Convention Emergence with Congested Resources

Priel Levy[1*] and Nathan Griffiths[2]

[1]Bar-Ilan University, Israel.
[2]University of Warwick,UK.

*Corresponding author(s). E-mail(s): priel.levy@biu.ac.il;
Contributing authors: nathan.griffiths@warwick.ac.uk;

Abstract

Norms and conventions enable coordination in populations of agents by establishing patterns of behaviour, which can emerge as agents interact with their environment and each other. Previous research on norm emergence typically considers pairwise interactions, where agents' rewards are endogenously determined. In many real-life domains, however, individuals do not interact with one other directly, but with their environment, and the resources associated with actions are often congested. Thus, agents' rewards are exogenously determined as a function of others' actions and the environment. In this paper, we propose a framework to represent this setting by: (i) introducing congested actions; and (ii) adding a central authority, that is able to manipulate agents' rewards. Agents are heterogeneous in terms of their reward functions, and learn over time, enabling norms to emerge. We illustrate the framework using transport modality choice as a simple scenario, and investigate the effect of representative initial, late and temporary manipulations on the emergent norms.

Keywords: Norm emergence, Conventions, Congestion games

1 Introduction

Norms and conventions enable populations of agents to interact in complex environments, by establishing patterns of behaviour that are beneficial, and

enabling coordination. Norms are viewed as equilibria, in which the interacting agents act in some expected way [1], either choosing the same action (in coordination games) or different actions (in anti-coordination games).[1] Existing norm emergence research often focuses on the population level phenomena that result from pairwise interactions between individual agents [2, 3].

In many real-life scenarios, however, individuals do not interact with one another through pairwise interactions, but instead select actions (which have a cost) according to some individual strategy, and receive rewards which are, at least in part, determined by the action choices of others. Thus, individuals interact with their environment, rather than directly with others. Furthermore, resources are often *congested*, meaning that an individual's valuation of a resource (and consequently their reward) is not endogenously determined, but rather depends on the number of others using the resource (i.e., it is a function of others' actions, not only in terms of agents' joint actions determining the outcome, but also the level of reward) [4]. This congestion effect manifests in many economic and social environments, where individuals 'compete' for some resource, with such congestion games being widely studied from a game theoretic perspective. However, in such environments, while it is often desirable to establish norms to facilitate coordination, the number of individuals who can simultaneously benefit from choosing a particular action may be limited. Examples of such environments include transport modality or route choices, bandwidth or compute allocation, and public service consumption, in which rewards are reduced if the capacity of the resource associated with an action is exceeded.

One important aspect of many economic and social environments, in addition to congested resources and which is not accounted for by traditional congestion games, is that of an authority figure with preferences about the distribution of individuals' action choices, and having some (but not complete) control over the payoffs they receive. While norms often emerge in the absence of an authority, the authority may manipulate the rewards of individuals or groups of individuals in order to 'nudge' the system towards a particular state [5]. Consider, for example, the scenario of commuters choosing a transport modality (e.g., car, bus or walk) and route. Such choices are made individually, but rewards are determined by the current state of the environment and are affected both by others' choices and the city authority. Many individuals choosing the bus may result in overcrowding and low rewards, but few individuals choosing the bus may cause the city authority to increase prices. Moreover, the city authority may have preferences in terms of reducing car use and increasing active travel, and so may impose charges for car use or offer rewards for walking. In London, for example, the city authority facilitates bike loans, encourages employers to offer financial or holiday incentives for employees who do not drive, and imposes charges for private car use.[2]

[1] Note that such equilibria are not necessarily Nash equilibria.
[2] See, for example, http://content.tfl.gov.uk/tfl-active-recovery-toolkit.pdf

The way that a norm is evaluated depends on the perspective and associated preferences, either that of an individual agent or of the authority. Behaviours and norms that are beneficial from one perspective may not be beneficial from another, a factor not typically considered in congestion games. Individual agents evaluate a norm by considering its impact on their own rewards, which are influenced by others' actions. Alternatively, the authority may evaluate it from a system level, considering whether it is appropriate for the system as a whole, potentially considering factors such as the long-term or indirect effects of a norm [2]. While the authority may aim to maximise social welfare (i.e., the sum of individuals' rewards), it may instead have its own preferences regarding resource utilisation, namely, which actions are selected and by what proportion of agents.

To illustrate this difference in perspectives, consider the example of transport modality choice and the possible norms of driving and walking. From an individual's perspective, choosing to drive may broadly result in two possible scenarios: either a high individual reward (if relatively few others choose to drive and so traffic is light), or a low individual reward (if many others choose to drive, resulting in congestion). Similarly, from an individual's perspective, walking may have a medium reward, regardless of others' transport choices. From the authority's perspective, assuming a preference of decreasing car use and encouraging active travel, the situation is independent of the individual rewards: there is higher value to the authority when fewer individuals choose to drive, and the highest value would be for all individuals to walk and not drive. Other norms may be more complex, for example, individuals may prefer to be on buses with a reasonable number of other passengers (for perceived personal safety), but not so many that they have to sit directly next to a fellow traveller, while the authority might prefer the bus to be full to capacity (i.e., full utilisation of the resource).

In this paper, we propose a framework that: (i) introduces congested actions into the norm emergence setting; (ii) adds a central authority to such congestion games, such that the authority is able to manipulate (but not fully control) the rewards of agents and groups of agents; and (iii) accounts for the different perspectives of the agents and authority in terms of their preferences.[3] We illustrate the framework using a simplified transport modality choice example, and show the impact of manipulations on the emergent norms in the population. We also investigate whether established norms can be destabilised and replaced with other norms. This situation may arise whenever the population converges to a sub-optimal (or non-optimal) but consistent norm from the authority's perspective, or when the established norm was previously optimal but no longer is due to some external conditions. In such cases, the authority may desire to encourage the population away from the established norm to a more beneficial one. We consider three types of interventions for achieving

[3]This paper is an extended version of [6], and adds consideration of late interventions and temporary interventions, in addition to adding depth to the discussion.

destabilisation, based on those explored previously in coordination game settings (see Section 2 for more details), namely: (i) initial interventions; (ii) late interventions; and (iii) temporary late interventions. Initial interventions take place at the start of the life of the system, late interventions are those which take place after a norm has emerged, and temporary late interventions occur for a limited period of time after a norm has emerged, enabling the cost of an intervention to be managed.

The reminder of this paper is structured as follows. In Section 2 we introduce the related work on norm emergence and problems similar to our unique setting. Section 3 presents our model of norm emergence with congested actions. We describe our experimental setting in Section 4, and present our results on initial, late and temporary late interventions in Section 5. Finally, in Section 6 we present our conclusions.

2 Related Work

Norm emergence has been widely studied in the context of agents who learn (or reproduce) based on the rewards received from their interactions with others. Such interactions typically take the form of an n-player m-action game, in which each agent's reward is a discrete function of others' actions and is determined according to a payoff matrix, which is typically common knowledge. Most literature on norm emergence either models cooperation using the Prisoner's Dilemma [7–9] or learning to choose common actions in a coordination game [2, 3, 10–14]. Such work typically focuses on pairwise interactions where agents select from two possible actions, i.e., $n = 2$ and $m = 2$. In this paper, we consider norm emergence from a more general perspective, in which individual agents select from a wider set of actions ($m > 2$) and receive rewards which are only partially determined by other agents' choices (i.e., interactions are not pairwise). While some studies have considered the more general setting of $n \geq 2$ and $m \geq 2$, they have typically focused on cases with small numbers of agents and actions per interaction [15, 16] and have not considered congested resources or the inclusion of an authority figure. Other work has considered the impact of large action spaces [17], but only from the perspective of agents learning common actions, rather than the more general setting.

Our setting is similar to the El-Farol Bar Problem (EFBP), a well-known congestion game which shares some characteristics with norm emergence [18–20]. In the EFBP, a group of n agents, representing people, independently decide whether to visit a bar on a certain evening, with the most enjoyable visits being when the bar is not too crowded, i.e., when the number of visitors is less than some (unknown) threshold. Choices are unaffected by previous visits and there is no communication or information on others' choices. Each agent only knows its own choice and the subsequent reward. The EFBP illustrates the key features of our setting, namely that agents compete for a resource (space in the bar), agents are rational (their rewards provide information on attendance, and they use this strategically), and there is limited information

(agents do not know others' strategies, but their rewards provide information on other agents' actions). In this paper, we introduce an authority figure that can influence agents' rewards, which can be viewed as adding a bar owner to the EFBP, who is able to change the available space in the bar (e.g., by opening or closing rooms), and considering rewards from both the owner and customer perspectives.

