Evolution in Social Dilemma Games, Social Pressure and Emergent Behaviours

Diego Calanzone
University of Trento

***Abstract*—In Reinforcement Learning literature, Social Dilemma Games model social dynamics to test multiple learning agents. In neural methods for Multi-Agent Reinforcement Learning (MARL), techniques such as direct punishment or shared experience actor-critic have been proposed to understand the emergence of cooperation [1][2]; however, behavior in biological systems could be influenced by intrinsic factors: a living being is driven by needs and limits such as aging, inherited genetic traits, starvation. This study introduces evolutionary and demographic dynamics to social dilemma games: with notions from crowd behavior and collective intelligence, we simulate learning through mating and mutation under different conditions and look for the presence of emerging strategies.**

***Index Terms*—evolution, social dilemma, cooperation, emergence**

I. Introduction

SOCIAL dilemma games involve multiple agents that act by balancing cooperation and competition. Depending on the conditions of the environment, prioritizing one behaviour or the other allows to maximize either a single-agent or a collective reward. From an evolutionary perspective, these factors can have an impact on different time spans: "selfish" players ensure their own survival while threatening the other players' one, only the fittest will generate the next offspring; contrarily, "cooperative" players aim at maximizing the outcomes for the entire group, which translates into diversity and homogenized rewards. These scenarios lead to populations with different distributions of ages and returns.

The goal of this study is to investigate the conditions under which collaboration emerges during evolution in a dynamic environment: level-based foraging (LBF), a free-roaming game where agents collect items with varying levels of difficulty. Different conditions, depending on food abundance and population density, are tested on 50 evolving individuals. This is based on the experimental framework set by Christianos, et al. [1].

II. Environment & assumptions

Multi-agent Markov Games. The interaction mechanisms in LBF are modeled as a Markov Game $\mathcal{M}$, defined by a finite set of states $\mathcal{S}$ over an environment matrix $E \in R^{M \times M}$; perception of the environment for each agent is modeled with an observation function $\mathcal{O} : \mathcal{S} \times \{1, ..., N_p\} \rightarrow R^{3 \times N \times N}$, with N_p the amount of players, $o \in \mathcal{O}$ a $3 \times N \times N$ observation tensor accounting for the local neighborhoods of players and items, respectively[1]; state transitions are modeled with the function $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_p \rightarrow \Delta(\mathcal{S})$ with $\Delta(\mathcal{S})$ a discrete distribution over the resulting states, $\mathcal{A}_p$ the set of allowable actions for each p player; state outcomes are represented by the agent reward function $r_p : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_p \rightarrow R$. Each player is represented by a policy function $\pi_p : \mathcal{O}(\mathcal{S}) \rightarrow a \in \mathcal{A}_p$ that given a partial observation, it returns one of the allowable actions.

Level-based foraging (LBF). The environment chosen for this study was introduced by Christianos et al. [1] for multi-agent reinforcement learning (MARL). The game is described by the following conditions:

- Agents P_i and items F_i are scattered around the map with a random level.
- An item is successfully collected if the sum of the levels of the involved agents is equal or greater than the item's level $\in [1, 10]$[1].
- When an item is collected, the individual reward is equal to the level of the item divided by the normalized level of each agent involved (relative contribution).
- Possible values for an action $a \in \mathcal{A}_p$ are in [0, 5], corresponding to: *do nothing, north, south, west, east, load*. The movements have 1 tile of stride.
- A partial observation $\mathcal{O}(\mathcal{S})$ is a tensor composed of two $R^{N \times N}$ matrices centered on the player's position, where $N = 2s + 1$ and s is the directional range of sight (in tiles). The first matrix is populated with values in [1, 10][1], where each cell is a player level, if present; similarly, the second matrix is populated with the same range of values corresponding to the presence of food items; the third matrix concerns binary accessibility values to access the tiles.
- In this study, each agent's sight reaches 20 tiles per direction (north, east, south, west). To reduce the amount of input variables, a mean pooling operator is applied direction-wise with a downsampling factor of 5. This results in partial observations $\mathcal{O}(\mathcal{S}) \in R^{2 \times 9 \times 9}$, that is 162 inputs.

