Self-Governance by Transfiguration: From Learning to Prescription Changes

Régis Riveret, Erivelton G. Nepomuceno, Jeremy Pitt
Dpt of Electrical and Electronic Engineering
Imperial College
London, United Kingdom
Email: {r.riveret,e.nepomuceno,j.pitt}@imperial.ac.uk

Alexander Artikis
University of Piraeus
NCSR demokritos
Athens, Greece
Email: a.artikis@unipi.gr

***Abstract*—Reinforcement learning is a widespread mechanism for adapting the individual behaviour of autonomous agents, while norms are a well-established means for organising the common conduct of these agents. Therefore, norm-governed reinforcement learning agents appear to be a powerful bio-inspired, as well as socio-inspired, paradigm for the construction of decentralised, self-adapting, self-organising systems. However, the convergence of learning and norms is not as straightforward as it appears: learning can 'misguide' the development of norms, while norms can 'stall' the learning of optimal behaviour. In this paper, we investigate the self-governance of learning agents, or more specifically the domain-independent (de)construction at run-time of prescriptive systems from scratch, for and by learning agents, without any agent having complete information of the system. Most importantly, because prescriptions may also misguide agents, we allow them to repeal any misguiding prescriptions that have previously been enacted. Simulations illustrate the approach with experimental insights regarding scalability and timeliness in the construction of prescriptive systems.**

I. Introduction

Reinforcement learning is a widespread mechanism for the adaptation of individual autonomous agents with incomplete information on their environment [1], while norms are a well-established means for guiding the common conduct of these agents, thereby easing their decision-making, coordination and thus organisation. When learning agents manage their own norms, for example when they self-govern, then norm-governed reinforced learning agents appear as a natural bio-inspired, as well as socio-inspired paradigm for the construction of decentralised, self-adapting, self-organising systems.

Norms in these systems may appear with different degrees of explicitness ranging from fully unambiguous written prescriptions to implicit unwritten norms and tacit emerging patterns. Computer scientists' studies mirror this polarity. Explicit norms are typically investigated in formal logics (e.g. deontic logics and argumentation, see e.g. [2], [3]) to represent and reason upon them, leading eventually to architecture for cognitive agents (see e.g. [4]) while implicit norms are accounted as patterns emerging from repeated interactions amongst agents (typically learning agents see e.g. [5]). This polarisation reflects the different treatments of norms by jurists and social researchers. While jurists concentrate on explicit norms as prescriptions promulgated by institutional powers and enforced by clear sanctions, social researchers study norms as tacit behavioural patterns emerging from expectations and enforced by entwined sanctions. Scholars have investigated the influence of social norms and prescriptions on each other, but the conceptual gap remains hardly explored in the community of computer scientists, c.f. [6].

Artificial population of norm-governed learning agents are thus systems where implicit and explicit norms and their mutual influences can be formally studied, with the opportunity to take each of their advantages for practical applications where 'ideal' behaviours shall be prescribed, ranging from protocols in autonomic intelligent networks to coordination of agents in business process management. In norm-governed systems, explicit norms are commonly advocated to facilitate their updates, and consequently system maintenance, improve system transparency and ease system governance. However, as stated by [7], 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. In contrast, implicit norms offer an "order without law" at no expensive cost and they are flexible so that they can change as the environment changes. Furthermore, if implicit social norms are understood as patterns emerging in the presence of learning agents, and assuming that learning agents are meant to maximise some (social) utility function, then it is arguable that implicit social norms are tightly connected to some (social) maxima.

To take advantage of explicit and implicit norms, we investigate the construction at run-time of prescriptive systems based on the emergence of behavioural patterns of learning agents. And since systems of multiple autonomous agents have their essence into decentralised control and computation, this construction at run-time shall occur in a distributed manner in the sense there is no entity with complete information taking the role of a central legislative body, c.f. [7].

However, the enactment of prescriptions without the possibility of repeal or abrogation may mislead the agents. This is particularly problematic with regard to a dissonance arousing in norm-governed reinforced learning agents: while learning agents are supposed to pursue a maximisation of individual utility by balancing the exploitation of promising strategies and

the exploration of other options, norms and entwined sanctions tend to imped opportunistic exploration. Norms stall learning, thereby agents' adaptation, eventually leading the system into normative traps: though prescriptions are commonly understood to guide agents, they may also misguide them. To get away from these misguidances, repeal must be possible so that agents can deconstruct prescriptive systems.

In this paper, we investigate thus the self-governance of learning agents, or more specifically the domain-independent (de)construction at run-time of prescriptive systems from scratch, for and by learning agents, without any agent having a complete information on the system.

