Inclusive Rewards as a First Step Towards Evolutionary Multi-Agent Autocurricula

Andries Rosseau
Vrije Universiteit Brussel
Brussels, Belgium
andries.rosseau@vub.be

Raphaël Avalos
Vrije Universiteit Brussel
Brussels, Belgium
raphael.avalos@vub.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

ABSTRACT

The competitive and cooperative forces of natural selection have driven the evolution of intelligence for many millions of years, eventually culminating in nature's vast biodiversity and the complexity of our human minds. In this paper, we present a novel multi-agent reinforcement learning framework, inspired by the process of evolution. We assign a genotype to each agent, and propose an *inclusive reward* that optimizes for the fitness of an agent's genes. Since an agent's genetic material can be present in other agents as well, our inclusive reward also takes genetically related individuals into account. We study the effect of inclusion on the resulting social dynamics in two network games, and find that our results follow well-established principles from biology. Furthermore, we lay the foundation for future work in a more open-ended 3D environment, where agents have to ensure the survival of their genes in a natural world with limited resources. We hypothesize the emergence of an arms race of strategies, where each new strategy will be a gradual improvement in response to an earlier adaptation from other agents, effectively creating a multi-agent autocurriculum similar to biological evolution. Our evolutionary autocurriculum provides a novel social dimension that features a non-stationary spectrum of cooperation due to the finite environmental resources and changing population distribution. It has the potential to create increasingly advanced strategies, where agents learn to balance cooperative and competitive incentives in a more complex and dynamic setup than previous works, where agents were often confined to predefined team setups that did not entail the social intricacies that biological evolution has. We argue this could be an important contribution towards creating advanced, general and socially intelligent agents.

KEYWORDS

Multi-agent; Reinforcement Learning; Evolution; Cooperation; Autocurriculum; Networks; Open-Endedness

1 INTRODUCTION

Creating intelligent agents with the ability to adapt to a diverse set of challenges and environments is a prominent goal of artificial intelligence research. In the past decades, the field of single-agent Reinforcement Learning (RL) [40] has made great progress in developing agents capable of completing tasks provided in the form of a reward signal [1, 15, 27, 28, 36]. However, in traditional single-agent RL, once an agent has learned to master its given task, learning stops, since there is no further incentive for improvement. This makes it difficult to create agents which can solve a wide variety of complex tasks; an important characteristic of general intelligence. Creating extra handcrafted tasks for an agent to train and generalize on (i.e., transfer learning [42]) can mitigate this, but this approach is resource intensive and still caps the final complexity that the agent can achieve. Recent work on procedurally generated tasks and environments [10, 38, 46] attempts to solve this problem, but creating adequate reward signals in the environments still remains costly.

Another problem is that an objective can be too complex to learn from scratch. In this case, the space of possible policies is too large to cover effectively with regular exploration strategies. This leads to an agent that cannot close the gap between its initial random behavior and solving the task, getting stuck in low-performing suboptima. Intermediate rewards (such as reaching checkpoints in a maze) can be created to help the agent learn the final overarching goal, but over-engineering the reward signal can lead to problems like specification gaming [2, 21] and potentially limits the range and originality of the learned strategies. A slightly different approach is the use of curriculum learning [5, 11, 29], in which an agent gradually progresses from an easy environment (e.g., a small maze) towards more difficult ones (the full maze), similar to how human education works. Curriculum learning can provide an efficient way of learning complex tasks, but the final complexity is limited by the most advanced task, and creating suitable progressions in the curriculum is resource intensive. Improved exploration strategies can help an agent to avoid getting stuck in suboptimal states. Work in the field of intrinsic rewards proposes ways of overcoming a sparse reward signal. For example, curiosity driven RL [8, 9, 33] provides a self-supervised reward signal which promotes the exploration of previously unknown environment dynamics, often leading to the development of useful skills solely by following the intrinsic reward.

Nature, however, is not a single-agent system, but a multi-agent world full of evolving organisms. The competitive and cooperative forces of natural selection have driven the evolution of intelligence for many millions of years, culminating in nature's great biodiversity and the richness of our human minds. In nature, when a strategy with an increased fitness emerges, it changes the environment dynamics for others, creating a new set of challenges to adapt to [48]. The agents that successfully adapt to these challenges have in turn improved their strategies, thereby again providing new challenges, and so forth. Less successful agents, which are unable to keep up, go extinct. Agents are therefore always at a similar level, and provide just the right amount of challenge for growth. This can be applied to reinforcement learning as well: learning agents continuously improve, thereby pushing the others to adapt, leading to the emergence of a multi-agent *autocurriculum* [23]. Multi-agent autocurricula provide a scalable way for agents to explore a large

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, da Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

strategy space by simply following the gradients of their experience, called "exploration by exploitation" [4, 23].

