The emergence of division of labour through decentralized social sanctioning

Anil Yaman1,†, Joel Z. Leibo2,†, Giovanni Iacca3 and Sang Wan Lee4

1Computer Science Department, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands
2DeepMind, London N1C 4AG, UK
3Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
4Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea

Human ecological success relies on our characteristic ability to flexibly self-organize into cooperative social groups, the most successful of which employ substantial specialization and division of labour. 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 thus 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 labour emerge in groups of self-interested lifetime-learning individuals? Here, we show that by introducing a model of social norms, which we regard as emergent patterns of decentralized social sanctioning, it becomes possible for groups of self-interested individuals to learn a productive division of labour 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.

1. Introduction

Human societies depend on division of labour. However, an individual’s role is not specified in their 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 would be possible with purely genetic evolution. However, many computational models of lifetime-learning formulate it as a process of maximization of individual payoffs [1–4]. Such a formulation cannot, on its own, account for the learning of division of labour. 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 labour 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 labour 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 [5–9].
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 & Gächter [8] designed experiments consisting of two stages. First, participants chose how much to contribute to the public good. 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 behaviour. 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.

Figure 1 illustrates the learning process used in our model. In Stage I (figure 1a), a group of self-interested lifetime-learners learn their roles. This process is modelled 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 modelled 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 [4]. Stage II (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.

In our model, social sanctioning occurs because individuals endorse normative rules of the form ‘ocial norm evolution, social sanctioning matrix values are optimized to produce optimum lifetime-learning of the social roles.
patterns of decentralized social sanctioning is supported by laboratory experiments [13–16] and ethnographic evidence [17,18]. 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 behaviourist psychology [19] and in algorithmic reinforcement learning [20]. Here groups can induce individuals by sanctioning to select any behaviour. 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 determined by rules that are computationally encoded in the form of a real-valued matrix with shape $k \times k$ (where $k$ is the number of roles). Entries in the matrix correspond to normative rules. Each entry defines how much reward or punishment an individual with role $i$ ought to impose on an individual of role $j$. Under our terminology, each such specific matrix corresponds 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 centres around theories of social structure which features this kind of intimate connection between ‘role psychology’ and normative behaviour. 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 [21–23]. 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. Note also that role-wise sanctioning—but not individual-targeted sanctioning—allows the implementation of impartial norms, like some human norms are [24,25]. It need not merely produce the self-serving resentment-driven sanctioning motivations that are sometimes observed in non-human primates [26].

Depending on the environment, a social norm may either facilitate high social welfare or court disaster. For instance, a norm that encourages many farmers to become soldiers could be important for survival in a hostile environment where soldiers are truly needed for defense. However, in a more benign environment, with no external threats for soldiers to defend against, then allocating too many individuals to the soldier role would be suboptimal, and it would be better to have more farmers instead. Typically the environment is a major factor in which specific norms are effective.

One implication is that norm evolution may grow more and more precisely optimized to local conditions over time. It is generally possible to achieve a better and better fit to the precise environmental conditions a community finds itself in by evolving more and more complex norms. However, we think there are at least two strong brakes on this process that prevent the evolution of greater and greater norm complexity. First, we must consider the fact that for cultural evolution to happen, norms must somehow be transmitted between generations. These conditions are thought to give rise to a simplicity bias (complexity regularization [27]) in the context of direct and indirect reciprocation rules for cooperation [28,29], and language learning [30]. We propose that the same mechanisms also constrain norm acquisition. So if norm learners (children or immigrants) find it easier to acquire simpler norms, then there will be an overall bias toward simplicity throughout the evolutionary process. There is additionally one other kind of brake on complexity growth: very high norm complexity may be problematic in itself if we assume that it implies a growing overhead cost. Tainter [31] argued that the tendency for states to accumulate complexity, with large attendant costs, has been a major causal force in the collapse of ancient empires. If true, this logic would also seem to put a brake on the benefits from increasing norm complexity.

In our model, social norms are optimized through a cultural evolution process (given in algorithm 3 in the supplementary material). We interpret its mechanism of norm change as cultural group selection [32–34]. As such, our model 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 [32]. 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 perturbing its constituent rules. 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-labour arrangement. In addition, we introduce complexity regularization [27,35] that imposes a penalty based on the complexity of the norms. The evolutionary process of the norms 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 in order 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-labour arrangements through lifetime-learning. Specifically, we show that the proposed method of simulating social norm evolution leads to higher collective payoff than groups of self-interested or altruist individuals. Moreover, complexity regularization promotes convergent evolution to simpler social norms in independent evolutionary processes.