The game is not trivial: agents have to balance between cooperation and competition to both increase the individual reward and to cooperate when they approach items difficult to collect. The presence of multiple agents in a region induces the creation of strategies, thus the choice of who to compete/collaborate with. We test these behaviors by simulating evolution different environments of two main types:

[1]Design choice for this study.

natural, with random food items levels; forced cooperation, with maximum food items levels. We examine the distribution of the average rewards as well as the portion of collected food.

Social pressure. Edward T. Hall [3] denoted with *Proxemics* the effects of population density in behaviour, communication and social interaction. Distances between individuals influence the perception of the relationships among them. Three regions of a subject's surrounding space are identified:

- Personal space: between 0.46m and 1.22m. The violation of this boundary is cause for the individual's stress.
- Social space: between 1.22m and 3.7m. Reserved to individuals the subject is interacting with, a small group of peers.
- Public space: beyond 3.7m. Generally regards large audiences and open crowded spaces.

The absolute distance D_a in meters is converted to the relative distance D_r in units, by assuming a subject size D_s = 55cm.

$$D_r = \frac{D_a}{D_s} \tag{1}$$

Hall's distances result in: *personal space* $\in (0.83, 2.22]$, *social distance* $\in (2.22, 6.73]$, *public distance* $\in (6.73, +\infty)$. In level-based foraging, a unit is one tile.

The density d, measured in players/tile is given by dividing the number of players spawned N_p by the total amount of tiles (positions in the environment) $M \times M$:

$$d = \frac{N_p}{M \times M} \tag{2}$$

In this study, we use Proxemics to build the *social pressure scale*: it measures population density with a sensitivity dictated by Hall's distances, which allows to describe the level of social pressure. We take the lower bound of Hall's *social distance* in relative units: it is rounded to 2 tiles of distance in all directions, resulting in a 5x5 squared area. The density of this space corresponds to $d = \frac{1}{5\times5} = 0.04$ players/tile. This denotes the *neutral level* of social pressure (SPL).

Level	Condition	Players/tile
1	Sparse	$x \leq 0.02$
2	Neutral	$0.02 < x \leq 0.04$
3	Stressful	$x > 0.04$

TABLE I: Social pressure levels: players per tile ratio.

Food abundance. The experimental settings of the environment used by Christianos et al. [1] are considered to derive three food abundance levels (FAL).

Level	Condition	Food/person
1	Scarcity	$x < 1$
2	Neutral	$x = 1$
3	Abundance	$x \geq 1$

TABLE II: Food abundance levels: items per person ratio.

III. Methodology

In this study the process of evolution of individuals is simulated with Evolutionary Algorithms (EA): a family of meta-heuristics for single and multiple objective optimization. In the context of social dilemma games, EAs simulate the evolution of a population of individuals given the expectation of the cumulate episode reward as fitness function (to maximize). Evolution and Reinforcement learning are conciliated with the use of Evolutionary Decision Trees [4][5], whose leaves are trained with $\mathcal{E}$-decay reinforcement learning.

Decision tree policies. Each player is a policy function π_p, modeled with a *Decision Tree (DT)*: the input state is a flattened observation tensor $\mathcal{O}(\mathcal{S}) \in R^{162}$; through a series of tests (DT nodes), given an input state the tree is traversed down to a specific leaf, which associates the available state actions with a predicted reward. Each DT node corresponds to a linear decision boundary in the feature space parallel or oblique[1] to the axis. Moreover, DTs with a contained branching factor and depth are human-interpretable [5]: this could provide insights on the agent's behavior.

Evolutionary Algorithms. EAs are iterative procedures that evolve a population of individuals through *selection, mutation, crossover, replacement, fitness evaluation*. In the context of Evolutionary DTs, each player is an individual $P_i \in R^k$ of the population and it is represented by a sequence of k genes, the **genome** or genotype. In this implementation, the genome is a list of integer values to translate into the *phenotype*: a series of if-then-else instructions generated according to the production rules of a defined grammar, as implemented by Custode, Iacca [5]. We adopt the following design decisions:

- A fixed population size of $\lambda = 50$.
- Tournament selection with a pool size of 2.
- For each environmental condition, the population is evolved for 50 generations.
- We apply uniform crossover with indipendent gene probability of 0.4 and general crossover probability of 0.5.
- A mutation probability of 0.9 for each gene.
- A genotype of length $|k| = 300$.
- The fitness function for each individual is the average cumulative return over all the episodes. The generation diversity is a matrix of pairwise Euclidean distances.