In theory, an autocurriculum enables the possibility of unbounded growth for innovation, limited only by the strategy space of the environment and the agents' learning capacities. Autocurricula have formed the backbone for some of the most advanced forms of artificial intelligence known to date. In the form of self-play, it has led to agents with superhuman skill levels in the two-player zero-sum games of Backgammon [43], Go, Chess and Shogi [37], and the continuous real-time strategy game of StarCraft II [44]. In team-based competitive environments, it has led to agents beating the world champions in the real-time strategy game of Dota2 [6], to human-level performance in a first-person 3D multiplayer game of capture-the-flag [20], and to an arms race in a 3D hide-and-seek game [4], where several distinct strategic phases emerged, each requiring increasingly sophisticated forms of cooperation and use of tools.

Our work fits in the tradition of the aforementioned works on multi-agent autocurriculum learning, aiming to create high levels of complexity starting from elegant, simple rules. However, the dynamics in previous work on multi-agent autocurricula are limited to either all competition, or in the case of predefined team setups, all cooperation with competition between teams. Humans – like many other organisms in nature – are not that binary, instead showing a range of cooperative behavior. As the main contribution of this work, we propose the first steps towards a novel multi-agent autocurriculum inspired by biological evolution, where we construct an evolutionary aligned reward based on the survival of an agent's genes. Since other agents potentially carry parts of the same genetic material, the reward function includes the fitness of others as well, weighted by a measure of genetic relatedness. Our reward structure therefore also leads to the emergence of a spectrum of cooperation, based on the relatedness of the genotypes present. This spectrum can shift over time: for example, when the total population size grows, resources become scarce, leading to a tension between helping closer relatives and relatives carrying less of your genetic material. At any given time, the actual levels of cooperation are therefore determined by the population of genotypes, but themselves influence the population of genotypes that will be present in the future. This cycle adds a novel social dimension to the multi-agent autocurriculum, which continuously challenges the agents to find new strategies that balance cooperation and defection appropriately. We argue that our multi-agent autocurriculum has the potential to show a continuous growth in agents' strategic complexity, only bounded by the strategy space of the environment and the agents' learning capabilities.

2 METHODS

2.1 Information stability

We propose a general definition of fitness as the stability of an *information state* with regards to its environment. The more stable an information state is, the longer it will keep existing. For example, galaxies or diamonds are information states with a high fitness in the realm of physics. When considering complex organic molecules on a primeval Earth, the strategy of producing a copy of oneself before being destroyed (i.e., replicating) turned out to be a particularly stable one, drastically lengthening the existence of one's information state. Yet, every so often, an error occurs in the copying process, known as a mutation. When a mutation is beneficial, it leads to an improvement in fitness, which is favored by natural selection. At the same time, however, a mutation also changes the information state. Mutations, coupled with natural selection, generally lead to a gradual shift in the population of replicators, towards increasingly stable (fit) information states. Replicators, which exist today in the form of DNA, have built an astonishing set of ingenious organisms around themselves to help them survive.

In biology, the parts of DNA that code for the observable traits (i.e., the behavior) of an organism are the genes, and together they form a genotype. The genotype represents the complete information state on which selection acts. To translate our definition of fitness into an evolutionary aligned reward function for reinforcement learning, we implement an abstract version of genetics in our agents. We assign an agent i with an abstract genotype $\boldsymbol{g}_i$, which is a sequence of n genes where each gene locus/index $k \in [1, n]$ contains a gene g_i^k. Different integer values for g_i^k then represent different gene variants, where in principle every gene locus can have an undetermined amount of gene variants .