2. The model

(a) 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 $p = \{p_1, p_2, \ldots, p_K\}$, 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 behaviour over time via value-based reinforcement learning [4]. On each step \( t \), each individual \( i \) selects its role \( A^{(t)}_i \) using its estimated value \( Q^{(t)}_i(p_j) \) with

\[
A^{(t)}_i = \arg\max_{\rho} Q^{(t)}_i(p_j),
\]

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

\[
Q^{(t+1)}_i(A^{(t)}_i) = Q^{(t)}_i(A^{(t)}_i) + \alpha \left( r^{(t)}_i - Q^{(t)}_i(A^{(t)}_i) \right),
\]

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 \( \epsilon \). 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 behaviours 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.

**(b) Incentivizing the lifetime-learning of division of labour via social norms**

We consider a reward function \( F \)

\[
[r^{(1)}_1, \ldots, r^{(m)}_t] = F(A^{(1)}_1, \ldots, A^{(m)}_t),
\]

where the reward \( r^{(t)}_i \) received by individual \( i \) at time step \( t \) depends on the roles selected by all individuals in the group, not just the individual’s 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.

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 in 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 neighbourhood 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 \( \ell \) is \( S_{k,\ell} \) (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, Stages I and II are performed consecutively for \( T \) iterations.

**(c) Evolution of social norms for incentivizing division of labour**

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 3 in the electronic supplementary material). The goal of the algorithm is to find a particular set of social sanctioning rules, represented as matrix \( S^* \), that can maximize

\[
S^* = \arg\max_S \text{Ft}(S),
\]

where \( \text{Ft}(S) \) is the fitness of a norm \( S \) composed of

\[
\text{Ft}(S) = R - \lambda ||S||_0,
\]

where \( R \) is the average group payoff received by the individuals as a result of the role distributions they learn, \( -\lambda ||S||_0 \) is the complexity regularization (also known as \( L_0 \) regularization [27,35]), where \( ||S||_0 \) denotes the number of non-zero values in the social sanctioning matrix and \( \lambda \) is a parameter to adjust the weight for the importance given to the complexity regularization during the optimization process. Given two norms that can achieve similar average reward, the one that consists of a smaller number of rules would be simpler in terms of their understandability and computational complexity. When \( \lambda > 0 \), complexity influences the overall fitness of social norms and favours those that are simpler. Higher \( \lambda \) values are expected to increase the selection pressure for favouring simpler social norms. The sensitivity analysis on \( \lambda \) for the evolutionary processes modelled in this paper is presented in the electronic supplementary material.

Other forms of complexity measures can be considered. For instance, the sum of the absolute values of the social sanctioning values (i.e. \( L_1 \) norm) could be a good measure for minimizing the total amounts of sanctioning values. On the other hand, this may not reduce the number of rules that constitute the social norms since there is still a possibility of the emergence of a large number of rules with very small sanctioning values. Another option may be the number of types of roles involved in the rules. In this case, there is also a possibility of the emergence of social norms that consist of a large number of rules but contain a small number of role types. Therefore, \( L_0 \) norm is a better fitting measure among these alternatives for modelling the complexity of the social norms that can facilitate the emergence of norms simpler in terms of their size.

Initially, the group is assigned a social sanctioning matrix where the amount of sanctions is randomly sampled within a certain range (i.e. uniformly in [−6, 6]). It is assumed that all individuals in the group use the same social sanctioning matrix. Then, the group can sample a new social norm by applying three possible operators to the existing one: (i) Gaussian perturbation: the values of non-zero cells perturbed by \( \mathcal{N}(0, \sigma) \), (ii) addition of a new rule: a randomly selected zero-valued cell initialized from \( \mathcal{N}(0, 1) \), and (iii) deletion of an existing rule: the value of a randomly selected non-zero cell is replaced by 0. Note that addition and deletion operators cause changes in the number of sanctioning rules, whereas Gaussian perturbation causes only a change in the sanctioning values of existing rules. In addition, we keep track of the usage frequency of each rule (i.e. how many times each rule is implemented by the individuals in a group), and if there are
rules that are not used, we remove them by assigning their amount to 0 in the social sanctioning matrix.

