Self-Governance by Transfiguration: from Learning to Prescriptions

Régis Riveret[1], Alexander Artikis[2], Dídac Busquets[1], and Jeremy Pitt[1]

[1] Imperial College London, United Kingdom
{r.riveret,didac.busquets,j.pitt}@imperial.ac.uk
[2] NCSR Demokritos, Greece
a.artikis@iit.demokritos.gr

Abstract. Norms are commonly understood as guides for the conduct of autonomous agents, thereby easing their individual decision-making and coordination. However, their study exhibits a polarity between (i) norms as behavioural patterns emerging from repeated agents' (inter)actions and (ii) norms as explicit prescriptions. In this paper, we attempt to build a bridge between these two conceptual poles of norms: it takes the form of a mental function for prescriptive transfiguration allowing reinforced learning agents to express their learning experiences into prescriptions. The population of transfigurative agents are then equipped with a consensus system to build and enforce prescriptive systems to self-govern on-line. Simple simulations suggest the pertinence of the approach and shows its weaknesses, in particular prescriptions stalling learning, and timeliness in norm construction.

1 Introduction

Norms are commonly understood as guides for the conduct of autonomous agents, thereby easing their individual decision-making and coordination. However, their study exhibits a polarity in their conception. On the one hand, jurists concentrate on norms as prescriptions promulgated by institutional powers and enforced by explicit sanctions. On the other hand, social researchers study norms as tacit behavioural patterns emerging from expectations and enforced by entwined sanctions. This polarisation is reflected by the treatment of norms by computer scientists. Prescriptions and legal reasoning are investigated in formal logics (typically deontic logics and argumentation, see e.g. [13]) to represent and reason upon explicit norms, leading eventually to architecture for cognitive agents (see e.g. [7]) while social norms are accounted as patterns emerging from repeated interactions amongst agents (typically learning agents see e.g. [14]).

Scholars have thus investigated the influence of social norms and prescriptions on each other, but the conceptual gap remains hardly explored by computer scientists, in particular with regard to applied systems, c.f. [12]. To address it, we propose a simple mental apparatus to perform a prescriptive transfiguration allowing reinforced learning agents to express their learning experiences into prescriptions.

As to the practical relevance, our proposal regards self-organising systems and in particular self-governing systems, specifically the on-line construction of prescriptive systems for and by reinforced learning agents. Indeed, while reinforcement learning is a prevalent mean for the adaptation of autonomous agents with incomplete information on their environment [16], norms are an attractive manner to guide the conduct of these agents. Furthermore, explicit norms are commonly advocated to facilitate their updates, and consequently system maintenance, improve system transparency and ease system governance. Unfortunately, as remarked by [8], the manual construction of prescriptive systems is often time-consuming and error prone, the construction at design time (i.e. off-line construction) is computationally complex, and both are unsuited for dynamic systems with unpredictable changes. Therefore we opt for on-line construction. Since systems of multiple autonomous agents have their essence into decentralised control and computation, this on-line construction shall occur in a distributed manner in the sense there is no entity with complete information taking the role of a central legislative body. We will focus on explicit primary norms and in particular regulative norms, i.e. those guiding the ideal behaviour of agents, leaving other primary constitutive norms (used to constitute institutions) and secondary norms (managing primary norms) for future work.

The practical challenge in this paper regards thus the self-governance of learning agents, or more specifically the domain-independent construction at run-time of explicit regulative norms from scratch, for and by learning agents, without any agent having a complete information on the system. Our solution, inspired by direct democracy, is a consensus system coupled with the mental function of prescriptive transfiguration so that every agent shall propose and vote for prescriptions meant to govern themselves. The overall system results thereby into a direct self-governance taking advantage of every agents' learning experiences.

Noting there is no obvious or immediate utility for a reinforced learning agent to share his own experiences to influence the construction of a prescriptive system (paradox of voting, also called Downs paradox), our proposal of direct self-governance is imposed to the agents (i.e. hard-coded). Nevertheless, as every agent is learning with respect to the qualities of behaviours, the construction of norms occurs in the same spirit. Every possible proposal and vote is associated with a probability reflecting a scalar potential and we assume that every agent is endowed with a mental apparatus described in this paper to compute these potentials. This apparatus is light so that it is compatible with the presumption of agents with bounded cognition.

The simulations of reinforced learning agents equipped with such legislative apparatus suggest the pertinence of such approach but also its weaknesses, in particular prescriptions stalling learning and timeliness in norm construction.

The remainder of this paper is organised as follows. In the next Section 2, we base our system of learning agents on the model of stochastic games and we define the problem of self-governance we are interested in. Our proposal for direct self-governance is given in Section 3. It is illustrate in Section 4 with some simulation results and related with other work in Section 5, before concluding.

