Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning

Fernández Domingos, Elias; Grujić, Jelena; Burguillo, Juan C.; Santos, Francisco C.; Lenaerts, Tom

Published in:
Simulation Modelling Practice and Theory

DOI:
10.1016/j.simpat.2021.102299

Publication date:
2021

License:
CC BY-NC-ND

Document Version:
Accepted author manuscript

Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning

Elias Fernández Domingos[a,b,c], Jelena Grujić[a,b], Juan C. Burguillo[c], Francisco C. Santos[d,b], Tom Lenaerts[a,b]

[a] *Artificial Intelligence Lab, Computer Science Department, Vrije Universiteit Brussel, Brussels, 1050, Belgium*
[b] *Machine Learning Group, Département d'Informatique, Université Libre de Bruxelles, Brussels, 1050, Belgium*
[c] *atlanTTic Research Center, Universidade de Vigo, Vigo, 36310, Spain*
[d] *INESC-ID & Instituto Superior Tècnico, Universidade de Lisboa, 2744-016 Porto Salvo, Portugal*

Abstract

From climate action to public health measures, human collective endeavors are often shaped by different uncertainties. Here we introduce a novel population-based learning model wherein a group of individuals facing a collective risk dilemma acquire their strategies over time through reinforcement learning, while handling different sources of uncertainty. In such an N-person collective risk dilemma players make step-wise contributions to avoid a catastrophe that would result in a loss of wealth for all players. Success is attained if they collectively reach a certain contribution level over time, or, when the threshold is not reached, they were lucky enough to avoid the cataclysm. The dilemma lies in the trade-off between the proportion of personal contributions that players wish to give to collectively reach the goal and the remainder of the wealth they can keep at the end of the game. We show that the strategies learned with the model correspond to those experimentally observed, even when there is uncertainty about either the risk of failing when the goal is not reached, the magnitude of the threshold to attain and the time available to reach the target. We furthermore confirm that being unsure about the time-window favors more extreme reactions and polarization, diminishing the number of agents that contribute fairly. The population-based on-line learning framework we propose is general enough to be applicable in a wide range

of collective action problems and arbitrarily large sets of available policies.

Keywords:
Public goods game, Population dynamics, Individual learning, Collective risk, Uncertainty

1. Introduction

Collective hunting, antibiotic abuse, vaccination hesitancy, climate action or even coordinating the population in order to comply to the covid19 regulations are some of the several examples of collective endeavours that entail a collective risk [1–13]. The collective-risk dilemma (CRD) [1] conveniently abstracts and operationalizes these global problems into a game theoretical framework. The resulting dilemma is a variant of a public goods game, wherein players aim to avoid a loss, instead of obtaining a benefit. The loss may only be avoided if the joint contributions of all participants surpasses a collective target [14–16]. In this N-player game, each participant receives an equal endowment (E), which they may use to provide contributions to a public account (g) over a fixed number of rounds (R). The more they contribute, the less they will keep for themselves once the game finishes. However, if the joint contributions of all members of a group at the end of the game do not surpass a certain threshold (T), then they risk loosing the remainder of their endowment, hence the dilemma. The risk or probability of loosing the endowment (p), models the uncertainty about the consequences or impact of not reaching the collective goal. During the game, individuals are able to observe the actions of the other members of the group, but, they will only know the outcome of the game once it has ended. This creates a game where returns are both time-delayed and uncertain.

Several experimental studies with human participants, mostly within the context of climate change negotiations, have developed on top of the initial CRD setup [1] (see Section 3.1 for the full definition of this game) in order to study the effects of inequality [2, 3], choice of representatives [17], communication [2, 9], threshold uncertainty [9] and timing uncertainty [11]. In Fig. 1, we summarize the results of a number of behavioral experiments relevant for the current work. It is important to note that there are some differences in the baseline setups of some of the extensions, which make their comparison difficult. These deviations from Milinski et al.[1] may have introduced other factors than those actually being studied, potentially also explaining

the differences in observations (see also Appendix A). Nonetheless, there are general patterns to derive from the experimental data.

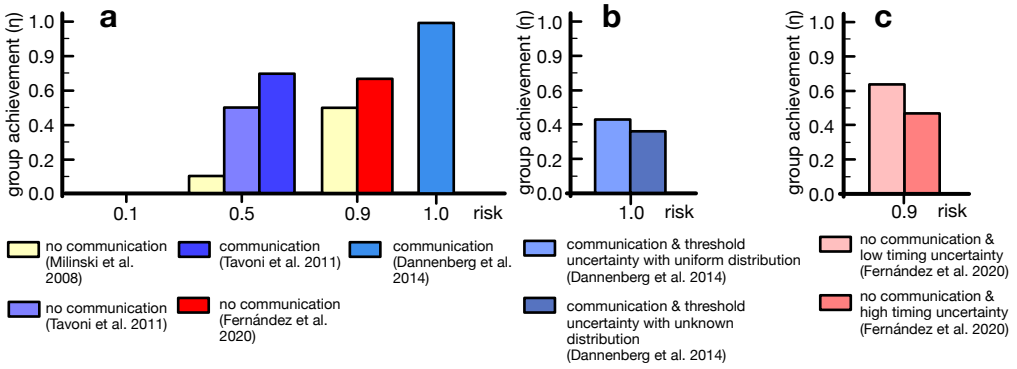

Figure 1: Summary of experimental results for the Collective Risk Dilemma. In the figure we plot the group achievement (η) for different levels of risk and uncertainty. Panel a) shows the summary of CRD experimental results for 4 different levels of impact uncertainty (risk), panel b) shows the experimental results for the CRD with only threshold uncertainty c) shows the results of CRD experiments with timing uncertainty and 90% of risk. These experimental results show that uncertainty is a very important factor in the CRD. High impact uncertainty (risk) increases group success, while coordination errors, due to uncertainty about the behavior of other players can reduce it. This latter can be overcome with communication, but other sources of uncertainty, such as threshold and timing uncertainty, can still considerably reduce group achievement.

First, as shown in Fig. 1a, the higher the risk (i.e., the higher the likelihood of losing the remainder of the endowment if the threshold is not reached) the higher the fraction of groups that achieve success (i.e., group achievement η in Fig. 1a). This effect of *impact uncertainty* appears to remain true both with and without communication, i.e., giving the participants the ability to make pledges or negotiate, and can be attributed to the change in the player's expected payoff. Interestingly, under high risk (90%), and when there is no possibility to communicate, only 50% to 65% of the experimental groups are successful [1, 11]. Errors in coordination appear to be the main cause for these results.

Second, as can also be observed in Fig. 1a, this coordination problem is alleviated (see 50% risk case) or even disappears fully (see 100% risk case) when participants can communicate, either about their own contributions [2] or what amount the group should achieve [9].

Third, as shown in Fig. 1b, even when communication is available, group

achievement may be significantly reduced when there is uncertainty about the amount (threshold) that needs to be collected at the end [9], a situation we will refer to as *threshold uncertainty*. Additionally, Barrett and Dannenberg showed analytically and experimentally that high threshold uncertainty, also transforms this game, in which players need to coordinate to avoid a catastrophe, into a classic prisoners dilemma, in which players prefer to defect [6, 18–21]. This type of uncertainty is present in many collective risk scenarios, such as climate negotiations, where the exact value of reduction in greenhouse emissions is not known [6, 9, 18, 20].

Finally, as visualized in Fig. 1c, group achievement also appears to be influenced by (high) *timing uncertainty* [11], i.e., when the number of rounds to achieve the target is uncertain. Yet the effect, is less strong than what was observed in the case of threshold uncertainty. However, Fernández Domingos et al. [11] find that participants in successful groups under timing uncertainty adopt reciprocal strategies, increasing (decreasing) their contributions if the rest of the group does the same [22], while the population becomes more polarized, with less participants adopting a fair behavior, i.e., contributing as well as gaining exactly $E/2$ (see Section 3.2).

There are still many questions that could be explored experimentally in the CRD to fully understand the impact of the different parameters that influence human decision-making (for instance, the number of actions from which participants can choose) or in order to explore novel mechanistic expansions (such as reward and punishment mechanisms [7, 23, 24]), and how they change the current experimental outcomes. Yet, as behavioral experiments are difficult and costly to organise, careful assessments are needed when trying to test some hypothesis. Analytical and computational models provide useful tools to understand the behavioral dynamics underlying each experimental outcome, together with preliminary assessment of novel hypotheses. As a minimal requirement, such theoretical models should be able to reproduce the main experimental observations for the baseline CRD as well as those that have been made for the different forms of uncertainty. Although a few theoretical works have been produced, especially in the context of evolutionary game theory (EGT) [4, 5, 7, 10, 23, 25–30] (see also Section 2), to our knowledge, there has been no model that captures simultaneously the baseline as well as all three forms of uncertainty and that is able to map both pure and mixed strategy profiles.

We present here a Population-based Learning (PBL) model with the capacity to achieve these goals, making it an useful model for further explo-

rations of, and predictions about, potential adaptations of the CRD (see Section 3.6 for details). This model allows agents, which are part of a larger population, to learn the most suitable behavior to "solve" a CRD in group, using a simple individual learning scheme inspired by Roth-Erev [31] and Q-learning [32] to update their behavior. The behaviors are represented as per-round probabilities for each action, enabling the exploration of a wide range of mixed strategies, which differentiates them from prior modelling work in the EGT community, where the focus is more often on pure strategies.It is worth noting that the training is not meant to reproduce the learning process that people go through when participating in a CRD experiment.

The training phase of this PBL model is followed by an evaluation phase that examines the performance of the learned behaviors outside of the context in which they were trained, i.e., they play the CRDs with participants trained in different populations. This independent evaluation process is necessary to ensure that the assessments about the quality and performance of the learned behaviors, is not biased by the presence of the other group members, and thus the mutual adaptations that are acquired in a group upon training. Just as in experiments, agents need to respond based on their experiences without having time to acquire the best way to coordinate their actions.

In the following we show that this PBL model not only reproduces the essential experimental observations, but also enables the generalization of results to different levels of risk, uncertainty, group size and number of available actions. Moreover, we explore how these parameters affect the behaviors learned by the population, and which strategies lead to success in function of the type of uncertainty the agents face.

The remainder of this article is structured as follows. In Section 2 more details are provide about the existing simulation models, which are always defined from an evolutionary perspective. Afterwards, in Section 3 the CRD, the types of uncertainty, the PBL model and the main parameter settings are introduced. The results section (i.e., Section 4) then reveals and discusses the main results generated by the PBL model. This includes a discussion on how the different forms of uncertainty affect the learned behavior, how these results match with the experimental observations made for the CRD, how uncertainty leads to the polarisation of contributions in the population, how it affects the space of learned behaviors and that increasing the granularity of choices may favour cooperation. Section 5 concludes this paper, followed by a number of additional details provided in the appendices of the paper.

