The emergence of division of labor through decentralized social sanctioning

Anil Yaman[1‡*], Joel Z. Leibo[2‡], Giovanni Iacca[3], Sang Wan Lee[4]

1 Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

2 DeepMind, London, UK

3 University of Trento, Trento, Italy

4 Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea

‡These authors also contributed equally to this work.

*Corresponding author (AY): a.yaman@vu.nl

Abstract

Human ecological success relies on our characteristic ability to flexibly self-organize in cooperative social groups. Successful groups employ substantial specialization and division of labor. Unlike most other animals, humans learn by trial and error during their lives what role to take on. However, when some critical roles are more attractive than others, and individuals are self-interested, then there is a social dilemma: each individual would prefer others take on the critical-but-unremunerative roles so they may remain free to take one that pays better. But disaster occurs if all act thusly and a critical role goes unfilled. In such situations learning an optimum role distribution may not be possible. Consequently, a fundamental question is: how can division of labor emerge in groups of self-interested lifetime-learning individuals? Here we show that by introducing a model of social norms, which we regard as patterns of decentralized social sanctioning, it becomes possible for groups of self-interested individuals to learn a productive division of labor involving all critical roles. Such social norms work by redistributing rewards within the population to disincentivize antisocial roles while incentivizing prosocial roles that do not intrinsically pay as well as others.

Significance statement

Division of labor allows individuals to specialize in different roles and working together to achieve greater welfare as a group than would be possible alone. However, such beneficial arrangements may be difficult to discover for groups of lifetime-learning individuals when these are driven only by self-interest. This is because some critical roles are less rewarding than other roles, so all individuals prefer someone else take them on. Here, we propose a social norm model, based on the idea that norms are patterns of social sanctioning. We show through simulation that such sanctioning can promote division of labor in groups of self-interested individuals. In our model, social sanctions are rewards and punishments that are imposed by one individual to another and may guide their lifetime-learning. We simulate how social norms may be learned through cultural evolution and show that they can help groups discover beneficial arrangements involving division of labor.

1 Introduction

Human societies depend on division of labor. However, an individual's role is not specified in his or her genes. Rather, human roles are learned during individual lifetimes. This is one of the reasons why human groups can achieve collective welfare more quickly than what would be possible with purely genetic evolution. However, many computational models of lifetime-learning formulate it as a process of maximization of individual payoffs [25, 24, 8, 40]. Such a formulation cannot, on its own, account for the learning of division of labor. When individuals are driven by self-interest they cannot learn to perform roles that do not pay as well as other roles, yet such roles are often necessary for the group to function and achieve a high overall welfare. Consequently, a fundamental question arises: how can division of labor emerge in groups of self-interested lifetime-learning individuals?

Here, we hypothesize that social norms, which we take to be patterns of social sanctioning, are sufficient to incentivize individuals in groups to select prosocial role choices, thereby enabling group-level division of labor to emerge from self-interested lifetime-learning. To test this hypothesis, we propose a model where lifetime-learning is shaped by social sanctioning. The specific social sanctioning mechanism presented here was inspired by studies that investigated the role of reward and punishment in laboratory-based social dilemmas such as the public goods game [13, 14, 33, 1, 6].

In these experiments, participants start with an initial endowment of tokens and can contribute to a public good by investing tokens in it. It is best for all if all individuals invest all their tokens, yet as individuals, all face an incentive to free-ride: holding back their own investment while still benefiting from the contributions of others. For instance, Fehr and Gächter [14] designed experiments consisting of two stages. First, participants chose how much to contribute to the public good. Then in a second stage, they received information about the choices of others, and then decided on that basis whether or not to pay a cost to punish others for their behavior. They found that participants were willing to engage in altruistic punishment: they would pay to punish free riders. Furthermore, the possibility of altruistic punishment in the case of repeated interactions led to an increase in the average cooperation level of the participants in the group.

Figure 1 illustrates the learning process used in our model. In **Stage I** (see Figure 1a), a group of self-interested lifetime-learners learn their roles. This process is modeled as a K-armed bandit problem where individuals learn the role (arm) that provides the highest payoff (optimum) from a finite set of roles. Each individual's lifetime-learning of their role is modeled by an ϵ-greedy algorithm: individuals select the role they judge to provide the maximum average payoff as estimated empirically by the payoffs they receive and explore other roles with a small probability ϵ [40]. **Stage II** (see Figure 1b) introduces the social sanctioning stage where individuals can monitor others, and impose sanctions based on their role choices. Subsequently, the payoffs received in Stage I are updated by the social sanctions imposed in Stage II.

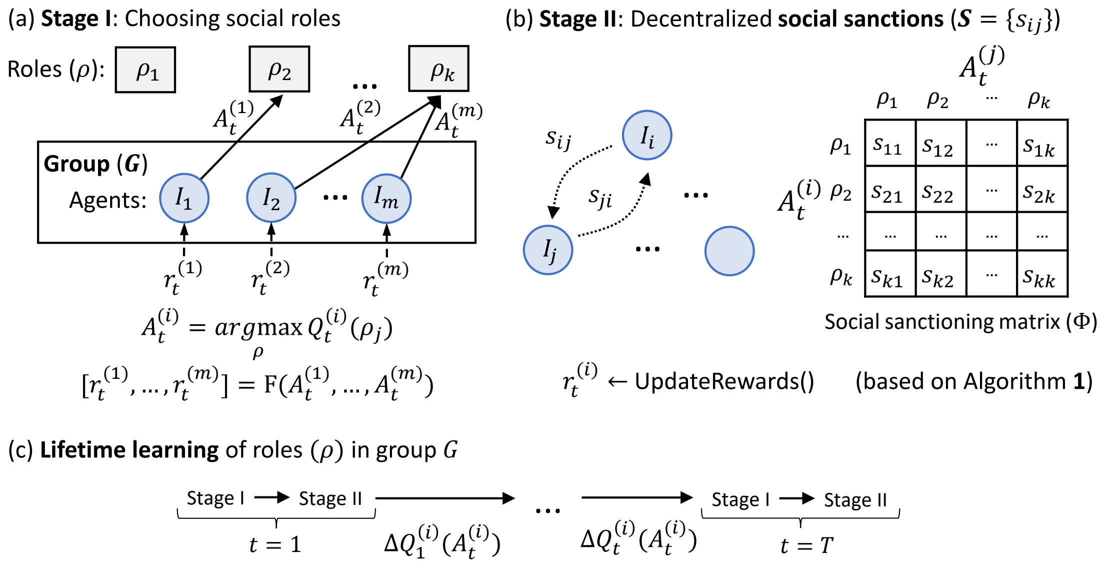


Figure 1: The decision process of the individuals consists of two stages: in (a) **Stage I**, individuals make a role selection based on the values of the roles ($Q_t^{(i)}(\rho)$) (see Equation (1)). The rewards received depend on a function (F) of the roles selected by all individuals in the group (see Equation (3)). In (b) **Stage II**, individuals impose decentralized social sanctions to shape the rewards of others to encourage the division of labor. Social sanctions can take forms of rewards and punishments (negative and positive values, or zero if no social sanction is imposed). An example social sanctioning matrix is shown here; each cell indicates the sanction imposed to individual j by individual i based on their roles corresponding to the columns and rows. The rewards are updated based on Algorithm 1. Such a pattern of decentralized social sanctioning constitutes a social norm ($\boldsymbol{S}$). The learning process consists of two levels: lifetime-learning and social norm evolution. (c) In lifetime-learning, groups of individuals perform Stage I and Stage II iteratively to learn their roles. Through social norm evolution, social sanctioning matrix values are optimized to produce optimum lifetime-learning of the social roles.