The proposed solution for the self-governance of learning agents is a development of [8] with the possibility of repeals: it is based on a transfiguration, enabling agents to express their learning experiences into prescription changes, and a coupled consensus system to enact these changes. The overall system results into a direct self-governance taking advantage of every agents' learning experiences. The proposal is a development of the framework presented in [8] so that learning agents have the possibility to repeal prescriptions without external intervention.

We focus thus on explicit primary norms and in particular regulative prescriptions, i.e. those guiding the ideal behaviour of agents, with normative changes about regulative prescriptions (catering for secondary prescriptions managing primary prescriptions) taking the simple forms of enactments (activation) and repeals (deactivation). Other primary norms such as constitutive norms and more sophisticated changes (in particular with regard to temporal aspects such a retroactive changes) are left for future investigations.

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 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 potential and we assume that every agent is endowed with a computational 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 show the pertinence of such approach, in particular with regard to the issue of prescriptions stalling learning, timeliness in norm construction and scalability.

The remainder of this paper is organised as follows. In Section II, 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 self-governance is given in Section III and is evaluated in Section IV with respect to some simulation results. Section V puts the work in context while Section VI presents our conclusions.

II. Problem setting

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 rewards. 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;
- $\mathcal{S} = \{s_1, \dots s_n\}$ 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 state s_r at time $t+1$ when attempting the joint action A in the state s_q at time t.
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^N$ is a payoff function, $R_i(s_q, A, s_r) = r_i(s_t = s_q, s_{t+1} = s_r)$ is the payoff of agent i upon transition from state s_q at time t to state s_r a 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 in a state at time t is to maximize the expected value of *the infinite horizon discounted return*:

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

that is,

$$R_{i,t} = \sum_{k=0}^{\infty} \gamma^k r_i(s_{t+k}, s_{t+1+k}) \tag{1}$$

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 [1] meant to pursue the best policies. At each time step, every agent senses its environment, and, given the observed state, every agent simultaneously selects a 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 of an agent 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 rewards associated to the individual action $a_{i,j}$ in state s. Let $a_{i,t}$ be an action selected in a state $s_t = s$ at time t, the quality $Q(s, a_{i,t})$ is updated as follows:

$$Q(s_t, a_{i,t}) \leftarrow Q(s_t, a_{i,t}) + \alpha_i \cdot \Delta_i \tag{2}$$

with

$$\Delta_i = r_{i,t+1} - Q(s_t, a_{i,t}) + \gamma_i . Q(s_{t+1}, a_{i,t+1}) \tag{3}$$

where α_i is a learning rate, and γ_i a discount factor trading off the importance of recent versus later rewards. 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:

$$\pi_i^t(s_t, a_{i,t}) = \frac{e^{Q(s_t,a_{i,t})/\tau_i}}{\sum_{a_{i,j}} e^{Q(s,a_{i,j})/\tau_i}} \quad (4)$$

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 objective 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. In this regard, when agents are guided by some prescriptions and in particular when these agents self-govern then we have to distinguish material utility to the normative sanctions. To distinguish the material payoff of a behaviour to eventual sanctions, we decompose the payoff $r_i(s_t, s_{t+!})$ into a material payoff denoted $r_{m,i}(s_t, s_{t+1})$ and a legal payoff $r_{s,i}(s_t, s_{t+1})$ (i.e. the sanctions) such that:

$$r_i(s_t, s_{t+1}) = r_{m,i}(s_t, s_{t+1}) + r_{s,i}(s_t, s_{t+1}) \quad (5)$$

Accordingly, we assume that any agent keeps track of the material qualities of behaviours:

$$Q_m(s_t, a_{i,t}) \leftarrow Q_m(s_t, a_{i,t}) + \alpha_i \cdot \Delta_{m,i} \quad (6)$$

where

$$\Delta_{m,i} = r_{m,i,t+1} - Q_m(s_t, a_{i,t}) + \gamma_i \cdot Q_m(s_{t+1}, a_{i,t+1}) \quad (7)$$

The distinction between material payoffs and sanctions allows us to define the *material* return of a prescriptive system $\mathcal{N}$ as the infinite horizon discounted material return:

$$R_{\mathcal{N},t} = \sum_{i \in \mathcal{G}} \sum_{k=0}^{\infty} \gamma^k r_{m,i}(s_{t+k}, s_{t+1+k}) \quad (8)$$

This return captures the global wealth of the agent population without taking into account other important measures, for examples those related to or inspired by social notions such as freedom or justice. For this reason, we may consider another social reward denoted r^* that we left unspecified and which is meant to account for the aggregation of other possible social measures. Accordingly, the objective of a prescriptive system N is to maximise the expected value of the *social* return defined as the following infinite horizon discounted social return:

$$R^*_{\mathcal{N},t} = \sum_{i \in \mathcal{G}} \sum_{k=0}^{\infty} \gamma^k r^*_i(s_{t+k}, s_{t+1+k}) \quad (9)$$

The problem we address here with regard to self-governance is the construction of a prescriptive system, i.e. a set of conditional obligations and prohibitions with the definition of their sanctions, for and by agents to maximize the social return. In this paper, management activities like the enforcement of prescriptions is assumed costless, leaving the integration of these costs for future investigations.