2 Stochastic games

We base our framework on the common model of stochastic games. A stochastic game can be considered as an extension of a Markov decision process with multiple agents with possibly conflicting goals, and where the joint actions of agents determine state transitions and payoffs. A stochastic game consists of a tuple $< \mathcal{G}, \mathcal{S}, \mathcal{A}, T, R >$ where:

- $\mathcal{G}$ is a set of N agents indexed by i;
- $\mathcal{S} = \{s_1, \ldots s_n\}$ is a finite non-empty set of global states;
- $\mathcal{A} = \prod_i A_i$ is a set of joint actions. A_i is a set of individual actions available to agent i.
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a function of transition, $T(s_r, A, s_q) = p(s_{t+1} = s_r | A, s_t = s_q)$ is the probability of resulting in a state s_r at time $t + 1$ when attempting the joint action A in a state s_q at time t.
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^N$ is a payoff function, $R_i(s_r, A, s_q) = r_i(s_{t+1} = s_r, s_t = s_q)$ is the payoff of agent i upon transition from a state s_q at time t to state s_r at time $t + 1$ under joint action A.

Though this setting implies that the possible states, transition and payoff functions are known by the investigator when specifying a game, it offers nevertheless a setting where we assume they are unknown by the agents.

The control of behaviours of agent i is described by a policy denoted π_i. It is a mapping from agent i's state history to individual behaviours. The objective of any agent i is to maximize *the infinite horizon discounted return*:

$$R_{i,t} = r_i(s_t, s_{t+1}) + \gamma . r_i(s_{t+1}, s_{t+2}) + \gamma^2 . r_i(s_{t+2}, s_{t+3}) + \ldots$$

where γ is a discount rate.

Since the probabilities and payoffs are unknown by the agents, and sanctions play an important role in normative multi-agent systems, we consider individual reinforced learning agents [16] meant to pursue the best policies. At each time step, every agent senses its environment, and, given the observed state, every agent simultaneously selects the best behaviour on the basis of past experiences (exploitation) and also by trying new options (exploration). No agent is informed about the actions performed and payoffs received by the other agents.

A behaviour j of an agent i, denoted by a pair state-action $(s, a_{i,j})$, is associated with a real number $Q(s, a_{i,j})$ representing the quality of this behaviour over time. The quality $Q(s, a_{i,j})$ is the discounted moving average of the payoffs associated to the individual action $a_{i,j}$ in state s. Let $a_{i,t} = a_{i,j}$ be an action j selected by agent i at time t in a state s_t, the quality $Q(s_t, a_{i,t})$ is updated as follows:

$$Q(s_t, a_{i,t}) \leftarrow Q(s_t, a_{i,t}) + \alpha_i . [\delta + \gamma_i . Q(s_{t+1}, a_{i,t+1})]$$

with $\delta = r_i(s_t, s_{t+1}) - Q(s_t, a_{i,t})$, where α_i is a learning rate, and γ_i a discount factor trading off the importance of recent versus later payoffs. For each

agent, the selection of a behaviour at time t, is simulated by a Gibbs-Boltzmann probability distribution over all the behaviours available for the agent i:

$$\pi_i^t(s_t, a_{i,t}) = \frac{e^{Q(s_t,a_{i,t})/\tau_i}}{\sum_{a_{i,j}} e^{Q(s_t,a_{i,j})/\tau_i}}$$

where τ_i is a positive real number balancing the exploitation and the exploration of behaviours.

A stochastic game can have diverse objectives. A very popular is to find a behavioural profile (a set of policies π_i) for which no agent can benefit from unilaterally changing its behaviour, i.e. a Nash equilibrium. Stochastic games can have several Nash equilibria thus those maximising social measures such as welfare or fairness shall be preferred.

The challenge addressed in this paper is twofold: firstly we investigate the problem of prescriptive transfiguration, that is the transfiguration of agents' policies (i.e. learning experiences and thereby behavioural patterns) into prescriptions (a prescription is a conditioned obligation or prohibition with an associated sanction), secondly we consider the problem of self-governance, that is the online construction of a set of prescriptions for and by agents to govern themselves. As a first approach, the objective of self-governing agents is to maximise a social return possibly defined as the sum of agents' infinite horizon discounted social returns:

$$\sum_i R^*_{i,t} = \sum_i r^*_i(s_t, s_{t+1}) + \gamma . r^*_i(s_{t+1}, s_{t+2}) + \gamma^2 . r^*_i(s_{t+2}, s_{t+3}) + \dots$$

where r^*_i is a payoff accounting for social measures, for example those catering for the notion of justice. In a simple case, the overall social return may only deal with the global wealth of the system, accordingly r^*_i shall be a material payoff disregarding the sanctions of violated prescriptions.

Since the essence of systems of multiple autonomous agents is to limit centralised control, we look at the problem in which there is no agent having complete information about the game to design the prescriptive system. So, we base our mechanism on the idea that every agent shall participate on the construction of prescriptions. Prescriptions shall be constructed for and by agents. Accordingly, we use a voting system: the set of messages are the possible motions (the explicit norm to be voted) and the votes; the results of social decisions are the enter of force (and thus reinforcement) of these explicit norms. Remark that though the implemented system implied a central entity implementing the voting system for accepting agent's motions, votes and for deliberations, other distributed consensus mechanisms could be employed.