We propose a metric of information similarity between genotypes to quantify their relatedness. In information theory, the Hamming distance $H(\boldsymbol{s_1}, \boldsymbol{s_2})$ [18] between two sequences $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ is the number of positions at which corresponding entries are different, measuring the amount of substitutions ('bit flips') needed to change one sequence back into the other. Starting from the normalized Hamming distance, we derive a similarity metric, expressing the genetic relatedness between two agents as a real number between 0 and 1, which we name the *Hamming similarity*. Considering two agents i and j, the Hamming similarity is defined as:

$$h(\boldsymbol{g}_i, \boldsymbol{g}_j) \equiv 1 - \frac{1}{n} H(\boldsymbol{g}_i, \boldsymbol{g}_j) = \frac{1}{n} \sum_{k=1}^{n} \delta(g_i^k, g_j^k), \tag{1}$$

where $\delta(\cdot, \cdot)$ is the Kronecker delta. Note that this metric is defined when both genotypes have the same length. In the case of different genotype lengths, we could use the Damerau-Levenshtein distance [7], an extension of the Hamming distance which takes into account information deletions and insertions as well.

2.2 Inclusion

2.2.1 Inclusive reward function. Since an agent's genetic material can be present in others as well, helping agents which are genetically related should also be promoted by our reward function. We therefore modify the reward of each agent i by adding the rewards of the other agents as well, multiplied by their Hamming similarity h (Eq. 1). We call this modified reward the *inclusive reward*, after the concept of inclusive fitness [16, 34], which posits that under the right circumstances, natural selection favors organisms that help their genetic relatives. We define the inclusive reward r^* as:

$$r_i^* \equiv \sum_j h(\boldsymbol{g}_i, \boldsymbol{g}_j) r_j = r_i + \sum_{j \neq i} h(\boldsymbol{g}_i, \boldsymbol{g}_j) r_j \tag{2}$$

An illustrative example of an inclusive reward is given in Figure 1, where we consider two agents playing a prisoner's dilemma [22]. Both agents can either cooperate (C), or defect (D). A rational agent will always choose to defect, since that action always gives more

	C	D
C	4, 4	0, 6
D	6, 0	2, 2

⟶

	C	D
C	7, 7	4.5, 6
D	6, 4.5	3.5, 3.5

Figure 1: A prisoner's dilemma played by two players with genotypes [1, 1, 1, 1] and [1, 1, 1, 0] becomes a harmony game under the inclusive reward.

payoff, regardless of the action of its opponent. However, when both agents defect, they are worse off than had they both cooperated, leading to the dilemma. However, the dynamics change drastically when we introduce genes with our inclusive reward function. If we consider that the genotypes of the row player and the column player are given by [1, 1, 1, 1] and [1, 1, 1, 0], respectively, and they therefore have a Hamming similarity of $\frac{3}{4}$. The direct payoff of an agent, which we will call the *individual payoff* P, indicates an agent's individual fitness, regardless of others. But the agent's action also influences the payoff of the opponent, which carries three of its own genes. The total inclusive reward of the row player then becomes $P_{\text{row}} + \frac{3}{4}P_{\text{column}}$, with a symmetric inclusive reward for the column player. Therefore, from the perspective of the genotypes, the prisoner's dilemma of Fig. 1 effectively becomes a harmony game [49], where the only Nash Equilibrium is cooperation for both agents.

2.2.2 General prisoner's dilemma. In a general prisoner's dilemma, b is the benefit provided to the other by cooperating, and c is the cost for cooperation. The payoff matrix is given by:

	C	D
C	$b-c, b-c$	$-c, b$
D	$b, -c$	0, 0

From this general payoff matrix, we can derive two inequalities that need to be satisfied for cooperation under the inclusive reward: $c < hb$ and $hc < b$, where h is the Hamming similarity between the two players. We do not consider the second inequality, which is simply a consequence of the first, since $h \in [0, 1]$. This first inequality turns out to be equivalent to Hamilton's rule [16] from biology, which posits that an cooperative trait can persist if the benefit b, multiplied by the relatedness r, exceeds the cost c.

3 COOPERATION ON NETWORKS

Our evolutionary aligned reward function should incentivize an agent to maximize the fitness of its genetic material, which can be present in other agents as well. Therefore, we defined an inclusive reward (Eq. 2) which adds the individual rewards by weighing them with the Hamming similarity defined in Eq. 1. In this section, we study the properties of this inclusive reward by focusing on two settings where independent Q-learners [47] play two-player prisoner's dilemmas on networks. Self-interested agents often fail to cooperate in prisoner's dilemmas due to the dominance of the defective strategy over cooperation. In nature, however, many organisms have evolved stable cooperative strategies [16]. The goal of these experiments is to show the emergence and stability of cooperation in environments where agents try to maximize the fitness of their genetic material.