To address this self-governance, we will investigate the transfiguration of agents' policies (i.e. learning experiences and thereby behavioural patterns) into prescriptions. In particular, 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 the prescriptive system.

Example 1: We will illustrate and evaluate the 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. [9]). 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.

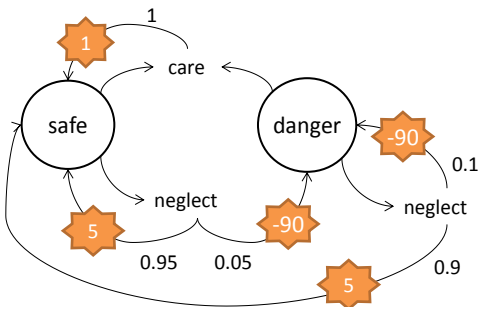

Fig. 1. The Markov decision process graph for a single agent. Each transition from an action to a state is represented by an arrow labelled with its probability and associated material payoff.

This example will be used in the remainder to illustrate our proposal for self-governance.

III. Self-governance model

To address self-governance, we endow agents with a cognitive apparatus to transform learning experiences into prescription changes and this apparatus is coupled with a consensus mechanism so that agents make a social choice on changes meant to govern themselves. The construction of the prescriptive system occurs in two phases:

1) *Individual prescriptive transfiguration*: every agent individually transfigures learning experiences into possible prescription changes,

2) *Consensus*: every agent submits a change and the most common proposal becomes the motion, then every agent votes for the motion.

The corresponding pseudo-code animating the population in its environment is given in Algorithm 1.