2. Related work

The CRD has been previously studied theoretically using evolutionary game theory (EGT) [4, 5, 10, 25–29], wherein imitation defines how behavior changes over time [33]. In [4] the authors find that high risk and smaller group sizes are able to increase the chances of coordination. They also show that a heterogeneous political networks may be beneficial for cooperation. The dependence on risk has been also confirmed through large-scale agent-based simulations [10]. In [8, 27], it is shown analytically how threshold uncertainty in a one-shot CRD affects negatively the chances of coordination, while providing a theoretical basis for the effects of inequality in players' endowments. In [29], Hilbe et al. show that, when strategic timing is allowed, players tend to delay their contributions towards the end of the game. Using an agent-based evolutionary model based on evolutionary simulations for the CRD, it was shown that the population is only able to coordinate for high risks [5]. Later on, the same authors studied the effect of timing uncertainty in a variant of the CRD modified to allow for continuous contributions, and conclude that, under this uncertainty, the best strategy is immediate action [10]. However, since the experimental results on the CRD resort to discrete contributions, it is difficult to compare empirical observations with these works. We resort to a similar CRD mimicking the behavioral experiments available, and, therefore, focus on discrete and bounded contributions, and force players to perform contribution over multiple rounds to be able to achieve the target. Not only will this allow us to compare our results to the current experimental observations, it will also allow for the emulation of the effect of time-distant public goods, i.e., those for which the benefits are not available immediately.

The effect of uncertain game lengths has also been previously studied in behavioral economics [34] on other games. For instance, in the iterated Prisoner's Dilemma it is also known as the shadow over the future [35], and it was shown to be an environmental component that enables cooperation to emerge. Moreover, other sources of uncertainty, such as uncertain returns have been studied previously, revealing that this uncertainty provides an increase in cooperation [36–39]. Additionally, [12, 40] give a good overview of the effects of uncertainty on climate governance and social dilemmas.

In our work, we opt for studying the dynamics of the CRD using a form of reinforcement learning in the PBL model to update players' behaviors, as it allows for the exploration of mixed strategies, and is widely accepted as

a technique for learning behaviors observed within behavioral economic experiments. Reinforcement learning has been for instance applied to different variations of 2-player games [31, 34, 41–46] as well as bargaining games [47–49], coordination games [50, 51] in well-mixed and structured populations, stochastic games [52, 53] and other social dilemmas [54, 55]. It provides a flexible and powerful framework for studying the dynamics and effects of different variables in the CRD, allowing for a large behavioral (strategic) space and mixed strategies. Differently from classical RL setups, here we employ RL in a population of agents to update their behaviors in each step of the process.

3. Models and Methods

3.1. The Collective-risk Dilemma

The collective-risk dilemma (CRD) was designed to represent the problem of environmental governance as a public goods game, wherein a group of N players aim to avoid a future loss, instead of trying to obtain a benefit. During the game, players are required to make discrete contributions $a_i \in A$ in every round, where A is a discrete set of possible contributions, out of their own private endowment, E, to a public good g. The number of rounds is defined as R and C_i is the sum of all the contributions made by player i over all the rounds. If at the end of the game the sum of all contributions of the group is higher or equal than a threshold T ($g \geq T$), then all members will keep the remainder of the endowment as a payoff. Otherwise, all players receive a payoff of 0 with probability p, which defines the risk – or *impact uncertainty* – of the game.

The payoff of a player i in a given game is either:

$$\pi_i = \begin{cases} E - \sum_{s=1}^{R} a_i(s) = E - C_i & \text{if } g \geq T \text{ or } rand > p \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Where $rand$ is a uniformly distributed real random number distributed in the interval $[0, 1)$. With this, we can define the expected payoff of a player i by Eq. 2, where P stands for the probability that the group achieves the collective target (threshold) T.

$$\Pi_i = (E - C_i)[P + (1 - p)(1 - P)] \tag{2}$$

When T is deterministic, the expected payoff of player i can be reduced to a function of its contribution C_i and the sum of contributions of the other players in the group excluding i, i.e., C_{-i} (see Eq. 3).

$$\Pi_i(C_i, C_{-i}) = (E - C_i)\{H(C_i + C_{-i} - T) + (1-p)[1 - H(C_i + C_{-i} - T)]\} \quad (3)$$

Where $H(x)$ in Eq. 3 is the Heaviside function for a discrete variable x: It takes the value 1 if $x \geq 0$ and 0 otherwise. This function indicates whether a group is successful, i.e., the group jointly contributes $g \geq T$, where $g = C_i + C_{-i}$.

3.2. Fairness in the CRD

A notion of *fairness* can be defined in the CRD. A fair contribution F is the accumulated contribution of every player such that the group achieves the threshold with an equal investment by each player. It's value corresponds to dividing the required amount for success, i.e., T, equally among all members of a group of size N, therefore $F = T/N$. With this definition, we are able to analyse how the frequency of players that contribute $C_i = 0$, $C_i < F$, $C_i = F$ and $C_i > F$, varies with uncertainty and game difficulty.

3.3. Cost of the dilemma and minimum number of contributors

We define the required effort, or cost, of the dilemma, assuming equal contributions, as $\sigma = F/E$, i.e., the ratio between the fair contribution F, and the endowment E. Therefore, $\sigma = 0$ if the game requires no contributions, and $\sigma = 1$ if players must contribute all their endowment in order to reach the target. This difficulty level is directly related to the collective target T that players must achieve, so that $\sigma = T/(NE)$. Moreover, there is a direct relationship between the σ and the minimum required number of contributors M in a game. For instance, if $N = 6$ and $\sigma = 0.5$, then $M = N/2 = 3$ players must contribute E, while if $\sigma = 0.1$, only $M = 1$ player needs to contribute $\sigma NE/M = 0.6E$. Therefore, in a game at least $M = \lceil \sigma N \rceil$ players in a group need to contribute a total of $g = \sigma NE$. Thus, altering σ, changes the number of required cooperators in a game, in a similar fashion as in [4].

3.4. Threshold uncertainty

Next to impact uncertainty, we examine in this paper the effect of uncertainty about the collective target T, defining it as in [27]: T is drawn

from a discrete uniform distribution of range $[\overline{T}(1-\delta), \overline{T}(1+\delta)]$, where $\overline{T}$ is the target in the absence of threshold uncertainty. We use δ to quantify threshold uncertainty, with δ taking any value from discrete interval $[0, 1]$ with increments of 0.1. Therefore, when $\delta = 0$, there is no uncertainty, and when $\delta = 1$, T may take any natural value in $[0, 2T]$.

Threshold uncertainty, changes the CRD into a game with incomplete information as it affects the ability of players to correctly assess the threshold of the game. Furthermore, it affects the probability P that the collective target is achieved for a particular group whose total contributions sum to g. Concretely,

$$Pr(g \geq T) \approx \sum_{T} \frac{1}{2\delta\overline{T}} H(g - T) \tag{4}$$

which will be 0 if $g < \overline{T}(1-\delta)$, 1 if $g \geq \overline{T}(1+\delta)$. When g takes values inside the threshold interval, players cannot know for certain whether the collective target will be achieved. The probability of achieving the collective target in this case is given by:

$$P \approx \frac{1}{2\delta\overline{T}}[g + \overline{T}(\delta - 1)] \tag{5}$$

This change in P affects the expected payoffs of players, and, as a consequence, their strategies.

3.5. Timing uncertainty

To model the effect of timing uncertainty, we represent the number of rounds of the game as a stochastic variable drawn from a geometric distribution with success probability $1-w$. As a result, $w \in [0, 1)$ thus quantifies the level of timing uncertainty of the CRD: the higher w, the higher the variance in the distribution of game lengths. Moreover, to guarantee that the collective goal of the game is always achievable and guarantee a fair comparison among different levels of uncertainty, we ensure that the game will have a minimum number of rounds R_0, and that the average game length, $\overline{R}$ is always the same. Additionally, to limit the size of the state-space, we introduce an additional parameter R_{max}, that defines the maximum number of rounds of the game. This constraint may introduce deviations in the mean of the distribution, so we have R_{max} set to a high enough value (50 rounds), so that this effect is minimized. Therefore, in any given game, the total number of rounds is drawn from:

$$R \sim min(G(1-w) + R_0, R_{max}) \tag{6}$$

For convenience, in all our results we consider $\overline{R} = 10$ and we will only mention the variations over w. R_0 can be calculated as $R_0 = (\overline{R} - 1/(1 - w)) + 1$.

3.6. Definition of the population-based learning model

We study the dynamics of the CRD in a population of $Z \geq N$ autonomous agents, that adapt their behavior over K learning steps following an individual learning model explained in the next section. At each learning step k, players are part of multiple groups (n_g in total) that will play a CRD. These groups are constructed by randomly sampling among the Z population members. Moreover, we always choose $n_g \gg Z$, therefore each agent in a given learning step will play on average $(n_g N)/Z$ games. In each game, the strategies followed by the N agents to play the game determine the outcome of the group. Therefore, we define the group achievement $\eta = n/n_g$ of the population, where n is the number of successful groups, as the average success of the CRD given the population composition. Finally, Eq. 7 gives the expected payoff of a player i in the population at a given learning step k, and can be calculated in function of η. Replacing P by η in Eq. 2 one gets:

$$\Pi_{i,k}(\eta) = (E - C_i)[\eta + (1 - p)(1 - \eta)] \tag{7}$$

3.7. Individual learning model

Each agent in the population updates its strategy synchronously at the end of a learning step. The strategy of each agent, $\boldsymbol{X_i}(s) : S \to A$ is a function that maps each state to a probability distribution over the actions. Concretely, the strategy or behavior of the agent is represented by a matrix of dimensions $|A| \times |S|$ containing the probabilities $x_i(a, s)$ of taking an action $a \in A$ at a state $s \in S$ (see Fig. 2), and where $\sum_a x_i(a, s) = 1$. This representation is able to capture both mixed strategies and up to $|A|^{|S|}$ pure strategies, making it more general than the one used in [10], as it does not make any assumptions abut thresholds used by participants in each round. As the round wherein an agent finds herself corresponds to a state, the set of states is given by $S = \{0, ..., R - 1\}$. Also, for convenience, the available actions are defined through the size of the set $|A|$, i.e., $A = \{0, 1, ..., |A| - 1\}$.