Reinforcement Learning. The DT leaves are trained a posteriori with *$\mathcal{E}$-greedy Q-Learning* [6] to associate a reward to each action in a given state. For each generation, the population interacts with a randomly generated environment for $N_{episodes}$ iterations, each N_{steps} long. Under this condition, the optimal policy consists in choosing the action that maximizes the reward. However, to nudge exploration a random action can be sampled instead, with probability $\epsilon \in [0, 1]$. We use the configuration:

- $\epsilon = 0.25$, constant over time.
- A learning rate $\alpha = auto$.
- $N_{episodes} = 1000$, $N_{steps} = 200$.

IV. Experiments

The original environments to evaluate RL agents in LBF [1] are considered as base for this study. We further introduce new parameters: the social pressure level (SPL) and the food abundance level (FAL). N_P denotes the number of players, while N_F the amount of food items.

Ref.	Tiles	N_P	N_F	N_P/tile	N_P/food
a	12x12	2	1	0.014	2
b	10x10	3	3	0.030	1
c	15x15	3	4	0.013	0.75
d	8x8	2	2	0.031	1

TABLE III: Experimental conditions from Christianos et al.[1]

We take in consideration the three most distinct cases (*a, b, c*), for which we compute the N_p/tile ratio and keep it fixed while scaling the environment matrix (tiles) to N_p = *50*:

Ref.	Tiles	N_P	N_F	N_P/tile	N_P/food	SPL	FAL
a	60x60	50	25	0.014	2	1	1
b	41x41	50	50	0.030	1	2	2
c	62x62	50	67	0.013	0.75	1	3
e	35x35	50	25	0.040	2	2-3	1
f	25x25	50	25	0.080	2	3	1

TABLE IV: Extended experimental conditions with the corresponding levels of social pressure and food abundance.

Conditions (*e*, *f*) have been introduced to experiment with a higher level of social pressure. For each condition, two types of environments are conducted:

- Natural setting (N): spawned food items have random levels $level(P_i) \in [0, 10]$, up to the max player level.
- Forced cooperation (FC): the levels of the spawned food items are fixed to the max player level, $level(P_i) = 10$. This forces multiple players to sum their contribution and collect items.

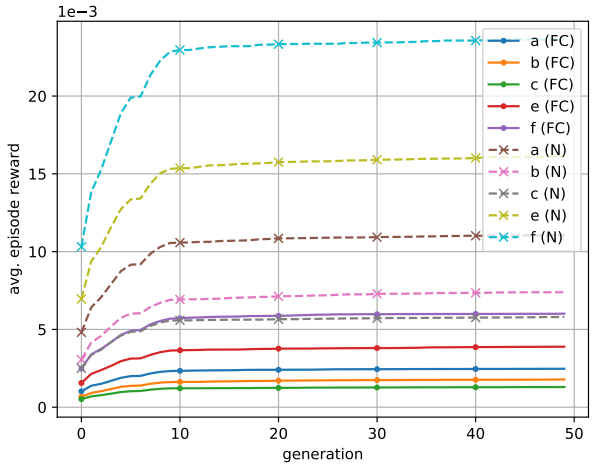

Fig. 1: Expected rewards at learning time for 10 experimental conditions over 50 generations.

V. Results

Figure 1 reports the learning curves of populations in *a, b, c, e, f* without (N) and with forced cooperation (FC, "coop"). Each population is evaluated for 10 episodes, 200 steps each. Two metrics are considered: C_{food}, the fraction of collected food; $\bar{r}$, the average episode reward over the individuals; mean and standard deviation of the scores over the episodes are reported in Table V.