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 rewards;
    Update quality of behaviours;
    Individual prescriptive transfiguration;
  end for
  Consensus (submission, motion selection and voting);
end for
```

In the remainder of this section, we describe the individual prescriptive transfiguration (generation and selection of prescription changes), and the consensus system (submission, motion selection and vote).

A. Individual prescriptive transfiguration

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

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

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

Example 2: For every agent, there are four possible prescriptions in each state:

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

Notice that we assume no equivalence between the obligation to act with care and the prohibition to act with negligence, but a kind of quantitative equivalence shall appear when we will introduce potentials for prescription changes (see below).
Possible self-prescriptions are not part of the prescriptive system in force, but every agent shall propose the most relevant amongst all them as a motion to the whole population before voting for its activation. When a prescription is active, then agents shall deactivate it. A prescription change is the activation or the deactivation of a prescription. A prescription change will be effective once there is consensus amongst the population of agents to do so at the end of a voting session. In each activity, every agent has to make choice about changes (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 prescription change with a flavour of reinforcement learning, as we will see in the remainder of this section.

The individual transfiguration of learning experiences into prescriptions is decomposed in two steps: first the agent decides or not to self-prescribe, then eventually, a submissible proscriptive change is drawn.

1) Generation of prescription changes: Before selecting prescription changes, every agent shall consider the possible sets of prescription activations and deactivations. A possible change with respect to prescription is conditioned with regard to (i) the current legal state of this prescription in terms of activation or deactivation (a prescription shall be activated if it is not active) (ii) the legal context of other prescriptions (e.g. a prohibition shall not be activated when the obligation of an alternative behaviour is activated), and (iii) an entropic measure of the set of alternative behaviours (one shall not prescribe a behaviour amongst behaviours with similar qualities). Let's detail these ideas.

First we define a threshold condition so that an agent will not transfigure learning policies when alternative behaviours have similar qualities. Indeed, there is no advantage to oblige or prohibit a behaviour with respect to 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 with respect to the material qualities. If $\mathcal{S}_i$ is less than a threshold τ_i^S then the agent will draw a prescription change. 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 and thus non-optimal changes may be considered. At the opposite, if the threshold is set to low, then the agent may have so much experiences that legal changes 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 material quality 4 and the negligent behaviour has material 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 $\tau_i^S = 0.5$, then Tom will consider alternative prescriptions to elevate one to the rank of submissible prescription. If the entropy was higher than this threshold, then Tom would consider no prescription for the safe state.

We assume that an obligation is activable if, and only if, it is not activated and the entropic threshold is attained and:

- there is no alternative action which is prohibited, and
- there is no alternative action which is obliged.

A prohibition is activable if it is not activated and the entropic threshold is attained and:

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

A prescription is deactivable if, and only if, it is activated.

A change of a prescription r is either an activation, denoted r^+, or a deactivation denoted r^-. A set of activation considered by agent i is denoted $\mathcal{R}_i^+$ while a set of deactivations is written $\mathcal{R}_i^-$. A set of prescription changes is denoted $\mathcal{R}_i^\pm$ is the union of activations and deactivations, i.e. $\mathcal{R}_i^\pm = \mathcal{R}_i^+ \cup \mathcal{R}_i^-$.