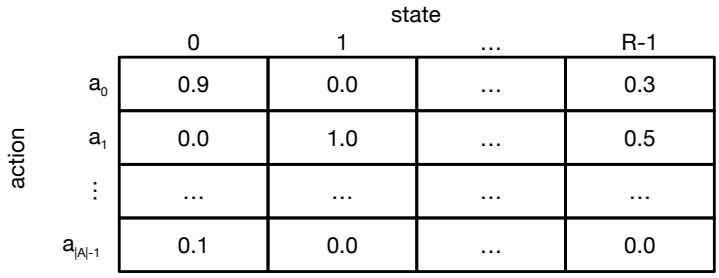

Figure 2: Strategy profile of an agent. This figure shows the structure of the strategy profile of a an agent i, i.e., $\boldsymbol{X_i}$. Each element in this matrix gives the probability of taking a certain action $a \in A$ at a given state $s \in S : s \in \{0, R-1\}$.

Agents use a variant of the well-known Q-learning algorithm [32] to update their strategy based on the outcomes of the games. The main change consists of performing batch updates over the Q-values of all actions after a game has finished, instead of updating them after every interaction. This is necessary since in the CRD players only obtain their payoff at the end of the game. Also, we do not consider any discount factors in order not to bias the importance of earlier or later actions in the game. There are off course more complex variations of Q-learning that could be used, yet our goal was to focus on a minimal one, sufficient to capture the observed behaviors. The algorithm defines a function $Q_{i,k}(a, s) : A \times S \rightarrow \mathbb{R}$ that gives the expected value/gain for an agent i to take action $a \in A$ in state $s \in S : \{0, ..., R-1\}$ in learning step $k \in K$. Each agent estimates this function by interacting with the environment throughout a number of steps, and performing the following update:

$$Q_{i,k+1}(a, s) = \begin{cases} Q_{i,k}(a, s) + \alpha(\Pi_i - Q_{i,k}(a, s)) & \text{if } a_i(s) = a \\ Q_{i,k}(a, s) & \text{otherwise} \end{cases} \tag{8}$$

With s the current state of the game, α the learning rate, $a_i(s)$ is the action taken in state s, and k the learning step wherein this occurs. It is important to notice that in our environment, agents only obtain their payoff at the end of the CRD. For this reason, during each game we store the actions taken by the agent at each learning step (i.e, the trajectory) in a history table (see also Roth-Erev model [31]). These action-value pairs are then used to update the Q-values according to Eq. 8. Finally, the strategy profile of each agent i, $X_{i,k}$ (see Fig. 2), is calculated in function of the Q-values at learning step k, according to the Gibbs-Boltzmann probability distribution in Eq. 9.

This process is repeated for K learning steps. The algorithm detailing the PBL model can be found in Algorithm 1.

$$x_{i,k}(a,s) = \frac{e^{Q_{i,k}(a,s)/\tau}}{\sum_l e^{Q_{i,l}(a,s)/\tau}} \tag{9}$$

Algorithm 1 Population-based learning.

for $l \leftarrow 1 : L$ **do** ▷ Repeat for each population
  create population of size Z.
  $Q_0 \leftarrow$ random initialization ▷ Initialize all agents
  for $k \leftarrow 1 : K$ **do** ▷ K learning steps
    $n \leftarrow 0$ ▷ initialize count
    for $l \leftarrow 1 : n_g$ **do** ▷ n_g games
      $group \leftarrow sample(population, N)$ ▷ sample randomly N players
      $g, \boldsymbol{\pi}, \boldsymbol{c} \leftarrow playGame(group, R)$ ▷ Eq. 1
      $n \leftarrow n + 1$ if $g \geq T$
      Save trajectory and payoff of each player in history table.
    end for
    save $\eta \leftarrow n/n_g$
    calculate $\Pi_{i,k}(\eta)$ ▷ Eq. 7
    update $\boldsymbol{Q_{i,k}}$ ▷ Eq. 8
    update $\boldsymbol{X_{i,k}}$ ▷ Eq. 9
  end for
  save population vector.
end for

3.8. Assessing learned behavior

After L populations have adapted independently in the learning phase (see Algorithm 1), agents go through an evaluation phase. In this phase, K_e evaluation games within different groups are played. To form each group, agents are randomly sampled from the set of all populations, i.e., groups are heterogeneous, with agents that might come from different populations (see Algorithm 2). At this point, agents no longer adapt, and only act based on the strategy learned while interacting within their own population. This step is important as essentially, even when the agents learned to play the

CRD successfully in their group, their acquired behavior may fail when being confronted with players that adapted in other groups. To assess thus correctly the learned strategy profile it needs to be evaluated against acquired behaviors outside of the group wherein the training occurred, just as other predictive machine learning methods need to be evaluated with an independent set not used for training the predictor.

Algorithm 2 Evaluation phase.

```
populations ← flatten(population1, ..., populationL)   ▷ gather agents of all populations in a flattened vector
for k_e ← 1 : K_e do                                    ▷ evaluation games
    group ← sample(populations, N)                      ▷ All individuals from any population have the same probability to be sampled
    playGame(group, R)
    save group success
end for
```

3.9. Clustering

To identify clusters in the behaviors learned under different uncertainty conditions, each player's strategy is encoded as the following vector:

$$\boldsymbol{b} = (\overline{a}_i(0), \overline{a}_i(1), ..., \overline{a}_i((N-1)a_{max})) \tag{10}$$

where $\overline{a}_i(\hat{a}_{-i})$ represents the average contribution made by a player i, averaged over all games and states s wherein the agent i was confronted with a particular cumulative contribution $\hat{a}_{-i}$ made by all the other players in that group in a previous state $s-1$. Thus $\hat{a}_{-i}$ can take any value in $\{0, 1, ..., (N-1)a_{max}\}$, with a_{max} representing the action with maximum value that a player can take. In other words, each agent's behavior is represented by it's contribution in function of the contributions of the rest of the group in the previous round. All vectors are stacked as rows of a matrix B.

The dimensionality reduction technique t-SNE [56, 57] is used to visualize the information in B. A t-SNE plot provides a 2-dimensional representation of the (multi-dimensional) behaviour in that matrix to reveal differences in

behavior resulting from the type of uncertainty to which individuals were exposed. Simultaneously, a k-means clustering [58, 59] is applied on the matrix B to identify subgroups. While looking for the best number of subgroups k, homogeneity within and separation between the clusters is carefully considered. The resulting optimal clustering is visualised using different colors in the t-SNE plot. Moreover, for each cluster we plot also $\overline{a}_i$ in function of $\hat{a}_{-i}$ as well as the average contributions in the first and second half of the game, always considering a maximum of 10 rounds so that all results are comparable with the timing uncertainty case (see Fig. 8). This cluster analysis reveals how behaviors changes under different uncertainty conditions.

3.10. Simulation parameters

In all simulations, we used a population of $Z = 50$ agents, however we did test other population sizes ($Z = 6, 12, 24, 48, 100, 200$), and verified that the results remained valid. Also, we have always considered $N = 6$ except for Figure 4a where group size N varies between 2 and 12. Moreover, in all results, excluding Fig. 9, $A = \{0, 1, 2\}$, so that we have the same number of actions as used in the Milinski et al. experiment [11]). Each populations adapted over $K = 10000$ learning steps. At each learning step, $n_g = 1000$ random groups played the CRD game. The difficulty of the game is set to $\sigma = 0.5$, which means that agents, on average, should contribute half of their endowment E to achieve the collective goal, unless specified otherwise. For instance, in Fig. C.10, displayed in the Appendix, the effect of varying σ is studied. In each game, unless otherwise indicated, the endowment is proportional to the length of the game $E = 2\overline{R}$, and the threshold is proportional to the group size $T = \sigma EN$. Also, $\overline{R} = 10$, which is also the exact number of rounds that the game takes in the absence of timing uncertainty. Finally, in all cases, the evaluation phase is performed with individuals from 10 independent populations, thus $L = 10$.

4. Results and Discussion

4.1. All uncertainties need to be mitigated to ensure coordination success.

We investigate the effect of three forms of uncertainty with our PBL model: impact uncertainty, threshold uncertainty and timing uncertainty. Impact uncertainty, is represented as the risk (probability) p that all players will lose the remainder of the endowment if the group does not achieve the collective target (see Section 3.1). In Fig. 3, the PBL model confirms the

baseline CRD behavioral experiments (see [1]): no group achieves the target when impact uncertainty is low, making defection (i.e., contributing zero) the dominant behavior over all rounds. This occurs for the current settings (see Section 3) when $p < 0.5$. On the contrary, when the risk is high, players are better off contributing (cooperating) and coordinating their contributions to avoid the collective loss, and the PBL model generates levels of group achievement close to 1.

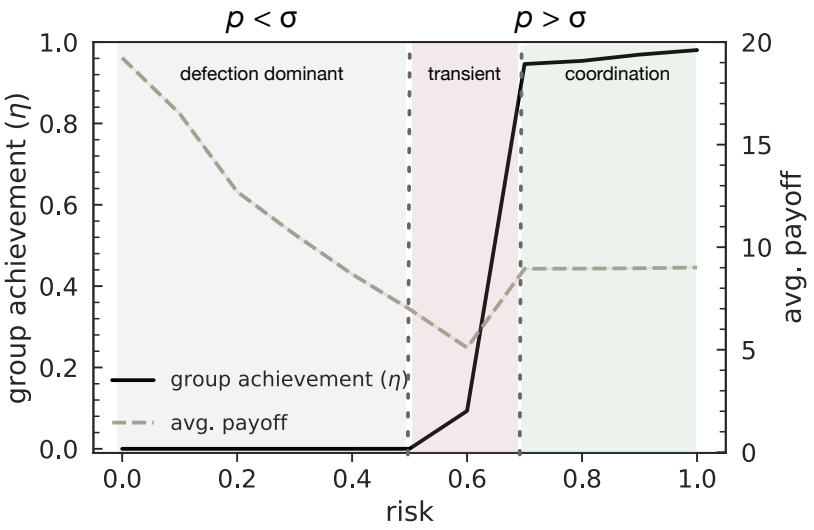

Figure 3: Effect of risk on group achievement and average normalized payoff. This figure shows the fraction of successful groups, or group achievement (η) in function of the level of risk p. For levels of risk $p < \sigma$, players are in a defection dominant zone, and cooperation does not emerge, since players may obtain a higher expected payoff by making less contributions than necessary to reach the target. In contrast when $p \gg \sigma$, the population is in a coordination zone, and agents are able to coordinate to achieve the target, thus, almost all groups are successful. Finally, for intermediate levels of risk, the population enters a transient region, where the minimum average payoff is achieved. This region defines the transition from a *failed* outcome to a *successful* one. Confidence intervals are < 0.1 for both group success and < 0.2 for avg. payoff. ($L = 10$, $Z = 50$, $R = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_t = 1000$, $K = 10000$, $K_e = 10000$)

There are multiple combinations of actions of the group members that may lead to a *successful* outcome, i.e., the situation in which the collective target is met. Yet, these action combinations are not stable, meaning that if a single player deviates from her actions, a change in the rest of the players' behaviors is generated that will either lead to a *failed* outcome, i.e., where

players do not meet the target, or a different *successful* outcome with similar contributions. Therefore, when the risk is sufficiently high, i.e., $p > 0.7$ in Fig. 3, the CRD is expected to become a coordination game wherein participants need to arrive somehow at one of the unstable coordination equilibria [4]. Note that the situation in which all players refrain from any contribution (full defection) is an equilibrium for all values of p, since in that case no agent may improve his or her situation by unilaterally giving more.