Our setting is also similar to the Multi-Armed Bandit (MAB) problem, where each individual sequentially pulls one of several arms (representing choosing actions), with each pull resulting in a reward from some distribution (which is unknown to the agent and may differ over time) [21–24]. While we could represent rewards in a MAB setting as being influenced by other agents' action choices through the use of a non-stationary distribution, such a representation is not intuitive. Moreover, it is less clear how an authority that can reward or penalise certain action choices might be introduced into the MAB setting.

Various forms of intervention have been considered to encourage norm emergence. One of the earliest studies was Axelrod's Norms Game [25], in which a population of agents repeatedly make decisions about whether to comply with a desired norm or defect, and whether to punish those who are seen to defect. More informed punishment methods, such as experience based punishment, have been developed for the Norms Game [26, 27], but these are peer-based and do not consider an authority figure. Other approaches have considered incentives and sanctions [28, 29], or the use of non-learning fixed strategy agents (i.e., that always choose the same action regardless of others' choices) [12, 16, 30, 31] to influence the emergence of norms, but typically in settings where rewards are a direct function of the choices of those involved in an interaction, and so can be represented as a simple payoff matrix. In this paper, we introduce an authority that is able to influence agents' rewards, by changing the costs of performing actions and the sensitivities of agents to the results. Similar manipulations have been considered in non-stationary MABs, for example adding constant noise [32], using adversarial (arbitrary) rewards [33–36], varying the expected values of the reward distributions [37], or assuming arms are contextual (i.e., no prior knowledge about the arms exists except for some historical data or some action features) [38–40]. To the best of our knowledge, we are the first to consider such a perspective in the context of norm emergence with congested actions, where rewards are partially determined by the actions of others.

Much of the existing literature investigating the impact of interventions on norm emergence considers the case where such interventions are applied at the beginning of interactions (i.e., before a norm emerges), and there has been relatively little investigation of interventions to change or destabilise already established norms or conventions. There has been some investigation of how fixed strategy agents, known as Intervention Agents (IAs), can be used to destabilise an already established convention and encourage emergence to an alternative desired strategy, by placing the IAs at topologically influential

locations, in a range of networks, both static [5, 41] and dynamic [42, 43]. Interventions using IAs have been shown to be able to influence norm and convention emergence, both when applied initially in a system and after a norm or convention has emerged. An important concept with respect to destabilisation in convention emergence is meta-stable subconventions, introduced by [44]. Meta-stable subconventions are conventions that exist within subsets of the population, that are persistent due to their stability. Several works describe methods for destabilising meta-stable subconventions by identifying and targeting particular topological structures [13, 44–46], showing that these meta-stable subconventions can hinder the emergence of a global convention. In this paper, we consider destabilisation in the context of a population selecting congested actions.

Interventions as a means to manipulate conventions have typically persisted indefinitely in a system once applied. There has been relatively little consideration of temporary interventions, i.e., inserting interventions within the population for a finite time rather than permanently. However, it has been shown in the context of coordination games that temporary interventions can be used to elicit the same level of change in a population as persistent interventions [5]. We explore, in this paper, whether such temporary interventions are also effective in the congested action case.

3 Modelling Norm Emergence with Congested Actions

We consider a population of n agents, or players, $P = \{p_1, ..., p_n\}$ and a single centralized authority, ψ. Agents are heterogeneous and may be of different types, or belong to different groups, which determine their preferences over actions and influence the rewards they receive. Agents interact with their environment by playing a repeated game in which they select an action, or option, o from a set of $m \geq 2$ alternatives, $O = \{o_1, ..., o_m\}$. At a given time t each agent p simultaneously interacts with the environment by choosing an action $o_{p,t}$ for which it receives a reward (which could be positive or negative). For simplicity, unless it is ambiguous, we assume that t refers to the current time and write o_p for the action selected by p.

Each action is viewed as requiring some resource, and we assume that resources are *congested*, meaning that there is a limit to how many agents can simultaneously use the resource and receive maximal individual reward. Let $\omega_o^* \in (0, 1]$ denote the maximum proportion of agents in the population who can simultaneously select o and receive maximal reward, i.e., ω_o^* represents the *capacity* of the resource associated with action o. Thus, if $\omega_o^* = 1/n$ then only a single agent can receive maximal reward for selecting the action at any time, while if $\omega_o^* = 1$ then all agents would receive the maximal reward if they simultaneously selected the action. We assume that agents are fully rational,

self-interested and act independently and simultaneously at any given time, without any knowledge of others' choices or strategies.

3.1 Actions and Congested Resources

The reward an agent receives for selecting an action is determined both by its individual preferences, represented by its type, and the value associated with the action, which is a function of the environment and others' action choices. We use the terms *value* and *reward* respectively to distinguish between the benefit resulting from an action in the current setting, independent of the agent's preferences, and the benefit an agent receives taking into account its preferences. Let $v_{o,t}$ denote the *value* that is associated with selecting action o in time step t. Again, unless it is ambiguous in the current context, we assume that t refers to the current time, and so we simply write v_o. For generality, we assume that the value of an action o is determined by some valuation function $\mathcal{V}_o(\omega_o)$ which maps the proportion of agents, ω_o, selecting action o to the value of the action,

$$v_o = \mathcal{V}_o(\omega_o). \tag{1}$$

The proportion of agents, $\omega_o \in [0, 1]$, who select action o in the current time step, is defined as

$$\omega_o = \frac{\mid \{p : p \in P \wedge o_p = o\} \mid}{\mid P \mid} \tag{2}$$

where o_p is the action selected by p.

Some actions might not be congested, or have sufficient capacity that $\omega_o^* = 1$, and therefore the value v_o associated with such actions is independent of the proportion of agents selecting that action, so

$$\mathcal{V}_o(\omega_o) = y \tag{3}$$

meaning that the value associated with o has a constant value of y.

For actions associated with congested resources, the capacity of the resource plays an important role in determining the value of such actions. There are two cases: the resource associated with action o is *in-capacity* if $\omega_o \leq \omega_o^*$, and it is *over-capacity* if $\omega_o > \omega_o^*$. For such actions, we assume that $\mathcal{V}_o(\omega_o)$ appropriately reflects the valuation function in both situations.

In the simplest case, the valuation function can be modelled in the same manner for the in- and over-capacity cases. For example, we might use a unimodal Normal function with mean $\mu_o = \omega_o^*$ and variance σ_o^2 (with σ_o being the standard deviation), such that agents receive the maximal value when the resource is at capacity, i.e., $\omega_o = \omega_o^*$,

$$\mathcal{V}_o(\omega_o) = \frac{1}{\sqrt{2\pi}\omega_o^*} e^{\frac{-(\omega_o - \omega_o^*)^2}{2\sigma_o^2}}. \tag{4}$$

In the more general case, the distributions defining the value for the in-capacity and over-capacity cases may be different. For example, if we use a uni-modal Normal distribution for both cases this could be represented using σ_o^2 and $\sigma_o'^2$ to represent the variance for the in- and over-capacity cases respectively (noting that the mean is fixed at the capacity, ω_o^*),

$$\mathcal{V}_o(\omega_o) = \begin{cases} \frac{1}{\sqrt{2\pi}\omega_o^*} e^{\frac{-(\omega_o-\omega_o^*)^2}{2\sigma_o^2}}, & \text{if } \omega_o \leq \omega_o^* \\ \frac{1}{\sqrt{2\pi}\omega_o^*} e^{\frac{-(\omega_o-\omega_o^*)^2}{2\sigma_o'^2}}, & \text{otherwise, i.e., } \omega_o > \omega_o^*. \end{cases} \tag{5}$$

We might also model the in- or over-capacity cases using constant values if they do not depend on ω_o.

3.2 Agent Types: Mapping Values to Rewards

Each individual agent's reward from choosing action o is a function of the value, v_o, of the action and the agent's type. We assume that the agents are partitioned into a set G of disjoint *types*, or *groups*, $G = \{g_1, g_2, \ldots, g_l\}$ such that $\forall g_i \in G, g_i \subseteq P$, $g_1 \cup g_2 \cup \ldots \cup g_l = P$ and $\forall g_i, g_j \in G, g_i \cap g_j = \emptyset$. We use agent types to represent that the cost and relative reward associated with a given action may vary for different agents. For example, in the context of selecting transport modalities in a city, the relative cost of a congestion charge for car use may be low for wealthy individuals compared to those on low incomes, while the relative reward of using a low polluting mode of transport may be higher for individuals who are concerned about environmental issues.