2) Selection of prescription changes : If an agent decides to transfigure learning experiences into prescription changes then it will draw a change that becomes a *submissible change*. To do so, every possible change is associated with a scalar measure that we call the submissible potential. The higher the material quality of a behaviour with respect to the material quality of other behaviours, the higher its potential to be associated to an obligation. At the opposite, the lower the material quality of a behaviour with respect to the material quality of other behaviours, the higher its potential to be associated to a prohibition. Let $\widehat{Q}_{m,i}$ denote the average of the material qualities of alternative behaviours in a state according to agent i. For the activation of an obligation $\mathsf{Obl}\,(B)$, its submissible potential according to agent i, denoted $\delta_i(r^+_{\mathsf{Obl}\,(B)})$, is the difference between the material quality for behaviour B and the average of the material qualities of alternative behaviours. The deactivation of an obligation, denoted $\delta_i(r^-_{\mathsf{Obl}\,(B)})$, has the opposite potential.

$$\delta_i(r^+_{\mathsf{Obl}\,(B)}) = Q_{m,i}(B) - \widehat{Q}_{m,i} \tag{10}$$

$$\delta_i(r^-_{\mathsf{Obl}\,(B)}) = \widehat{Q}_{m,i} - Q_{m,i}(B) \tag{11}$$

and for changes regarding prohibitions, we have the opposite:

$$\delta_i(r^+_{\mathsf{Forb}\,(B)}) = \widehat{Q}_{m,i} - Q_{m,i}(B) \tag{12}$$

$$\delta_i(r^-_{\mathsf{Forb}\,(B)}) = Q_{m,i}(B) - \widehat{Q}_{m,i} \tag{13}$$

Consequently the following equalities hold:

$$\delta_i(r^+_{\mathsf{Obl}\,(B)}) = -\delta_i(r^+_{\mathsf{Forb}\,(B)}) \tag{14}$$

$$\delta_i(r^+_{\mathsf{Obl}\,(B)}) = -\delta_i(r^-_{\mathsf{Obl}\,(B)}) \tag{15}$$

$$\delta_i(r^+_{\mathsf{Forb}\,(B)}) = \delta_i(r^-_{\mathsf{Obl}\,(B)}) \tag{16}$$

On this basis, every agent i shall consider the submissibility of possible changes with a probability p_i^δ using a Gibbs-Boltzmann distribution over the potentials:

$$p_i^\delta(r^\pm) = \frac{e^{\delta(r^\pm)/\tau_i^\delta}}{\sum_{r^\pm \in \mathcal{R}_i^\pm} e^{\delta(r^\pm)/\tau_i^\delta}} \tag{17}$$

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 change with the highest submissible potential. The probability of a possible activation will be written $p_i^{\delta^+}$ and the probability of a possible deactivation is written $p_i^{\delta^-}$.

Example 4: Table 1 illustrates Tom's measure of submissible potentials and the associated probabilities when there is no active prescriptions. 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. 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.

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

TABLE I
ILLUSTRATION OF SUBMISSIBLE QUALITIES δ_i AND ASSOCIATED PROBABILITIES $p_{\delta,i}$ TO CONSIDER THE CHANGE AS SUBMISSIBLE.

B. Consensus

The consensus about the enactment of changes is decomposed in two phases: (i) submission and motion selection, and (ii) vote.

1) Submission and motion selection: Once some agents have transfigured some learning experiences into a set of prescription changes, these agents shall submit each a change. A submitted change is a submission. The most common submission becomes a motion, and agents shall vote for it. Every agent will draw a submission from the set of submissible changes using again a Gibbs-Boltzmann distribution. Let $\mathcal{S}_i^\pm$ be the set of submissible changes of agent i, the agent i will draw a change $r^\pm$ from this set with a probability $p_i^D(r^\pm)$ from a Gibbs-Boltzmann distribution over the potentials $\delta_i(r^\pm)$:

$$p_i^D(r^\pm) = \frac{e^{\delta_i(r^\pm)/\tau_i^D}}{\sum_{r^\pm \in \mathcal{S}_i^\pm} e^{\delta_i(r^\pm)/\tau_i^D}} \tag{18}$$

where τ_i^D balances 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.

Example 5: Between the activation of 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 activation of the obligation to act with care. We further assume that the most common proposals by the population is the activation of the prohibition to act with negligence when the state is safe. Consequently, this proposal becomes a motion.