As shown in Fig. 3, these two regimes — defection dominance, and coordination — emerge naturally from the PBL model. In between, a transition region connects the defection dominant zone and the coordination zone. This region marks a tipping point for coordination, i.e., when the number of cooperative players surpasses a threshold that leads the population to coordinate into contributions that lead to a successful outcome. The location of this tipping point in function of impact uncertainty is important, since our model shows that the average payoff of the population reaches its minimum before it is crossed.

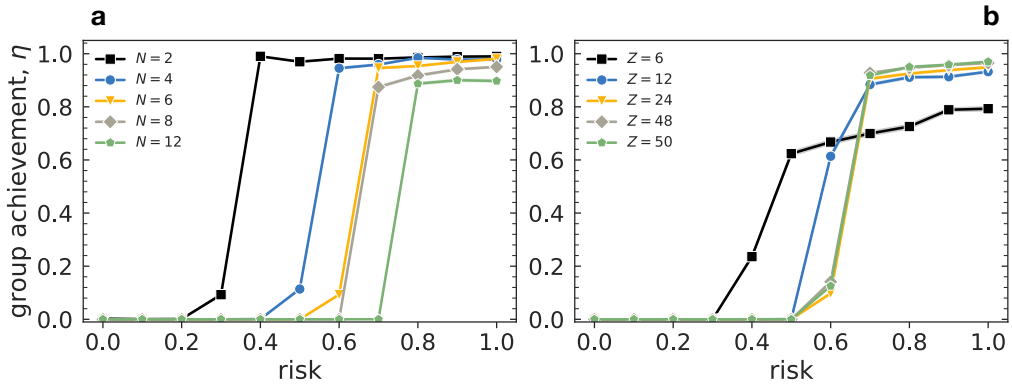

Figure 4: Effect of group and population size on group achievement. Results in panel a) suggest that group size is an important factor for coordination. Bigger groups require higher levels of risk to successfully coordinate and achieve the collective target. These results were obtained evaluating $L = 10$ populations of size $Z = 50$. Panel b) shows that group achievement is only significantly reduced when $Z = 6$, indicating that players in small populations adopt strong biases towards behaviors that are positive in their environment, yet do not adapt well when interacting with individuals of other populations. These results were obtained by evaluation $L = 100$ independent populations with $N = 6$. Confidence intervals are < 0.1. ($R = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$)

One can formally identify this tipping point between the defection and

coordination zone by considering the following simplification : First, let us define the investment cost σ required to achieve the threshold T in terms of the ratio between the fair contribution F and the total endowment E that each agent receives, i.e., $\sigma = F/E$ (see Section 3.2). Given this definition, the expected payoff of players that do not contribute (i.e., they defect - D) while also being in a group that does not reach the target is $\pi_D = (1-p)E$. In contrast, the expected payoff of a player that contributes overall F (i.e., they cooperate - C) and participates in a group that achieved the target is $\pi_C = (1-\sigma)E$. In order for cooperative behavior to become preferred over defective behavior the following condition needs to be met: $\pi_C > \pi_D \implies \sigma < p$. Fig. 3 shows these for $\sigma = 0.5$.

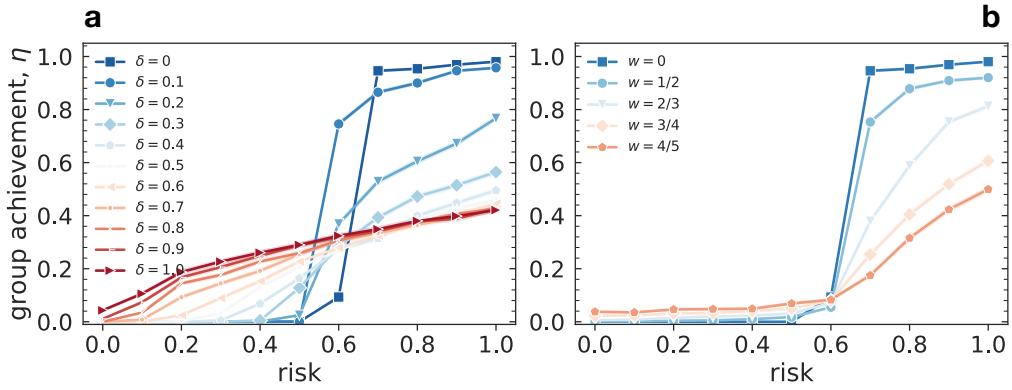

Figure 5: Effect of threshold and timing uncertainty on group achievement. Panel a) shows the variation of the fraction of successful groups, or group achievement (η) in function of risk and threshold uncertainty. Higher δ values indicate a larger range of possible values that the threshold may adopt (see Section 3.4) and, therefore, a higher level of uncertainty. Results show that, even for a very high risk, η decreases as δ increases. We also observe an increase in group achievement for lower risk levels. However, this increase in success is due to a higher chance that participants may avoid the collective loss with a lower threshold. Panel b) shows the level of group achievement in function of risk and timing uncertainty. In this case, w indicates the level of uncertainty over the final round, and the game has exactly $R = 10$ rounds when $w = 0$. Timing uncertainty is defined as a geometric distribution $R \sim G(1-w)$ (see Section 3.5), and therefore the variance of the distribution increases with w. Moreover, in all cases the mean of the distribution is at $\overline{R} = 10$, and we impose a minimum number of rounds, so that the target is always achievable (see Section 3.5). The results show that the η in the coordination zone diminishes with timing uncertainty. Confidence intervals are < 0.1. ($L = 10$, $Z = 50$, $N = 6$, $\overline{R} = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$)

The size and placement of the three regions observed in Fig. 3 is influenced by the size of the group in which the agents learn as well as the number of different participants with whom they are confronted. As shown in Fig. 4a on the one hand, smaller group sizes lead to an earlier transition between defection dominant and coordination zones. This outcome shows that it is easier to attain a successful outcome for lower impact uncertainty when playing the CRD in small groups. This result is in line with previous findings from the behavioral experiment performed in [60], and also obtained with evolutionary games [4, 7], showing that multiple agreements at a local level will increase the odds of collective success, in contrast to large-scale climate summits, where all parties participate simultaneously. On the other hand, as is visualized in Fig. 4b, if learning occurs in restricted groups where one often encounters the same participants ($Z = N$) then learned behaviors are specialised to coordinate with their group members, reducing significantly the success of coordinating, even for high risk, when encountering other players independently trained in one of the other L populations. This result, therefore, suggests that smaller populations over-fit their environment, developing strong biases towards certain behaviors, and adapt poorly when encountering members from a different population, producing coordination errors.

As said, the risk of collective failure provides just one possible source of uncertainty. When threshold uncertainty (i.e., when collective goals are ill defined) and timing uncertainty (i.e., when the time-frame to coordinate efforts is uncertain) are included, the PBL model reveals that both uncertainty types significantly influence group success (see Section 3 for a discussion how both uncertainty types were incorporated into the PBL model). Fig. 5a shows that increasing threshold uncertainty reduces success, a result found also in Danneberg's behavioral experiments [9] and two evolutionary game theoretical models [5, 27]. Additionally, our PBL model shows an increase of η for risks lower than 0.5, when δ (the level of threshold uncertainty) is high. This increase in success is due to a higher chance for participants to avoid the collective loss with a lower threshold T (see Section 3). For timing uncertainty, Fig. 5b shows that as in the experiments η is reduced in the coordination zone. The experimental results obtained for low (high) timing uncertainty by [11], and shown in Fig. 1, correspond to the setting $w = 2/3$ ($w = 4/5$)[1]. These results for threshold and timing uncertainty reveal that in

[1]As indicated in Section 3.5, the parameter w represents the amount of timing uncer-

order to achieve collective success, especially in the coordination zone where cooperation is preferred, all forms of uncertainty need to be mitigated so that coordination is more likely to be achieved.

4.2. Uncertainty increases polarization and reduces coordination

As there are differences in group success associated with uncertainty, the PBL model may also help us to understand subtle behavioral differences observed in lab experiments. A first coarse-grained approach to unravel the behaviors is to consider the fairness of each strategy and study how the proportion of fair individuals varies in the population with risk and uncertainty (see Section 3.2 for the definition of fairness). The individual behaviors learned in the PBL model either contribute exactly ($C = F$), more ($C > F$), less ($C < F$) in total. Since there is no incentive to contribute in the defection dominant zone, we also consider $C = 0$.

In Fig. 6 we visualize the fraction of each of these behaviors as a function of risk and uncertainty. In Fig 6a, which corresponds to the baseline CRD with impact uncertainty, the population is in a defection dominant zone when $p < 0.5$, with most agents contributing less than fair, i.e., $C < F$. In the transient zone, where $0.5 < p < 0.7$, the proportion of $C < F$ decreases quickly with risk, while the more generous behavior ($C > F$) increases, until their proportions intersect and $C > F$ becomes dominant. Once it reaches stability, the coordination zone starts. The proportional difference between $C > F$ and $C < F$ defines the likelihood of the population succeeding in reaching the threshold T. In Fig, 6a where only impact certainty is in effect, fair players ($C = F$) surpass also the fraction of $C < F$. However, when either timing (Fig. 6b) or threshold uncertainty (Fig. 6c) is present next to impact uncertainty, the intersection point between the fractions of $C < F$ and $C > F$ moves to higher levels of risk, enlarging the transient zone. Moreover, the fraction of $C = F$ players no longer surpasses that of the less generous ones ($C < F$). As a consequence, a behavioral polarization emerges in the population, favouring generous ($C > F$) and non-generous ($C < F$) behaviors over fair ones ($C = F$), which also leads, in this case, to a reduction in group success.

tainty, thus, higher values indicate that there is less certainty about which is going to be the final round of the game.