We model such differences by associating each agent type g with a *cost*, $c_{o,g}$, and *sensitivity*, $s_{o,g}$, for each action o. For simplicity, we assume that the cost and sensitivity of each action for each agent type is predetermined and fixed over time, unless subject to manipulation by the authority. Rewards are defined at the group level, such that any agent in group g will receive reward $r_{o,g}$ for selecting action o, determined by multiplying the value v_o of the action by the corresponding sensitivity and subtracting the corresponding cost,

$$r_{o,g} = s_{o,g} \cdot v_o - c_{o,g} \tag{6}$$

where $s_{o,g}$ represents the sensitivity for action o, and $c_{o,g}$ the cost of action o for group g. Thus, the reward $r_{o,p}$ that an individual agent p receives for selecting action o is determined by p's group g, namely,

$$r_{o,p} = r_{o,g} \ni p \in g. \tag{7}$$

3.3 Authority Preferences and Influence on Rewards

We assume that the system contains an authority that has preferences over the state of the system in terms of the proportions of agents selecting each of the actions, and is able to influence the rewards agents receive by manipulating

the sensitivities and costs of agent types. The authority's preferences with respect to a given action o are determined by the overall proportion of agents it desires to choose the action, $\hat{\omega}_o$, along with a utility function $\mathcal{U}_o(\omega_o)$ mapping the proportion of agents who choose the action, ω_o, to the utility from the authority's perspective. We assume that the utility function accounts for the potentially different distributions in the cases where $\omega_o \leq \hat{\omega}_o$ and $\omega_o > \hat{\omega}_o$. For clarity, we use the term *utility* for the authority's perspective to distinguish from individual agent's *rewards* and action *values*. The utility u_o of action o from the authority's perspective is therefore,

$$u_o = \mathcal{U}_o(\omega_o). \tag{8}$$

The authority's utility function for a given action can be modelled in a similar manner to the valuation functions in Equations 3, 4 and 5, using a constant value ($\hat{y}$), a variance ($\hat{\sigma}_o^2$), or a pair of variance values ($\hat{\sigma}_o^2$ and $\hat{\sigma}_o'^2$) for the in- and over-capacity cases.

The overall utility u to the authority of the current action choices of the population is simply the aggregation of the utility (from the authority's perspective) of each individual agent's choice,

$$u = \sum_{p \in P} u_{o,p} \tag{9}$$

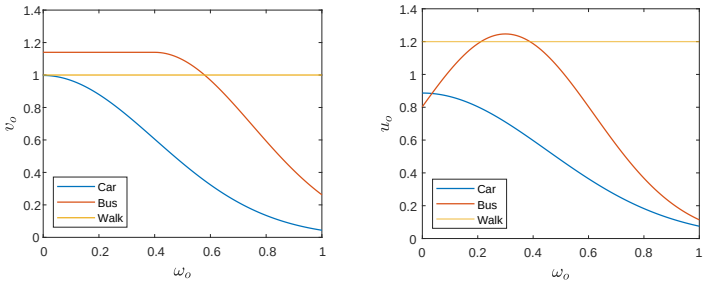
where $u_{o,p}$ represents the utility to the authority of agent p selecting action o.

While the authority is not able to directly control the action choices of agents, or the values associated with those actions, we assume that it is able to exert influence over the rewards agents receive, which in turn may cause agents to adopt different strategies. There are two methods we consider through which the authority can affect rewards, namely, modifying the cost or modifying the sensitivity associated with an action for a group of agents. Thus, the authority is able to replace the default sensitivity $s_{o,g}$ or cost $c_{o,g}$, with respect to group g for action o, with modified values $\tilde{s}_{o,g}$ and $\tilde{c}_{o,g}$ respectively. The reward is then calculated using these updated values in Equation 6, i.e.,

$$r_{o,g} = \tilde{s}_{o,g} \cdot v_o - \tilde{c}_{o,g}. \tag{10}$$

3.4 Agent Learning

Norms can emerge through social learning [47], such that an individual's estimate of the desirability of each possible action is affected by others' actions in the environment. To illustrate our framework, we assume that agents use Q-learning [48], since this has been shown as effective for norm emergence [12, 15, 49–52], although other methods such as HCR [1] or WoLF-PHC [53] can also be used. For each action $o \in O$, each agent maintains a Q-value that estimates the benefit of choosing that action. The Q-values are initially set to zero and are updated based on the rewards received. Whenever



(a) Agents (b) Authority

Fig. 1: Value functions of (a) agents and (b) authority.

agent p selects action o and receives reward $r_{o,p}$, it will update its Q-value for o using,

$$Q(o) \leftarrow (1-\alpha)Q(o) + \alpha\Big(r_{o,p} + \gamma \max_{o'} Q(o')\Big) \tag{11}$$

where $0 < \alpha \leq 1$ is the learning rate and γ is the discount factor. We assume that agents use ϵ-greedy action selection ($0 < \epsilon < 1$), such that an agent selects a random action with probability ϵ, and with probability $1 - \epsilon$ selects the action with the highest Q-value.

4 Experimental Methodology

In this section, we describe our simulation and experimental methodology using transport modality choice as an illustrative example. The environment contains $n = 3000$ agents who select from actions $O = \{Car, Bus, Walk\}$, i.e., $m = 3$, representing the available transport modalities. In each iteration (i.e., time step) every agent selects an action, receives a reward, and updates its Q-values.[4] We ran the simulation for $10,000$ iterations and averaged our results over 10 runs. We used $\epsilon = 0.05$, $\alpha = 0.1$ and $\gamma = 0.75$ as representative values for the exploration rate, learning rate and discount factor respectively.

We assume that the *Walk* action is not associated with a congested resource, and so for simplicity we define its value as $\mathcal{V}_{Walk}(\omega_{Walk}) = 1$, while both the *Car* and *Bus* actions are assumed to be congested. We define $\mathcal{V}_{Car}(\omega_{Car})$ using a uni-modal Normal function with $\omega^*_{Car} = 1/n$ and $\sigma'^2_{Car} = 0.4$ (see Equation 4), which represents that an agent obtains the highest value when no other agents choose *Car*. We represent $\mathcal{V}_{Bus}(\omega_{Bus})$ as

[4] We have also run experiments with SARSA to determine whether on-policy learning has an impact. The results have the same form as those discussed in Section 5, and so for reasons of space and to avoid repetition we focus on Q-learning in this paper.

Table 1: Action costs and sensitivities for each group and each action.

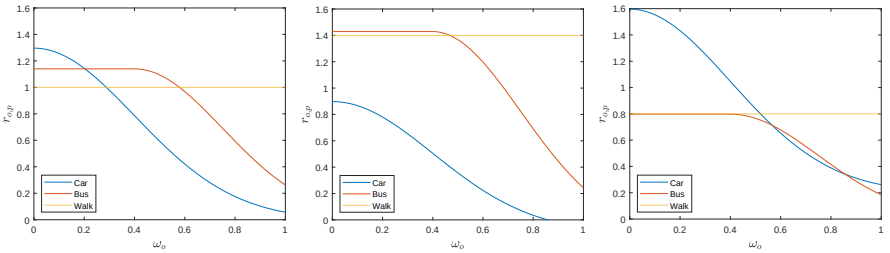
	g_1		g_2		g_3	
	c_{o,g_1}	s_{o,g_1}	c_{o,g_2}	s_{o,g_2}	c_{o,g_3}	s_{o,g_3}
$o = Car$	0	1.3	0.1	1	−0.2	1.4
$o = Bus$	0	1	0.11	1.35	0	0.7
$o = Walk$	0	1	0	1.4	0	0.8

$$\mathcal{V}_{Bus}(\omega_{Bus}) = \begin{cases} y_{Bus}, & \text{if } \omega_{Bus} \leq \omega^*_{Bus} \\ \frac{1}{\sqrt{2\pi}\omega^*_{Bus}} e^{\frac{-(\omega_{Bus}-\omega^*_{Bus})^2}{2\sigma'^2_{Bus}}}, & \text{otherwise, i.e., } \omega_{Bus} > \omega^*_{Bus}. \end{cases} \tag{12}$$

where $\omega^*_{Bus} = 0.4$, $y_{Bus} = 1.14$, and $\sigma'^2_{Bus} = 0.35$, meaning that for the in-capacity case the value is constant, while the over-capacity case is modelled as a uni-model Normal function. The parameters of these value functions are for illustration, and the resulting value functions are shown in Figure 1(a), which shows that: (i) the value of choosing *Walk* is independent of others' choices; (ii) the value of choosing *Car* reduces as more agents select the *Car* action, representing an increase in traffic and journey time; and (iii) an agent obtains the highest reward when choosing *Bus*, provided that only a moderate proportion of others make the same choice, with the value reducing when higher proportions cause the resource to be over-capacity. From the agents' perspective, the highest (social) utility possible is for 60% of the agents to choose *Bus* and 40% to choose *Walk*.