We model social norms as patterns of social sanctioning adopted by all individuals in a group. That is, every individual in the group applies the same sanctioning scheme. This is consistent with conceptions of what it is for a collective behavior pattern to be a social norm that rely on most agents in the group to conform [5]. This kind of sanctioning mechanism is said to be decentralized since it does not require a centralized enforcement mechanism like a government-run police force. Rather, the individuals in the group perform all the sanctioning actions themselves. However, their participation is not voluntary. They do not decide on their own whether to sanction or not. At each point in time, whether they sanction another individual, and if so, how much, is entirely determined by the group's social sanctioning matrix, i.e., the social norm itself. This may be justified by assuming the existence of a metanorm that demands individuals sanction in accord with their group's overall pattern, an assumption also shared by other models where norms are seen as public goods and metanorms are consequently necessary to evade the second-order free-rider problem [2, 19]. The existence of the requisite metanorm to stabilize patterns of decentralized social sanctioning is supported by laboratory experiments [50, 21, 43, 49] and ethnographic evidence [28, 12]. In our model, one implication of involuntary social sanctioning is that, given the right norm, it is possible to shape learning in any direction. This works for the same reason that reward shaping techniques are effective both in behaviorist psychology [34] and in algorithmic reinforcement learning [11]. Here groups can induce individuals by sanctioning to select any behavior. This is consistent with other models in the evolution of cooperation literature where sanctioning may stabilize behaviors regardless of whether or not they are adaptive [6].

As illustrated in Figure 1b, social sanctions are represented for our model in the form of a real-valued matrix with shape $k \times k$ (where k is the number of roles) which defines how much reward or punishment each individual with role i imposes on each individual of role j. Each specific such matrix is considered to correspond to a distinct social norm. Sanctioning in this model is role-wise. The amount of sanctioning applied by one individual to another at time t is a function of both of their roles at that time. There is a rich and cross-disciplinary tradition that centers around theories of social structure which features this kind of intimate connection between "role psychology" and normative behavior. Intuitively, norms are deeply entangled with social roles. For instance, it would be inappropriate (i.e. sanctionable) for a student to behave like a teacher or for a judge to behave like a legislator [39, 4, 10]. The errant student would likely face discipline from an academic administrator—another role, and one for which such sanctioning is part of its job description. Notice also that role-wise sanctioning—but not individual-targeted sanctioning—allows the implementation of *impartial* norms, like some human norms are [35], and does not merely produce the kind of self-serving unfairness-resentment-driven sanctioning motivations that are sometimes observed in non-human primates [7].

In our model, social norms are optimized through a cultural evolution process (given in Algorithm 2). This process is based on the principles of cultural group selection [48, 20, 36]. As such, it depends critically on all the assumptions necessary for the strength of group selection to outweigh that of individual selection e.g. sufficient separation between groups [20]. Note though that in the present work we do not explicitly model other groups beyond the focal group. Formally, our model considers only a single representative group where the effect of evaluating and evolving social norms can be computed iteratively and independently. Each social norm evolution experiment starts from a randomly initialized norm, obtained by sampling the values of the sanctioning matrix from a uniform distribution over a certain range. In each iteration, a new variant of the social norm is generated by adding Gaussian noise to the values in the previous norm's matrix representation. The newly generated norm is then evaluated, and its success measured by average group payoff achieved on a task demanding the learning of a division-of-labor arrangement. This process starts from an initial condition that does not incorporate any prior knowledge of the task at hand. If the norm achieves a higher average group payoff then it replaces the status quo.

Overall, we find that the social norms that emerge from evolution involve redistribution mechanisms where individuals periodically pay others to incentivize them to perform beneficial roles for the group that they would not otherwise select. Consequently, these mechanisms allow groups of self-interested individuals to discover effective division-of-labor arrangements through lifetime-learning. Specifically, we show that the proposed method of simulating social norm evolution leads to higher collective return than other approaches based on (a) directly evolving roles without lifetime-learning and (b) a model based on lifetime-learning with an additional assumption of altruistic social preferences.

2 Results

2.1 Lifetime-learning of individuals' roles

In our model, groups consist of individuals. Individuals select and re-select roles for themselves throughout their lifetime. Individuals learn from this experience which of their role choices are the most rewarding and usually select the role they expect will provide the most reward. That is, individuals face a K-armed bandit problem $\rho = \{\rho_1, \rho_2, \ldots, \rho_k\}$ where K is the number of roles. They select a new role on each step, so it is best to think of the simulation steps as corresponding to some substantial period of time like a week or a month.

The lifetime-learning process is illustrated in Figure 1a. Individuals change their behavior over time via value-based reinforcement learning [40]. On each step t, each individual i selects its role $A_t^{(i)}$ using its estimated value $Q_t^{(i)}(\rho_j)$ with

$$A_t^{(i)} = \arg\max_{\rho} Q_t^{(i)}(\rho_j) \tag{1}$$

and updates value estimates using the reward received on the previous step, as shown below:

$$Q_{t+1}^{(i)}(A_t^{(i)}) = Q_t^{(i)}(A_t^{(i)}) + \alpha \left[r_t^{(i)} - Q_t^{(i)}(A_t^{(i)}) \right] \tag{2}$$

where $0 < \alpha \leq 1$ is the learning rate parameter and $r_t^{(i)}$ is the reward received after selecting $A_t^{(i)}$. Individuals aim to maximize their rewards and thus usually select the role they estimate to have the highest value, occasionally also exploring other roles with a small probability ϵ. We measure an individual's performance after T iterations.

In this model, individuals learn to select roles in order to maximize their personal reward. However, this could lead them to select roles containing selfish behaviors that gain personal reward at the expense of the wider group. Individualistic reinforcement learning has the effect of discouraging agents from "taking one for the team", resulting in lower joint performance in environments that require such cooperation.