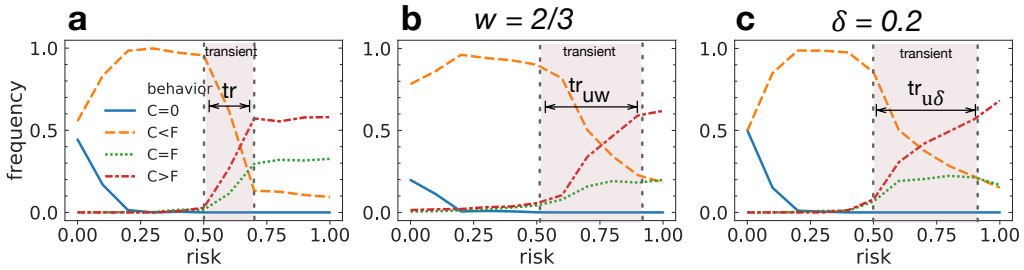

Figure 6: Effect of risk and uncertainty on the population's behavior distribution. This figure shows the variation of the fraction of agents in the population which make a total contribution of $C = 0$, $C < F$, $C = F$ or $C > F$, with $F = T/N$. If all players in a group contribute F, the sum of their contributions is exactly the target T, therefore we consider agents that adopt $C = F$ to be fair. Agents that contribute $C > F$ can be considered altruists, while agents that contribute $C = 0$ or $C < F$ are free-riders. Panel a) shows how these behavioral profiles change in the population in the presence of different levels of risk. When $p < 0.5$, the population is in a defection dominant zone, and most agents contribute $c < F$. For $0.5 < p < 0.7$, the population enters a transient region, and we observe how the proportion of $c < F$ decreases linearly with risk, while $C > F$ increases, until their proportions intersect. Afterwards, $C > F$ becomes dominant, and reaches stability, indicating the beginning of the coordination zone. The difference in proportion of $C > F$ and $C < F$ defines the likelihood of the population succeeding in avoiding a disaster, since these altruist agents compensate for the free-riders. In this coordination region, we also observe an increase of fair players, surpassing the fraction of $C < F$. However, when either timing (panel b)) or threshold uncertainty (panel c)) is also present, the intersection point between free-riders and altruists moves to higher levels of risk, which enlarges the transient zone. Moreover, the fraction of fair players no longer surpasses that of free-riders. This results in a polarized population mostly composed of $C > F$ and $C < F$ players. This behavioral change, may explain the reduced levels of group achievements in the presence of uncertainty. ($L = 10$, $Z = 50$, $N = 6$, $\overline{R} = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$)

4.3. Behaviors dealing with timing uncertainty differ from those learned for impact and threshold uncertainty

To better grasp the differences in learned behaviors and how they are associated with each uncertainty type, we investigate more closely the agent strategies that are attaining a successful outcome. As explained in more detail in the Section 3.9, the decision-process of each agent is encoded as a vector which contains the average contributions of an agent for each possible cumulative contribution made by the other agents in the same group in the previous round. Fig. 7 shows 2-dimensional representation of this n-dimensional information using a t-SNE plot (see also Section 3.9). Each point

in this figure represents such a vector and the square, cross or circle markers highlight the type of uncertainty under which the agent was trained — uncertainty (the risk, r, in this case) is present in all cases; squares (crosses) agents additionally experience threshold (timing) uncertainty. Using k-means clustering [58, 59] (see Section 3.9) six homogeneous and optimally separating groups were identified, as can be seen in Fig. 8. The results reveal that the learned behaviors under timing uncertainty are highly different from those acquired under impact and threshold uncertainty (see also panel b in Figure 7): they are mostly situated left to the zero on the x-axes. Fig. 8 further clarifies these differences: While panels a-d represent mostly unconditional behavior with differences in the mean contribution per round or the starting and ending contributions, panels e and f display a clear conditional, i.e reciprocal, behavior. The agents in those clusters decrease (increase) their contributions when the rest of the group gave more (less) than a fair, round-level donation. Moreover, in these clusters, agents tend to contribute more in the first part of the game then in the rest. Interestingly, these results were also observed experimentally by Fernández Domingos et al. [11], confirming again the differences in displayed behavior when timing uncertainty is present in the CRD.

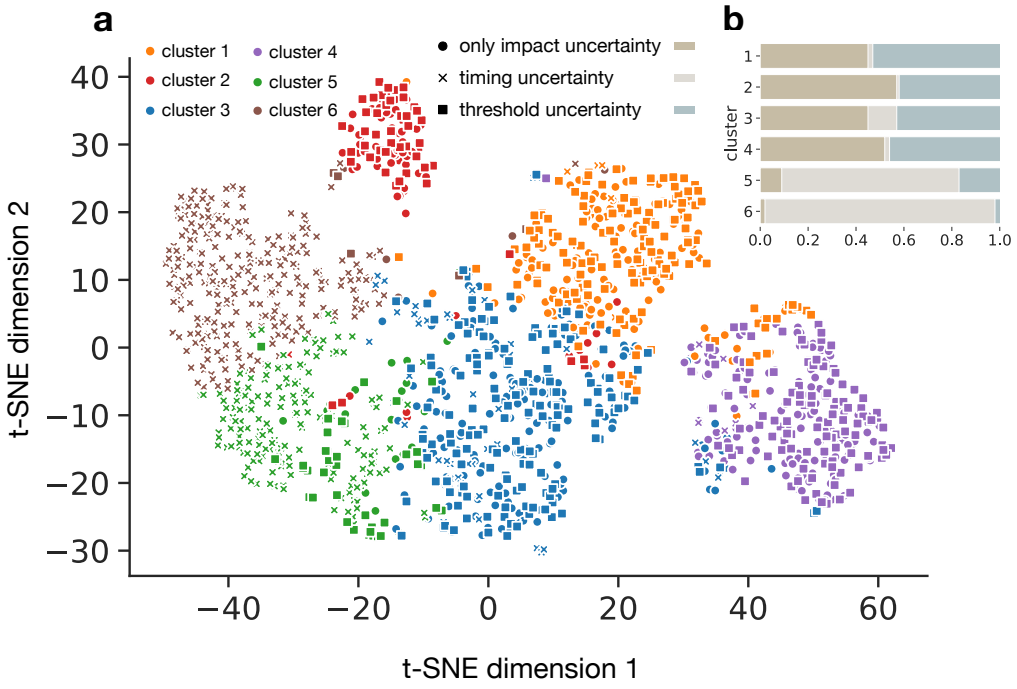

Figure 7: Behavioral distribution of a population of agents that adapted their choices under different types of uncertainty. This t-SNE plot (panel a) shows a 2-dimensional representation of the populations of agents that have adapted in different uncertainty environments. In all cases the risk of the CRD was $p = 0.9$. Players' behavior is encoded into vectors that represent the average contribution an agent makes relative to the cumulative contributions made by the other players in the previous round (see Section 3.9). Here we only look at results from successful groups, to understand what behaviour is adopted in these cases. Each point represents an agent, and the markers indicate the type of uncertainty under which the agent has trained. The groups are identified with k-means clustering (see Section 3.9). Each point in the plot is colored according to its cluster. Panel b shows the frequency of agent types in each cluster: whereas cluster 5 and 6 are dominated by agents trained for impact and timing uncertainty, the others are trained for impact and/or threshold uncertainty. ($L = 10$, $Z = 50$, $N = 6$, $\overline{R} = 10$, $\sigma = 0.5$, $p = 0.9$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$)

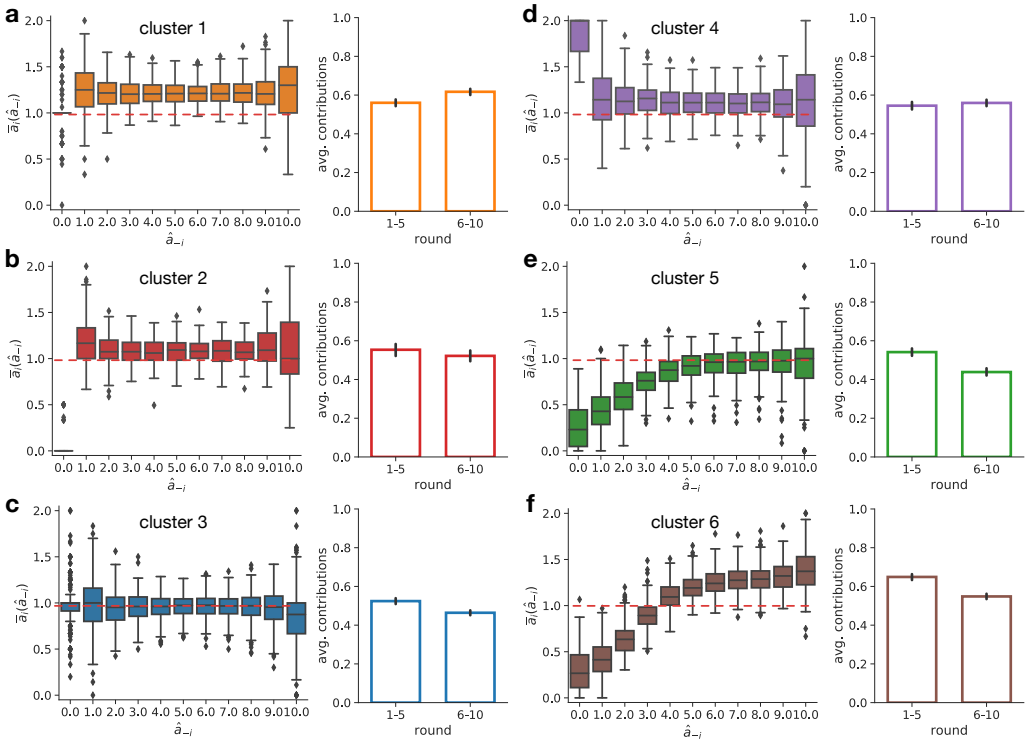

Figure 8: Conditional and time dependent behavior per cluster. In each panel, the figure on the left column shows the distribution of average contributions of the agents in the cluster in function of the contributions of the other group members in the previous learning step (see Section 3.9). The figure on the right column shows the avg. contribution in the first and second halves of the game, normalized by the maximum possible contribution in each half. The clusters in panels a-c display unconditional behavior with respect to the rest of the group, and are distinguished only by the average level of contributions and whether they happen mostly in the first or second half of the game. Panel d shows a slightly compensatory behavior, where agents increase their contributions when the group-mates decrease theirs. In contrast, the behavior shown in e and f are clearly conditional, with agents in those clusters decreasing (increasing) their contributions if the rest of the group also decrease (increase) theirs. Moreover, in these cluster, agents tend to contribute more in the first part of the game then in the rest. Interestingly, these clusters are composed by agents from the populations that adapted under timing uncertainty, which suggests that this type of uncertainty promotes such reciprocal behavior. ($Z = 50$, $N = 6$, $\overline{R} = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$)

4.4. Increasing the number of available actions facilitates cooperation in the transient zone

Finally, we look at the effect of increasing the number of available contribution levels, or actions, ($|A|$) per round as a possible driver for coordination. In Fig. 9, we show how the group achievement changes depending the number of actions that can be taken by each participant and the types of uncertainty. Fig. 9a shows how altering the number of available actions affects group achievement in function of risk in the absence of other sources of uncertainty. These results show that the number of actions do not influence the outcome when the risk is high (the game is in the coordination zone). Thus, for most parameters, in the absence of uncertainty, the results shown in the previous sections are robust to changes on the number of actions available. However, for intermediary levels of risk ($p = 0.5$) decreasing the contribution options from 3 to 2 affects negatively the fraction of successful groups. In fact, the simulation results show that when only 2 actions are available per round, a higher risk is required for the agents to cooperate.