We introduce four parameters to adjust the amount of change these operators can lead to. These parameters: mutation probability (mp) and mutation rate (mr) for Gaussian perturbation, rule addition probability (rap) for rule addition and rule deletion probability (rdp) for rule deletion. Gaussian perturbation is applied with probability mp and rate mr, where σ = mr, and rule addition and rule deletion operators are applied based on the probabilities rap and rdp, respectively. Sensitivity analysis for these parameters is provided in the electronic supplementary material.

After the perturbation step, if the perturbed social sanctioning rules S provides better performance (as measured by Ft), it is selected by the group. It then replaces the social sanctioning rules currently in use. After a certain number of iterations of this process we expect to find social sanctioning rules that are better at incentivizing division of labour.

The average reward of the social sanctioning rules R is found as follows:

$$R = \text{EVAL}(S) = \frac{1}{M} \sum_{m=1}^{M} \sum_{i \in \xi} \sum_{j = 0}^{T} r_{i,j}^{(m)},$$

(2.4)

where EVAL() is the evaluation function of social sanction S, M is the number of individuals in the group, and ξ is the 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.

(d) 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 neighbours (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 neighbours (i.e. their Moore neighbourhood). 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 is interpreted here as a model for small-scale human societies [36,37]. Payoffs associated with each role are shown in figure 2b. In this game, individuals can obtain payoffs by performing foraging and hunting roles for 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 neighbourhood then the adverse effect of waste accumulation is mitigated. Furthermore, there is an 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 neighbourhood.

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 may occur [38,39]. 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 behaviour, we model a pasture environment with a certain amount of starting resources in each location (e.g. grass for the herders’ herds). These resources can be appropriated by the herders either greedily (C) or considerately (C). The resources regenerate over time unless fully depleted.

Figure 2. (a) Types of roles, colour coded for visualization, can be performed by the individuals that are spatially distributed on a two-dimensional toroidal grid environment and can impose social sanctions on the individuals within their neighbourhood. 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 (b) settlement maintenance and (c) common pasture games. The settlement maintenance has four and common pasture has three role choices.
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 in 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 [40–42]. It differs from the purely destructive-negative-sum social sanctioning schemes considered by Köster et al. [43] and Vinitsky et al. [44]. 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 behaviour. 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 one’s own reward to increase the other’s (positively sanctioning them). Symmetrically, one may take resources away from another agent, gaining them oneself (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 2 (see electronic supplementary material).

(i) Settlement maintenance
In the settlement maintenance game, individuals can select one of the roles in each time step \( t \) as: \( A_i^t \in \{ \text{cleaner, forager, hunter, soldier} \} \). We assume there is a level of waste accumulation \( d_{i+1}^{(j)} = \min(w_{a}, d_{i}^{(j)} + \tau) \) in each cell location \((i, j)\) with a constant rate \( \tau \) reaching a maximum level. 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:

\[
r_{t}^{(j)} = \max(0, r_{t}^{(j)} - d_{t}^{(j)}),
\]

where \( r_{t}^{(j)} \) 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 neighbourhood 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 the attack probability \( P_{a} \), which is assumed to be the same for all cells. 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 neighbourhood, the attacks are defended and payoffs are protected.

(ii) Common pasture
In common pasture, individuals can select one of the roles as: \( A_i^t \in \{ \text{considerate, greedy, worker} \} \). The amount of resources in each cell is represented as \( D^{(j)} \). At the beginning of the game, a certain amount of resources is allocated in each cell. Herders, represented as considerate or greedy, consume the resources in their neighbourhood either considerately or greedily, as shown in figure 2c: (payoffs they receive are the resources consumed in the environment). They can use available resources in their neighbouring 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 modelled by the logistic growth model as

\[
\frac{\partial D^{(j)}}{\partial t} = (c + w)(1 - \frac{D^{(j)}}{K}) D^{(j)},
\]

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

3. Results
(a) Social norms enable lifetime-learning of division of labour
In both games, two environmental parameters are used for testing the emergence of learning proper role distributions in five different environmental variations. In the settlement maintenance game (figure 3a), these parameters are based on waste accumulation and adversarial attack probability, and initialized as: \( (w_{a}, p) = (0, 0), (6, 0), (0, 1), (3.2, 0.5), (6, 1) \). 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 games, we chose maximum 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 environmental instances multiple times (i.e. 30 runs each) with complexity regularization \( \lambda = 0.2 \) to find social norms that could provide optimum learning in each one. We observed that, owing to the complexity regularization, multiple runs of the evolutionary processes converged to similar social norms that can provide (near-)optimum group welfare. Without the complexity regularization, we observed the emergence of various social norms. The analysis of the effect of complexity regularization is provided in \( \S 3 b \), and examples of the social norms and the spatial role distributions they converged on are provided in the electronic supplementary material.

