



In contrast to prior work on norm emergence [5, 19, 20], this paper investigates the norm emergence phenomenon in more realistic situations where the agents are physically distributed over a grid space and can use different learning algorithms. In physical environments, e.g., real-life physical interactions between humans in the society, agents are much more likely to interact with those in close physical proximity compared to others located further away. Such physical or spatial interaction constraints or biases have been well-recognized in social sciences [12] and, more recently, in the multiagent systems literature [18]. In this paper, we first focus on agents located in a grid world where they interact predominantly with agents in their physical neighborhood. The goal is to evaluate the effects of neighborhood sizes on the rate and pattern of norm emergence. Secondly we evaluate the effects of the following factors on the speed and success of emergence of norms in the agent societies.

1. Homogeneous Vs heterogeneous society of learners.
2. Uniform selection Vs non-uniform selection of opponents in neighborhood.

## 2. RELATED WORK

The need for effective norms to control agent behaviors is well-recognized in multiagent societies [3, 5]. In particular, norms are key to the efficient functioning of electronic institutions [9]. Most of the work in multiagent systems on norms, however, has centered on logic or rule-based specification and enforcement of norms [6]. Similar to these research, the work on normative, game-theoretic approach to norm derivation and enforcement also assumes centralized authority and knowledge, as well as system level goals [2, 3]. While norms can be established by centralized diktat, norms in real-life often evolve in a bottom-up manner, via “the gradual accretion of precedent” [23]. We find very little work in multiagent systems on the distributed emergence of social norms. We believe that this is an important niche research area and that effective techniques for distributed norm emergence based on local interactions and utilities can bolster the performance of open multiagent systems.

In our formulation, norms evolve as agents learn from their interactions with other agents in the society using multiagent reinforcement learning algorithms [14]. Most multiagent reinforcement learning literature involve two agents iteratively playing a stage game and the goal is to learn policies to reach preferred equilibrium [16]. Another line of research considers a large population of agents learning to play a cooperative game where the reward of each individual agent depends on the joint action of all the agents in the population [21]. The goal of the learning agent is to maximize an objective function for the entire population, the world utility.

The social learning framework we use to study norm emergence in a population [19] is somewhat different from both of these lines of research. This framework considers a potentially large population of learning agents. At each time step, however, each agent interacts with a single opponent agent chosen from the population, and the opponent changes at each interaction. The payoff received by an agent for a time step depends only on this interaction as is the case when two agents are learning to play a game. In the two-agent case, a learner can adapt and respond to the opponent’s policy. In our framework, however, the opponent changes

at each interaction. It is not clear *a priori* if the learners will converge to useful policies in this situation. Other work with similar interaction assumptions either use deterministic adaptation schemes or assume knowledge of local state of other agents [5].

## 3. SOCIAL LEARNING FRAMEWORK

The specific social learning situation for norm evolution that we consider is that of learning “rules of the road”. In particular, we will consider the problem of which side of the road to drive in and who yields if two drivers arrive at an intersection at the same time from neighboring roads <sup>2</sup>. We will represent each interaction between two drivers as a  $n$ -person,  $m$ -action stage game. These stage games typically have multiple pure strategy equilibria. In each time period, each agent is paired with another agent from the population to interact according to some interaction bias. An agent is randomly assigned to be the row or column player in any interaction. We assume that the stage game payoff matrix is known to both players, but agents cannot distinguish between other players in the population. Hence, each agent can only develop a single pair of policies, one as a row player and the other as a column player, to play against any other player from the agent population. The learning algorithm used by an agent is fixed, i.e., an intrinsic property of an agent.

When two cars arrive at an intersection, a driver will sometimes have another car on its left and sometimes on its right. These two experiences can be mapped to two different roles an agent can assume in this social dilemma scenario and corresponds to an agent playing as the row and column player respectively. Consequently, an agent has a private bimatrix: a matrix for when it is the row player, one matrix for when it is the column player. Each agent has a learning algorithm and learns independently to play. An agent is randomly assigned as the row or the column player in every interaction. Each agent develops a pair of policies, one for its role as a row player and another for its role as a column player. An agent does not know the identity of its opponent, nor its opponent’s payoff, but it can observe the action taken by the opponent (perfect but incomplete information).