Fig. 9b and Fig. 9c show the effect of increasing the number of available actions in the presence of timing $w = 2/3$ and threshold $\delta = 0.2$ uncertainty, respectively. In both cases, when the risk levels are intermediary ($p = 0.5$), there is an increase in group achievement with a higher number of available actions per round. This positive effect appears to be stronger in the presence of threshold uncertainty, where it can be observed that more actions lead to an earlier transition from the defection dominant to the coordination region. In the presence of also timing uncertainty (middle panel) or threshold uncertainty (right panel), more actions alter the slope of the transition between both regions, leading to more collective success (η) than when there are only a few actions to choose from. Therefore, our results suggest that increasing the number of available actions may facilitate consensus and coordination under uncertainty, being beneficial for collective action in the mid range of risks (transient region), and it should be an important factor to take into account when designing public policy that aims to address CRD situations in this region. Apart from these transients, all previous results are shown to be robust to variations on the number of actions available to players.

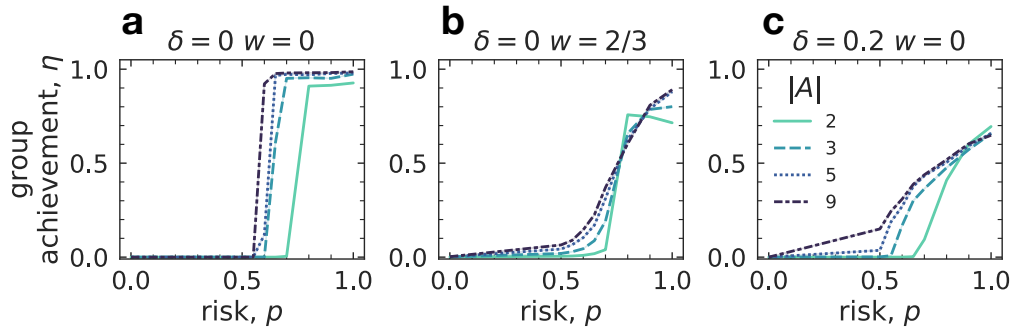

Figure 9: Group achievement as a function of risk and the number of available actions. Here we show how the fraction of successful groups (group achievement) changes if agents can select among more contribution levels at each round. Panel a) shows how this variation affects the outcome of the CRD for different risk levels and no threshold or timing uncertainty. The results show that decreasing the number of available actions from 3 to 2 reduces group achievement for intermediary risk levels ($p = 0.5$). Panels b) and c) show the results in the presence of timing ($w = 2/3$) and threshold ($\delta = 0.2$) uncertainty, respectively. In both cases, increasing the number of actions has a positive effect for intermediary levels of risk, and this positive effect appears to be stronger when threshold uncertainty is present. In general, varying the number of actions (or options) only affects the transient zone (p 0.5); in all cases the coordination zone (higher levels of risk) remains unaltered. Confidence intervals are < 0.1. ($L = 20$, $Z = 50$, $N = 6$, $\overline{R} = 10$, $\sigma = 0.5$, $\alpha = 0.09$, $\tau = 0.1$, $n_t = 10000$, $K = 10000$, $K_e = 100000$).

5. Conclusions

Many human endeavours as well as computational problems can be characterised by non-linear, uncertain and delayed rewards. These constraints conform to difficult dilemmas that require the weighing not only of personal and public interests but also that of short and long-term rewards. Climate action is a prominent example of such a situation, since the consequences of failing to prevent it might only be observed in the future, and, potentially, by the next generations. Therefore, failing to consider future consequences, or, in other words, if the perception of collective risk is not high enough, might drive the participants in this global dilemma to a *tragedy of commons* [61]. Likewise, unfortunately, the COVID-19 pandemics has provided another strong case that exemplifies such a social dilemma. When individuals do not perceive risk to be high enough, they might be tempted to bypass the confinement measures and recommendations, however by doing so, they not only increase their chances of getting infected, but also the risk of infecting

others, increasing also the likelihood the local health system will collapse. As result, this has a future effect that they might not have been perceived: if they do get ill, they might not be treated. Thus, avoiding free-riding and promoting cooperation is essential for group success, and this situation is very well abstracted by the collective risk dilemma (CRD). Under this scenario, we present a PBL model that is both simple and flexible, yet powerful enough to reproduce results in a variety of behavioral experiments. Additionally, our model can be easily extended to other n-player social dilemmas. As mentioned before, it is important to know that the PBL is not modelling the learning process faced by humans during the reported experiments. Instead, the system is only used to reproduce the behavioral patterns they have at the start of the CRD, which they have learned before arriving to the laboratory.

We find through our PBL model that while the perception of risk can drive the emergence of cooperation, threshold and timing uncertainty have a negative impact on group achievement, confirming the results observed experimentally [1, 9, 11]. Moreover, Fig. 3 indicates that there is a transient region in which the global payoff of the population is minimized (see Fig. 6). Our PBL model indicates that both threshold and timing uncertainty widen this region. Therefore, it is important to find strategies that may increase group success in these conditions. In Fig. 9 we show that increasing the number of available actions may ease consensus, providing an escape from *failed* outcomes in the transient regions, while it hardly has any effect for very high risks. However, when looking at the coordination zone (high risk) in all three uncertainty scenarios, we observe that both threshold and timing uncertainty increase the polarization in the population, reducing the number of fair players. Also, when looking at the distribution of successful behaviors learned in the coordination zone (high risk), we see that there is not much difference between only impact and threshold uncertainty, which highlights that the reduction in group achievement in the presence of threshold uncertainty is mostly due to the increase in the probability that achieving the target will require players to make greater effort. In contrast, under timing uncertainty, players appear to develop reciprocal behaviors, i.e., when co-adapting through the same population they learn to contribute less (more) when there is a higher chance that the other players will contribute less (more). This result was also observed experimentally by Fernández et al. [11].

The sources of uncertainty that we study in this paper, can only be reduced, but not eliminated. Yet, when the level of risk is lower and the CRD is in a transient zone, increasing the number of available actions can provide

an escape from the *tragedy of commons* and increase cooperation. Additionally, we find that uncertainty can considerably reduce group success and increase polarization, but, under timing uncertainty, reciprocation promotes cooperation and group success in the PBL. Finally, the perception of risk is key in this dilemma, which brings perhaps even more attention to the importance of awareness within the global population that is under a collective risk scenario.

Acknowledgements

E.F.D. is supported by an F.W.O. (Fonds Wetenschappelijk Onderzoek) Strategic Basic Research (SB) PhD grant (nr. G.1S639.17N) and J.G. is supported by an F.W.O. postdoctoral grant. T.L. is supported by the F.W.O. project with grant nr. G.0391.13N, the F.N.R.S. project with grant number 31257234 and the FuturICT2.0 (www.futurict2.eu) project funded by the FLAG-ERA JCT 2016. E.F.D., J.G. and T.L. are supported by the Flemish Government through the AI Research Program. F.C.S acknowledges support from FCT-Portugal (grants UIDB/50021/2020, PTDC/MAT-APL/6804/2020, and PTDC/CCI-INF/7366/2020)). T.L. and F.C.S. acknowledge the support by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. J.C.B is supported by Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2019-2022) and the European Union (European Regional Development Fund - ERDF).

Appendix A. Results from behavioral experiments

In Fig. 1, we show a summary of the group achievement obtained in different CRD experiments that are related to our paper. All of this results have been extracted from the papers referenced in the figure. However, in the case of threshold uncertainty [9], the authors only indicate in the experimental groups' collective contributions instead of how many groups were successful. With this information they extrapolate their results showing how many groups would have been successful had the threshold been 120, 100, 80, 60, 40 and 20. In order to calculate the expected group success from these results, we simply multiplied the probability of each of these scenarios by the fraction of groups that would have achieved the target in each of them, i.e., $\sum_{\tau \in T} P(x = \tau) * n(\tau)/n_g$, where T represent the range of possible threshold values, $n(\tau)$ the number of successful groups for this collective target, and n_g the total number of groups.

Moreover, in the 2 experiments on threshold uncertainty [9], participants only knew that the threshold (collective target) of the experiment would be picked randomly within a discrete range of [0, 240] with steps of 20 units. In one case, they also knew that the stochastic distribution from which the threshold would be picked was uniform, while in the other case they were not informed about the distribution. To simplify our calculations, and given that the authors do not specify which distribution was used for each group, we assume a uniform distribution in this case as well. Therefore, these results should be taken only as a reference value, and the actual group achievement might be slightly higher or lower.

The authors indicate that for the "uniform distribution" treatment, also named Risk, only 2/10 groups contributed at least 120 units (half of the threshold range), and 6/10 contributed 100 units, 9/10 at least 80, 9/10, and 10/10 at least 40. For the case in which the distribution is unknown, also named Ambiguity, 0/10 contributed at least 120 units, 4/10 at least 100, 6/10 at least 80, 8/10 at least 60, 9/10 at least 40 and 10/10 at least 20. Therefore, according to our calculations, the group achievement for the Risk treatment is 0.43 and for the Ambiguity treatment is 0.36.

Appendix B. Equilibria

The CRD is a threshold public goods game [4, 14, 16, 25, 62, 63], whose equilibria depend on the impact of not reaching the threshold T. First, we

define this impact in terms of the expected payoff of a player, and therefore it is a function of the risk p, assuming there are no other sources of uncertainty. When $p = 0$, players maximise their payoffs by not contributing to the public good, and since no player has any incentive to move away from this behavior, this is also a Nash equilibrium. The same applies to all values of p where $(1 - p)E > T/N$. On the limit, where $(1 - p)E = T/N$, $C_i = (1 - p)E$ is a weak symmetric Nash equilibrium.

For all other cases, where $(1-p)E < T/N$, there are two sets of equilibria: i) a collectively good set, that include all trajectories that lead exactly to achieve $g = T$; ii) and a degenerate or negative set, where players do not contribute, and therefore do not achieve the threshold. Both cases are Nash equilibria, since players do not have any incentive to deviate unilaterally. Furthermore, the equilibrium where all players contribute exactly $F = T/N$ is Pareto efficient.