Example 1. We will illustrate and evaluate the prescriptive transfiguration and the proposed self-governance of learning agents with an example inspired by accident law (we do not aim at legal precision, c.f. [10]). Consider a population of agents acting in two possible global states: one is safe and the other is dangerous. In any state, every agent can act with care or with negligence. Whatever the

state, if all the agents act with care then the next state will be safe. If an agent acts with negligence then there is a risk of an accident and the next state is dangerous. The probability of an accident is higher when the negligent act is performed in a dangerous state. Hence it suffices that only one agent acts with negligence and that an accident occurs to bring the population in a dangerous state. The Markov decision problem graph is drawn in Figure 1 for a system populated by a single agent. The unique Nash equilibrium takes place when all

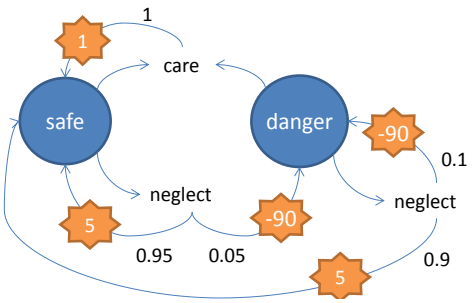

Fig. 1: The Markov decision process graph for a system populated by a single agent. Each transition from an action to a state is represented by an arrow labelled with its probability and associated payoff.

agents act with care. Reinforced learning agents may or may not learn to act with care, in any case we will investigate a system of self-governance where agents construct prescriptions to guide themselves.

3 Direct self-governance

To address the problem of transfiguration and self-governance as presented in the previous section, we endow agents agents with a mental apparatus to transform learning experiences into prescriptions and this apparatus is coupled with a consensus mechanism so that agents make a social choice on those prescriptions meant to govern themselves. The pseudo-code animating the population in its environment is given in Algorithm 1. In the remainder of this section, we describe the step regarding transfiguration, and the steps concerning self-governance, i.e. submissions, motion selection and voting.

3.1 Individual prescriptive transfiguration

Prescriptive transfiguration is based on a mapping from a learning policy to prescriptions. In practice any behaviour B in a state s resulting in an action a is associated with two prescriptive counterparts that we call possible self-prescriptions and that we represent with the following rules:

$$r_{\mathsf{Obl}\,(B)} : s \Rightarrow \mathsf{Obl}\, a \qquad\qquad r_{\mathsf{Forb}\,(B)} : s \Rightarrow \mathsf{Forb}\, a$$

Algorithm 1 Animation of self-governed learning agents for an episode.

```
Initialise the system;
for each step of an episode do
  for each agent do
    Choose an action amongst alternatives;
  end for
  Compute the environment;
  for each agent do
    Observe the individual payoffs;
    Update quality of behaviours;
    Individual prescriptive transfiguration;
  end for
  Submissions, Motion selection and Voting;
end for
```

where $r_{\mathsf{Obl}\,(B)}$ ($r_{\mathsf{Forb}\,(B)}$) is an identifier of the self-obligation (self-prohibition), s represents the conditions and $\mathsf{Obl}\,a$ ($\mathsf{Forb}\,a$) is the consequent. The identifier may be dropped when its omission does not raise any ambiguity.

Example 2. For every agent, there are four possible self-prescriptions:

$$\begin{array}{ll} \text{safe} \Rightarrow \mathsf{Obl}\,\text{care} & \text{danger} \Rightarrow \mathsf{Obl}\,\text{care} \\ \text{safe} \Rightarrow \mathsf{Forb}\,\text{care} & \text{danger} \Rightarrow \mathsf{Forb}\,\text{care} \\ & \\ \text{safe} \Rightarrow \mathsf{Obl}\,\text{neglect} & \text{danger} \Rightarrow \mathsf{Obl}\,\text{neglect} \\ \text{safe} \Rightarrow \mathsf{Forb}\,\text{neglect} & \text{danger} \Rightarrow \mathsf{Forb}\,\text{neglect} \end{array}$$

Notice that we assume no equivalence between the obligation to act with care and the prohibition to act with negligence. The adopted logic is thus light on this aspect. Nevertheless a kind of quantitative equivalence shall appear when we will introduce potentials to prescriptions (see below).

Possible self-prescriptions are not active: every agent shall propose the most relevant amongst all them as a motion to the whole population before voting for its enforcement. The construction of the prescriptive system occurs in three activities:

1. Individual prescriptive transfiguration: every agent shall individually transfigure learning experiences into (self-)prescriptions,
2. Submissions and Motions: every agent shall submit a prescription and the most common proposal becomes the motion,
3. Voting: every agent votes for the motion with respect to its self-prescriptive background (the set of agent's self-prescriptions).