We consider the agents are distributed over space where each agent is located at a grid point (see Figure 1). Each agent has a fixed location on the grid and hence a static set of neighbors. In this grid world, an agent can interact only with agents located within its neighborhood. The neighborhood of an agent is composed of all agents within a distance  $D$  of its grid location. We have used the Manhattan distance metric, i.e.,  $|x_1 - x_2| + |y_1 - y_2|$  is the distance between grid locations  $(x_1, y_1)$  and  $(x_2, y_2)$ . Different  $D$  values are used to represent different neighborhood sizes.

In each time period, each agent interacts with another agent in the society. The selection of opponents follow either of two modes:

**Uniform Selection:** Agents are randomly selected from the neighborhood of the learner. So every agent within the neighborhood is selected with uniform probability for the interaction.

<sup>2</sup>It might seem to the modern reader that “rules of the road” are always fixed by authority, but historical records show that “Society often converges on a convention first by an informal process of accretion; later it is codified into law.” [23].

---

**Algorithm 1:** Non-uniform selection of learners

---

```
initialization : neighborhood distance =  $D$ ;  
for Each player  $i \leftarrow 1$  to  $|G|$  do  
   $Sum\_dist^i = 0$ ;  
  for Each neighbor  $j$  with  $dist\ d_j^i < D$  do  
     $Sum\_dist^i = sum_{j=1}^{|Nb_i|} \frac{1}{d_j^i}$ ;  
  for Each neighbor  $j$  with  $dist\ d_j^i < D$  do  
     $Pr_j^i = \frac{\frac{1}{d_j^i}}{\sum_{j=1}^{|Nb_i|} \frac{1}{d_j^i}}$ ;
```

---

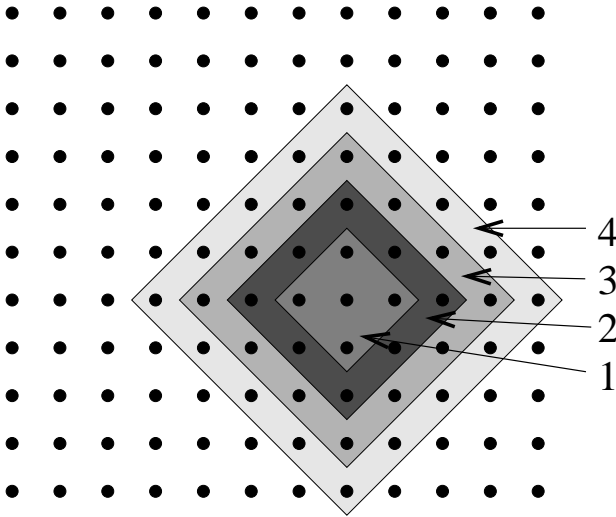


Figure 1: Agents located on a grid and allowed to interact only in a limited neighborhood.

**Non-uniform Selection:** Agents located closer to the learners are selected for an interaction with higher probability. The probability of selection  $Pr_j^i$  is computed from Algorithm 1, where  $d_j^i$  is the distance between agent  $i$  and agent  $j$  and  $|G|$  is the grid size. The physical proximity acts as a bias in the selection process.

## 4. RESULTS

We present experiments with a society of  $N$  agents placed in a  $\sqrt{N} \times \sqrt{N}$  grid. For the experiments in this paper, we use 225 agents placed on a 15 by 15 grid. We run experiments using the two-action coordination game, where agents receive high payoff for using the same action and otherwise receive a low-payoff (see Table 1). It can model the situation where agents are deciding which side of the road to drive on. Note that either action combinations (0,0) or (1,1) would work equally well. The goal is then for all agent to develop a norm of choosing the same action consistently.

	left	right
left	4, 4	-1, -1
right	-1, -1	4, 4

Table 1: Payoff in a coordination game.

A payoff of 1.5 is achieved when the agents use a uniform

distribution over their actions when playing the game. The maximum reachable payoff for this game is 4 and is obtained when the players play joint actions (L,L) or (R,R). However, as the learners use  $\epsilon$ -greedy scheme, the maximum payoff value cannot be reached. We recognize that a norm has emerged when the average payoff reaches 3.5.