At this stage, the activation of a prescription 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 $\delta(r^+)$ of the change proposal 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 activation motion r^+ with a potential $\widehat{\delta}(r^+)$ which is the average of the potentials of the proposals unifying with r^+. The average potential of an activation is meant to feature the value of a scalar sanction. Accordingly, we choose in this paper to define the sanction as $\widehat{\delta}(r^+).\mu$ where μ is a positive real number (typically set superior to 1).

Example 6: Suppose that three 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 activation motion r^+ is $\widehat{\delta}(r^+) = 3$. Assuming $\mu = 10$ the associated scalar sanction associated to this prohibition is 30.

2) Vote: Once there is a motion, 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 $r^\pm$ is measured by its average potential $\widehat{\delta}(r^\pm)$ (featuring the sanction of the associated prescription - see previous section). Since agents have to vote about the motion about a prescription and its associated sanction $|\widehat{\delta}(r^\pm)|.\mu$, then we suppose that agents are communicated $\widehat{\delta}(r^\pm)$. We further assume that an agent shall vote in favour or against a motion by comparing the average potential of this motion $\widehat{\delta}(r^\pm)$, with the potential of this motion according to this agent $\delta_i(r^\pm)$. The lower the difference between the potential $\delta_i(r^\pm)$ of the motion and the average potential $\widehat{\delta}(r^\pm)$, the higher the probability for agent i to vote for the motion.

Accordingly, we capture these considerations with a scalar measure called the *individual potential of activation*, c.f. [8]. The agent i's individual potential of a change $r^\pm$, denoted $\Delta_i(r^\pm)$, is a linear function of $\delta_i(r^\pm)$:

$$\Delta_i(r^\pm) = \frac{\delta_i(r^\pm) - \mu^\Delta}{s^\Delta} \tag{19}$$

where μ^Δ is a threshold such that

$$\mu^\Delta = \frac{\tau_i^\Delta - 1}{\tau_i^\Delta} \cdot \hat{\delta}_i(r^\pm) \tag{20}$$

so that the potential can be more intelligibly rewritten to highlight the difference $\delta_i(r^\pm) - \hat{\delta}_i(r^\pm)$:

$$\Delta_i(r^\pm) = \frac{[\delta_i(r^\pm) - \hat{\delta}_i(r^\pm)] + \hat{\delta}_i(r^\pm)/\tau_i^\Delta}{s^\Delta} \tag{21}$$

Notice that $\Delta_i(r^\pm) = 0$ when

$$\delta_i(r^\pm) = \hat{\delta}_i(r^\pm) - \hat{\delta}_i(r^\pm)/\tau_i^\Delta \tag{22}$$

An agent i will vote in favour of a change $r^\pm$ with a probability $p_i^\Delta(r^\pm)$ using a logistic function akin to the Gibbs-Boltzmann family:

$$p_i^\Delta(r^\pm) = \frac{1}{1 + e^{\Delta_i(r^\pm)}} \tag{23}$$

Thus, when the parameter τ_i^Δ verifies Eq. 22 then $p_i^\Delta(r^\pm) = 1/2$. The lower τ_i^Δ, the higher the probability for agent i to vote for the motion, at the risk of being ruled by a minority. We will experiment the influence of this parameter in the next section with simulations.

Example 7: Recall the most common submitted change by the population is the activation of a 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, $\widehat{\delta}(r^+) = 3$. The individual potential of Tom for this motion is thus:

$$\Delta_{Tom}(r^+) = \frac{1 - 1.1 \times 3}{10} \sim -0.23 \tag{24}$$

Tom will vote in favour of this activation with a probability $p_{Tom}^\Delta(r^+) = 1/(1 + e^{-0.23}) \sim 0.56$.

Interestingly, it is easy to see that the symmetry of individual potentials entails a coherence in the probability of a vote with respect to prescription changes. Let r a prescription, r_{Obl} the obligation and r_{Forb} the prohibitive counterpart. Suppose four scenarios where the motion is the (de)activation of either the obligation or the prohibition. By definition:

$$p_i^\Delta(r^+_{\mathsf{Obl}}) = \frac{1}{1 + e^{\Delta_i(r^+_{\mathsf{Obl}})}} \tag{25}$$

$$p_i^\Delta(r^+_{\mathsf{Forb}}) = \frac{1}{1 + e^{\Delta_i(r^+_{\mathsf{Forb}})}} \tag{26}$$

$$p_i^\Delta(r^-_{\mathsf{Obl}}) = \frac{1}{1 + e^{\Delta_i(r^-_{\mathsf{Obl}})}} \tag{27}$$

$$p_i^\Delta(r^-_{\mathsf{Forb}}) = \frac{1}{1 + e^{\Delta_i(r^-_{\mathsf{Forb}})}} \tag{28}$$

Suppose that in each scenario the potentials $\delta_i(r)$ are the same, and that the change has been submitted by the same set of agents, from the Eq. 14, 15 and 16 we have:

$$\hat{\delta}_i(r^+_{\mathsf{Obl}}) = -\hat{\delta}_i(r^+_{\mathsf{Forb}}) \tag{29}$$

$$\hat{\delta}_i(r^+_{\mathsf{Obl}}) = -\hat{\delta}_i(r^-_{\mathsf{Obl}}) \tag{30}$$