In each activity, every agent has to make a choice about (self-)prescriptions (self-prescribe, make a proposal, vote for the motion). Since we have learning agents, every agent will make its choice by taking into consideration the quality or potential of the (self-)prescriptions with a flavour of reinforcement learning, as we will see in the remainder of this section.

Once and every time an agent observes the current state and considers the set of alternative behaviours, this agent shall individually transfigure learning experiences into submissible prescriptions. This phase is decomposed in two steps: the agent decides or not to self-prescribe, then eventually, a submissible self-prescription is drawn.

Self-prescribe or not This step is meant to avoid an agent to transfigure learning policies when alternative behaviours have similar qualities. Indeed, there is no advantage to oblige or prohibit a behaviour with respect to the others when they all result in similar payoffs. There are many manners to avoid the prescription of behaviours with similar qualities. We chose to do so by using an entropic threshold. Every agent i computes the entropy $\mathcal{S}_i$ of the distribution of the alternative behaviours in a state. If $\mathcal{S}_i$ is less than a threshold τ_i^S then the agent will draw a self-prescription. We propose no calculus here to set up this threshold τ^S, but we can give some basic considerations. If it set to high, then the agent i may not gain enough experiences before considering prescriptions and thus non-optimal prescriptions may be selected in the next phase. At the opposite, if the threshold is set to low, then the agent may have so much experiences that prescriptions shall appear useless.

Example 3. Suppose the agent named Tom is in a safe state. Tom has two behavioural alternatives: either behave with care or behave with negligence. Assume that the careful behaviour has a quality 4 and the negligent behaviour has quality 2, thus their respective probability is:

$$p(\text{care}|\text{safe}) = \frac{e^4}{e^4 + e^2} \sim 0.88 \qquad p(\text{neglect}|\text{safe}) \sim 0.12$$

The entropy is $\mathcal{S}_{Tom} \sim -0.88 \cdot \ln(0.88) - 0.12 \cdot \ln(0.12) \, (\sim 0.37)$. Consider a threshold $\tau^S_{Tom} = 0.5$, then Tom will consider alternative prescriptions to elevate one to the rank of submissible prescription (see below). If the entropy was higher than this threshold, then Tom would consider no prescription for the safe state.

Selection of submissible prescriptions If an agent decides to transfigure learning experiences into self-prescriptions then it will draw a self-prescription that becomes a *submissible prescription*. To do so, every possible self-prescription is associated with a scalar measure that we call the submissible potential. The higher the quality of a behaviour with respect to the quality of other behaviours, the higher its potential to be considered as an obligation. At the opposite, the lower the quality of a behaviour with respect to the quality of other behaviours, the higher its potential to be considered as a prohibition. Let's capture formally these ideas. Let $\widehat{Q}_i$ denote the average quality of alternative behaviours in a state according to agent i. For a self-obligation $r_{\mathsf{Obl}\,(B)}$, its submissible potential according to agent i, denoted $\delta_i(r_{\mathsf{Obl}\,(B)})$, is the difference between the quality for behaviour B and the average quality of alternative behaviours. For a self-prohibition, we have the opposite:

$$\delta_i(r_{\mathsf{Obl}\,(B)}) = Q_i(B) - \widehat{Q}_i \qquad \delta_i(r_{\mathsf{Forb}\,(B)}) = \widehat{Q}_i - Q_i(B)$$

Consequently $\delta_i(r_{\mathsf{Obl}\,(B)}) = -\delta_i(r_{\mathsf{Forb}\,(B)})$.

At every step, every agent will consider a set of self-prescriptions compatible with the prescriptions in force regulating the states. A self-obligation is compatible with the prescriptions in force if:

- there is no prohibited alternative,
- there is no obliged alternative.

A self-prohibition is compatible if:

- there is another alternative not being prohibited,
- there is no obliged alternative.

As a matter of compactness of the prescriptive system, the above items assume we won't oblige an action and explicitly prohibit one of its alternative. On this basis, every agent i shall draw a self-prescription r amongst n compatible self-prescriptions $\{r_1, \ldots, r_n\}$ with a probability $p_i^\delta(r)$ using a Boltzmann-Gibbs distribution over the submissible potentials:

$$p_i^\delta(r) = \frac{e^{\delta_i(r)/\tau_i^\delta}}{\sum_{i=1}^{n} e^{\delta_i(r_i)/\tau_i^\delta}}$$

where τ_i^δ is a parameter balancing the exploitation and exploration for submissions. If this parameter tends to 0, then the agent shall pick up the prescription with the highest submissible potential. In this case, the potential of the selected prescription shall be positive, $0 \leq \delta_i(r)$. The choice of this distribution is meant to pave the way for learning prescriptive agents, in particular for frameworks where the repeal of prescriptions is possible.