Ref.	Type	C_{food}	$\bar{r}$
a	N	0.3840 ± 0.1061	0.0073 ± 0.0021
b	N	0.5020 ± 0.0878	0.0095 ± 0.0018
c	N	0.3448 ± 0.0679	0.0066 ± 0.0014
e	N	0.5640 ± 0.1193	0.0110 ± 0.0027
f	N	**0.7720 ± 0.0781**	**0.0149 ± 0.0018**
a	FC	0.1720 ± 0.0840	0.0034 ± 0.0017
b	FC	0.2560 ± 0.0320	0.0051 ± 0.0006
c	FC	0.1403 ± 0.0373	0.0028 ± 0.0007
e	FC	0.3480 ± 0.1132	0.0070 ± 0.0023
f	FC	**0.5040 ± 0.1120**	**0.0101 ± 0.0022**

TABLE V: Average population returns in collected food and rewards over 10×200 steps.

With or without the enforcement of cooperation, we observe the highest C_{food} and $\bar{r}$ are achieved in *e* and *f*, these scenarios feature high social pressure (SPL ≥ 2) and food scarcity (FAL = 1). Condition *b* follows with neutral SPL and FAL. This favors our hypothesis: competition due to high population density and limited items nudges agents to collect more food within the same timespan.

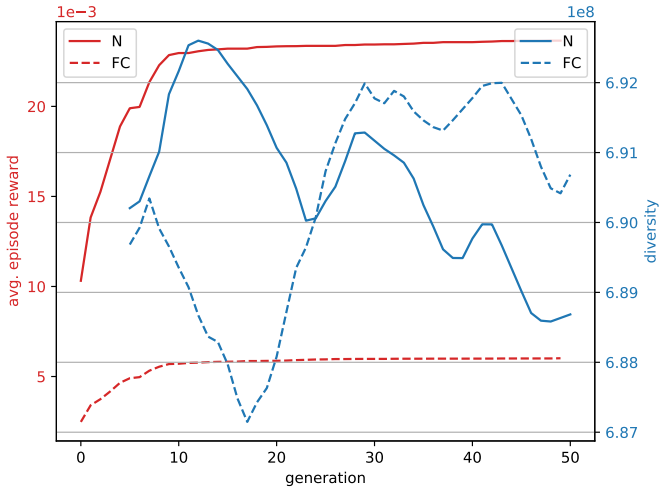

Fig. 2: Expected episode reward against diversity in (*f*), with and without forced collaboration (FC). Diversity is smoothed with a moving average of 5 steps.

In Figure 2, we focus on condition (*f*) to observe the relationship between diversity and fitness: when the former increases, the population encodes a larger variety of sub-solutions that can be exploited, which leads to significant improvements in returns. The population diversity achieves a peak around generation 10, where the average reward starts to saturate. Subsequently, as fitness achieves a plateau we observe an alternation of peaks and valleys in diversity, with an overall

decreasing trend, which could be a symptom of premature convergence. This phenomenon may be caused by multiple factors: from the evolutionary perspective, longer genomes and more sophisticated crossover operators could be tested; overall, in condition (f) the agents explore the environment for 10.000.000 steps and collect $77.20 \pm 7.81\%$ of food items within a generation, this is comparable with the performance achieved Shared Experience Actor-Critic [1].
Forced collaboration lowers the returns in all the experiments: with a reduced ability to collect any item on their own, rewards are more sparse; this is reflected in a less steep learning curve. A peak in diversity is followed by the achievement of a plateau in returns; with respect to the (N) natural condition, by the end of the evolutionary process individuals keep a higher variety (Figure 2), which supports our hypotheses on cooperation and genetic preservation.

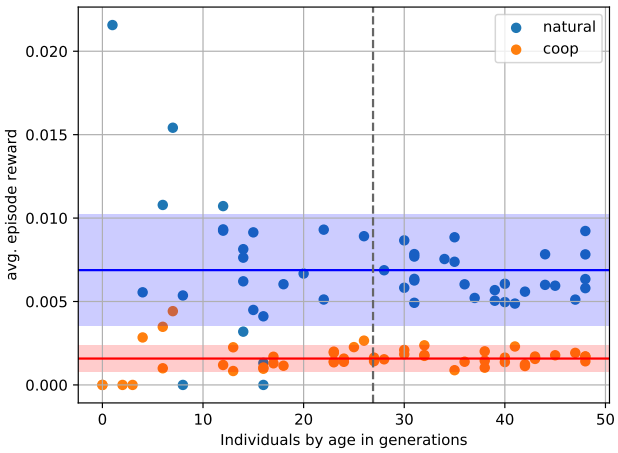

Fig. 3: Distribution of returns over individuals, sorted by age in conditions *f(N)* and *f(FC)*. Horizontal and vertical lines correspond to the axis averages over the population.

