choice and Welfare*, 50(2):213–245, 2018.

Martin A Nowak. Five rules for the evolution of cooperation. *Science*, 314 (5805):1560–1563, 2006.

Frans A Oliehoek and Christopher Amato. A concise introduction to decentralized POMDPs, volume 1. Springer, 2016.

Fabio Panizzo, Nicola Gatti, and Marcello Restelli. Evolutionary dynamics of Q-learning over the sequence form. In *Conference on Artificial Intelligence (AAAI)*, pages 2034–2040, 2014.

Christos Papadimitriou and Georgios Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*, 16(2):53–63, 2019.

Ronald Parr and Stuart Russell. Approximating optimal policies for partially observable stochastic domains. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1088–1094, 1995.

Roger Permann, Yue Ma, James McGilvray, and Michael Common. *Natural resource and environmental economics*. Pearson Education, 2003.

Chetan Prakash, Kyle D Stephens, Donald D Hoffman, Manish Singh, and Chris Fields. Fitness beats truth in the evolution of perception. *Acta Biotheoretica*, pages 1–23, 2020.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Yuzuru Sato and James P Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, January 2003.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

Sara J Shettleworth. *Cognition, evolution, and behavior*. Oxford University Press, 2009.

Satinder P. Singh, Tommi Jaakkola, and Michael I Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier, 1994.

Linda B Smith. Cognition as a dynamic system: Principles from embodiment. *Developmental Review*, 25(3-4):278–298, 2005.

Matthijs TJ Spaan. Partially observable Markov decision processes. In *Reinforcement Learning: State-of-the-Art*, pages 387–414. Springer, 2012.

Felix M Strnad, Wolfram Barfuss, Jonathan F Donges, and Jost Heitzig. Deep reinforcement learning in world-earth system models to discover sustainable management strategies. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):123122, 2019.

Steven H Strogatz. *Nonlinear dynamics and chaos*. CRC press, 2018.

Qi Su, Alex McAvoy, Long Wang, and Martin A Nowak. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences*, 116(51):25398–25404, 2019.

R. S. Sutton, E. Rafols, and A. Koop. Temporal abstraction in temporal-difference networks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1313–1320, 2006.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, 2018.

Attila Szolnoki and Xiaojoie Chen. Environmental feedback drives cooperation in spatial social dilemmas. *EPL (Europhysics Letters)*, 120(5):58001, 2018.

Alessandro Tavoni, Maja Schlüter, and Simon Levin. The survival of the conformist: social pressure and renewable resource management. *Journal of Theoretical Biology*, 299:152–161, 2012.

Andrew R Tilman, Joshua B Plotkin, and Erol Akçay. Evolutionary games with environmental feedbacks. *Nature Communications*, 11(1):1–11, 2020.

Peter M Todd and Gerd Ed Gigerenzer. *Ecological rationality: Intelligence in the world*. Oxford University Press, 2012.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for Q-learning in multi-agent systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 693–700, 2003.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8 (3-4):279–292, 1992.

Joshua S Weitz, Ceyhun Eksin, Keith Paaporn, Sam P Brown, and William C Ratchff. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences*, 113(47):E7518–E7525, 2016.

John K. Williams and Satinder P. Singh. Experimental results on learning stochastic memoryless policies for partially observable Markov decision processes. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1073–1079, 1998.

Yuxuan Xie, Jilles Dibangoye, and Olivier Buffet. Optimally solving two-agent decentralized POMDPs under one-sided information sharing. In *International Conference on Machine Learning (ICML)*, pages 10473–10482, 2020.

Lenka Zdeborová. Understanding deep learning is also a job for physicists. *Nature Physics*, 16(6):602–604, 2020.