In figure 3a, examples of learned spatial role distributions in five environmental variants in both games are shown. The individuals converged on different role distributions depending on the environmental conditions. Interestingly, the role distributions that emerged show clear signs of neighbourhood patterns where the roles benefit their neighbours, distributed in various neighbourhoods to minimize their overlap.

In the settlement maintenance game, since there is no need for cleaner and soldier roles when there are no waste accumulation and adversarial attacks \( (w_{a} = 0, p = 0) \), 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, which prevents additional farmers in the environment, and therefore avoids 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 converge to select the forager role since it is easier to learn because it does not depend on coordination with others. The hunter role, which can provide higher rewards, depends on coordination within the neighbourhood, so it is more difficult to discover by chance since at least five individuals would have to select it
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.

(b) Complexity regularization promotes convergent evolution of simpler social norms

Figure 4 shows the relationship between the average group payoffs and the complexity of the norms for different complexity regularization ($\lambda$). To facilitate a better comparison, average group payoff shows only $R$ (see equation (2.4)); therefore, it does not include the complexity regularization ($-\lambda ||S||_0$). The complexity of the norms is assessed by the size of the social norms: the social norms that are composed of larger number of rules are more complex.

We observe that a larger $\lambda$ value leads to the emergence of simpler rules in all environments. Furthermore, there is no statistical difference (based on the Wilcoxon rank-sum test [45] on $\alpha = 0.05$ significance level) between the average payoff achieved with different $\lambda$ values in each environment except in Environments 3 and 4 in common pasture. For the exception cases, cultural evolution processes with complexity regularization (when $\lambda > 0$) were able to produce norms that can successfully promote the division of labour only in a limited number of independent processes (i.e. 4 and 7 out of 30 independent cultural evolutionary processes in Environments 3 and 4, respectively). Therefore, on average, the results are lower in those cases. If only cultural evolution processes that led to a successful division of labour are considered, then the results are statistically similar to the case without complexity regularization ($\lambda = 0$).

Overall, complexity regularization is able to promote the emergence of norms that can achieve similar average group
payoffs. Moreover, the complexity regularization promoted convergent evolution by means of producing similar norms at the end of independent cultural evolution processes (see the electronic supplementary material for examples of norms that emerged as a result of the independent cultural evolution processes for \( \lambda \) values 0 and 0.2). These results are further elaborated in the next two sections.

(i) Settlement maintenance

In Environment 1, independent of \( \lambda \) value, social norms allow groups to achieve maximum average payoff of 6 (achieved when all individuals learn to choose hunter role). When there is no importance given to the complexity regularization (\( \lambda = 0 \)), the average number of rules of evolved social norms is more than 8. Only a few of these evolved social norms converged to a similar rule set. The rest each converged to a unique and complex set of rules. On the other hand, when the importance of complexity is introduced, all the norms converge to two possible types that consist of a single rule: either, ‘Hunter ought to discourage forager by amount \( X \)’ or ‘Forager ought to encourage hunter by amount \( Y \)’. Here, \( X \) and \( Y \) can take values from 0.03 to 2.8, though discouragement is associated with a negative sign. This is quite intuitive, and in this case, encouragement/discouragement between these two roles allows individuals to learn to select hunter to receive a higher payoff by cooperating with others.

Environments 2 to 5 are more complex in that they require additional roles in the groups to achieve a higher average payoff. In the case of Environment 2, cleaner, in Environment 3, soldier, and in 4 and 5, both cleaner and soldier roles are needed. In these environments, increasing \( \lambda \) leads to the emergence of similar rules in multiple independent cultural evolutionary processes. In these environments, we note the emergence of norms that are concerned with these roles in addition to the baseline norm that emerged in Environment 1. An example of this is a norm that emerged for Environment 2, when \( \lambda = 0.2 \): ‘(i) hunter ought to discourage forager by \( -1.45 \), and (ii) hunter ought to encourage cleaner by \( 0.6 \)’. The first part of this norm is the same as the norms that emerged for Environment 1; in addition, the second part emerged to encourage uptake of the cleaner role in the group. Another example is for Environment 4: ‘(i) hunter ought to discourage forager by \( -2.28 \), (ii) hunter ought to encourage cleaner by \( 0.57 \), and (iii) hunter ought to encourage soldier by \( 0.42 \)’. Here, we observe appearance of two rules regarding the cleaner and soldier roles in addition to the norm that emerged in Environment 1. Similar patterns can be observed in the results for other environments provided in the electronic supplementary material in detail.