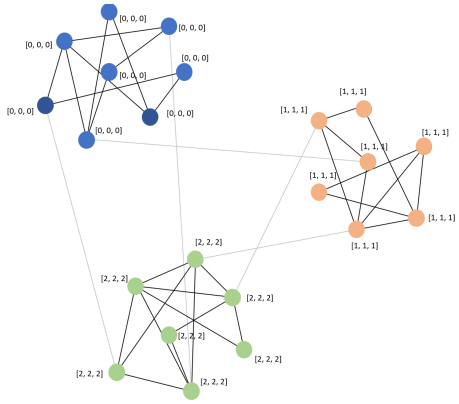

Figure 2: Example network with community structure. Here, the probability of a connection between agents inside a community is $p_{in} = 0.9$, while the probability for agents between communities is $p_{out} = 0.1$. All three communities represent a separate genotype. Figure adapted from [14].

3.1 Experiments

3.1.1 Opponent discrimination. A first experiment considers fully connected networks where agents can recognize each other (opponent discrimination). This means that an agent knows which opponent it is playing, but it does not know what genotype the opponent has, nor does it remember anything of what the agent did in the past; it only bases its action on a learned behavior for that opponent (Q-table). The setup is based on the evolution of sensing organs, which provide an organism the ability to observe the *phenotype* of other organisms in the environment, but not directly its genotype. Senses such as vision are crucial for animals, and over many generations led to the intuitive recognition of offspring, or the avoidance of predators [48], examples which we intend to capture with our setup. Opponent discrimination is implemented in our agents as a Q-table where every state corresponds to a different opponent on the network, and the agents receive each time step as an observation which opponent they are playing.

3.1.2 Limited dispersal. Our second experiment gives agents no opponent discrimination, which means agents have only one strategy for all interactions. Instead, we look at the effect of limited dispersal (also called population viscosity [16, 17]) on the emergence of cooperation between independent Q-learners under our inclusive reward. Under the limited dispersal hypothesis, it is assumed that organisms do not disperse far from their birth place, making them more likely to interact with genetic relatives.

To model limited dispersal, we move from fully connected networks to random partition networks [12] which have community structure [14]. Random partition networks are constructed starting from predefined groups of nodes that form (still unconnected) communities. Nodes that belong to the same community are connected with probability p_{in}, and nodes between communities with p_{out}. We define a *dispersal coefficient* $\eta \equiv p_{out}/p_{in} \in [0, 1]$, denoting the

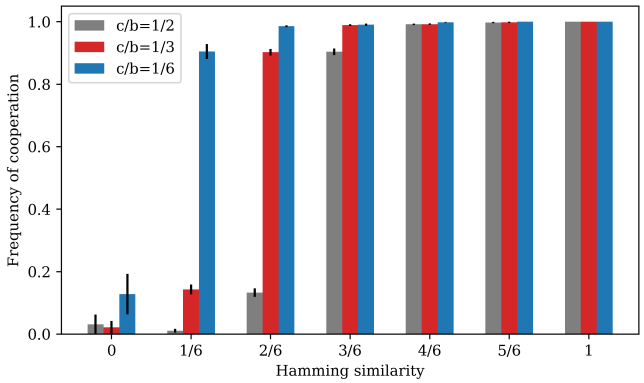

Figure 3: Frequency of cooperation in function of the Hamming similarity with the opponent on a fully connected network (genotype length 6, 2 variants per gene) with 64 agents, one per unique genotype. The cost-benefit payoff ratio c/b influences the Hamming similarity threshold at which agents start cooperating, matching Hamilton's rule [16].

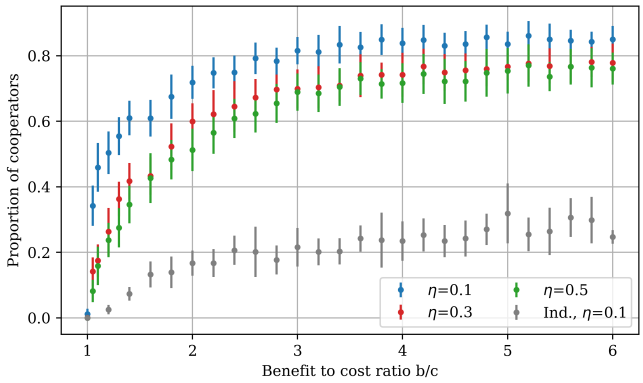

Figure 4: Proportion of cooperators in the (converged) population in function of the benefit to cost ratio b/c in a random partition network (genotype length 3, 2 variants per gene). Each community represents one of 8 unique genotypes, with 8 agents per genotype/community. η is the network dispersal coefficient. $\langle k \rangle = 9$ for all experiments, which together with η determines p_{in} and p_{out} (Eq. 4). Blue, red and green values show the proportion of cooperators under the inclusive reward for different dispersal coefficients. Gray values show results without inclusiveness. Cooperation increases under the inclusive reward and under small dispersal coefficients.