It is important to notice however, that, given a single group of players, the equilibrium $g^* = T$ is not stable, since any deviation from it, will also provoke a change in the behavior of all players. In [25], the authors show that, in a one-shot version of the CRD with only 2 possible contributions, a stable cooperative equilibrium appears for certain levels of difficulty of the game.

In Milinski's experiments [1], the cost of cooperation, assuming equal contributions, is half of the endowment $(E/2)$, and the expected payoff of defecting (contributing 0) for a risk p is $\Pi_D = (1-p)E$. Therefore, if $p < 1/2$, defecting is risk dominant.

Appendix C. Reducing the cost of the collective risk promotes cooperation in the transient region

The PBL model shows that both timing and threshold uncertainty widen the transient zone of the CRD, reducing cooperation and preventing players from coordinating efforts effectively. Additionally, uncertainty appears to increase polarization in terms of overall contributions per player in the population, which results in payoff heterogeneity and inequality. Therefore, public policy that aims to address CRD situations, such as climate action or pandemics, such as the COVID19, must aim to reduce uncertainty. Yet, often sources of scientific uncertainty are unavoidable, and other measures must be taken to increase cooperation, but also to reduce wealth inequality and polarization. One effective way to tackle this problem is to act over

the difficulty of the dilemma, by reducing the cost that fair players must pay to avoid the collective loss. Thus, we define the cost of the dilemma $\sigma = T/(NE)$ as the fraction of the endowment that should be contributed by all members of the group in order to reach the collective target (see Methods). Hence, if $\sigma = 0$, no contributions are required and if $\sigma = 1$, players must contribute all their endowment. In Fig. C.10 we show how the average payoff of the population of agents varies for different levels of difficulty and risk over 4 uncertainty scenarios: a) only impact uncertainty (risk); b) timing uncertainty; c) threshold uncertainty; and d) both timing and threshold uncertainty. These results suggest that varying σ may cause non trivial effect. For instance, both in the presence of only impact uncertainty (Fig. C.10a) and also threshold uncertainty (Fig. C.10c), for high risks, lowering σ reduces the average wealth of the population, instead of increasing it. Contrarily, in in the presence of timing uncertainty (Fig. C.10b) the region of fair payoffs (when the population wealth is on average $E/2 = 10$) becomes wider. Moreover, in all cases, it is possible to increase the global wealth of the population in the transient region by only decreasing σ slightly. Which suggests, that public policy should aim to keep the cost of cooperation low in this region of risk. Finally, in each plot of Fig. C.10, we draw dashed red and black lines. These lines represent the threshold for cooperation, i.e., the group achievement on all points below the red line is $\eta > 0.1$ and $\eta > 0.5$ for all points below the black line. In the figure, the separation between these two thresholds grow with uncertainty, which suggests the space for coordination and cooperation is reduced in the presence of timing and threshold uncertainty, and reinforces the conclusion that uncertainty should be reduced in a CRD.

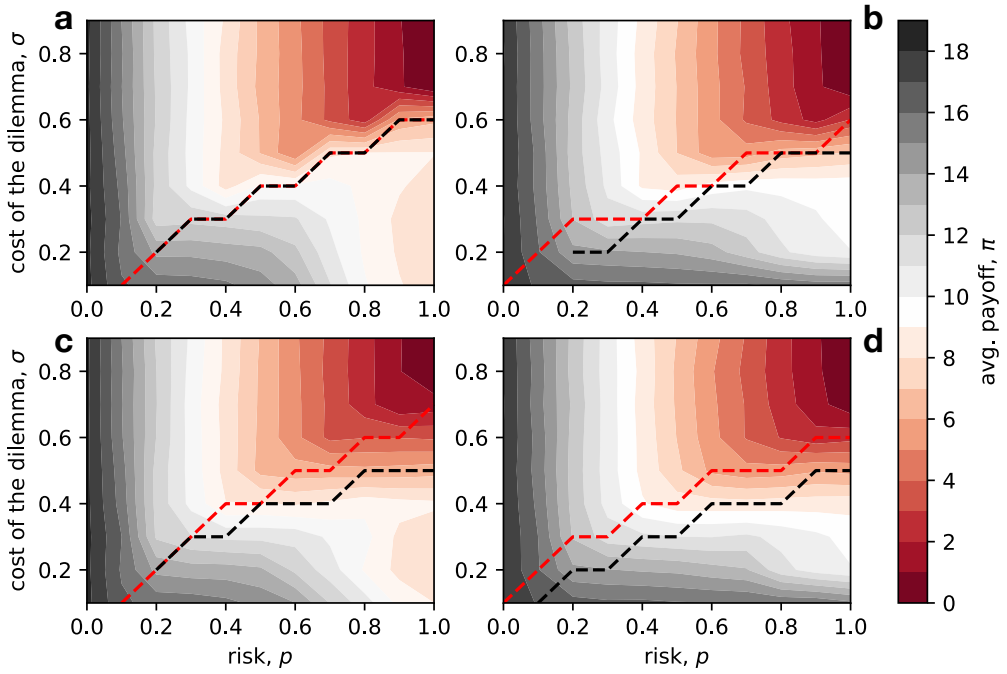

Figure C.10: Effect of the cost of the dilemma and risk on population wealth for different types of uncertainty. The plots show how the average payoff of the population depends on the level of risk, uncertainty and the dilemma difficulty, σ. Panel a) shows the results when only impact uncertainty (risk) is present. Panel b) shows the results in the presence of low levels of timing uncertainty ($w = 2/3$). Panel c) shows the same with threshold uncertainty ($\delta = 0.2$). Finally Panel d) shows the effect of the combination of the three types of uncertainty. In all panels the dashed red and black lines indicate the threshold for cooperation. The group achievement in all the points below the red line is $\eta > 0.1$ and $\eta > 0.5$ for all the points below the black line. The figure shows how the separation between these two thresholds grow with uncertainty, which suggests the space for coordination and cooperation is reduced in the presence of timing and threshold uncertainty. ($L = 10$, $Z = 50$, $N = 6$, $\overline{R} = 10$, $\alpha = 0.09$, $\tau = 0.1$, $n_g = 1000$, $K = 10000$, $K_e = 10000$).

Appendix D. List of parameters

Table D.1: List of model parameters and the range analysed in the manuscript.

Symbol	Parameter	Range analysed
Z	population size	$\{6, 12, 24, 48, 50\}$
N	group size	$\{2, 4, 6, 8, 12\}$
K	number of learning step	10000
K_e	number of evaluation games	10000
L	number of independent populations	$\{10, 20\}$
R_0	minimum number of rounds	$\{6, 7, 8, 9, 10\}$
$\overline{R}$	mean number of rounds	10
R_{max}	maximum number of rounds	50
E	individual endowment	$2\overline{R}$
T	collective target or threshold	σEN
p	risk	$[0, 1]$, steps of 0.1
w	level of timing uncertainty	$\{0, 1/2, 2/3, 3/4, 4/5\}$
δ	level of threshold uncertainty	$[0, 1]$, steps of 0.1
n_g	total number of groups in a learning step	1000
σ	cost of the dilemma	$[0, 0.9]$, steps of 0.1
$\|A\|$	number of actions per round	$\{2, 3, 5, 9\}$

Table D.2: List of symbols used in the manuscript. Here we list the symbols that are used throughout the manuscript and their meaning. The parameters already listed in Table D.1 are excluded.

Symbol	Explanation
A	discrete set of possible actions/contributions per round
S	set of states
$a_i(s)$	action of player i at round/state s
$\hat{a}_{-i}$	group contribution excluding player i in the previous round
$\overline{a}_i(\hat{a}_{-i})$	average contribution of player i when confronted with $\hat{a}_{-i}$
C_i	sum of all contributions made by player i over all the rounds
F	fair contribution
π_i	payoff of player i
Π_i	expected payoff of player i
n	number of successful groups in a learning step
η	group achievement/fraction of successful groups
$\boldsymbol{X_i}$	strategy profile
$x_{i,k}(a,s)$	probability that i takes action a at state s in learning step k
$Q_{i,k}(a,s)$	Q value of action a at state s for player i in learning step k
$\boldsymbol{b}$	encoding of a player's strategy for the clustering analysis

References

[1] M. Milinski, R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, J. Marotzke, The collective-risk social dilemma and the prevention of simulated dangerous climate change., Proceedings of the National Academy of Sciences of the United States of America 105 (7) (2008) 2291–2294. doi:10.1073/pnas.0709546105.

[2] A. Tavoni, A. Dannenberg, G. Kallis, A. Loschel, A. Löschel, A. Loschel, Inequality , communication , and the avoidance of disastrous climate change in a public goods game, Proceedings of the National Academy of Sciences 108 (29) (2011) 1–10. arXiv:arXiv:1408.1149, doi:10.1073/pnas.1102493108.

[3] M. Milinski, T. Röhl, J. Marotzke, Cooperative interaction of rich and poor can be catalyzed by intermediate climate targets, Climatic Change 109 (3-4) (2011) 807–814. doi:10.1007/s10584-011-0319-y.

[4] F. C. Santos, J. M. Pacheco, Risk of collective failure provides an escape from the tragedy of the commons., Proc. Natl. Acad. Sci. USA 108 (26) (2011) 10421–10425. doi:10.1073/pnas.1015648108.

[5] M. Abou Chakra, A. Traulsen, Evolutionary Dynamics of Strategic Behavior in a Collective-Risk Dilemma, PLoS Computational Biology 8 (8) (2012) 1–7. doi:10.1371/journal.pcbi.1002652.

[6] S. Barrett, Climate treaties and approaching catastrophes, Journal of Environmental Economics and Management 66 (2) (2013) 235–250. doi:10.1016/j.jeem.2012.12.004.

[7] V. V. Vasconcelos, F. C. Santos, J. M. Pacheco, A bottom-up institutional approach to cooperative governance of risky commons, Nat. Clim. Chang. 3 (9) (2013) 797–801.

[8] V. V. Vasconcelos, F. C. Santos, J. M. Pacheco, S. A. Levin, Climate policies under wealth inequality, Proc. Natl. Acad. Sci. USA 111 (6) (2014) 2212–2216. doi:10.1073/pnas.1323479111.

[9] A. Dannenberg, A. Löschel, G. Paolacci, C. Reif, A. Tavoni, On the Provision of Public Goods with Probabilistic and Ambiguous Thresholds, Environmental and Resource Economics 61 (3) (2014) 365–383. doi:10.1007/s10640-014-9796-6.

[10] M. A. Chakra, S. Bumann, H. Schenk, A. Oschlies, A. Traulsen, M. Abou Chakra, S. Bumann, H. Schenk, A. Oschlies, A. Traulsen, Immediate action is the best strategy when facing uncertain climate change, Nature Communications 9 (1) (2018) 2566—-2574. doi:10.1038/s41467-018-04968-1.