Example 4. Table 1 illustrates Tom's measure of submissible potentials and the associated probabilities. We suppose in the remainder that Tom has selected two submissible prescriptions: the obligation to act with care when the state is safe, and the prohibition to act with negligence when the state is dangerous.

3.2 Submissions and Motions

Once some agents have transfigured some learning experiences into a set of submissible prescriptions, these agent shall submit each a prescription. The most common submission becomes a motion, and agents vote for its enforcement. A submitted prescription is a submission. Every agent will draw a submission from the set of submissible prescriptions using again a Boltzmann-Gibbs distribution. Let $\{r_1, \ldots, r_n\}$ be the set of submissible prescriptions of agent i (drawn in the previous step), the agent i will draw a submission r from this set with a probability $p_i^D(r)$ from a Gibbs-Boltzmann distribution over the potentials $\delta_i(r)$ with a temperature τ_i^D balancing the exploitation and exploration of submissions amongst submissible self-prescriptions. Amongst all the submissions within a population of agents, the most common submission becomes a motion, and in the next phase every agent will vote or not for this motion.

Prescription	Q_{Tom}	δ_{Tom}	p^{δ}_{Tom}
safe $\Rightarrow$ Obl care	4	1	0.5
safe $\Rightarrow$ Forb care	4	- 1	0
safe $\Rightarrow$ Obl neglect	2	-1	0
safe $\Rightarrow$ Forb neglect	2	1	0.5

Table 1: Illustration of submissible qualities δ_i and associated probabilities p^{δ}_i to consider the prescription as submissible. The qualities of corresponding behaviours (Q_{Tom}) are arbitrary given (its average is 3) and the parameter τ^{δ}_{Tom} balancing the exploitation and exploration for submissions is set at 0.1.

Example 5. Amongst the obligation to act with care when the state is safe, and the prohibition to act with negligence when the state is dangerous, we assume that Tom draws the obligation to act with care. We further assume that the most common proposals by the population is the prohibition to act with negligence when the state is safe. Consequently, this proposal becomes a motion.

At this stage, the prescription of a motion is not associated to any sanction. There is a well-accepted principle in retributive justice according which the level of the sanction should be scaled relative to the severity of the offending behaviour. In our framework, a simple mean to evaluate the severity of an offending behaviour is to consider the potential δ_i of the proposals meant to guide this behaviour. Thus, the higher the potential of a proposal, the higher the severity of a violation, the higher the sanction.

So, we associate any motion m with a potential $\widehat{\delta}(m)$ which is the average of the potentials of the proposals unifying with m. This average potential is meant to feature the value of a scalar sanction. Accordingly, we choose in this paper to define the sanction as $\widehat{\delta}(m).\mu$ where μ is a positive real number (typically set superior to 1).

Example 6. Suppose that 3 agents proposed the prohibition to act with negligence when the state is safe (the motion), and they proposed it with the potential 2, 3 and 4. The average potential is 3 and thus the quality of the motion m is $\widehat{\delta}(m) = 3$. Assuming $\mu = 10$ the associated scalar sanction associated to this motion is 30.

3.3 Voting

Once there is a motion about a prescription with its sanction, every agent is invited to vote for it. The cognitive process resulting in a vote against or in favour is not trivial to model. In a utilitarian setting, we could argue that an agent shall vote for a globally useful motion and vote against a useless motion. We assumed that the 'global potential' of a motion m is measured by its average

potential $|\widehat{\delta}(m)|$ (featuring its associated sanction - see previous section). Since agents have to vote about the motion and the associated sanction $|\widehat{\delta}(m)|.\mu$, then we suppose that agents are communicated $\widehat{\delta}(m)$. We further assume that an agent shall vote in favour or against a motion by comparing the average potential of this motion $\widehat{\delta}(m)$, with the potential of this motion according to this agent $\delta_i(m)$. The lower the difference between the potential $\delta_i(m)$ of the motion and the average potential $\widehat{\delta}(m)$, the higher the probability for agent i to vote in favour of the motion m. In punishment terms, an agent shall vote in favour of a motion if the associated sanction corresponds to a sanction "as it should be" according to this agent. Furthermore, an agent shall vote in favour of the motion m only if its corresponding individual potential $\delta_i(m)$ matches the positive or minus sign of the average potential $\widehat{\delta}(m)$, in other words an agent will not vote in favour of a motion with a positive average potential if this agent holds that this motion has a negative potential. Accordingly, we capture these considerations with a scalar measure called the *individual potential of the motion*. The agent i's individual potential of the motion m is denoted $\Delta_i(m)$:

$$\Delta_i(m) = \frac{|\delta_i(m) - \widehat{\delta}(m)|}{\tau_i^{\Delta}.|\widehat{\delta}(m)| + \epsilon} \cdot \frac{2}{1 + sgn(\delta_i(m) \cdot \widehat{\delta}(m)) + \epsilon}$$

where ϵ tends towards 0 and τ_i^{Δ} is a strictly positive real number. An agent i will vote in favour of a motion m with a probability $p_i^{\Delta}(m)$ using a folded sigmoid function:

$$p_i^{\Delta}(m) = \frac{1}{1 + \Delta_i(m)}$$

The higher τ_i^{Δ}, the higher the probability for agent i to vote in favour of the motion m. If τ_i^{Δ} is large then agents shall vote for any motion (the most common proposal) at the risk of being ruled by a minority.