strength of the network dispersal. Every node in a community has the same genotype. This means that agents with similar genotypes are more likely to be connected than others (Fig. 2). The influence of network structure on the emergence of cooperation has been well-studied in evolutionary game theory [30, 32, 35, 41]. Here, we provide an alternative approach of modelling the strategies with reinforcement learning, where we study the resulting dynamics under an evolutionary aligned (inclusive) reward.

3.1.3 Reward. The payoff matrix of the prisoner's dilemma provides the individual fitnesses that each agent receives under their combined actions. We again use these individual fitnesses to construct our inclusive reward, according to Eq. 2. After every interaction, a player i uses its individual payoff P_i and the opponent's payoff P_j to determine its inclusive reward r_i^*:

$$r_i^* = P_i + h(\boldsymbol{g}_i, \boldsymbol{g}_j) P_j \,. \tag{3}$$

In both the opponent recognition and the limited dispersal experiment, our Q-learners try to optimize their myopic inclusive reward (meaning a bandit-like discount factor of 0), similar to how generational fitness is often defined in evolutionary game theory [13, 31, 32, 41]. Players pick and update their Q-values according to an ϵ-greedy scheme with exponentially decaying exploration.

3.2 Results

3.2.1 Opponent discrimination. We use a fully connected network for opponent discrimination, where each node represents an agent. We consider genotypes of length 6, with 2 gene variants per gene locus. We create one agent for every possible genotype, thereby making the network symmetric for all agents, for a total of $2^6 = 64$ combinations. All Q-tables are initialized to zero. Initial populations of all defectors, all cooperators, and mixtures were tried as well, but did not influence the results. Each time step, agents play one prisoner's dilemma against all of their opponents simultaneously, including itself, where the individual payoffs are defined by the benefit b and the cost c (c is fixed at 1, while we vary b). Figure 3 shows the resulting frequencies (averaged over the agents) that agents cooperate against opponents with respect to the Hamming similarity, for three values of c/b. The results match with Hamilton's rule [16], which as noted in section 2.2.2 predicts the spread of a cooperative strategy when $c/b < h$. Results without inclusive rewards (not shown in Figure 3) led to all defection.

3.2.2 Limited dispersal. We consider a random partition network, where agents have genotypes of length 3, with 2 variants per gene. A community of 8 nodes is created per unique possible genotype, i.e. 8 communities of 8 nodes. In figure 4, we consider three values of the dispersal coefficient η, and measure the proportion of cooperators that emerge in the network after convergence, with respect to the benefit to cost ratio b/c. Although we vary η, we keep the average degree fixed at $\langle k \rangle = 9$ to avoid that a variation in η leads to a variation in average degree, since it is known that varying the average network degree can drastically influence the spread of cooperation [30, 32, 35]. We keep $\langle k \rangle$ steady by deriving p_{in} and p_{out} through the following relation for $\langle k \rangle$, η and p_{in}:

$$\langle k \rangle = 9 = 7p_{in} + 56p_{out} = (7 + 56\,\eta)\, p_{in}. \tag{4}$$

The results from Figure 4 with inclusive reward show higher proportions of cooperation than with individual rewards, even though some level of cooperation can emerge without inclusiveness. Small dispersal coefficients and large benefit-to-cost ratios also lead to higher levels of cooperation, which matches with our prediction based on limited dispersal theory.

4 MARKOV GAMES

In future work, we intend to move beyond the limitations of networks and abstract matrix games, into temporally and spatially