Without complexity regularization, while the difficulty of the environments increases, the complexity of the emergent norms also increases. For example, complexity remains around 8 in Environment 1. This shows that through the cultural evolution processes the size of the norms remains stable (considering expected norm size during the initialization is 8). However, in other environments, the complexity of the norms increases throughout the processes. This indicates that larger number of rules may be required to perform better in those environments. Nevertheless, when the complexity regularization is introduced, especially when \( \lambda = 0.2 \), the average sizes of norms are reduced around 2.5 in Environments 2 and 3, and around 3 and 3.5 in Environments 4 and 5, respectively.

In Environment 2, the simplest norms consist of two rules. Interestingly, these norms are composed of the same rules that emerged for Environment 1 to promote the cooperation between hunter roles, and an additional rule that promotes learning cleaner roles. A similar outcome is observed for Environment 3, where soldier roles are needed as well as the cooperation between hunter roles to achieve the optimum average payoff. Finally, in Environments 4 and 5, the lengths of the rules that emerged range from 3 to 5. Similarly, here, we observe appearance of the conjunction of the norms that emerged in Environments 1, 2 and 3, since both cleaner and soldier roles are needed in these environments.

(ii) Common pasture

It is not possible to achieve any payoff in Environment 1 because of the parameters of the environment since they do not allow natural and worker growth rate for recovering the environmental resources. Therefore, the norms that are found at the end of the cultural evolutionary processes are random. However, when the complexity regularization is introduced, the length of the rules converges to 0. That is because, if no norm is helpful, having no norm is the best.

In Environment 2, we observe convergent evolution of the same type of social norm that is composed of two rules where greedy (herder) encourages considerate (herder) and considerate (herder) encourages considerate (herder). This norm facilitates the emergence of role distributions that consist mainly of considerate (herders) with some greedy (herders) that are spread around. This pattern leads to the exploitation of the environmental resources by greedy (herders) without disrupting the sustainability.

In Environment 3, only four independent cultural evolution processes produced social norms that led to sustainable use of environmental resources. These social norms are composed of partially different rules although they have similar length and produce similar division of labour patterns except in one case (provided in the electronic supplementary material). Similarly, in Environment 4, only seven of the processes produced social norms that promote sustainable use of resources. These rules are composed of partially the same rules. In Environment 5, in many cases, the norms are composed of two rules and converged to the same rules.

4. 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 labour 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 labour in groups of self-interested lifetime-learning individuals. We modelled social norms as decentralized social sanctioning patterns and studied how they can encourage individuals to cooperate. We also proposed a regularized cultural group evolution model to study how such norms that enable the learning of pro-social 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 [11,46–48]. Some authors, like us, studied the emergence of cooperation in spatial game settings [49–51]. However, this line of work usually does not model lifetime-learning. We modelled 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.

Most evolutionary game theory-based models share a common assumption that individuals make one choice and stick with it throughout their lives [53]. Others model the evolution and spread of successful strategies [54,55] through social learning over a population of individuals. By contrast, the cultural evolutionary process in our model operates on the social norm level. Additionally, individuals are endowed with the capability of changing their behaviour based on their own experience. Therefore, we can assume that individuals can re-select what role to play over and over again throughout their lives. As a result, our model is more about deciding what one wants to do today than it is about deciding what to do with one’s life. It should correspond best to situations where roles may be revisited each day. There are likely many such roles in a foraging society. For instance, one could hunt one day and gather the next day. By contrast, the evolutionary game theory approaches where role definition is specified genetically becomes established depends on its guiding the learning of all individuals to a more socially advantageous outcome.

In this line of work, the norm is regarded as a piece of public knowledge that individuals may condition their behaviour on in order to rationally coordinate with one another [60–63]. For instance, a focal point is a ‘default choice’ from which it would be irrational to deviate when one knows others are trying to coordinate [64]. Many models incorporate aspects of both mechanisms (e.g. [21,65]). Our approach resonates more strongly with approach A. The two games we studied, settlement maintenance and common 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 labour [32].