2.2 Incentivizing the lifetime-learning of division of labor via social norms

We consider a reward function F in the form shown in Equation (3) where the reward $r_t^{(i)}$ received by individuals i at time step t depends on the roles selected by all individuals in the group, not just its own choice. Individuals' rewards may be interdependent. In situations of interdependence, self-interested optimization of personal reward often does not converge to a socially optimum role distribution since some roles that are critical for optimal group welfare are not as individually rewarding as other roles, so no individuals learn to select them.

$$\left[r_t^{(1)}, \ldots, r_t^{(m)}\right] = F(A_t^{(1)}, \ldots, A_t^{(m)}) \tag{3}$$

Social norms in our model are regarded as decentralized patterns of social sanctioning. In our model, sanctions are rewards and punishments imposed by one individual on another individual. Amounts of sanctioning are a function of the roles of the sanctioning and the sanctioned player. The social sanctioning stage (Stage II as shown Figure 1b) is applied every time step after Stage I (shown in Figure 1a). Here, the individuals in the group monitor other individuals (based on a certain neighborhood function that defines the connectivity of their social network) and impose sanctions. The amount of the sanction provided by an individual taking role k to an individual taking role ℓ is $s_{k,\ell}$ (see Figure 1b). The rewards received after the role selection (Stage I) are then updated based on the social sanctioning scheme proposed in Stage II. As shown in Figure 1c, Stage I and II are performed consecutively for T iterations.

2.3 Evolution of social norms for incentivizing division of labor

We formalize the process of social norm evolution with an optimization algorithm that models cultural evolution from the cultural group selection point of view (see Algorithm 2). The goal of the algorithm is to find a particular social sanctioning matrix $\boldsymbol{S^*}$ that can maximize R as: $\boldsymbol{S^*} = \arg\max_{\boldsymbol{S}} R$, where R is the average group payoff received by the individuals from performing their roles.

Initially, the group is assigned a social sanctioning matrix where the amount of sanctions are randomly sampled within a certain range (i.e., uniformly in $[-5, 5]$). It is assumed that all individuals in the group use the same social sanctioning matrix. Then, the group can sample new social norms by applying the following exploration operator: Gaussian perturbation for the real-valued part and bit-flip for the binary part, see Methods section for details. If the perturbed social sanctioning matrix $\boldsymbol{S'}$ provides better performance (as measured by R), it is selected by the group. It then replaces the social sanctioning matrix currently in use. After a certain number of iterations of this process we expect to find social sanctioning matrices that are better at incentivizing division of labor and achieve higher R as a result.

The *fitness* of social sanctioning matrices R are found as follows:

$$R = \text{EVAL}(\boldsymbol{S}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t=\zeta}^{T} r_t^{(m)} \tag{4}$$

where EVAL() is the evaluation function of social sanction $\boldsymbol{S}$, M is the number of individuals in the group, and ζ is cutoff for computing the rewards. The problems the groups tackle require the individuals to learn various roles so that R can be maximized. The evaluation function is concerned with the lifetime-learning process of individuals and repeated for T steps as illustrated in Figure 1. The lifetime-learning process is stochastic since it depends on the role choices (i.e., random exploration of roles and the order of sanctions). Therefore, to rule out randomness, the evaluation is performed multiple times and the average is accepted as the final result.

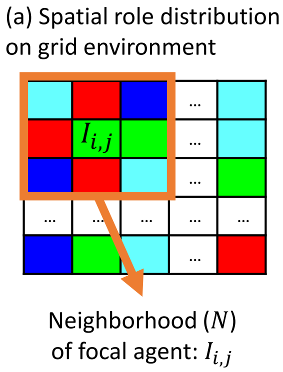
2.4 Spatial games

We investigate the evolution of cooperative norms on two spatially structured games called *settlement maintenance* and *common pasture*. In both games, each individual occupies a position in space and directly interacts only with their neighbors (see Figure 2a). Each individual selects a role to perform, and receives a payoff as a result. Some of the roles do not provide individuals who select them with any reward at all, yet it still may be necessary that some agent perform them for the group as a whole to be successful.

We indicate the locations (cells) that individuals are assigned by (i, j) coordinates. They are constrained to interact with their closest eight neighbors (i.e., their Moore neighborhood). Furthermore, the environment is toroidal so that it allows the individuals located at the first and last columns/rows to interact with the individuals located at the last and first columns/rows.

The settlement maintenance game was inspired by research on ant colonies but interpreted here as a model for small-scale human societies [16, 26]. Payoffs associated to each role are shown in Figure 2b. In this game, individuals can obtain payoffs by perform foraging and hunting roles in form of food. Hunting provides a higher payoff but requires coordination with a sufficient number of other hunters. In addition, there are cleaner and soldier roles that do not directly provide any payoffs but are still required for maintaining and protecting the settlement respectively. In each time step, there is a constant rate of waste accumulation in each cell of the environment which negatively affects the rewards received by the individuals. However, if an individual performs the cleaner role in the neighborhood then the adverse effect of waste accumulation mitigated. Furthermore, there is an constant adversarial attack probability for each cell that causes the reward of an individual to be stolen (reduced to 0). However, when individuals take on the soldier role they then provide protection against adversarial attack within their neighborhood.

The payoffs of the roles in the common pasture game are provided in Figure 2c. This game models situations of resource appropriation where a bad outcome known as the tragedy of the commons



(b) Settlement maintenance

Roles (ρ)	$r_t^{(i,j)}$
Cleaner (C)	0
Forager (F)	3
Hunter (H)	6, if 5 or more hunters, 0, otherwise.
Soldier (S)	0

Role functions
- C: mitigates waste accumulation
- S: protects against adversarial attacks

(c) Common pasture

Roles (ρ)	$r_t^{(i,j)}$
Worker (W)	0
Herder (C - considerate)	5
Herder (G - greedy)	10

Role functions
- W: increases speed of resource recovery

Figure 2: (a) Types of roles, color coded for visualization, can be performed by the individuals that are spatially distributed on a 2-dimensional toroidal grid environment and can impose social sanctions to the individuals within their neighborhood. The goal is to learn a role distribution (a role in each location) that can maximize the average group payoff. The payoffs of the roles are shown in Tables (b) and (c) in settlement maintenance and common pasture games respectively. The settlement maintenance has four and common pasture has three role choices.

may occur [18, 32]. Here, the tragedy is when resources in the environment become depleted as a result of individuals pursuing their own selfish interests without regard to mounting social cost. To avoid the tragedy of the commons, a critical fraction of individuals must learn to act sustainably, restraining their use of resources. To simulate this behavior, we model a pasture environment with a certain amount of starting resources in each location (e.g. grass for the herders' herd). These resources can be appropriated by the herders either greedily (G) or considerately (C). The resources regenerate over time unless fully depleted. In addition, we introduce another role: worker (W) which can increase the regeneration speed of the resources (e.g. speed of grass growth). However, if the resource on a cell is fully depleted, it cannot be recovered within the present lifetime. Therefore, the individuals cannot receive any reward when the resources are depleted. The environment resets in terms of its resources and recovery ability between generations.