We assume that the authority prefers fewer agents to select *Car*, more agents to select *Walk*, and that there is some ideal preferred proportion of agents who select *Bus*. This represents a desire to reduce car use, increase active travel, and ensure that investment in providing a bus service is fully utilised (e.g., such a service may be required to cater for groups of individuals with restricted mobility, who might have very high costs associated with walking, meaning that $c_{Walk,g}$ has a high value). From the authority's perspective we represent the utility for the actions as: $\mathcal{U}_{Walk}(\omega_{Walk}) = 1.2$, with $\mathcal{U}_{Car}(\omega_{Car})$ and $\mathcal{U}_{Bus}(\omega_{Bus})$ being uni-modal Normal functions with $\hat{\omega}_{Car} = 0$, $\hat{\sigma}'^2_{Car} = 0.45$, $\hat{\omega}_{Bus} = 0.3$, and $\hat{\sigma}^2_{Bus} = 0.32$, which are illustrated in Figure 1(b). For the authority, the highest utility occurs when 17% of the agents choose *Bus* and the remainder choose *Walk*.

We divide the agents into three equal size groups, $G = \{g_1, g_2, g_3\}$, where g_1 corresponds to a baseline agent type, g_2 has strong environmental concerns, and g_3 represents affluent agents. For simplicity, since groups are disjoint, we do not consider affluent agents who also have strong environmental concerns, as this would require an additional group to be defined. An agent's group defines the cost and sensitivity associated with each action which, along with the proportion of other agents choosing the action (in the case of congested actions), determines the reward received for selecting an action.



(a) g_1 (b) g_2 (c) g_3

Fig. 2: Reward functions for each group.

The costs and sensitivities associated with each action for each group in our simulation are given in Table 1 where $c_{o,g}$ and $s_{o,g}$ denote the cost and sensitivity associated with action o for group g, respectively. These costs and sensitivities determine the shape of the reward function (see Equations 6 and 7) for each group, as illustrated in Figure 2. These reward functions are not intended to be realistic models of the costs and sensitivities associated with the actions for each group, but rather are intended to illustrate how our framework models congested resources. The baseline group g_1 (Figure 2(a)) receives the highest reward when choosing *Car* if few other agents make the same choice. When more agents choose *Car* the reward decreases. If a high proportion of agents choose *Car*, then agents of type g_1 can obtain a high reward by choosing *Bus*, again provided that not too many others make the same choice. The reward associated with the *Walk* action does not depend on other agents' choices, and if a high proportion of agents choose *Car* or *Bus*, then *Walk* provides the highest reward. Group g_2 (Figure 2(b)) receives significantly higher rewards from the *Walk* and *Bus* actions, and lower rewards from *Car*, reflecting their environmental concerns. Again, for group g_2 the reward of *Bus* decreases if a high proportion of agents choose *Bus*. Finally, the affluent agents in g_3 (Figure 2(c)) receive the highest reward from selecting *Car*, as they have both higher sensitivity and lower (relative) costs associated with this action, provided that only a moderate number of others choose *Car*, otherwise, if many agents choose *Car* then *Walk* gives the highest reward.

To illustrate our framework, we performed several experiments, the results of which are presented in the following section. As a baseline, we start by considering the effect of the different costs and sensitivities associated with each group, with no interventions from the authority. We then consider three kinds of initial interventions. First, we introduce fixed strategy agents into the population at the beginning of the simulation, and show that this is not an effective intervention in our setting. Second, we investigate the impact of manipulating the sensitivity to different actions, allowing us to model, for example, the effect of a targeted behavioural change intervention. Third, we investigate the effect

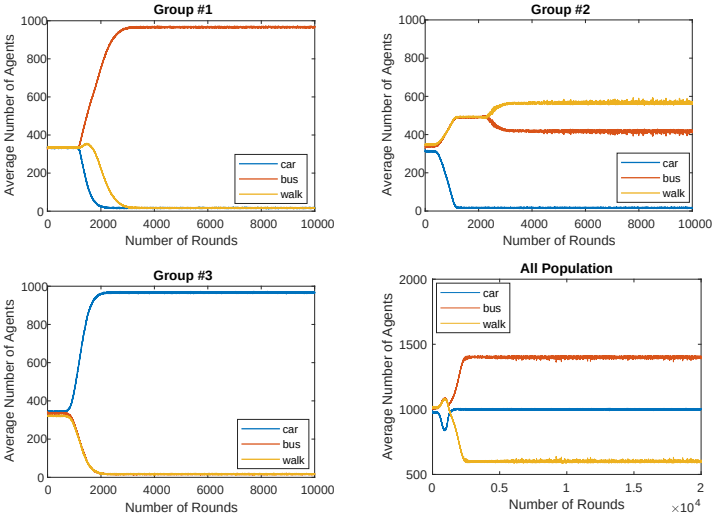


Fig. 3: Average number of agents choosing each of the actions against time, in the baseline setting.

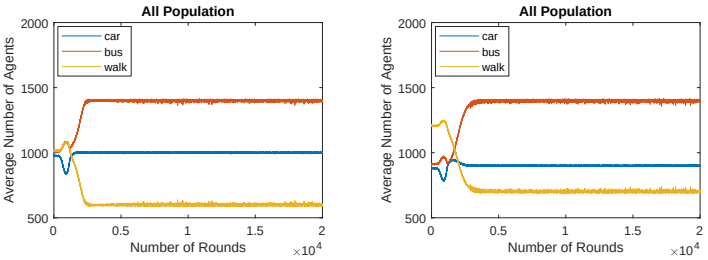
of modifying the cost associated with different actions, i.e., we consider different values of $\tilde{c}_{o,g}$. This allows us to model interventions such as means-tested charging for private car use, or exemption from congestion charges for certain groups. We then consider these interventions (namely, fixed strategy agents along with manipulating sensitivity and cost) in the cases of late interventions and temporary late interventions.

5 Results

All results are averaged over 10 runs for each configuration, with $n = 3{,}000$ agents. In this section, we consider agents' behaviour under different interventions assuming, for illustration, that the authority's aim is to encourage the population to choose *Walk*. To illustrate our framework, we discuss representative manipulations, however other alternative manipulations are possible. We begin by considering the baseline setting with no interventions (Section 5.1). We then consider three types of interventions, namely: (i) fixed strategy agents; (ii) manipulating agents' sensitivities; and (iii) manipulating agents' costs. These interventions are applied either at the start of the simulation (Section 5.2), after a norm has emerged (Section 5.3), or for a temporary period after a norm has emerged (Section 5.4).

5.1 Baseline Setting

As a baseline, we begin by considering agents' behaviour without any intervention. Figure 3 shows that agents' choices in each group are in accordance



(a) Baseline setting (b) Fixed-strategy agents

Fig. 4: Average number of agents choosing each of the actions against time in (a) our baseline setting and (b) with 10% fixed-strategy agents.

with their corresponding reward functions given in Figure 2. A common measure of norm emergence is the Kittock criteria [54], where a norm is considered to emerge if some proportion of the population (often 90%) adopts a particular action. While this is an effective measure in populations of homogeneous agents playing coordination or Prisoner's Dilemma games, in our setting of heterogeneous agents groups and congested actions, we do not typically expect to see such large, population wide, adoption. Therefore, for simplicity, rather than specifying a convergence threshold we consider any dominant action in a group (or the population) as being a norm. Thus, we see the norms emerge of choosing *Bus* in group g_1, *Walk* in g_2, and *Car* in g_3, with *Bus* being the overall population norm.

5.2 Initial Interventions

5.2.1 Fixed-Strategy Agents

Fixed-strategy agents perform the same action regardless of others' choices, and small numbers of such agents can cause particular norms to emerge [15]. Fixed strategy agents have been shown to be effective in coordination and Prisoner's Dilemma games, and so it is natural to explore whether they are effective in our setting. Figure 4 shows that introducing 300 fixed-strategy agents (10% of each group selected at random, i.e., 10% of the population) who always select *Walk*, is not sufficient to cause a norm of *Walk* to emerge. Introducing such a set of fixed-strategy agents (Figure 4(b)) gives similar results to the baseline setting (Figure 4(a)). In the context of coordination games, Airiau *et al.* [15] have shown that it is sufficient for only 1% of the population to be fixed strategy agents in order to influence the whole population [15]. However in the congested action setting, our results show that even a large number of such agents (10% of the population) is not sufficient to achieve a population-level change. We therefore conclude that in our setting there is a need for new interventions, such as manipulating agents' sensitivities or costs (for one or more groups).