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 [58,59] 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 them 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 behaviour on in order to rationally coordinate with one another [60–63]. For instance, a focal point is a ‘default choice’ from which it would be irrational to deviate when one knows others are trying to coordinate [64]. Many models incorporate aspects of both mechanisms (e.g. [21,65]). Our approach resonates more strongly with approach A. The two games we studied, settlement maintenance and common 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 labour [32].

**Ethics.** This work did not require ethical approval from a human subject or animal welfare committee.

**Data accessibility.** Data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.dv41ns24r [66].

Supplementary material is available online [67].

**Declaration of AI use.** We have not used AI-assisted technologies in creating this article.

**Authors’ contributions.** A.Y.: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; J.Z.L.: conceptualization, investigation, methodology, validation, writing—original draft, writing—review and editing; G.I.: methodology, validation, writing—review and editing; S.W.L.: conceptualization, formal analysis, funding acquisition, investigation, methodology, validation, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed herein.

**Conflict of interest declaration.** We declare we have no competing interests.

**Funding.** S.W.L. was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (no. RS-2023-0023251, System3 Reinforcement learning with high-level brain functions, and no. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST), and by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2019M3E5D2A01066267).
References

1. Dayan P, Niv Y. 2008 Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 18, 185–196. (doi:10.1016/j.conb.2008.08.003)
2. Lee D, Seo H, Jung MW. 2012 Neural basis of reinforcement learning and decision making. Annu. Rev. Neurosci. 35, 287. (doi:10.1146/annurev-neuro-062111-150512)
3. Lee SW, Seymour B. 2019 Decision-making in brains and robots—the case for an interdisciplinary approach. Curr. Opin. Behav. Sci. 26, 137–145. (doi:10.1016/j.cobeha.2018.12.012)
4. Sutton RS, Barto AG. 2018 Reinforcement learning: an introduction. Cambridge, MA: MIT Press.
5. Albrecht F, Kube S. 2018 Peer-punishment in a cooperation and a coordination game. Games 9, 54. (doi:10.3390/g9030054)
6. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation in sizable groups. Ethol. Sociobiol. 13, 171–195. (doi:10.1016/0162-3095(92)90022-Y)
7. Fehr E, Gächter S. 2002 Altruistic punishment in public goods experiments. Am. Econ. Rev. 90, 988–994. (doi:10.1257/aer.90.4.988)
8. Fehr E, Gächter S. 2002 All-inclusive punishment in humans. Nature 415, 137–140. (doi:10.1038/415137a)
9. Perry L, Shrestha MD, Vose MD, Gavrilets S. 2018 Measuring the complexity of human cooperation under indirect reciprocity. Phil. Trans. R. Soc. B 376, 20202991. (doi:10.1098/rstb.2020.0291)
10. Wang WZ, Beliaev M, Bıyık E, Lazar DA, Pedarsani R, Sadigh D. 2021 Emergent prosociality in multi-agent games through gifting. arXiv:2105.06593v1 [cs.MA]. (doi:10.48550/arXiv.2105.06593)
11. Willis R, Luck M. 2023 Resolving social dilemmas through reward transfer commitments. In Proc. Adaptive and Learning Agents Workshop (ALA2023) 29–30 May 2023, London, UK (eds F Cruz, CF Hayes, C Wang, C Yates).
12. Köster R, Hadfield-Menell D, Everett R, Weidinger L, Köster R, Hadfield-Menell D, Everett R, Weidinger L, Brosnan SF, Waal FD. 2003 Monkeys reject unequal pay. Nature 425, 297–299. (doi:10.1038/nature01963)
13. Hadfield GK, Leibo JZ. 2022 Spurious normativity in agent games through gifting. arXiv:2307.10124v3 [cs.LG]. (doi:10.48550/arXiv.2307.10124)
14. Köster R, Hadfield-Menell D, Everett R, Weidinger L, Hadfield GK, Leibo JZ. 2022 Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. Proc Natl Acad Sci USA 119, e8106028118. (doi:10.1073/pnas.2106028118)