Example 7. Recall the most common submitted prescription by the population is the prohibition to act with negligence when the state is safe. Hence every agent is invited to vote about this motion. We computed that the average agents' quality over this motion is 3, $\widehat{\delta}(m) = 3$. Let $\tau_{Tom}^{\Delta} = 0.1$, the individual potential of Tom for this motion is thus: $\Delta_{Tom}(m) \sim |1 - 3|/0.1 \cdot 3$. Tom will vote in favour of this motion with a probability $p_{Tom}^{\Delta}(m) \sim 0.01$.

The consensus can take many different forms, it can be distributed or centralised for example, but for our purposes we arbitrary considered a majority rule. Accordingly a prescription and its enforcement voted by the majority enters in force. The abrogation of a prescriptions shall be possible but we reserve its presentation in another work. Any prescription in force is enforced by applying its associated sanction to any non-compliant agent (modifying thus the payoffs of the underlying stochastic game).

4 Simulation results

To evaluate and get more insights into the proposed prescriptive transfiguration and associated self-governance, we animated the stochastic game of Example 1 with a homogeneous population of reinforced learning agents with no initial prescriptions. The environment, the agents, their interactions and the prescriptions were implemented as a development of the platform based on a probabilistic rule-based argumentation and machine learning [11], so that the system specifications were directly executed. The results are averaged over 100 runs of 250 time steps of a population of 50 agents.

The probability of careful behaviours in the safe and dangerous state with or without self-governance is shown in the figures 2 and 3. When self-governance is deactivated, agents learn to behave with care in both states, but the convergence is slower in the safe state as the careful and negligent behaviours in this state have closer expected utilities. When self-governance is activated, the enforcement of careful behaviours guided the agents towards careful behaviours with a higher speed of convergence in both states. The possible prescriptions and

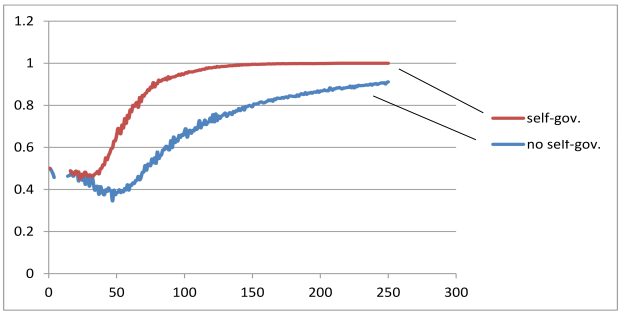

Fig. 2: Average probability of careful behaviours in the safe state with self-governance (red) and without self-governance (blue) vs. time.

their empirical probability of enforcement with respect to the parameter τ_i^{Δ} (see Section 3.3) are shown in Table 2. Remark that the probabilities with respect to a state may not add up to one as few simulations did not end up with prescribed states. The reason holds in the choice of a low value (see e.g. $\tau_i^{\Delta} = 0.1$) so agents appeared quite picky in their vote. At the opposite, when this parameter was set large, e.g. $\tau_i^{\Delta} = 1$, all the simulations ended with prescribed states. The benefits of the system are thus illustrated by these simulations: an increase of global wealth (since careful agents shall accumulate more wealth when behaving with care) while addressing the problem of (i) prescriptive transfiguration, that is the transfiguration of agents' learning experiences and thereby behavioural patterns into prescriptions, and (ii) self-governance, that is the on-line construction of prescriptions for and by agents to govern themselves.

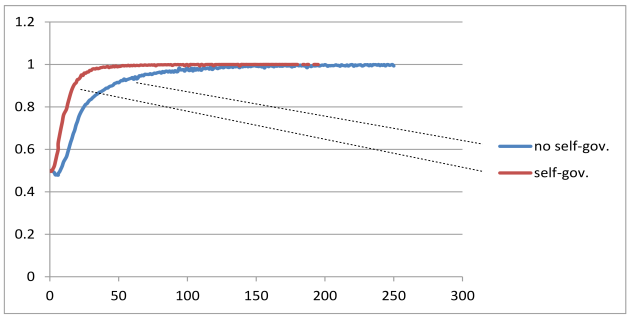

Fig. 3: Average probability of careful behaviours in the dangerous state with self-governance (red) and without self-governance (blue) vs. time.