Though some aspects of results from our simulated agent society can be transferred to human situations (with additional mechanisms such as empowering agents with sanction schemes), our results are targeted towards a better understanding of how to develop self-adaptive agent societies. Accordingly, we make no claims about using our results to predict human social behavior.

### 4.1 Learning Algorithms Used

We use four different learning algorithms for norm emergence: Q-Learning [22] with  $\epsilon$ -greedy exploration with learning rate  $\alpha = 0.1$  and probability of exploration  $\epsilon = 0.1$ , WoLF-PHC (Win or Learn Fast-Policy Hill Climbing) [4] with learning rate  $\alpha = 0.1$ , Fictitious Play (FP) [8] with rate of learning 0.1 and Highest Cumulative Reward (HCR) [5, 20]. Q-Learning is well suited for repeated games against unknown opponents and is widely used in multiagent systems. WoLF-PHC can learn mixed strategies and is guaranteed to converge to a Nash equilibrium of the repeated game in 2-person 2-action games. Fictitious Play (FP) is the basic learning approach widely studied in the game theory literature. An FP player uses the historical frequency count of its opponents' past actions and tries to maximize expected payoff by playing the best response to the observed mixed strategy. HCR is a deterministic scheme that uses finite memory of size  $M$  and chooses the action that fetched the maximum cumulative value over the last  $M$  interactions. We will also present some experiments when a small minority of the agent population are non-learners, i.e., they play fixed strategies.

### 4.2 Effect of neighborhood size

For the first set of experiments, all agents use the WoLF-PHC learning algorithm. We have experimented by varying the neighborhood size and observed the corresponding effects on the rate of convergence of the learning agents. We present results from experiments with both uniform and non-uniform selection to understand the effect of the neighborhood size on learning of agents is observed (see Figure 2). We have tested with four neighborhood distances,  $D$  (1, 5, 10, and 15), for each agent. When  $D = 1$  only an adjacent agent is a neighbor (there are 4 neighbors in that case). The computation of number of neighbors should follow the recurrence relation  $D_i = 4 \cdot i + D_{i-1}$ , where  $D_1 = 4$ . When the distance is 15, every agent is a neighbor of every agent.

We present in Figures 2(a) and 2(b) the dynamics of the average payoff over a run of populations of Q-learning and WoLF-PHC learners respectively when all agents are learning concurrently. We observe that the smaller the neighborhood distance, the faster the emergence of a norm. It is also interesting to note from Figure 2(a) and 2(b) that the learning rate for non-uniform opponent selection falls in between the smallest ( $D = 1$ ) and the larger group ( $D = 5, 10, 15$ ) of neighborhood sizes. Norm emergence in society with Non-uniform selection does not depend on neighborhood  $D$ .

When an agent has four neighbors ( $D = 1$ ), the agents

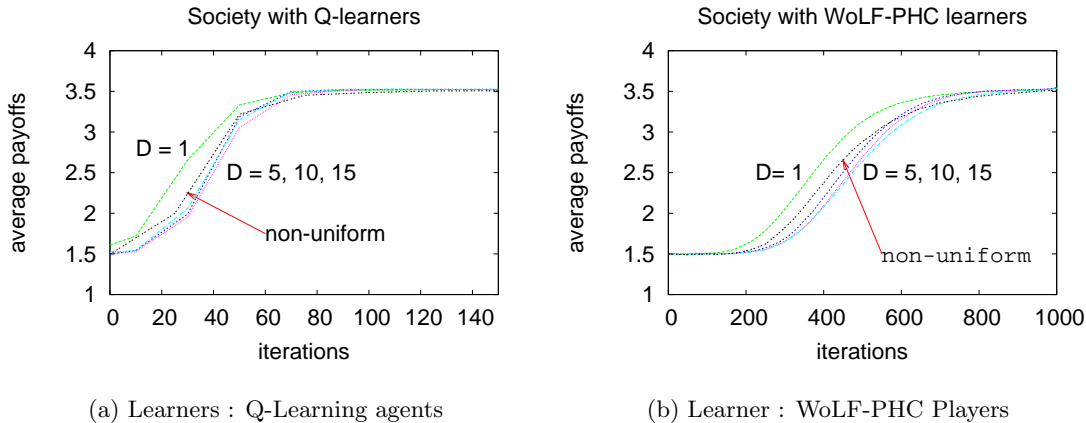


Figure 2: Learning of homogeneous agents.