$$\hat{\delta}_i(r^+_{\mathsf{Forb}}) = \hat{\delta}_i(r^-_{\mathsf{Obl}}) \tag{31}$$

and thus:

$$\Delta_i(r^+_{\mathsf{Obl}}) = -\Delta_i(r^+_{\mathsf{Forb}}) \tag{32}$$

$$\Delta_i(r^+_{\mathsf{Obl}}) = -\Delta_i(r^-_{\mathsf{Obl}}) \tag{33}$$

$$\Delta_i(r^+_{\mathsf{Forb}}) = \Delta_i(r^-_{\mathsf{Obl}}) \tag{34}$$

Since

$$\frac{1}{1 + e^x} + \frac{1}{1 + e^{-x}} = 1 \tag{35}$$

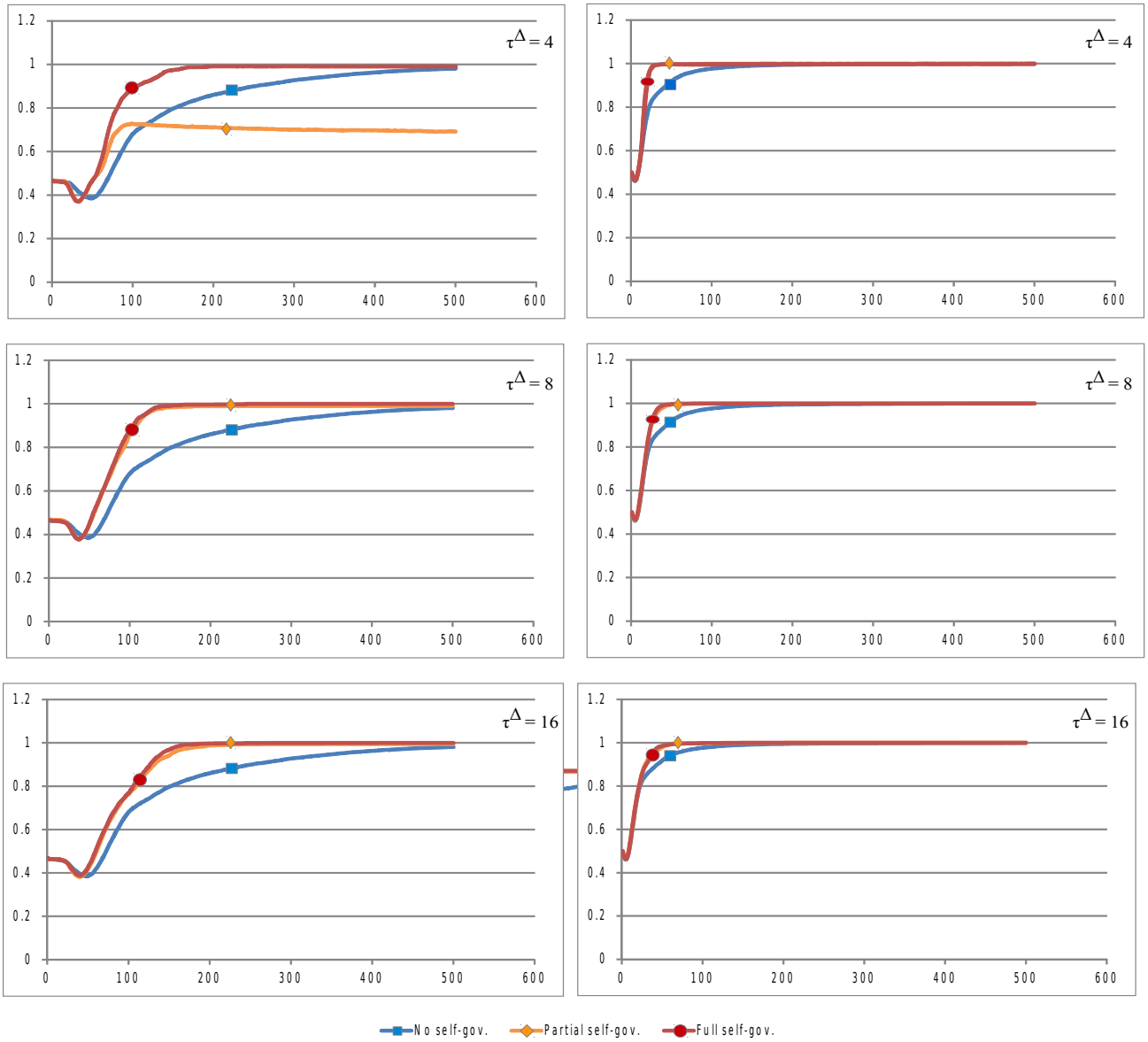

Fig. 2. Probability of careful behaviours in the safe state (left graphs) and in the dangerous state (right graphs) with respect to three values of the parameter τ^{Δ} (4, 8, and 16) in three regimes: self-governance, partial self-governance (no possible repeals) and no self-governance vs. time.

we have:

$$p_i^{\Delta}(r^{+}_{\mathsf{Obl}}) = 1 - p_i^{\Delta}(r^{+}_{\mathsf{Forb}}) \tag{36}$$

and

$$p_i^{\Delta}(r^{+}_{\mathsf{Obl}}) = 1 - p_i^{\Delta}(r^{-}_{\mathsf{Obl}}) \tag{37}$$

In words, if an agent was to vote in favour of the activation of an obligation with a probability $p_i^{\Delta}(r^{+}_{\mathsf{Obl}})$, then this agent would vote in favour of the activation of the prohibitive counterpart with a probability $1 - p_i^{\Delta}(r^{+}_{\mathsf{Obl}})$, or following the symmetry, this agent would vote in favour of its deactivation with a probability $1 - p_i^{\Delta}(r^{+}_{\mathsf{Obl}})$.

The consensus can take many different forms, it can be distributed or centralised. For our purposes we arbitrary considered a majority rule. Accordingly an activated prescription and its enforcement voted by the majority enters in force while a deactivated prescription is abrogated. 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).

IV. Simulation results

To evaluate and get more insights into the proposed 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 [10], so that the

τ_i^Δ	2	4	8	16		2	4	8	16
r^+ : safe $\Rightarrow$ Obl care	51	40	53	57	r^+ : danger $\Rightarrow$ Obl care	47	55	43	59
r^+ : safe $\Rightarrow$ Forb care	44	14	1	0	r^+ : danger $\Rightarrow$ Forb care	3	0	0	0
r^+ : safe $\Rightarrow$ Obl neglect	51	18	1	0	r^+ : danger $\Rightarrow$ Obl neglect	13	1	0	0
r^+ : safe $\Rightarrow$ Forb neglect	49	59	47	43	r^+ : danger $\Rightarrow$ Forb neglect	53	45	57	42
r^- : safe $\Rightarrow$ Obl care	0	0	0	0	r^- : danger $\Rightarrow$ Obl care	0	0	0	0
r^- : safe $\Rightarrow$ Forb care	44	14	1	0	r^- : danger $\Rightarrow$ Forb care	3	0	0	0
r^- : safe $\Rightarrow$ Obl neglect	51	17	1	0	r^- : danger $\Rightarrow$ Obl neglect	13	1	0	0
r^- : safe $\Rightarrow$ Forb neglect	0	0	0	0	r^- : danger $\Rightarrow$ Forb neglect	0	0	0	1

TABLE II
NUMBER OF PRESCRIPTION CHANGES FOR 100 RUNS.

system specifications were directly executed. The results are averaged over 100 runs of a population of 100 agents animated for 500 time steps.

The number of changes with respect to the parameter τ_i^Δ (see Section III-B2) are shown in Table 2. When τ_i^Δ was set low then agents had a tendency to vote in favour of motion in a hasty manner resulting into undesirable enactments, but they quickly repealed all them before the ends of the simulation (exception to one obligation to behave with negligence in the safe state for $\tau_i^\Delta = 4$). When τ_i^Δ was increased agents became more picky in their vote, this effect resulted in less enactments of undesirable prescriptions but also with the possibly that they did not get repealed. When τ_i^Δ was set sufficiently high, then no undesirable enactments occurred. Whatever the value of τ_i^Δ, simulations indicate that the approach exhibits stability in the sense that there is no back and forth of enactments and repeals concerning the prescriptions.

The probability of careful behaviours in the safe and dangerous state is shown in Fig. 2 with respect to different values of the parameter τ^Δ (4, 8 and 16) and three regimes: with or without self-governance, and with partial self-governance for which the repeal of prescriptions was not possible. When self-governance was deactivated, agents learned to behave with care in both states, but the convergence was slower in the safe state as the careful and negligent behaviours in this state have closer expected utilities. In the regime of full self-governance, the enforcement of careful behaviours guided the agents towards careful behaviours with a higher speed of convergence in both states. The value $\tau^\Delta = 8$ showed a higher speed of convergence compared to the runs with $\tau^\Delta = 16$ because agents voted in a less sluggish manner. When $\tau^\Delta = 4$, agents were hastier to enact prescriptions: the comparison of the regimes of partial self-governance (no possible repeals) and full self-governance in the safe state demonstrates the importance of enabling agents to repeal prescriptions.