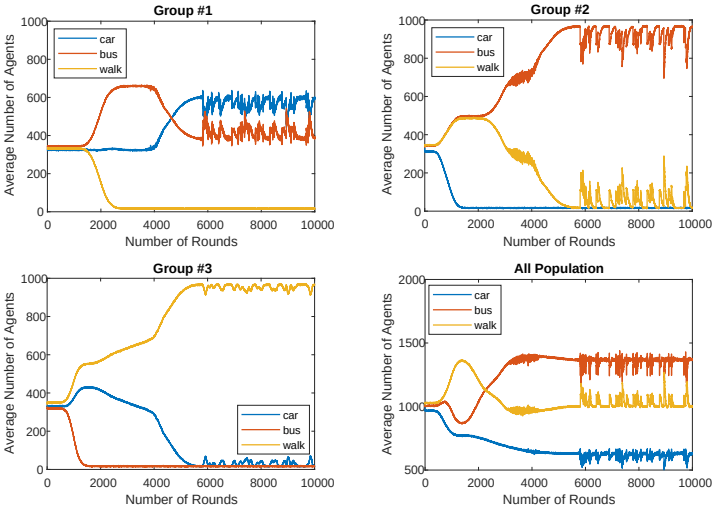


Fig. 5: Average number of agents choosing each of the actions against time, with an intervention of: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$.

5.2.2 Manipulating Agents' Sensitivities

We now consider the effect of manipulating agents' sensitivities, which for example, can model the impact of an advertising campaign on the health benefits of walking. Suppose that the aim of the manipulation is to increase the proportion of agents from g_3 that choose *Walk*, and receive high reward from doing so, by reducing the proportion of agents who can choose *Car* and receive maximal reward from 53% in the baseline setting to 19%. We do this by modifying the sensitivity of g_3 to *Walk* by setting $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$.[5] The resulting behaviour, depicted in Figure 5, shows that agents from g_3 converge to the norm of *Walk* (i.e., fewer agents choose *Car*), but that this leads to an increase in the number of agents from g_1 choosing *Car* and a decrease in those choosing *Bus*. This consequently leads to more agents in g_2 choosing *Bus* and fewer choosing *Walk*. However at the population level, *Walk* increases overall but does not become the overall norm.

Although this basic manipulation causes a shift in behaviour towards *Walk*, it is not enough to cause the whole population to adopt a norm of *Walk*. For this reason we look at two alternative manipulations, each one is a combination of the basic intervention (depicted in Figure 5) with an additional measure.

First, suppose that the authority manipulates the rewards such that agents from g_1 are able to choose *Walk* and receive a higher reward than *Bus*. This can be modelled by decreasing the sensitivity of g_1 towards *Bus*, i.e., $\tilde{s}_{Bus,1} = 0.86 \cdot s_{Bus,1}$. As can be seen in Figure 6, applying this intervention, together with reducing the proportion of agents from g_3 who can choose *Car* and receive

[5] A similar manipulation (with similar effect) is decreasing the sensitivity of g_3 towards *Car*.

maximal reward (i.e., the existing manipulation of $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$), shifts g_1 towards *Car* (instead of *Bus*), leading agents from g_2 to shift towards *Bus* instead of *Walk*. Overall, the whole population changes its preferences, with many choosing *Walk*.

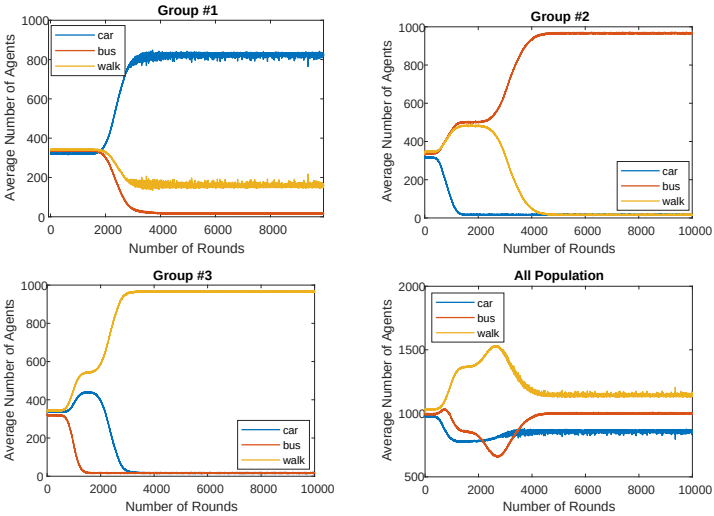


Fig. 6: Average number of agents choosing each of the actions against time, with the interventions of: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ and $\tilde{s}_{Bus,1} = 0.86 \cdot s_{Bus,1}$.

The second manipulation is to decrease the proportion of agents from g_1 that receive a small reward when choosing *Walk*, by increasing their sensitivity such that $\tilde{s}_{Walk,1} = 1.14 \cdot s_{Walk,1}$. This second manipulation, applied alongside reducing the proportion of agents from g_3 who can choose *Car* and receive maximal reward, causes agents from g_1 and g_3 to change their behaviour, while g_2 continues to choose *Bus*. Overall, the population shifts towards *Walk*, as can be seen in Figure 7.

5.2.3 Manipulating Agents' Costs

We now consider manipulating agents' costs, which models interventions such as charging individuals who have polluting vehicles or subsidising the costs of electric vehicles. Suppose that the authority increases the cost of *Car* for g_3 such that other actions have a lower cost (by setting $\tilde{c}_{Car,3} = 0.8 + c_{Car,3}$). As can be seen in Figure 8, this results in g_3 adopting a norm of *Walk*, while the population overall still adopts *Bus* as most common action choice.

In order to achieve population wide shift, we consider an additional manipulation, namely, increasing the cost of g_2 when choosing *Bus* such that a higher reward is associated with *Walk* (by setting $\tilde{c}_{Bus,2} = 0.03 + c_{Bus,2}$). This manipulation, applied alongside increasing the cost of g_3 from *Car*, gives

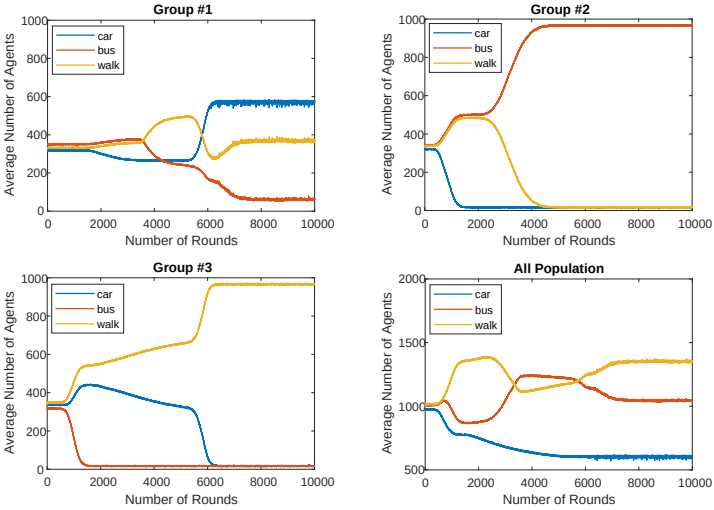


Fig. 7: Average number of agents choosing each of the actions against time, with the interventions of: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ and $\tilde{s}_{Walk,1} = 1.14 \cdot s_{Walk,1}$.

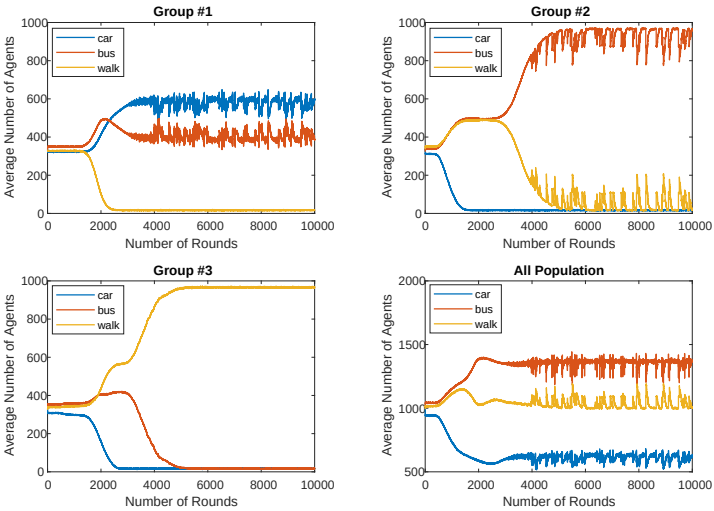


Fig. 8: Average number of agents choosing each of the actions against time, with the intervention of: $\tilde{c}_{Car,3} = 0.8 + c_{Car,3}$.

a significant change in individuals' preferences with *Walk* emerging as a norm in the population overall, as depicted in Figure 9.