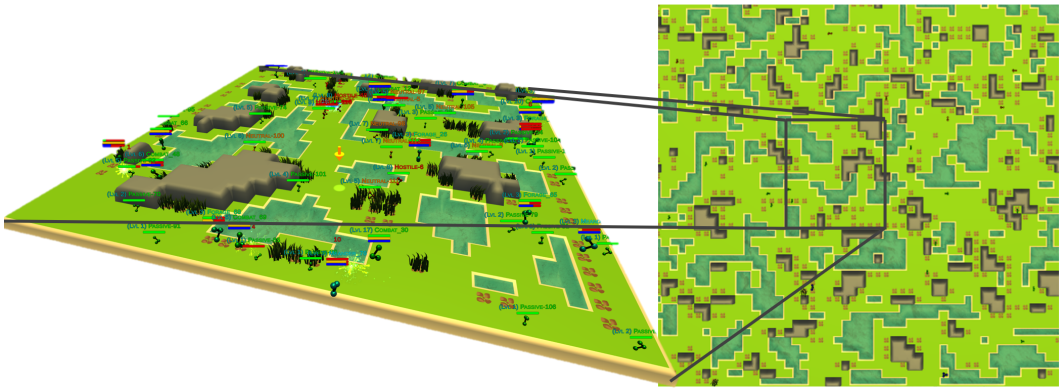

Figure 5: Impression of the Neural MMO environment [38]. Rocks, water, grass and forest cover the area. Agents, represented by three connected nodes, gather food and water, and engage in combat with each other or with environmental threats.

extended Markov games [25]. Our main focus will be on the Neural MMO environment [38] (Fig. 5). Neural MMO is a multi-agent video game environment in which agents survive by gathering resources such as water and food in a rugged environment. The agents can also engage in combat against other agents or with threats from the environment itself. Neural MMO is open-source, and serves as a customizable platform for multi-agent intelligence research, where agent configurations, reward functions, environment layout, resources, etc. can be rewritten and adjusted to suit the purposes of our research.

The goal of the previous network experiments was to examine the evolution of cooperation under a genetic inclusive reward function. Now, by moving to more open-ended environments like Neural MMO, we will test our hypothesis that our autocurriculum can lead to the emergence of increasingly complex and socially intelligent strategies. In Neural MMO, agents can move around in the environment, which leads to a dynamic 'network' structure. Where matrix games previously provided strategies as directly accessible atomic actions, the agents now need to learn to implement high-level strategies with policies executing a sequence of many low-level actions. Rewards cannot be matched with one clear, causal action anymore, resulting in the credit-assignment problem [39], and the concept of a general well-defined interaction between two agents ceases to exist, since an agent's actions can now influence events removed in distance as well as time. Moreover, the most rewarding and innovative strategies are not directly accessible anymore, but have to be discovered. Strategies of cooperation and defection can take many forms, and decisions do not always need to occur at the same moment: some information about what a player is starting to do can change another player's actions in the process.

4.1 Evolutionary aligned rewards

In the previous games, fitness was simply provided by the payoff matrix, and then used to create the inclusive reward. Here, computing the fitness of an agent is not as straightforward anymore, and we need to propose a mathematical expression of fitness to create our evolutionary aligned reward, in line with our definition of fitness as a measure of stability or longevity of an information state (section 2.1).

Longevity reward. To optimize for longevity, an agent receives a reward of +1 for every time step that at least one copy of its genotype is still alive. Since other agents can share its genes, the agent receives an additional inclusive reward of +1 for every related genotype alive, multiplied by the Hamming similarity. This 'longevity' reward r^L for agent i at time step t then becomes

$$r^L_{i,t} := \sum_{\boldsymbol{g}_j \in G_t} h(\boldsymbol{g}_i, \boldsymbol{g}_j), \tag{5}$$

where G_t is the set of *unique* genotypes alive at t.

This reward optimizes for the long term survival of an agent's genes, while leaving the strategy of how to accomplish this long-term survival open for the agents to discover.

Replication reward. In the field of biology, fitness is often defined as the expected number of offspring an organism produces [24, 26, 45]. Therefore, we also propose a reward function which directly promotes the maximization of the *amount* of shared genetic material that is present in the environment. Now, an agent i receives a reward of +1 for every newborn, and -1 for every agent that dies, which we again weigh by the Hamming similarity to include the potential presence of an agent's genes in other agents as well. This 'replication' reward r^R for agent i at time step t is then given by

$$r^R_{i,t} = \sum_{j \in J_t} h(\boldsymbol{g}_i, \boldsymbol{g}_j) - \sum_{j \in J_{t-1}} h(\boldsymbol{g}_i, \boldsymbol{g}_j), \tag{6}$$

where the first sum is over the group of agents J_t, alive at time step t, and the second sum over the agents from the previous time step $t-1$, where every agent is weighted by its Hamming similarity.