Finally, in condition (f) we compute the genetic age of each individual in the final population: in inverse chronological order, we count the amount of generations in sequence where the genome appears (Figure 3); individuals and the respective avg. episode reward are sorted by age. In the natural setting, outliers are mostly recent individuals with competitive behaviors, this is supported also by the width of the distribution and a concentration of older individuals below the average return value. Conversely, in the collaborative setting the distribution of returns is concentrated: at the cost of an overall smaller return, the majority of the population achieves similar rewards; fluctuations could be explained also by the randomness of the player levels, which are proportional to the return.

VI. Discussion

One major challenge in this project was the computational cost: the limited budget to test large-scale environments required finding a tradeoff between the length of the experiments (in generations of evolution, episodes, steps) and the variety of environmental conditions (social pressure, food scarcity, forced collaboration and natural competition). Both dimensions were crucial in obtaining good conditions for learning. Firstly, we have reduced the complexity of the inputs with a mean pooling operator: the agent perceives in a discretely large neighborhood $\mathcal{O}(\mathcal{S}) \in R^{3\times41\times41}$, but these information are summarized by computing the direction-wise mean each 5 tiles, leading to $\mathcal{O}(\mathcal{S}) \in R^{2\times9\times9}$– 2 channels are kept as the last one, regarding physical accessibilities, is discarded.

Originally, the experiments conducted by Christianos et al. [1] involved up to 3 agents (an insufficient number from an evolutionary perspective) trained for over 30,000,000 iterations (unfeasible for this study). By formalizing the environmental conditions with Proxemics, we scaled local social dynamics to a population of 50 individuals. We found $N_{episodes} \geq 1000$ to be an acceptable lower bound for the training environments; we extended N_{steps} from originally 25 to 200 in order to allow further exploration of the scaled environment space and strategies in a wider time span. This resulted in an increased rate of collected food wrt. experiments with more episodes (3000) and less steps (25).

We experienced the known problem of *sparse rewards*: within an episode, few actions lead to rewards, which is linked to small average episode returns encountered in all the experiments. We adopted two methods to tackle this issue: with evolution, we set a high mutation probability (e.g. 0.8) and a small tournament pool size (e.g. 2) in order to push the exploration of behaviors; with RL, we set a higher amount of steps per episodes to allow agents to accumulate more rewards within an episode.

We suspect some limitations in our approach: in all of our experiments, we found no significant improvements in returns after 10-15 generations, suggesting further exploration could be necessary by introducing a decaying ϵ parameter (initial "warmup" with more random actions), by increasing the amount of episodes or by adopting Evolutionary Strategies [7].

Finally, from our experiments we hypothesize that cooperative behavior can be learned through evolution, especially under extreme environmental conditions. This comes at the cost of overall population returns, while it can ensure higher diversity whose effects in the longer term could be further investigated in more complex social dilemmas.

References

[1] S. V. A. Filippos Christianos, Lukas Schäfer, "Shared experience actor-critic for multi-agent reinforcement learning," 2021.
[2] M. M. Nayana Dasgupta, "Investigating the impact of direct punishment on the emergence of cooperation in multi-agent reinforcement learning systems," 2023.
[3] B. B. P. B. A. R. D. J. M. D. M. S. E. J. L. F. D. H. S. T. K. W. L. B. F. L. S. J. J. E. M. D. S. M. G. B. M. H. B. S. G. L. T. Edward T. Hall, Ray L. Birdwhistell and A. P. Vayda, "Proxemics," 1968.
[4] S. N. L. Z. Yashesh Dhebar, Kalyanmoy Deb and D. Filev, "Interpretable-ai policies using evolutionary nonlinear decision trees for discrete action systems." 2020.
[5] G. I. Leonardo Lucio Custode, "Interpretable ai for policy-making in pandemics," 2022.
[6] C. J. C. H. Watkins, "Learning from delayed rewards." 1989.
[7] X. C. S. S. I. S. Tim Salimans, Jonathan Ho, "Evolution strategies as a scalable alternative to reinforcement learning," 2017.