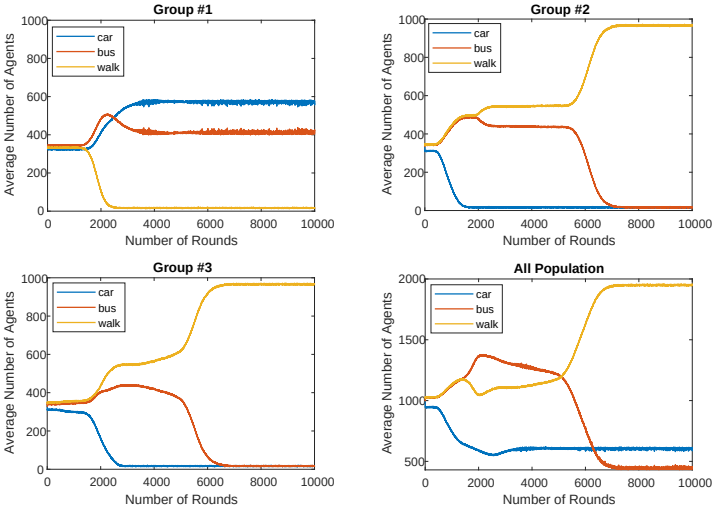


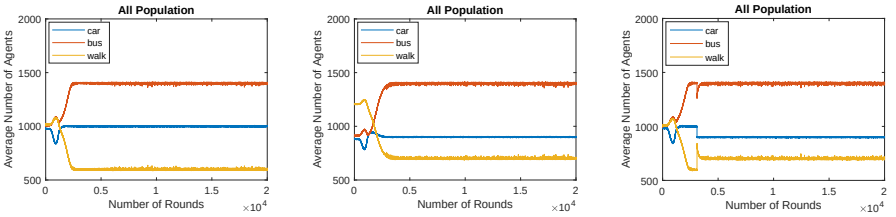
Fig. 9: Average number of agents choosing each of the actions against time with the interventions of: $\tilde{c}_{Car,3} = 0.8 + c_{Car,3}$ and $\tilde{c}_{Bus,2} = 0.03 + c_{Bus,2}$.

5.3 Late Interventions

We previously studied the case where the interventions were placed at time $t = 0$ and left within the system for its duration. We now change our focus to considering destabilising an existing norm by manipulating agents' sensitivities and costs *after* convergence has occurred. In this section, we consider agents' behaviour under the interventions illustrated in Section 5.2, but with the interventions only being applied after the population has converged, and using the same setting as detailed in Section 4.

5.3.1 Fixed-Strategy Agents

The use of non-learning fixed strategy agents (known as influencer agents, or IAs) has been investigated in literature as a means to destabilise an already established convention, as discussed in Section 2. As demonstrated in Section 5.2.1 (Figure 4), initially setting 10% of the population to be fixed-strategy agents who always select *Walk* results in a similar behaviour as in the baseline setting. Figure 10(c) shows the effect of introducing 300 agents from the population (i.e., 10%) that always select *Walk*, from time $t = 3{,}100$ after the population has converged. We see that there is a small immediate effect, but very quickly the population returns to the same norm as in the baseline setting. Interestingly, an intervention of changing 300 agents to choose always *Walk* does not result in 300 more agents choosing that action overall. Instead, compared to the baseline setting, as a result of the reward functions the late introduction of 300 fixed-strategy agents leads to 100 more agents choosing *Walk* and 100 less agents choosing *Car* (the same average numbers



(a) Baseline setting (b) Initial intervention (c) Late intervention

Fig. 10: Average number of agents choosing each of the actions against time in (a) our baseline setting, (b) initial intervention with 10% fixed-strategy agents, and (c) late intervention with 10% fixed-strategy agents.

of agents as with the initial intervention). Therefore, placing fixed-strategy after convergence is not sufficient to cause a new norm to emerge in our congested action setting. This is in contrast to previous research on coordination games [55, 56] which showed the effectiveness of such interventions for destabilisation. Note that to confirm that the magnitude of the intervention is not the primary factor, we also considered a similar late intervention but with 30% of the population being fixed-strategy agents, and obtained a similar result to that discussed above. This indicates that in our setting it is much harder to cause a change in adopted behaviour (using fixed strategy agents) than in coordination games.

5.3.2 Manipulating Agents' Sensitivities and Costs

In this section, we show the effect of late interventions in which we manipulate agents' sensitivities and costs. Figure 11 shows the results of a late intervention at time $t = 5{,}800$ in which we set $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ (equivalent of Figure 5). As can be seen, the manipulation is effective almost immediately with the average number of agents from g_3 choosing *Car* falling, and the number choosing *Walk* increasing. When destabilisation is sufficient to allow changes in other groups, this consequently leads to an increase in the number of agents from g_1 choosing *Car* and a decrease in those choosing *Bus*, while in g_2 more agents choose *Bus* and fewer choose *Walk*.

Figure 12 shows the result of two late manipulations applied at time $t = 7800$, namely: (i) increasing the proportion of agents from g_3 that choose *Walk* and receive high reward; and (ii) decreasing the proportion of agents from g_1 that choose *Walk* and receive a small reward (equivalent of the initial intervention shown in Figure 7). We observe that the time taken for the system to converge after the late intervention is much longer than that with no intervention. This demonstrates that it is much slower to take a system that is already converged and influence it to take a different state, than it is to allow the system to converge initially without any intervention.

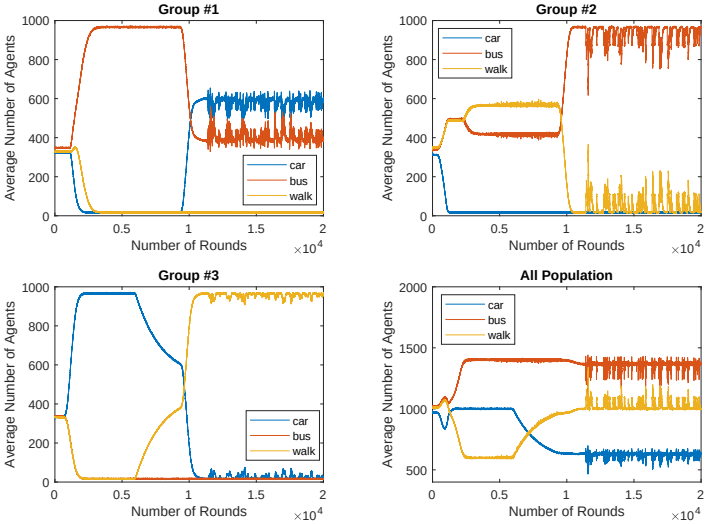


Fig. 11: Average number of agents choosing each of the actions along time. Intervention applied after convergence: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ (the equivalent of Figure 5).

Overall, we see that with the introduction of late interventions the initial norm adopted by each of the groups (i.e., before applying the intervention, as depicted in our baseline setting) becomes sufficiently destabilised for another norm to overtake it, resulting in the same outcome as with an initial intervention. This is in line with other research (e.g., [43]) using fixed strategy agents to encourage convention emergence and destabilisation in coordination games. Of particular interest is the speed of the change, since while there is an immediate change in the manipulated group(s), other groups are stable until the destabilisation is sufficient to allow the emergence of another norm (namely, that which would be adopted by agents if the intervention was applied at the beginning of simulation). Interestingly, while the population changes its behaviour due to the manipulation, it takes roughly the same number of iterations to fully stabilise, compared to the time it takes to stabilise when the manipulation is applied at the start of the simulation.

5.4 Temporary Late Interventions

In the previous section, interventions remained in the system for its duration, and in this section we explore whether temporary interventions are effective and sufficient to influence norm emergence. As before, we consider agents' behaviour under the same interventions as described in Section 5.2, but in this case with interventions only being applied after population's behaviour has stabilized and for a fixed duration, using the same setting as detailed in Section 4.