Combined reward. Given that the longevity reward – excluding the possibility of immortal organisms – will eventually lead to the discovery of replication as a crucial part of an agent's strategy, we can extend our longevity reward with the concept behind the replication reward, providing a reward signal that combines the best of both. We define this 'combined' reward r^C for agent i at time step t as

$$r^C_{i,t} = \sum_{j \in J_t} h(\boldsymbol{g}_i, \boldsymbol{g}_j), \tag{7}$$

where we sum over the group of agents J_t that are alive at time step t. The combined reward simply gives a positive reward (the Hamming similarity) when an agent which carries the same genetic material is alive for one more time step. The difference between the original longevity reward and the combined reward is subtle. In the former, we sum over the unique *genotypes* alive, for which all agents that carry a specific genotype only count for one positive reward, since only one copy is required for the information state to be alive. In the latter, however, we also take into account how many copies of those unique genotypes there are. Also, the replication reward is equal to the difference of the combined reward over two time steps.

When performing our experiments, we intend to try all three reward functions and study their properties.

4.2 Rules of the game

The rules of the game in our intended adaption of the Neural MMO environment are as follows. One health point is subtracted every time step from an agent's total health. To replenish health, an agent can consume water or food, which has to be gathered by standing near a pool or walking through a forest area. Besides foraging, agents can engage in combat, and attack each other when they are in striking distance for high damage, or with projectiles from afar, which deal less damage but are safer. The action space of an agent consists of movement actions, attack actions, and very importantly, an action that makes the agent reproduce. For the reproductive process, we propose to implement a simple system: when an agent decides to reproduce, it gives 1/4th of its health and resources to its offspring. More elaborate schemes are of course possible. Here, a strategy that reproduces fast will have more offspring, but produces agents that are in general weaker and are therefore more susceptible to getting killed. Either our agents will learn one optimal reproductive strategy, or niches will develop, where some agents focus on multiplication while others focus on individual strength (e.g., predator-prey dynamics).

The resources in Neural MMO are not endless. When resources are plentiful, each reward function will promote cooperation between any two agents that share at least one gene, and in principle induce indifference for agents that share none. However, when the population size grows towards the carrying capacity of the environment, resources become scarce, and a tension arises between helping closer relatives and more distant ones. Helping agents with a lower genetic similarity would mean the consumption of resources that could be used to help agents with higher similarity instead, so we expect to see a non-stationary spectrum of cooperation emerge.

4.3 Experimental setup

4.3.1 Training. To train our agents, we will move from tabular Q-learning to Deep Reinforcement Learning (Deep RL) with Proximal Policy Optimization (PPO) [36] and Long Short-Term Memory (LSTM) layers [19] to enable our agents to reach the necessary strategic complexity. We use a parameter-sharing scheme where all agents share the same neural network weights, but where the policy is conditioned on a unique genotype identifier, provided to the network as a dedicated part of the observation state. This is a common strategy in multi-agent RL [3, 4] which allows for specialization of strategies, where all the specializing strategies are condensed in one (large) neural network (i.e., a function approximation of different function approximators). The identifiers therefore allow the neural network to learn different policies for each agent.

Using one policy network for all agents leads to a large computational benefit, where the alternative would be to store and update separate policy networks for every agent, which can be extremely taxing for large-scale multi-agent simulations. Another benefit is that using the same network will act as a learning stabilizer by promoting a shared understanding of the environment dynamics, and by having more experiences to update the weights, thereby reducing the variance.

4.3.2 Evolving the world. We start the game with one agent, carrying a single genotype. Once this agent learns to reproduce, its offspring carries the same genotype. However, with a probability μ, each of its genes can mutate to another gene variant. This creates a new species, which also carries a new unique policy identifier and can therefore grow into a new strategy. So far, genes only influence the behavior of agents through the reward function, but in future experiments, we intend to try whether genes can also express properties of the agents themselves, such as in-game statistics like maximal health or combat strength. There are no predefined generations; each agent can reproduce at any time step, making the world and the population in it grow organically.

5 DISCUSSION

Our initial experiments on networks with opponent discrimination and limited dispersal match well-established biological principles, such as Hamilton's rule and limited dispersal theory. The results hint at the the potential of our inclusive reward function for the emergence of dynamic social structures not limited to only full cooperation or competition. We believe that the Neural MMO environment will allow us to present an empirical proof that our evolutionary aligned reward functions (section 4.1) can provide a continuous incentive for progress towards increasingly complex strategies within non-stationary social structures. We only expect to observe a *direction* towards increasingly complex strategies; no specific strategic properties are hypothesized, except the maximization of genetic fitness. Still, the Neural MMO setting will likely lead to behavior that is interpretable in an evolutionary context. Our reward is of course not limited to Neural MMO, and has the potential to be applied in many multi-agent reinforcement learning environments where inclusiveness and dynamic coalitions are important characteristics.