τ_i^{Δ}	0.1	0.5	1		0.1	0.5	1
safe ⇒ Obl care	0.54	0.42	0.46	danger ⇒ Obl care	0.46	0.50	0.50
safe ⇒ Forb care	0	0	0	danger ⇒ Forb care	0	0	0
safe ⇒ Obl neglect	0	0.02	0.02	danger ⇒ Obl neglect	0	0	0
safe ⇒ Forb neglect	0.40	0.50	0.52	danger ⇒ Forb neglect	0.40	0.50	0.50

Table 2: Prescriptions with their empirical probability of enforcement.

Weaknesses exist as well. For instance, remark that an obligation to act with negligence was voted in one simulation: its enforcement occurred at the time step 41 when the probability of careful behaviour in the safe state was low enough to let a minority of negligent agents to pass this obligation. This shows a weakness of the present framework regarding the difficulty to appropriately prescribe behaviours with close qualities at voting time. There is indeed a risk of a consensus for policies enforcing undesirable behaviours when the quality of these behaviours is close to desirable behaviours. This occurs when the expected utilities of alternatives are close or when the dynamics is such that undesired behaviours appear with relatively high quality for a period of time during which a vote occurred. This later unfortunate condition emphasis the importance of timeliness in norm construction. In conditions where accidents are sparse but very harmful, if a vote occurs too early then there is risk that agents vote for policies enforcing undesired behaviours. At the opposite, a late vote may imply new explicit policies enforcing a well-established social norm; in this case policies shall be nevertheless useful to newcomers. The good timeliness shall necessary occur between the 'too early' and the 'could have been earlier', but the probabilistic setting implies that the vote of optimal policies cannot be ensured. This is particularly annoying when one reckons the difficulties to get rid of policies impeding opportunistic exploration of better behaviours.

Another weakness regards the dissonance arousing from reinforcement learning agents and norm-governed agents. On the one hand, learning agents are

supposed to pursue a maximisation of individual wealth by balancing the exploitation of promising strategies and the exploration of other options. On the other hand, norms tend to imped opportunistic exploration. Norms stall learning, and thereby agents may get trapped into suboptimal prescriptive systems.

5 Related work

Social norms are often studied in two extremes: in game theoretical settings of strategic agents and in simulation of thoughtless agents like evolutionary game theoretical investigations. In both types of approaches, the convergence to an equilibrium is interpreted as the emergence of a social norm: norms are not explicitly represented and agents do not have a mental representation of them.

On the contrary, formal logics (typically deontic logics and argumentation, see e.g. [13]) are commonly investigated to represent and reason upon explicit norms, leading eventually to architecture for cognitive agents (see e.g. [7,2]). These architectures are usually based on a BDI template and without learning abilities, while our agents are logic-based and reinforced learners but they have no explicit desires or intentional features (their implicit desire is to maximize the accumulation of payoffs). BDI frameworks usually assume that prescriptions are built-in whereas our agents have to ability to learn best behaviours and thereby generate new prescriptions (though prescriptions could be also built-in). The limitation of BDI architecture with regard to norm recognition has been addressed by Conte at al. in [12] where BDI agents recognise norms by observing other agents, c.f. [3]. Our agents transfigure individual experiences into prescriptions without the need to observe other agents, and the utility of these individual experiences are the results of the (inter)actions with other agents.

Multi-agent learning is an active field of research where agents are meant to coordinate by learning joint actions, typically using individual reinforcement learning or its extensions to collective tasks. Partalas et al. proposed in [9] to combine reinforcement learning with voting. Their agents learn predefined strategies (joint actions) while our agents learn individual actions. When their agents are in a strategic state they vote for a common strategy: there is no transfiguration and no construction of prescriptions.

With regard to norm-synthesis, the problem was pioneered by the work of Shoham and Tennenholtz [15]. Fitoussi and Tennenholtz [6], for example, described the synthesis of 'minimal' and 'simple' prohibitions. The rationale for minimality is that a minimal norm provides the agents more freedom in choosing their behaviour (that is, it prohibits fewer actions) while ensuring that they conform to the system specification. The rationale for simplicity is that a simple norm relies less on the agents capabilities rather than a non-simple one. Agotnes and Wooldridge [1] included the implementation costs of norms and multiple design goals with different priorities. Christelis and Rovatsos [4] proposed a first-order planning approach to better cope with the size of the state-space. The approaches mentioned above are typically applied off-line. However, off-line design is not appropriate for coping with open systems, that are inherently dynamic

and the state space may change over time. To address this issue, Morales et al [8] proposed a mechanism called IRON for the on-line synthesis of norms. IRON employs designated agents, often called 'institutional agents' [5], representing a norm-governed system/institution, and observing the interactions of the members of the system in order to synthesise conflict-free norms without lapsing into over-regulation. Our work is fundamentally different: we target multi-agent systems without designated agents receiving updates about the system interactions and the authority to enforce norms.