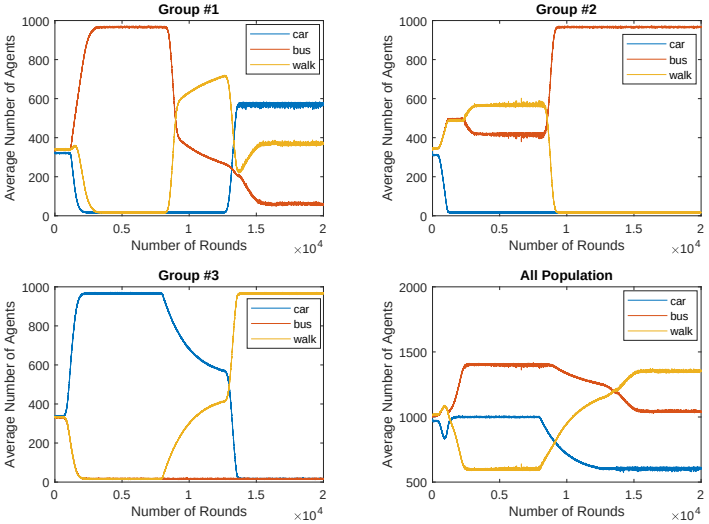


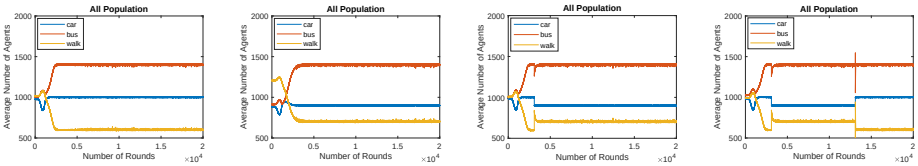
Fig. 12: Average number of agents choosing each of the actions along time. Interventions applied after convergence: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ and $\tilde{s}_{Walk,1} = 1.14 \cdot s_{Walk,1}$ (the equivalent of Figure 7).

As discussed in previous sections, we have thus far assumed that no restrictions (e.g., costs to the authority) exist when inserting the interventions. In real-life scenarios however, using interventions for shifting the population towards some desired norm is likely to have a cost. In this section, we investigate the effect of such a cost, which can be thought of as resulting in a fixed time intervention, and its relation to the efficiency of intervention. A real-life example for a fixed-time intervention is the temporary license-plate based driving bans Paris imposed on drivers in 2017 in order to reduce car use. According to this scheme, cars were banned from circulation based on whether their license plates ended with odd or even numbers.[6]

5.4.1 Fixed-Strategy Agents

Figure 13 shows the results from a temporary late intervention, in which we add 300 agents that always select *Walk* between times $t = 3{,}100$ and $t = 13{,}100$. Similarly to the initial and (permanent) late interventions, we see 100 more agents choosing *Walk* and 100 less agents choosing *Car* when the intervention is active. However, we also see that when the intervention is removed the population immediately goes back to our baseline setting. This shows that while a late intervention is able to change the behaviour of a population, and cause a different norm to emerge, the changes do not persist after the intervention is removed.

[6]For details see: https://www.reuters.com/article/us-france-pollution-idUSKBN13W2EQ



(a) Baseline (b) Initial (c) Late (d) Temporary

Fig. 13: Average number of agents choosing each of the actions against time in (a) the baseline case, and with 10% fixed-strategy agents with (b) initial intervention, (c) late intervention, and (d) temporary late intervention.

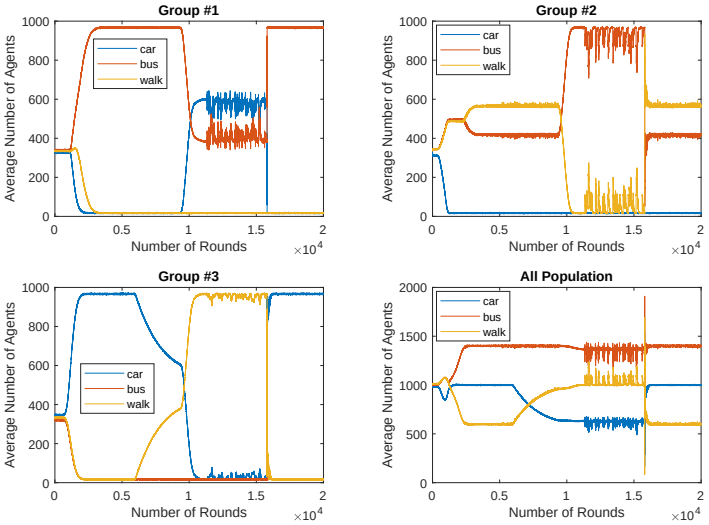


Fig. 14: Average number of agents choosing each of the actions along time. Temporary late intervention applied: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ (the equivalent of Figures 5 and 11).

5.4.2 Manipulating Agents' Sensitivities and Costs

Figures 14 and 15 show the results of temporary late interventions, corresponding to temporary versions of the interventions shown above in Figures 11 and 12, where we remove each late intervention 10,000 time steps after its application. As in the case of fixed-strategy agents, we also observe that once we remove the intervention the population immediately goes back to the baseline setting. Thus, temporary late interventions are only effective at destablising an established norm for the duration of the intervention. This is in contrast to results seen in coordination games, in which as long as some minimum number of fixed strategy agents for some minimum duration are used, then a permanent change in norm is observed [5, 41]. We believe that the reason

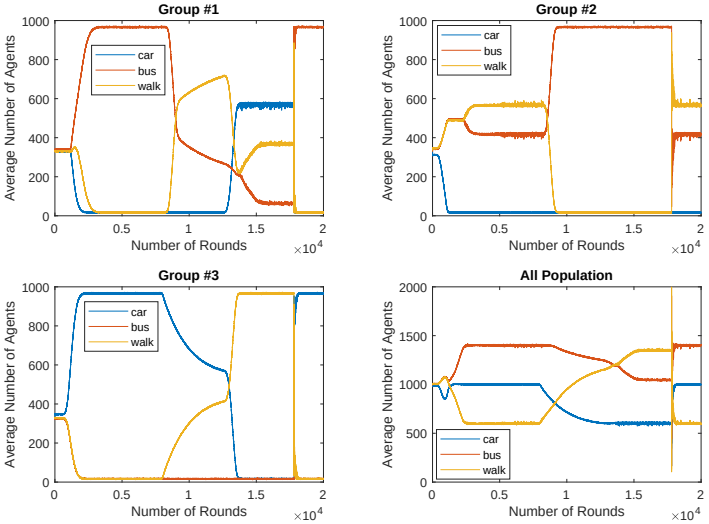


Fig. 15: Average number of agents choosing each of the actions along time. Temporary late interventions applied: $\tilde{s}_{Walk,3} = 1.8 \cdot s_{Walk,3}$ and $\tilde{s}_{Walk,1} = 1.14 \cdot s_{Walk,1}$ (the equivalent of Figures 7 and 12).

such temporary interventions do not have permanent impact here is due to the characteristics of the congested action setting.

6 Conclusions and Future Work

In this paper, we presented a framework for modelling norm emergence where actions are associated with congested resources. We considered a general setting in which agents are heterogeneous, and comprised of groups differing in their preferences regarding actions. Unlike previous research on norm emergence, which typically assumes pairwise interactions, we introduced congested actions with rewards determined exogenously. We also introduced an authority figure which is able to manipulate agents' rewards. Using a simplified transport modality choice illustration, we demonstrated the impact of manipulations on the emergent norms in the population, showing that in the presence of heterogeneous agents, different interventions may be required, targeted to the different groups. We showed that unlike interventions in coordination games, temporary late interventions are not sufficient to achieve a permanent change in norms, and so in the context of congested actions, alternative interventions and targeting strategies are required. There are several directions for future work, including relaxing assumptions about the knowledge available to agents and further exploring agent heterogeneity. We also plan to investigate dynamic populations, and situating agents on an underlying network.