In conclusion, this paper laid the first foundations for a translation of evolutionary processes into multi-agent reinforcement learning, thereby providing a viable approach for creating generally capable and socially intelligent agents, starting from elegant and simple rules. We propose that in sufficiently rich environments, our evolutionary aligned reward has the potential to lead to a high strategic complexity, where agents will learn to balance cooperative and competitive incentives.

REFERENCES

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas,

et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113* (2019).

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[3] Raphaël Avalos, Mathieu Reymond, Ann Nowé, and Diederik M Roijers. 2022. Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning. *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems* (2022).

[4] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2020. Emergent Tool Use From Multi-Agent Autocurricula. In *International Conference on Learning Representations*.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[6] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[7] Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th annual meeting of the association for computational linguistics*. 286–293.

[8] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355* (2018).

[9] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by random network distillation. In *International Conference on Learning Representations*.

[10] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*. PMLR, 2048–2056.

[11] Wojciech Czarnecki, Siddhant Jayakumar, Max Jaderberg, Leonard Hasenclever, Yee Whye Teh, Nicolas Heess, Simon Osindero, and Razvan Pascanu. 2018. Mix & match agent curricula for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1087–1095.

[12] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

[13] Drew Fudenberg, Martin A Nowak, Christine Taylor, and Lorens A Imhof. 2006. Evolutionary game dynamics in finite populations with strong selection and weak mutation. *Theoretical population biology* 70, 3 (2006), 352–363.

[14] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.

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

[16] William D Hamilton. 1964. The genetical evolution of social behaviour. *Journal of theoretical biology* 7, 1 (1964), 17–52.

[17] William D Hamilton. 1972. Altruism and related phenomena, mainly in social insects. *Annual Review of Ecology and systematics* 3, 1 (1972), 193–232.

[18] Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell system technical journal* 29, 2 (1950), 147–160.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364, 6443 (2019), 859–865.

[21] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog* (2020).

[22] Steven Kuhn. 2008. Prisoner's dilemma. *Stanford Encyclopedia of Philosophy* (2008).

[23] Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. 2019. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742* (2019).

[24] Isadore Michael Lerner et al. 1958. The genetic basis of selection. *The genetic basis of selection.* (1958).

[25] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[26] Susan K Mills and John H Beatty. 1979. The propensity interpretation of fitness. *Philosophy of Science* 46, 2 (1979), 263–286.

[27] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. 2017. Learning to navigate in complex environments. *International Conference on Learning Representations* (2017).

[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[29] Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research* 21, 181 (2020), 1–50.

[30] Martin A Nowak and Robert M May. 1992. Evolutionary games and spatial chaos. *Nature* 359, 6398 (1992), 826–829.

[31] Hisashi Ohtsuki. 2010. Evolutionary games in Wright's island model: kin selection meets evolutionary game theory. *Evolution: International Journal of Organic Evolution* 64, 12 (2010), 3344–3353.

[32] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 7092 (2006), 502–505.

[33] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*. PMLR, 2778–2787.

[34] Kara Rogers. 2021. Inclusive fitness. In *Encyclopedia Britannica.* https://www.britannica.com/science/inclusive-fitness

[35] Francisco C Santos and Jorge M Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical review letters* 95, 9 (2005), 098104.

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[37] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[38] Joseph Suarez, Yilun Du, Clare Zhu, Igor Mordatch, and Phillip Isola. 2021. The Neural MMO Platform for Massively Multiagent Research. *Advances in Neural Information Processing Systems* 34 (2021).

[39] Richard Stuart Sutton. 1984. *Temporal credit assignment in reinforcement learning.* Ph.D. Dissertation. University of Massachusetts Amherst.

[40] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[41] György Szabó and Gabor Fath. 2007. Evolutionary games on graphs. *Physics reports* 446, 4-6 (2007), 97–216.

[42] Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10, 7 (2009).

[43] Gerald Tesauro et al. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[44] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[45] Conrad Hal Waddington. 1968. Towards a theoretical biology. *Nature* 218, 5141 (1968), 525–527.

[46] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. 2019. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv preprint arXiv:1901.01753* (2019).

[47] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

[48] George Christopher Williams. 2018. *Adaptation and natural selection.* Princeton university press.

[49] Daniel Zizzo. 2002. On the measurement of harmony in normal form games. (2002).