This model of social sanctioning is broadly compatible with other redistributive (zero-sum) sanctioning mechanisms in the reinforcement learning literature [27, 47]. It differs from the purely destructive negative-sum social sanctioning schemes considered by [45, 23]. In the present work, the payoffs received in Stage I stand in for physical objects such as resources like the food received as a result of foraging behavior. This kind of sanctioning is called zero-sum because it respects a conservation law: resources are neither created nor destroyed by sanctioning. Consequently, the function of sanctioning can be considered as a decentralized resource redistribution scheme. Losing resources is punishing while gaining resources is rewarding. One may "gift" resources to another, reducing your own reward to increase theirs (positively sanctioning them). Symmetrically, one may take resources away from another agent, gaining them yourself (negatively sanctioning them). Importantly, once reward has been reduced to zero it cannot be reduced any further since one cannot take away resources that do not exist. One implication is that individuals can only positively sanction others when they have enough resources to do so. The sanctioning process is shown in detail in Algorithm 1.

2.4.1 Settlement maintenance

In settlement maintenance game, individuals can select one of the roles in each time step t as: $A_t^{(i)} \in \{cleaner, forager, hunter, soldier\}$. We assume there is a level of waste accumulation $d_{t+1}^{(i,j)} = \min(wa, d_t^{(i,j)} + \tau)$ in each cell location (i, j) with a constant rate of τ, reaching a maximum level of ma. The ideal living conditions of the individuals are affected by the waste accumulation. Thus, their payoffs are updated based on the waste accumulation as follows:

$$r_t^{(i,j)} = \max(0, r_t'^{(i,j)} - d_t^{(i,j)}). \tag{5}$$

where $r_t'^{(i,j)}$ is the payoff received by the individuals. On the other hand, the negative effect of waste accumulation is mitigated by the cleaner individuals by setting the rate of waste accumulation to 0 within the neighborhood of the cleaner individuals.

In addition, we assume that the environment is subject to random adversarial attacks that reduce the payoffs received by the affected individuals. The frequency of these attacks is controlled by an independent attack probability $P_{i,j}$ in all cell locations. If there is an attack on a cell location, the payoff of the individual is reduced to 0. On the other hand, if there is a soldier in the neighborhood, the attacks are defended and payoffs are protected.

2.4.2 Common pasture

In common pasture, individuals can select one of the roles as : $A_t^{(i)} \in \{considerate, greedy, worker\}$. The amount of resources in each cell represented as $D^{(i,j)}$. In the beginning of the game, certain amount of resources allocated in each cell. Herders, represented as roles as: considerate or greedy, consume the resources in their neighborhood either considerately or greedily as shown in Figure 2c (payoffs they receive are the resources consumed in the environment). They can use available resource in their neighboring cells selected randomly.

The resources replenish with a constant natural growth rate of c. In addition, the growth rate can be increased by w if a cell is occupied by a worker (otherwise $w = 0$). Consequently, the growth of the resource in each cell is modeled by the logistic growth model as:

$$\frac{\partial D^{(i,j)}}{\partial t} = (c + w)(1 - \frac{D^{(i,j)}}{K})D^{(i,j)} \tag{6}$$

where K is the upper limit (carrying capacity) of the resources. Note that, when the resource in a cell is depleted $D^{(i,j)} = 0$, it cannot be recovered.

2.5 Social norms enable lifetime-learning of division of labor

In both games, two environmental parameters are used for testing the emergence of learning proper role distributions in five different environment variations. In the settlement maintenance game (see Figure 3a), these parameters are based on waste accumulation and adversarial attack probability, and initialized as: $(wa, P) = \{(0,0),(6,0),(0,1),(3.2,0.5),(6,1)\}$. In the common pasture game, these parameters are natural growth rate and worker rate, and are initialized as $(w, c) = \{(0,0),(0.5,0),(0,0.5),(0.27,0.27),(0.5,0.5)\}$. For the two environmental parameters in both of games, we chose max value ranges and initialized five parameter assignment cases, four on the limits, and one in the middle of the parameter ranges. We ran the evolutionary processes independently and separately for each of these environment instances multiple times (i.e., 30 runs each) to find social norms that could provide optimum leaning in each one. We observed that multiple runs of the evolutionary processes can produce distinct social norms, however, they can still converge on a similar spatial role distributions that can provide (near-)optimum group welfare. Examples of these social norms and the spatial role distributions they converged are provided in Supplementary Material (see Figures 4 and 5) in both games.

In Figure 3a, examples of learned spatial role distributions in five environment variants in both games are shown. The individuals converged on different role distributions depending on the environmental conditions. Interestingly, the emerged role distributions show clear signs of neighborhood patterns where the roles that benefit their neighbors, distributed in various neighborhoods for minimizing their overlap.

In the settlement maintenance game, since there is no need for cleaner and soldier roles when there is no waste accumulation and adversarial attacks ($wa = 0, P = 0$), all individuals converged to hunter role to achieve the maximum payoff. When waste accumulation or adversarial attack probabilities is increased, some individuals converge to cleaner or soldier roles respectively to mitigate the adverse effects of these environments. When both waste accumulation and adversarial attack probability are increased, we see the emergence of both cleaner and soldier roles.

In the common pasture game, when the natural growth and worker rates are low, there is no possibility of sustaining the resources in the environment. Therefore, we see random role selection in this case (when $w = 0$ and $c = 0$). When one or both of these factors increase, various role distributions can be learned to collect as much resources as possible while sustaining the environment. For instance, when the natural growth rate c is high, the number of (considerate) herders increases. Interestingly, an increase in the number of worker roles can also be observed even though they do not have any function when $w = 0$. This is due to the fact that workers can function as empty placeholder cells that prevents additional farmers in the environment, and therefore avoid excess use of resources. When the natural growth rate is maximum, we can observe the appearance of some greedy herders.