6 Conclusion and future directions

We tackled the challenge regarding prescriptive transfiguration and self-governance. We proposed a simple cognitive apparatus with which learning agents can transfigure learning policies (and thus behavioural patterns) into prescriptions. This apparatus was coupled to a consensus system so that agents can submit prescriptions for a vote and vote eventually for their enforcement.

The simulations of a self-governed population of learning agents suggested the benefits of our approach with regard to the convergence to desirable behaviours. However, simulations with large stochastic games have to confirm these benefits. Timeliness in run-time construction with learning agent appeared of the most importance. A vote may indeed occur when there is a risk agents consider inadequate prescriptions, or when useless prescriptions shall enforce behaviours already adopted by agents. Nevertheless, these useless prescriptions shall ease the decision-making and coordination of newcomers.

In practice, our proposal illustrates an alternative of off-line construction of prescriptive systems: a domain-independent construction at run-time of explicit primary regulative prescriptions from scratch, for and by learning agents, without any agent having a complete information on the system.

Future directions can be multiple. They include learning of joint actions and the transfiguration of these collective into complex prescriptive systems, distributed consensus systems (possibly in network) to avoid a central body collecting the votes. An important point regards learning of norms modifications so that agents can escape from unfortunate prescriptive systems. But how could agents change prescriptions without having the possibility to explore and without jeopardizing the coherence and the temporal stability of the overall system? A solution holds in agents simulating the system to explore "without moving" but it implies computational resources a priori incompatible with bounded agents. Maverick agents on whose payoffs santions have a less significant effect may be an interesting line of research.

Eventually, we hope the reader found inspiration in a manner to bridge the gap between social norms and prescriptions, and its use for run-time constructions of prescriptive systems in a population of learning agents and thereby for self-organisation and in particular self-governance.

Acknowledgements The authors would like to thank the anonymous reviewers. Part of this work is supported by the Marie Curie Intra-European Fellowship PIEF-GA-2012-331472.

References

1. T. Ågotnes and M. Wooldridge, 'Optimal social laws', in *AAMAS*, pp. 667–674, (2010).
2. J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre, 'The boid architecture: Conflicts between beliefs, obligations, intentions and desires', in *Proceedings of the Fifth International Conference on Autonomous Agents*, AGENTS '01, pp. 9–16, New York, NY, USA, (2001). ACM.
3. C. Castelfranchi, F. Giardini, E. Lorini, and L. Tummolini, 'The prescriptive destiny of predictive attitudes: from expectations to norms via conventions', in *In: Proceedings of CogSci 2003, 25th Annual Meeting of the Cognitive Science Society*, (2003).
4. G. Christelis, M. Rovatsos, and R. P. A. Petrick, 'Exploiting domain knowledge to improve norm synthesis', in *AAMAS*, pp. 831–838, (2010).
5. M. Esteva, J. Rodriguez-Aguilar, J. Arcos, C. Sierra, and P. Garcia, 'Institutionalising open multi-agent systems', in *Proceedings of the International Conference on Multi-agent Systems (ICMAS)*, ed., E. Durfee, 381–382, IEEE Press, (2000).
6. D. Fitoussi and M. Tennenholtz, 'Choosing social laws for multi-agent systems: minimality and simplicity', *Artificial Intelligence*, **119**(1-2), 61–101, (2000).
7. G. Governatori and A. Rotolo, 'Bio logical agents: Norms, beliefs, intentions in defeasible logic', *Autonomous Agents and Multi-Agent Systems*, **17**(1), 36–69, (2008).
8. J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, Michael Wooldridge, and Wamberto Vasconcelos, 'Automated synthesis of normative systems', in *AAMAS*, pp. 483–490, (2013).
9. I. Partalas, I. Feneris, and I. P. Vlahavas, 'Multi-agent reinforcement learning using strategies and voting', in *ICTAI (2)*, pp. 318–324. IEEE Computer Society.
10. *Handbook of Law and Economics*, eds., A. Mitchell Polinsky and S. Shavell, volume 1, Elsevier, 1 edn., 2007.
11. R. Riveret, A. Rotolo, and G. Sartor, 'Probabilistic rule-based argumentation for norm-governed learning agents', *Artif. Intell. Law*, **20**(4), 383–420, (2012).
12. *Minding Norms: Mechanisms and dynamics of social order in agent societies*, eds., G. Andrighetto R. Conte and M. Campennl, Oxford Scholarship, 2013.
13. G. Sartor, *Legal Reasoning: A Cognitive Approach to Law*, Springer, 2005.
14. S. Sen and S. Airiau, 'Emergence of norms through social learning', in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 1507–1512, San Francisco, CA, USA, (2007). Morgan Kaufmann Publishers Inc.
15. Y. Shoham and M. Tennenholtz, 'On social laws for artificial agent societies: off-line design', *Artificial Intelligence*, **73**(1-2), 231–252, (1995).
16. R.S. Sutton and A.G. Barto, *Reinforcement learning: An introduction*, volume 116, Cambridge Univ Press, 1998.