15. Vinitsky E, Köster R, Agapiou JP, Duéñez-Guzmán EA, Veizheevets AS, Leibo JZ. 2023 A learning agent that acquires social norms from public sanctions in centralized multi-agent settings. Collect. Intell. 2, 26339137231162025. (doi:10.1177/2471582X22117037)
16. Wilcoxon F. 1992 Individual comparisons by ranking methods. In Breakthroughs in statistics (eds S Kotz, NL Johnson), pp. 196–202. New York, NY: Springer.
17. Axelrod R. 1986 An evolutionary approach to norms. Am. Polit. Sci. Rev. 80, 1095–1111. (doi:10.2307/1960858)
18. Heckathorn DD. 1989 Collective action and the second-order free rider problem is solved in a small-scale society. Am. Econ. Rev. 107, 578–581. (doi:10.1257/aer.p20171090)
19. Peterson GB. 2004 A day of great illumination: BF Skinner’s discovery of shaping. J. Exp. Anal. Behav. 82, 317–328. (doi:10.1901/jeab.2004.82-317)
20. Erez T, Smart WD. 2008 What does shaping mean for computational reinforcement learning? In Proc. 7th IEEE Int. Conf. Development and Learning, Monterey, California, 9–12 August 2008, pp. 215–219. New York, NY: IEEE.
21. Bičieri C 2006 The grammar of society: the nature and dynamics of social norms. Cambridge, UK: Cambridge University Press.
22. Eichks G. 2022 Coordinating behaviors: is social interaction scripted? J. Theory Social Behav. 53, 85–99. (doi:10.1111/jtsb.12357)
23. Sunstein CR. 1996 Social norms and social rules. Columbia Law Rev. 96, 903. (doi:10.2307/1123430)
24. Hadfield GK, Weingast BR. 2014 Microfoundations of the rule of law. Annu. Rev. Political Sci. 17, 21–42. (doi:10.1146/annurev-polisci-100711-135226)
25. RAFIOT P. 2017 Moral learning: conceptual foundations and normative relevance. Cognition 167, 172–190. (doi:10.1016/j.cognition.2016.08.015)
26. Brosnan SF, Waal FD. 2003 Monkeys reject unequal pay. Nature 425, 297–299. (doi:10.1038/nature01963)
27. Hastie T, Tilghman R, Friedman JH, Friedman JH 2009 The elements of statistical learning: data mining, inference, and prediction, vol. 2. New York, NY: Springer.
28. Rubinstein A. 1986 Finite automata play the repeated prisoner’s dilemma. J. Econ. Theory 39, 83–96. (doi:10.1016/0022-1902(86)90076-8)
29. Santos FP, Pacheco JM, Santos FC. 2021 The complexity of human cooperation under indirect reciprocity. Phil. Trans. R. Soc. B 376, 20202991. (doi:10.1098/rstb.2020.0291)
30. Tamariz M, Kirby S. 2016 The cultural evolution of cooperation. Curr. Opin. Behav. Sci. 9, 43–47. (doi:10.1016/j.cobeha.2015.09.003)
31. Tainter J. 1988 The collapse of complex societies. Cambridge, UK: Cambridge University Press.
32. Henrich J. 2004 Cultural group selection, coevolutionary processes and large-scale cooperation. J. Econ. Behav. Org. 53, 3–35. (doi:10.1016/j.jebo.2003.06.004)
33. Richerson P et al. 2016 Cultural group selection plays an essential role in explaining human cooperation: a sketch of the evidence. Behav. Brain Sci. 39, e30. (doi:10.1017/S0140525X1400106X)
34. Waring T, Acheson J. 2018 Evidence of cultural group selection in territorial lobstering in Maine. Sustain. Sci. 13, 21–34. (doi:10.1007/s11625-017-0501-x)
35. Louizos C, Welling M, Kingma DP. 2018 Learning sparse neural networks through Lp regularization. arXiv:1712.07312v2 [stat.ML]. (doi:10.48550/arXiv.1712.07312)
36. Gordon DM. 1996 The organization of work in social insect colonies. Nature 380, 121–124. (doi:10.1038/380121a0)
37. Lichocki P, Tarapore D, Keller L, Floreano D. 2012 Neural networks as mechanisms to regulate division of labor. Am. Nat. 179, 391–400. (doi:10.1086/664079)
38. Hardin G. 1968 The tragedy of the commons. Science 162, 1243–1248. (doi:10.1126/science.162.3859.1243)
39. Ostrom E 1990 Governing the commons: the evolution of institutions for collective action. Cambridge, UK: Cambridge University Press.
40. Lupu A, Precup D. 2020 Gifting in multi-agent reinforcement learning. In Proc. 19th Int. Conf. Autonomous Agents and MultiAgent Systems (AAMAS2020), 9–13 May 2020, Auckland, NewZealand (eds B An, N Yorke-Smith, A El Fallah Seghrouchni, G Sukthankar), pp. 789–797. International Foundation for Autonomous Agents and MultiAgent Systems.
49. Nowak MA, Bonhoeffer S, May RM. 1994 Spatial games and the maintenance of cooperation. *Proc. Natl Acad. Sci. USA* **91**, 4877–4881. (doi:10.1073/pnas.91.11.4877)