In both games, self-interested lifetime-learning individuals without social sanctions, referred to as "selfish" from now on, do not obtain good performance. In the settlement game, all the individuals

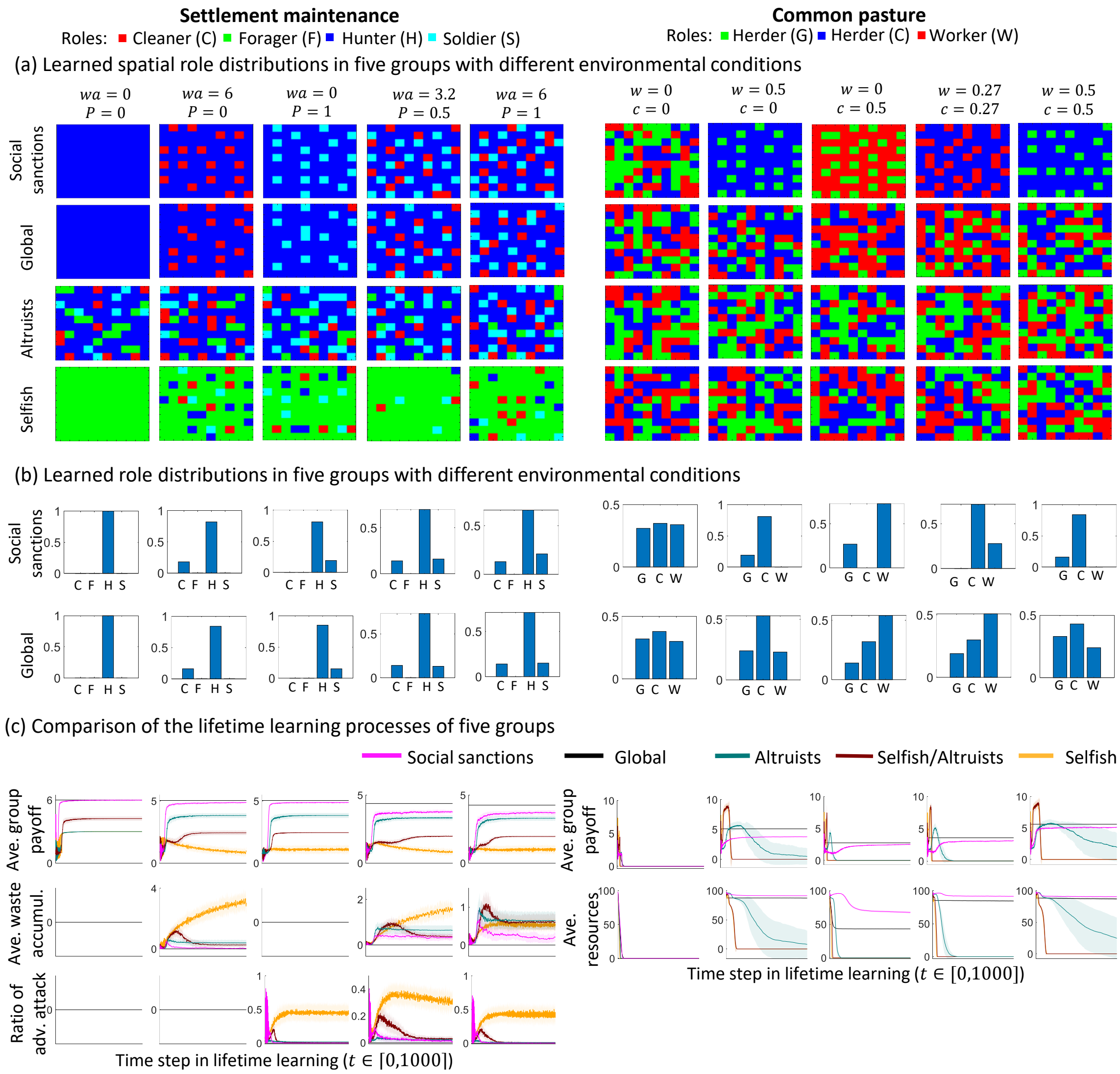


Figure 3: Social sanctions facilitate learning role distributions in self-interested lifetime-learning individuals various environment instances in settlement maintenance and common pasture games. The results shown in each row corresponds to a distinct environment parameter settings specified on top of each column, wa (waste accumulation) and P (adversarial attack probability) in settlement maintenance, and w (worker growth rate) and c (natural growth rate) in common pasture game. Shown in (a) and (b), different role distributions are learned locally (enforced by the decentralized social sanctions) to maximize group payoffs. (c) shows the change in the average group payoff and other environmental properties during the learning processes with social sanctions, and other three compared approaches.

converge to select the forager role since it is easier to learn because it doesn't depend on coordination with others. The hunter role, which can provide higher rewards, depends on coordination within the neighborhood, so it is more difficult to discover by chance since at least five individuals would have to select it simultaneously. In the common pasture game, selfish individuals learn the greedy herder role because it has the highest immediate payoff. However, when they do, they deplete all the resources of the environment quite quickly. Both games have critical roles that do not provide any payoff on their own: cleaner and soldier in settlement maintenance and worker in common pasture. Selfish individuals will not select these roles since they do not provide any payoff. Thus, groups of selfish individuals cannot establish a socially advantageous division of labor.

The learning performance of self-interested lifetime-learning individuals with social sanctions, referred as "social sanctions" from now on, and selfish individuals were compared with three additional approaches. These were: global, altruists and selfish/altruists. Briefly, in the global approach, the role distributions of the groups are directly optimized by evolutionary algorithms where the role of each individual in each location is defined when they are initialized, and remains fixed throughout their lifetime. In selfish/altruists, the individuals aim to maximize the average payoffs of self and neighboring individuals. In case of altruists, they aim to maximize the average payoffs of neighboring individuals, not including themselves. The details of these approaches are provided in Methods section.

Figure 3c compares all five approaches to one another. In both games, the global approach provides the upper bound. This is as expected since it can take into account global knowledge of the problem which allows making improvements on some roles independently of others in different locations by keeping them constant. In addition, this approach does not involve the costs of lifetime-learning that arise from trial and error. Results from the social sanctions approach were the closest to this upper bound. The selfish individuals learn to perform the role that maximizes their payoff. However, without the other regulatory roles in the groups, the environment degrades quickly leading to the worst performance. Altruists and selfish/altruists improve over selfish individuals but do not perform as well as social sanctions.

3 Discussion

Human ecological success has been underpinned by our species' ability to overcome the challenges of group collaboration. In these situations, the individuals who compose a group often need to divide labor and perform various specialized roles collaboratively. However, in groups consisting of self-interested individuals that aim to maximize their own payoff, establishing such collaboration may not be straightforward. In this work, we focused on the problem of learning such a division of labor in groups of self-interested lifetime-learning individuals. We modeled social norms as decentralized social sanctioning patterns and studied how they can encourage individuals to cooperate. We also proposed a cultural group evolution model to study how such norms that enable the learning of prosocial roles can be established. We demonstrated this learning problem, and how our model of social norm evolution resolves it, for two different spatial games. In both cases the emergent social norms discovered by our cultural group evolution model were successful in incentivizing individuals to learn to collaborate with one another.

There is an extensive literature on the evolution of cooperative strategies in populations of self-interested individuals [3, 2, 29, 37]. Some, like us, studied the emergence of cooperation in spatial game settings [31, 30, 41]. However, this line of work has not modeled lifetime-learning. We modeled individual-level lifetime-learning in conjunction with group-level norm evolution. In our model, the two levels were linked because group-level norms guide the lifetime-learning of individuals. At the same time, whether a newly innovated norm undergoing testing makes the cut and becomes established depends on its guiding the learning of all individuals to a more socially advantageous outcome. Wang et al. (2019) took an analogous two-level strategy, integrating individual reinforcement learning with an evolutionary process, which in their case, determined the individuals' reward functions [46]. They showed that group selection was needed to get reward functions that incentivize altruistic behavior to evolve, a result that supports a key assumption of our model: that social norms evolve on the level of groups.