[11] E. Fernández Domingos, J. Grujić, J. C. Burguillo, G. Kirchsteiger, F. C. Santos, T. Lenaerts, Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization, iScience 23 (12) (2020) 101752. doi:10.1016/j.isci.2020.101752.

[12] A. E. Camacho, Adapting governance to climate change: managing uncertainty through a learning infrastructure, Emory LJ 59 (2009) 1.

[13] W. Barfuss, J. F. Donges, V. V. Vasconcelos, J. Kurths, S. 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 USA 117 (23) (2020) 12915–12922.

[14] T. Offerman, A. Schram, J. Sonnemans, Quantal response models in step-level public good games, European Journal of Political Economy 14 (1) (1998) 89–100. doi:10.1016/S0176-2680(97)00044-X.

[15] C. B. Cadsby, E. Maynes, Voluntary provision of threshold public goods with continuous contributions: Experimental evidence, Journal of Public Economics 71 (1) (1999) 53–73. doi:10.1016/S0047-2727(98)00049-8.

[16] J. M. Pacheco, F. C. Santos, M. O. Souza, B. Skyrms, Evolutionary dynamics of collective action in N-person stag hunt dilemmas., Proc. Royal Society B: Biological Sciences 276 (1655) (2009) 315–21. doi:10.1098/rspb.2008.1126.

[17] M. Milinski, C. Hilbe, D. Semmann, R. Sommerfeld, J. Marotzke, Humans choose representatives who enforce cooperation in social dilemmas through extortion, Nature communications 7 (2016) 10915.

[18] S. Barrett, A. Dannenberg, Climate Negotiations under Scientific Uncertainty, Proc. Natl. Acad. Sci. USA 109 (43) (2012) 17372–17376. doi:10.1073/pnas.1208417109.

[19] S. Barrett, A. Dannenberg, Sensitivity of collective action to uncertainty about climate tipping points, Nature Climate Change 4 (1) (2014) 36–39. doi:10.1038/nclimate2059.

[20] S. Barrett, Collective Action to Avoid Catastrophe: When Countries Succeed, When They Fail, and Why, Global Policy 7 (May) (2016) 45–55. doi:10.1111/1758-5899.12324.

[21] S. Barrett, Coordination vs. voluntarism and enforcement in sustaining international environmental cooperation, Proceedings of the National

Academy of Sciences 113 (51) (2016) 201604989. doi:10.1073/pnas.1604989113.

[22] S. Van Segbroeck, J. M. Pacheco, T. Lenaerts, F. C. Santos, Emergence of fairness in repeated group interactions, Phys. Rev. Lett. 108 (15) (2012) 158104.

[23] A. R. Góis, F. P. Santos, J. M. Pacheco, F. C. Santos, Reward and punishment in climate change dilemmas, Sci. Rep. 9 (1) (2019) 1–9.

[24] M. C. Couto, J. M. Pacheco, F. C. Santos, Governance of risky public goods under graduated punishment, Journal of Theoretical Biology (2020) 110423.

[25] F. C. Santos, V. V. Vasconcelos, M. D. Santos, P. N. B. Neves, J. M. Pacheco, Evolutionary Dynamics of Climate Change Under Collective-Risk Dilemmas, Mathematical Models and Methods in Applied Sciences 22 (supp01) (2012) 1140004. doi:10.1142/S0218202511400045.

[26] J. M. Pacheco, V. V. Vasconcelos, F. C. Santos, Climate change governance, cooperation and self-organization, Physics of Life Reviews 11 (4) (2014) 573–586. doi:10.1016/j.plrev.2014.02.003.

[27] V. V. Vasconcelos, F. C. Santos, J. M. Pacheco, Cooperation dynamics of polycentric climate governance, Mathematical Models & Methods in Applied Sciences 25 (13) (2015) 2503–2517. doi:10.1142/S0218202515400163.

[28] M. Abou, A. Traulsen, Under high stakes and uncertainty the rich should lend the poor a helping hand, Journal of Theoretical Biology 341 (2014) 123–130. doi:10.1016/j.jtbi.2013.10.004.

[29] C. Hilbe, M. Abou Chakra, P. M. Altrock, A. Traulsen, The Evolution of Strategic Timing in Collective-Risk Dilemmas, PLoS ONE 8 (6) (2013) 1–7. doi:10.1371/journal.pone.0066490.

[30] K. Hagel, M. A. Chakra, B. Bauer, A. Traulsen, Which risk scenarios can drive the emrgence of costly cooperation?, Nature Scientific Reports (2016). doi:10.1038/srep19269.

[31] A. E. Roth, I. Erev, Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term, Games and economic behavior 8 (1) (1995) 164–212.

[32] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[33] J. Grujić, T. Lenaerts, Do people imitate when making decisions? evidence from a spatial prisoner?s dilemma experiment, Royal Society Open Science 7 (7) (2020) 200618.

[34] D. Fudenberg, F. Drew, D. K. Levine, D. K. Levine, The theory of learning in games, Vol. 2, MIT press, 1998.

[35] R. Axelrod, D. Dion, The Further Evolution of Cooperation, Science 242 (4884) (1988) 1385–1390.

[36] M. Perc, Coherence resonance in a spatial prisoner's dilemma game, New Journal of Physics 8 (2) (2006) 22.

[37] Y. Wang, K. Sycara, P. Scerri, Towards an understanding of the value of cooperation in uncertain world, in: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02, IEEE Computer Society, 2011, pp. 212–215.

[38] Y. Wang, S. Luo, J. Gao, Uncertain extensive game with application to resource allocation of national security, Journal of Ambient Intelligence and Humanized Computing 8 (5) (2017) 797–808.

[39] J. E. Harrington Jr, A non-cooperative bargaining game with risk averse players and an uncertain finite horizon, Economics Letters 20 (1) (1986) 9–13.

[40] E. Van Dijk, A. Wit, H. Wilke, D. V. Budescu, What we know (and do not know) about the effects of uncertainty on behavior in social dilemmas, Contemporary psychological research on social dilemmas (2004) 315–331.

[41] T. Börgers, R. Sarin, Learning through reinforcement and replicator dynamics, Journal of Economic Theory 77 (1) (1997) 1–14.

[42] M. W. Macy, A. Flache, Learning dynamics in social dilemmas., Proceedings of the National Academy of Sciences of the United States of America 99 Suppl 3 (2002) 7229–36. doi:10.1073/pnas.092080099.

[43] T. Ezaki, Y. Horita, M. Takezawa, N. Masuda, Reinforcement Learning Explains Cooperation and Its Moody Cousin, PLoS Computational Biology 12 (7) (2016) e1005034. doi:10.1371/journal.pcbi.1005034.

[44] Y. Horita, M. Takezawa, K. Inukai, T. Kita, N. Masuda, Reinforcement learning accounts for moody conditional cooperation behavior: experimental results, Scientific reports 7 (1) (2017) 1–10.

[45] D. Bloembergen, K. Tuyls, D. Hennes, M. Kaisers, Evolutionary dynamics of multi-agent learning: A survey, Journal of Artificial Intelligence Research 53 (2015) 659–697.

[46] S. Tanabe, N. Masuda, Evolution of cooperation facilitated by reinforcement learning with adaptive aspiration levels, Journal of theoretical biology 293 (2012) 151–160.

[47] S. De Jong, S. Uyttendaele, K. Tuyls, Learning to reach agreement in a continuous ultimatum game, Journal of Artificial Intelligence Research 33 (2008) 551–574.

[48] F. P. Santos, F. C. Santos, F. S. Melo, A. Paiva, J. M. Pacheco, Dynamics of fairness in groups of autonomous learning agents, in: International Conference on Autonomous Agents and Multiagent Systems, Springer, 2016, pp. 107–126.

[49] E. F. Domingos, J. C. Burguillo-rial, T. Lenaerts, Reactive Versus Anticipative Decision-Making in a Novel Gift-Giving Game, in: 31st AAAI Conference on Artifitial Intelligence, 2017, pp. 4399–4405.

[50] S. Van Segbroeck, S. de Jong, A. Nowé, F. C. Santos, T. Lenaerts, Learning to coordinate in complex networks, Adaptive Behavior 18 (5) (2010) 1–17. doi:10.1177/1059712310384282.

[51] T. Ezaki, N. Masuda, Reinforcement learning account of network reciprocity, PloS one 12 (12) (2017) e0189220.

[52] W. Barfuss, J. F. Donges, S. J. Lade, J. Kurths, When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nature Communications 9 (1) (2018) 1–10. doi:10.1038/s41467-018-04738-z.

[53] W. Barfuss, J. F. Donges, J. Kurths, Deterministic limit of temporal difference reinforcement learning for stochastic games, Physical Review E 99 (4) (2019) 1–18. arXiv:arXiv:1809.07225v2, doi:10.1103/PhysRevE.99.043305.

[54] C. F. Camerer, T. Ho, J.-K. Chong, Sophisticated Experience-Weighted Attraction Learning and Strategic Teaching in Repeated Games, Journal of Economic Theory 104 (1) (2002) 137–188. doi:10.1006/jeth.2002.2927.

[55] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, T. Graepel, A multi-agent reinforcement learning model of common-pool resource appropriation, in: Advances in Neural Information Processing Systems, 2017, pp. 3643–3652.

[56] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.

[57] M. Wattenberg, F. Viégas, I. Johnson, How to use t-sne effectively, Distill 1 (10) (2016) e2.

[58] K. Alsabti, S. Ranka, V. Singh, An efficient k-means clustering algorithm, Working paper, College of Engineering and Computer Science, Syracuse University (1997).
URL https://surface.syr.edu/eecs/43

[59] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE transactions on pattern analysis and machine intelligence 24 (7) (2002) 881–892.

[60] Z. Wang, M. Jusup, H. Guo, L. Shi, S. Geček, M. Anand, M. Perc, C. T. Bauch, J. Kurths, S. Boccaletti, et al., Communicating sentiment and outlook reverses inaction against collective risks, Proceedings of the National Academy of Sciences 117 (30) (2020) 17650–17655.

[61] G. Hardin, The Tragedy of the Commons, Science (New York, N.Y.) 162 (1968) 1243–1248.

[62] R. T. a. Croson, M. B. Marks, Step returns in threshold public goods: A meta- and experimental analysis, Experimental Economics 2 (3) (2000) 239–259. doi:10.1007/BF01669198.

[63] F. P. Santos, F. C. Santos, A. Paiva, J. M. Pacheco, Evolutionary dynamics of group fairness, Journal of Theoretical Biology 378 (2015) 96–102.