learn to coordinate faster by driving on the same side of the road than when it has 35 or 99 neighbors ( $D = 5$  and 10 respectively). For a given number of iterations, the agents interact more often with a particular neighbors for smaller neighborhoods. This means that the impact an agent has on another agent is larger when the neighborhood size is small. In addition, an agent with few neighbors will encounter few different behaviors from its neighbors, and it is *a priori* easier to coordinate with a small set of agents rather than a larger one. As the neighborhood distance increases, an agent has to coordinate with many other agents, and in addition, interactions between two particular neighbors in the network become less frequent. This decreasing interaction frequency between pairs of learners increases the time for exploration of the behavior space and thereby influences the learning patterns of the agents in the network. This problem is exaggerated when every agent is everyone’s neighbor ( $D = 15$ ) which further reduces the rate of learning.

When the entire population uses the same learning algorithm, from Figures 3(a), 3(b) and 3(c) it is clearly observed that population of Q-Learners is fastest to evolve a convention ( $\approx 100$  iterations), followed by the society of WoLF ( $\approx 1000$  iterations) and FP ( $\approx 50000$  iterations) for selected values of the neighborhood distance  $D$ .

Figure 4 represents, for largest ( $D = 15$ ) and smallest ( $D = 1$ ) neighborhoods, the policy of each agent in the population at different iterations in a single run. Each cell represents the policy of an agent: the darker it is, the higher the probability of driving on the left, whereas lighter colors denote higher probability of driving on the right. When a cell is completely dark, or white, it means that the learning algorithm of the agent has converged. In the particular runs we present, the norm of “driving on the right” emerges (over different runs “driving on the left” and “driving on the right” norms were evolved in roughly the same number of runs). At iteration 145, the agents are exploring and are receiving low payoff (see corresponding payoff dynamics in Figure 2). At iteration 355, for  $D = 1$ , we are close to the inflection point for the curve of the payoff dynamics: the agents start to favor one norm over the other. For  $D = 15$ , however, there is a lesser bias favoring one action. We can

see that, on the average, the snapshot for  $D = 1$  is lighter than that with  $D = 15$ . At iteration 480, we can see that many more agents have converged for the smallest compared to the largest neighborhood. So smaller neighborhoods induce faster learning among agents on a grid.

The above effect of agent neighborhood size on learning rate was somewhat surprising. A priori, it was unclear whether smaller neighborhoods will engender divergent norms to initially form over the agent space, which would subsequently delay the convergence of the population to a consistent norm. Such effects, however, were overshadowed by the effects of increased interaction frequencies between neighbors in our framework.

### 4.3 Influence of non-learning agents

So far, we have observed that all norms with equal payoffs were evolved roughly with the same frequency over multiple runs. This is expected because the payoff matrix for the coordination game (Table 1) has no preference for one norm over the other. Extraneous effects, however, can bias a society of learners towards a particular norm. For example, some agents may not have learning capabilities and always choose a predetermined action. We now study the influence of agents playing a fixed pure strategy (FPS agent) on the emergence of a norm. We are interested in the effect of multiple pure strategy players with the same or different fixed strategies.

We do not preclude the possibility of multiple coexistent norms in sufficiently isolated populations [19]. Without sufficient isolation, stochastic biases introduce enough differential to lead to norm conformance. The norm adopted with larger number of FPS agents is more likely to emerge. Even with a few FPS agents, for a given agent, most of its neighbors are learners and influences this agent’s eventual norm selection.

#### 4.3.1 Non-learners use same strategy

In the first experiment, we replace some learning agents by FPS agents and we study the effect of the speed of emergence of a norm. When there are no FPS agents, as the learners explore early in the run, they should encounter each joint ac-