The game-theoretic literature on social norms has pursued two broad categories of mechanisms by which norms may affect cooperation. They are (A) transforming the payoffs (via sanctioning or equivalent) into a new effective game with different Nash equilibria, e.g. making mutual cooperation an equilibrium in a transformed game derived from the Prisoners' Dilemma [42, 22], and (B) equilibrium selection. When social norms are regarded as equilibrium selection devices, the assumption is that many Nash equilibria are possible and would be stable if achieved, and that some of which are quite bad (e.g. tragedy of the commons situations). Society must navigate its way to the good equilibria while avoiding the bad ones. In this line of work, the norm is regarded as a piece of public knowledge that individuals may condition their behavior on in order to rationally coordinate with one another [44, 15, 17]. For instance, a focal point is a "default choice" from which it would be irrational to deviate when you know others are trying to coordinate [38]. Many models incorporate aspects of both mechanisms (e.g. [4]). Our approach resonates more strongly with approach (A). The two games we studied, settlement maintenance and commons pasture, are social dilemmas. That is, they contain socially deficient Nash equilibria, and purely egoistic, utility-maximizing agents do not learn to cooperate; they never take on the unpaid-but-critical roles. In our model, social norms are patterns of social sanctioning that affect the rewards individuals experience. We can see them as transforming the game into a different one where it is profitable to select the unpaid-but-critical roles because other individuals positively sanction that choice, effectively paying the agents who choose them for the service they provide to the group. It is thus critical in our model that the norms themselves do not evolve within individuals, and individuals cannot unilaterally decide whether to sanction or not. Norms must evolve on the level of the cultural group, using a "fitness" that incorporates the well-being of all the group's individuals, otherwise they could not incentivize individuals to choose dominated strategies, as they must to achieve a stable and socially advantageous division of labor [20].

4 Methods

4.1 Social norm representation and evolution

Social norms are represented by the social sanctioning matrix Φ that is a Hadamard product of two $k \times k$ matrices as $\Phi = \Phi_1 \odot \Phi_2$. The first matrix Φ_1 encodes the real-valued social sanctioning values within a specified range (i.e., $[-5, 5]$). The second $k \times k$ matrix Φ_2 encodes binary values and has a one-to-one correspondence with the cells encoded in the first matrix. The binary matrix is used to indicate whether a social sanction value in the first matrix is used or not (1: used, 0: not used). As a result of the Hadamard product, the social sanctioning value can be zero when the cells of Φ_2 encode zero.

The rationale behind this representation is to aid the search process to find the cases where social sanction is not needed. Otherwise, converging to the value of 0 using only real-valued representation may not be possible. The use of the second binary matrix allows easy switch between the use/non-use of the social sanctioning values.

For the cultural evolutionary process, two matrices are concatenated and stored as $1 \times 2k^2$ vectors to simplify the representation. New norms are explored by perturbing the values of these vectors. However, since different parts use different representations (i.e., real-valued and binary), we use different perturbation operators for different representations: for real-values, random Gaussian noise (i.e., $x\prime = x + \mathcal{N}(0, 0.1)$) is added; for binary values, the not operator is used with a small probability (i.e., the value of a cell is set to $x\prime = \neg x$ with a probability of 0.05).

4.2 Social sanctioning procedure

The update procedure of the rewards based on decentralized social sanctioning mechanism is shown in Algorithm 1. Here, for each neighbor of each individual (in random order), the social sanction (defined in the social sanctioning matrix) is imposed based on the role selection of the focal individual and its neighbor. We used also a version where the social sanctions are imposed based on a specified order of the neighbors (such as starting always from the neighbors on the upper right to the neighbor on the lower left). However, the random order is preferred to avoid the dependency to this ordering during the sanctioning process.

4.3 Social norm evolution

Algorithm 2 provides the evolutionary process of the social norms. The algorithm starts with a randomly initialized norm (by randomly initializing the cells of the social sanctioning matrix) and aims to find the norm that provides the highest average group payoff. This is achieved by making a small change in the norm (via Gaussian perturbations of the cell values of the social sanctioning matrix) and testing whether the new norm provides better performance (learning of better role distributions that can maximize the group payoff). If so, the previous norm is replaced by the new,

Algorithm 1 The update procedure of the individual rewards based on the decentralized social sanctioning.

procedure UpdateRewards
    for each $i \in I$ **do** ▷ In random order
        $neighbors$ =getNeighbors(i)
        for each $j \in neighbors$ **do** ▷ In random order
            s_{ij} =sanction($role(i,t), role(j,t)$)
            if $s_{ij} > 0$ and $s_{ij} \geq r_t^{(i)}$ **then** ▷ Individual i is rewarding j
                $r_t^{(i)} = r^{(i)} - s_{ij}$
                $r_t^{(j)} = r^{(j)} + s_{ij}$
            else if $s_{ij} < 0$ and $|s_{ij}| \geq r_t^{(j)}$ **then** ▷ Individual i is punishing j
                $r_t^{(i)} = r_t^{(i)} + |s_{ij}|$
                $r_t^{(j)} = r_t^{(j)} - |s_{ij}|$
            end if
        end for
    end for
end procedure

better performing norm. This process is repeated until a maximum number of iteration is reached. The social norm at the final iteration is then provided as the result of the algorithm.

Algorithm 2 Pseudocode for the evolution of social norms.

procedure NormEvolution
    $\boldsymbol{S} \leftarrow$ randomSocialNorm()
    $R \leftarrow$ EVAL($\boldsymbol{S}$)
    $i = 1$ ▷ counter for the number of iterations
    while $i < maxIter$ **do** ▷ until the max iterations
        $\boldsymbol{S}' \leftarrow \boldsymbol{S} + \mathcal{N}(0, \sigma)$
        $R' \leftarrow$ EVAL($\boldsymbol{S}'$)
        if $R' > R$ **then**
            $\boldsymbol{S} \leftarrow \boldsymbol{S}'$
            $R \leftarrow R'$
        end if
        $i \leftarrow i + 1$
    end while
end procedure

4.4 Other compared learning approaches

4.4.1 Global

For the global approach, instead of optimizing the social norms, we optimize the spatial role distributions directly using standard genetic algorithms [9]. The pseudocode for the global approach is provided in Algorithm 3.

For both games, we aim to find a 10 by 10 matrix where each cell can take one of the roles defined for each game. In this case, since there are 4 and 3 possible role selections in each cell, there are 4^{100} and 3^{100} possible spatial role distributions in the settlement maintenance and common pasture games, respectively.