References

[1] Shoham, Y., Tennenholtz, M.: On the emergence of social conventions: modeling, analysis, and simulations. Artificial Intelligence **94**(1-2), 139–166 (1997)

[2] Haynes, C., Luck, M., McBurney, P., Mahmoud, S., Vítek, T., Miles, S.: Engineering the emergence of norms: A review. The Knowledge Engineering Review **32**, 1–31 (2017)

[3] Morris-Martin, A., De Vos, M., Padget, J.: Norm emergence in multiagent systems: A viewpoint paper. Autonomous Agents and Multi-Agent Systems **33**, 706–749 (2019)

[4] Malialis, K., Devlin, S., Kudenko, D.: Resource abstraction for reinforcement learning in multiagent congestion problems. In: Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 503–511 (2016)

[5] Marchant, J., Griffiths, N., Leeke, M., Franks, H.: Destabilising conventions using temporary interventions. In: International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, pp. 148–163 (2014). Springer

[6] Levy, P., Griffiths, N.: Convention emergence with congested resources. In: European Conference on Multi-Agent Systems, pp. 126–143 (2021). Springer

[7] Arce, D.G.: Stability criteria for social norms with applications to the prisoner's dilemma. Journal of Conflict Resolution **38**(4), 749–765 (1994)

[8] Heckathorn, D.D.: Collective sanctions and the creations of prisoner's dilemma norms. The American Journal of Sociology **94**(3), 535–562 (1988)

[9] Helbing, D., Johansson, A.: Cooperation, norms, and revolutions: A unified game-theoretical approach. PloS One **5**(10), 1–15 (2010)

[10] Hu, S., Leung, H.-F.: Achieving coordination in multi-agent systems by stable local conventions under community networks. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), pp. 4731–4737 (2017)

[11] Sen, O., Sen, S.: Effects of social network topology and options on norm emergence. In: International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, pp. 211–222 (2009). Springer

[12] Sen, S., Airiau, S.: Emergence of norms through social learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1507–1512 (2007)

[13] Villatoro, D., Sabater-Mir, J., Sen, S.: Social instruments for robust convention emergence. In: Proceedings of the 22th International Joint Conference on Artificial Intelligence (IJCAI), pp. 420–425 (2011)

[14] Yu, C., Lv, H., Sen, S., Ren, F., Tan, G.: Adaptive learning for efficient emergence of social norms in networked multiagent systems. In: Pacific Rim International Conference on Artificial Intelligence (PRICAI), pp. 805–818 (2016)

[15] Airiau, S., Sen, S., Villatoro, D.: Emergence of conventions through social learning. Autonomous Agents and Multi-Agent Systems **28**(5), 779–804 (2014)

[16] Marchant, J., Griffiths, N.: Convention emergence in partially observable topologies. In: Autonomous Agents and Multiagent Systems, pp. 187–202 (2017)

[17] Salazar, N., Rodriguez-Aguilar, J.A., Arcos, J.L.: Robust coordination in large convention spaces. AI Communications **23**, 357–371 (2010)

[18] Arthur, W.B.: Inductive reasoning and bounded rationality. The American economic review **84**(2), 406–411 (1994)

[19] Farago, J., Greenwald, A., Hall, K.: Fair and efficient solutions to the santa fe bar problem. In: Proceedings of the Grace Hopper Celebration of Women in Computing (2002)

[20] Schlag, K.H.: Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. Journal of Economic Theory **78**(1), 130–156 (1998)

[21] Auer, P., Cesa-Bianchi, N., Fischer, P.: Finite-time analysis of the multiarmed bandit problem. Machine Learning **47**(2-3), 235–256 (2002)

[22] Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The nonstochastic multiarmed bandit problem. SIAM journal on computing **32**(1), 48–77 (2002)

[23] Granmo, O.-C., Glimsdal, S.: Accelerated bayesian learning for decentralized two-armed bandit based decision making with applications to the goore game. Applied intelligence **38**(4), 479–488 (2013)

[24] Kuleshov, V., Precup, D.: Algorithms for multi-armed bandit problems.

Journal of Machine Learning Research **1**, 1–48 (2000)

[25] Axelrod, R.: An evolutionary approach to norms. The American Political Science Review **80**(4), 1095–1111 (1986)

[26] Mahmoud, S., Griffiths, N., Keppens, J., Luck, M.: Overcoming omniscience for norm emergence in axelrod's metanorm model. In: International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, pp. 186–202 (2011). Springer

[27] Mahmoud, S., Griffiths, N., Keppens, J., Luck, M.: Efficient norm emergence through experiential dynamic punishment. In: Proceedings of the 20th European Conference on Artificial Intelligence (ECAI), pp. 576–581 (2012)

[28] de Pinninck, A.P., Sierra, C., Schorlemmer, M.: Distributed norm enforcement via ostracism. In: Proceedings of the 4th International Workshop on Coordination, Organization, Institutions and Norms (2007)

[29] Savarimuthu, B.T.R., Purvis, M., Purvis, M.: Social norm emergence in virtual agent societies. In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1521–1524 (2008)

[30] Franks, H., Griffiths, N., Jhumka, A.: Manipulating convention emergence using influencer agents. Autonomous Agents and Multi-Agent Systems **26**(3), 315–353 (2012)

[31] Griffiths, N., Anand, S.S.: The impact of social placement of non-learning agents on convention emergence. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), vol. 3, pp. 1367–1368 (2012)

[32] Granmo, O.-C., Berg, S.: Solving non-stationary bandit problems by random sampling from sibling kalman filters. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 199–208 (2010). Springer

[33] Amin, K., Kale, S., Tesauro, G., Turaga, D.: Budgeted prediction with expert advice. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 2490–2096 (2015)

[34] Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: Gambling in a rigged casino: The adversarial multi-armed bandit problem. In: Proceedings of IEEE 36th Annual Foundations of Computer Science, pp. 322–331 (1995). IEEE

[35] Kale, S.: Multiarmed bandits with limited expert advice. In: Conference on Learning Theory, pp. 107–122 (2014)

[36] Seldin, Y., Bartlett, P.L., Crammer, K., Abbasi-Yadkori, Y.: Prediction with limited advice and multiarmed bandits with paid observations. In: Proceedings of the 30th International Conference on Machine Learning (ICML), pp. 280–287 (2014)

[37] Zeng, C., Wang, Q., Mokhtari, S., Li, T.: Online context-aware recommendation with time varying multi-armed bandit. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 2025–2034 (2016)

[38] Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th International Conference on World Wide Web (WWW), pp. 661–670 (2010)

[39] Shivaswamy, P., Joachims, T.: Multi-armed bandit problems with history. In: Artificial Intelligence and Statistics, pp. 1046–1054 (2012)

[40] Yang, A., Yang, G.H.: A contextual bandit approach to dynamic search. In: Proceedings of the ACM International Conference on Theory of Information Retrieval (SIGIR), pp. 301–304 (2017)

[41] Marchant, J., Griffiths, N., Leeke, M.: Destabilising conventions: Characterising the cost. In: 2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems, pp. 139–144 (2014). IEEE

[42] Marchant, J., Griffiths, N.: Manipulating conventions in a particle-based topology. In: International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, pp. 242–261 (2015). Springer

[43] Marchant, J., Griffiths, N., Leeke, M.: Convention emergence and influence in dynamic topologies. In: AAMAS, pp. 1785–1786 (2015)

[44] Villatoro, D., Sen, S., Sabater-Mir, J.: Topology and memory effect on convention emergence. In: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 233–240 (2009). IEEE

[45] Toivonen, R., Castelló, X., Eguíluz, V.M., Saramäki, J., Kaski, K., San Miguel, M.: Broad lifetime distributions for ordering dynamics in complex networks. Physical Review E **79**(1), 016109 (2009)

[46] Epstein, J.M.: Learning to be thoughtless: Social norms and individual computation. Computational economics **18**(1), 9–24 (2001)

[47] Conte, R., Paolucci, M.: Intelligent social learning. Journal of Artificial Societies and Social Simulation **4**(1) (2001)

[48] Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning **8**, 279–292 (1992)

[49] Beheshti, R., Ali, A.M., Sukthankar, G.: Cognitive social learners: an architecture for modeling normative behavior. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 2017–2023 (2015)

[50] Mukherjee, P., Sen, S., Airiau, S.: Norm emergence under constrained interactions in diverse societies. In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 779–786 (2008)

[51] Vouros, G.A.: The emergence of norms via contextual agreements in open societies. In: Koch, F., Guttmann, C., Busquets, D. (eds.) Advances in Social Computing and Multiagent Systems, pp. 185–201. Springer, Cham (2015)

[52] Yu, C., Zhang, M., Ren, F.: Collective learning for the emergence of social norms in networked multiagent systems. IEEE Transactions on Cybernetics **44**(12), 2342–2355 (2014)

[53] Bowling, M., Veloso, M.: Multiagent learning using a variable learning rate. Artificial Intelligence **136**(2), 215–250 (2002)

[54] Kittock, J.E.: Emergent conventions and the structure of multi-agent systems. In: Proceedings of the 1993 Santa Fe Institute Complex Systems Summer School, vol. 6, pp. 1–14 (1993)

[55] Marchant, J., Griffiths, N.: Limited observations and local information in convention emergence. In: AAMAS, pp. 1628–1630 (2017)

[56] Marchant, J., Griffiths, N.: Convention emergence in partially observable topologies. In: International Conference on Autonomous Agents and Multiagent Systems, pp. 187–202 (2017). Springer