50. Nowak MA, May RM. 1992 Evolutionary games and spatial chaos. *Nature* **359**, 826–829. (doi:10.1038/359826a0)

51. Szolnoki A, Perc M. 2012 Conditional strategies and the evolution of cooperation in spatial public goods games. *Phys. Rev. E* **85**, 026104. (doi:10.1103/PhysRevE.85.026104)

52. Wang JX, Hughes E, Fernando C, Czarnecki WM, Duéñez-Guzmán EA, Leibo JZ. 2019 Evolving intrinsic motivations for altruistic behavior. In *Proc. 18th Int. Conf. Autonomous Agents and MultiAgent Systems* (ALA2023), 9–10 May 2023, London, UK (eds N Agmon, ME Taylor, E Elkind, M Veloso), pp. 683–692. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

53. Weibull JW. 1997 *Evolutionary game theory*. Cambridge, MA: MIT Press.

54. Traulsen A, Hauert C, Silva HD, Nowak MA, Sigmund K. 2009 Exploration dynamics in evolutionary games. *Proc. Natl Acad. Sci. USA* **106**, 709–712. (doi:10.1073/pnas.0808450106)

55. Yaman A, Bredeche N, Cayla O, Leibo JZ, Lee SW. 2022 Meta-control of social learning strategies. *PLoS Comput. Biol.* **18**, e1009882. (doi:10.1371/journal.pcbi.1009882)

56. Henrich J, Boyd R. 2008 Division of labor, economic specialization, and the evolution of social stratification. *Curr. Anthropol.* **49**, 715–724. (doi:10.1086/587889)

57. Lundberg S, Startz R. 1998 On the persistence of racial inequality. *J. Labor Econ.* **16**, 292–323. (doi:10.1086/209890)

58. Kelley HH, Holmes JG, Kerr NL, Reis HT, Rusbult CE, Van Lange PAM. 2003 *A n a t a l o g y s e n t e r p e r s o n a l i s t i o n s*. Cambridge, UK: Cambridge University Press.

59. Ullmann-Margalit E. 1977/2015 *The emergence of norms*. Oxford, UK: Oxford University Press.

60. Axtell RL, Epstein JM, Young HP. 2001 The emergence of classes in a multi-agent bargaining model. In *Social dynamics* (eds SN Durlauf, HP Young), pp. 177–195. Cambridge, MA: MIT Press. (doi:10.7551/mitpress/6294.003.0009)

61. Gintis H. 2010 Social norms as choreography. *Polit. Philos. Econ.* **9**, 251–264. (doi:10.1177/1470594X0934574)

62. Hadfield GK, Weingast BR. 2012 What is law? A coordination model of the characteristics of legal order. *J. Legal Anal.* **4**, 471–514. (doi:10.1009/jla/laa008)

63. Vanderschraaf P. 1995 Convention as correlated equilibrium. *Erkenntnis* **42**, 65–87. (doi:10.1007/BF01668112)

64. Schelling TC. 1960 *The strategy of conflict*. Cambridge, MA: Harvard University Press.

65. Boyd R, Mathew S. 2021 Arbitration supports reciprocity when there are frequent perception errors. *Nat. Hum. Behav.* **5**, 596–603. (doi:10.1038/s41562-020-01008-1)

66. Yaman A, Leibo JZ, Iacca G, Wan Lee S. 2023 Data from: The emergence of division of labor through decentralized social sanctioning. Dryad Digital Repository. (doi:10.5061/dryad.dv47ms24)

67. Yaman A, Leibo JZ, Iacca G, Wan Lee S. 2023 The emergence of division of labour through decentralized social sanctioning. Figshare. (doi:10.6084/m9.figshare.c.6879576)