Algorithm 3 Pseudocode for optimizing the spatial role distributions directly using the global approach.

procedure Global
  $\boldsymbol{X} \leftarrow$ randomSetRoleDistributions()
  $F \leftarrow$ Evaluate($\boldsymbol{X}$)
  $g = 1$ ▷ counter for the number of iterations
  while $g < maxGen$ **do** ▷ until the max number of generations
    $\boldsymbol{X}' \leftarrow$ Crossover($\boldsymbol{X}$)
    $\boldsymbol{X}'' \leftarrow$ Mutate($\boldsymbol{X}'$)
    $F' \leftarrow$ Evaluate($\boldsymbol{X}''$)
    $[\boldsymbol{X}, \boldsymbol{F}] \leftarrow$ Select($\boldsymbol{X} \cup \boldsymbol{X}'', F \cup F'$)
    $g \leftarrow g + 1$
  end while
end procedure

Using genetic algorithms, we represent a set of randomly initialized role distributions $\boldsymbol{X}$ where each role distribution in the set represents a candidate solution. We encode each candidate solution as a vector form (1 by 100), flattening the corresponding 10 by 10 spatial role distributions for computation convenience. The goal is to search this solution space to find the solution (i.e., the role distribution) that maximizes the average group payoff. We do that by applying evolutionary inspired operators (*crossover* and *mutation*) to produce an offspring candidate solution set $\boldsymbol{X}''$ using the current set of candidate solutions. Then, we create a new set of solutions by selecting the best solutions from the union of the current $\boldsymbol{X}$ and offspring sets $\boldsymbol{X}''$. This process is performed iteratively until a certain number of generations. At the end, the algorithm provides the best solution that is the spatial role distribution that provides the maximum average group payoff encountered during the search process. The evaluation process of each candidate solution does not involve lifetime-learning since the roles of the individuals in each cell are directly specified. To evaluate the solution, we run the game process for a certain number of action steps, record the rewards achieved by each individual, and then find the average reward of the group.

4.4.2 Selfish

Selfish individuals are modelled based on a reinforcement learning approach, known as the ϵ-greedy algorithm [40], that aims to maximize their own payoffs based on the Equations (1) and (2) [40]. This is the basic form depicted as the **Stage I** in Figure 1 without applying the social sanction part shown as the **Stage II**.

4.4.3 Selfish/altruists and altruists

Selfish/altruists individuals aim to maximize the average of the payoffs of their own and the average payoffs of the other individuals in the neighborhood, and altruist individuals aim to maximize the average payoffs of the individuals in their neighborhood. To model these behaviors, we modified the equation used by selfish individuals (given in Equation (2)) as follows:

$$Q_{t+1}^{(i)}(A_t^{(i)}) = Q_t^{(i)}(A_t^{(i)}) + \alpha \left[r - Q_t^{(i)}(A_t^{(i)}) \right] \tag{7}$$

$$r = \Omega \frac{1}{N} \sum r_t^{(N)} + (1 - \Omega) r_t^{(i)} \tag{8}$$

where $\Omega \in \{0.5, 1\}$ is a parameter that specifies the weight given to maximizing the average payoffs of the individuals that are in the neighborhood (denoted as $\frac{1}{N} \sum r_t^{(N)}$, where N is the number of individuals within the neighborhood), rather than an individual's own payoff. For selfish/altruist behavior, we set $\Omega = 0.5$, while for altruist behavior $\Omega = 1$. Note that, when $\Omega = 0$, we achieve selfish behavior.

References

[1] Felix Albrecht and Sebastian Kube. Peer-punishment in a cooperation and a coordination game. *Games*, 9(3):54, 2018.

[2] Robert Axelrod. An evolutionary approach to norms. *The American political science review*, pages 1095–1111, 1986.

[3] Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[4] Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2006.

[5] Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. Social Norms. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edition, 2018.

[6] Robert Boyd and Peter J Richerson. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology*, 13(3):171–195, 1992.

[7] Sarah F Brosnan and Frans De Waal. Monkeys reject unequal pay. *Nature*, 425(6955):297–299, 2003.

[8] Peter Dayan and Yael Niv. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2):185–196, 2008.

[9] Agoston E Eiben, James E Smith, et al. *Introduction to evolutionary computing*, volume 53. Springer, 2003.

[10] Gen Eickers. Coordinating behaviors: Is social interaction scripted? *Journal for the Theory of Social Behaviour*, 2022.

[11] Tom Erez and William D Smart. What does shaping mean for computational reinforcement learning? In *2008 7th IEEE international conference on development and learning*, pages 215–219. IEEE, 2008.

[12] Kimmo Eriksson, Pontus Strimling, Michele Gelfand, Junhui Wu, Jered Abernathy, Charity S Akotia, Alisher Aldashev, Per A Andersson, Giulia Andrighetto, Adote Anum, et al. Perceptions of the appropriate response to norm violation in 57 societies. *Nature communications*, 12(1):1–11, 2021.

[13] Ernst Fehr and Simon Gachter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

[14] Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415(6868):137–140, 2002.

[15] Herbert Gintis. Social norms as choreography. *politics, philosophy & economics*, 9(3):251–264, 2010.

[16] Deborah M Gordon. The organization of work in social insect colonies. *Nature*, 380(6570):121–124, 1996.

[17] Gillian K Hadfield and Barry R Weingast. What is law? a coordination model of the characteristics of legal order. *Journal of Legal Analysis*, 4(2):471–514, 2012.

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

[19] Douglas D Heckathorn. Collective action and the second-order free-rider problem. *Rationality and society*, 1(1):78–100, 1989.

[20] Joseph Henrich. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, 53(1):3–35, 2004.

[21] Christine Horne and Anna Cutlip. Sanctioning costs and norm enforcement: An experimental test. *Rationality and Society*, 14(3):285–307, 2002.

[22] Harold H Kelley, John G Holmes, Norbert L Kerr, Harry T Reis, Caryl E Rusbult, and Paul AM Van Lange. *An atlas of interpersonal situations.* Cambridge University Press, 2003.

[23] Raphael Köster, Dylan Hadfield-Menell, Richard Everett, Laura Weidinger, Gillian K Hadfield, and Joel Z Leibo. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3):e2106028118, 2022.

[24] Daeyeol Lee, Hyojung Seo, and Min Whan Jung. Neural basis of reinforcement learning and decision making. *Annual review of neuroscience*, 35:287, 2012.

[25] Sang Wan Lee and Ben Seymour. Decision-making in brains and robots—the case for an interdisciplinary approach. *Current Opinion in Behavioral Sciences*, 26:137–145, 2019.

[26] Paweł Lichocki, Danesh Tarapore, Laurent Keller, and Dario Floreano. Neural networks as mechanisms to regulate division of labor. *The American Naturalist*, 179(3):391–400, 2012.

[27] Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 789–797, 2020.

[28] Sarah Mathew. How the second-order free rider problem is solved in a small-scale society. *American Economic Review*, 107(5):578–81, 2017.

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

[30] Martin A Nowak, Sebastian Bonhoeffer, and Robert M May. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences*, 91(11):4877–4881, 1994.

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

[32] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action.* Cambridge university press, 1990.

[33] Logan Perry, Mahendra Duwal Shrestha, Michael D Vose, and Sergey Gavrilets. Collective action problem in heterogeneous groups with punishment and foresight. *Journal of Statistical Physics*, 172(1):293–312, 2018.

[34] Gail B Peterson. A day of great illumination: BF Skinner's discovery of shaping. *Journal of the experimental analysis of behavior*, 82(3):317–328, 2004.

[35] Peter Railton. Normative guidance. *Oxford studies in metaethics*, 1(3):2–33, 2006.

[36] Peter Richerson, Ryan Baldini, Adrian V Bell, Kathryn Demps, Karl Frost, Vicken Hillis, Sarah Mathew, Emily K Newton, Nicole Naar, Lesley Newson, et al. Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, 39, 2016.

[37] Rick L Riolo, Michael D Cohen, and Robert Axelrod. Evolution of cooperation without reciprocity. *Nature*, 414(6862):441–443, 2001.

[38] Thomas C Schelling. *The Strategy of Conflict.* Harvard university press, 1960.

[39] Cass R Sunstein. Social norms and social roles. *Colum. L. Rev.*, 96:903, 1996.

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

[41] Attila Szolnoki and Matjaž Perc. Conditional strategies and the evolution of cooperation in spatial public goods games. *Physical Review E*, 85(2):026104, 2012.

[42] Edna Ullmann-Margalit. *The emergence of norms.* Oxford University Press, 1977/2015.

[43] Amrisha Vaish, Esther Herrmann, Christiane Markmann, and Michael Tomasello. Preschoolers value those who sanction non-cooperators. *Cognition*, 153:43–51, 2016.

[44] Peter Vanderschraaf. Convention as correlated equilibrium. *Erkenntnis*, 42(1):65–87, 1995.

[45] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, and Joel Z Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *arXiv preprint arXiv:2106.09012*, 2021.

[46] Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 683–692, 2019.

[47] Woodrow Z Wang, Mark Beliaev, Erdem Bıyık, Daniel A Lazar, Ramtin Pedarsani, and Dorsa Sadigh. Emergent prosociality in multi-agent games through gifting. *arXiv preprint arXiv:2105.06593*, 2021.

[48] Timothy Waring and James Acheson. Evidence of cultural group selection in territorial lobstering in Maine. *Sustainability science*, 13(1):21–34, 2018.

[49] Till O Weber, Ori Weisel, and Simon Gächter. Dispositional free riders do not free ride on punishment. *Nature communications*, 9(1):1–9, 2018.

[50] Toshio Yamagishi. The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology*, 51(1):110, 1986.

A Supplementary material

A.1 Evolved social norms

We found completely different social sanctioning matrices after multiple runs of the evolutionary processes. Nevertheless, these social sanctioning matrices converged on a similar spatial role distributions that can maximize the group welfare. Figures 4 and 5 show two examples of the evolved social sanctioning matrices and the role distributions learned through the social sanctions. Note that in the first row of Figure 5, labeled as “Random”, no result is provided because for this parameter settings of the common pasture game ($w = 0$ and $c = 0$) it was not possible to find a sustainable solution. Therefore, the emerging role distributions would be random.

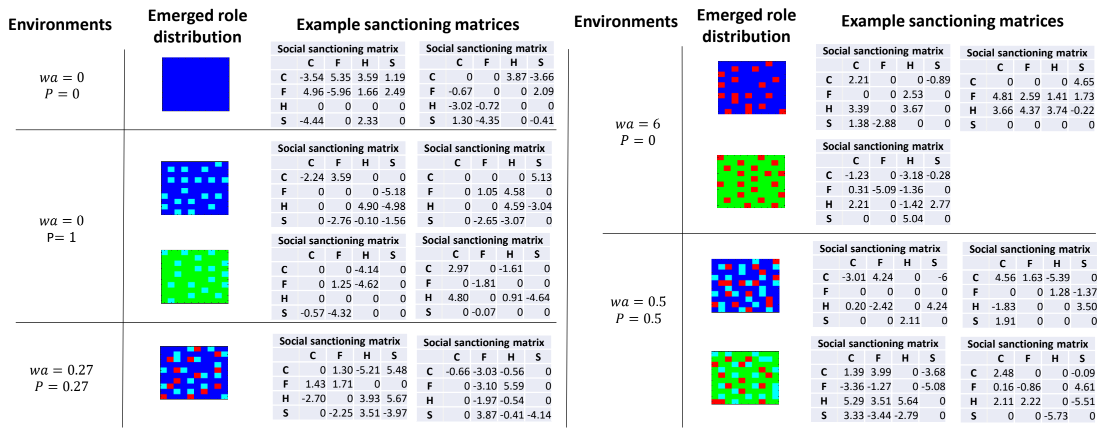


Figure 4: For settlement maintenance game: under column header “Emerged role distributions”, the spatial role distributions that are learned through the social sanctions are shown. Under ”Example social sanctioning matrices”, the evolved social sanctioning matrices found through the evolutionary processes are shown.

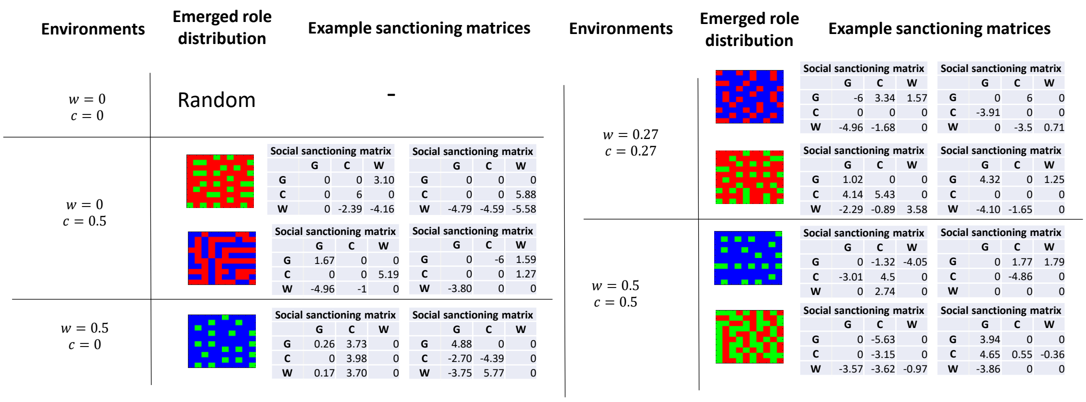


Figure 5: For common pasture game: under column header “Emerged role distributions”, the spatial role distributions that are learned through the social sanctions, are shown. Under ”Example social sanctioning matrices”, the evolved social sanctioning matrices found through the evolutionary processes are shown.