The enactments of misguiding prescriptions (and in particular one without its repeal before the end of the simulation run) suggest 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 prescriptions 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 emphasises the importance of timeliness in norm construction. If agents vote in a hasty manner then a vote occurs too early and there is risk that agents activate prescriptions enforcing undesired behaviours. At the opposite, if agents are sluggish to vote, then a late vote may imply prescriptions enforcing a well-established social norm; in this case these prescriptions shall be nevertheless useful to newcomers. The good timeliness necessary occurs between the 'too early' and the 'could have been earlier': investigations of optimal regimes will be looked for in future work.

In order to get more experimental insights, in particular with respect to scalability, we run stochastic games extending the example with up to 50 states populated by 50 agents. The games were randomly generated as follows. Each state s is associated with a degree $d(s)$ of safeness from 1 to 50, $d(s) \in [1, 50]$, such that the safest state has degree 1. In any state, every agent can act with care or with negligence. If all the agents act with care in a state s_t with a degree strictly more than 1 then the next state s_{t+1} will be safer with a random degree, $d(s_{t+1}) < d(s_t)$. If an agent acts with negligence in a state s_t with degree strictly less than 50 then there is a risk of an accident and the next state s_{t+1} is more dangerous with a random degree, $d(s_t) < d(s_{t+1})$. The higher the safeness of the state s_t, the lower the probability p of an accident ($p = 1/\sqrt{d(s_t)}$), and the lower the individual payoff of careful or dangerous behaviours, $\sqrt{d(s_t)}$ and $2\sqrt{d(s_t)}$ respectively. If there is no accident then the individual payoff of a dangerous behaviour is therfore higher than the payoff of a safe behaviour. If there is an accident then the payoff of the faulty agent is diminished by $-2.5d(s_t)$, in this case the individual payoff of a dangerous behaviour is therefore lower than the payoff of a safe behaviour. For every occuring accident the payoff of every agent is diminished by $-\sqrt{d(s_t)}$. Finally, whatever the agents behaviours in the safest state, if there is an accident then the next state is the most dangerous state, and if there is no accident then the the next state is more dangerous (without necessarily being the most dangerous).

The global wealth (i.e. the sum of payoffs accumulated

by agents) with or without self-governance (averaged over 50 games) over time is shown in Fig. 3, for $\tau^{\Delta} = 10$. On average, the self-governed runs allow for faster wealth accumulation compared to the non self-governed runs. The reason is that agents enacted the obligation to act with care (or the prohibition to act with negligence), resulting in less accidents.

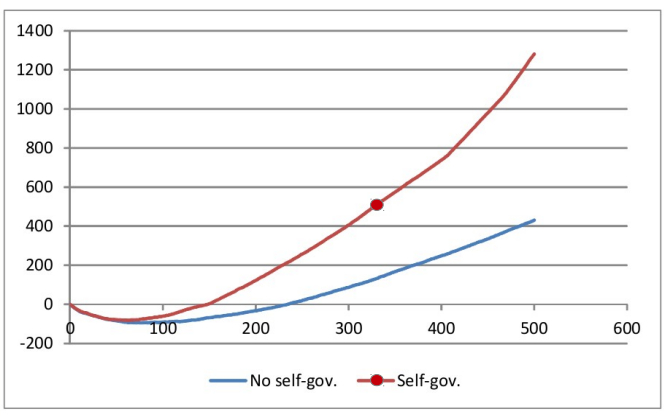

Fig. 3. Average of the accumulated global wealth per agent with self-governance and without self-governance vs. time.

The different parameters of the system (ranging from those of reinforcement learning to those of transfiguration and voting) are fixed in the current work, and some settings shall bring the agents into undesirable dynamics. For example, if the 'temperature' parameter τ_i^{δ} is set very high, then agents shall submit changes in a random manner (see Eq. 17). Questions regarding optimal or self-adaptive settings are left for future investigations.

V. 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. A particular case regards conventions emerging from repeated interactions amongst agents. In these settings a convention is often understood as a restriction on agents behaviour to one particular action [11]. Another field for investigating social norms are multi-agent learning and voting settings where agents are meant to coordinate by learning joint actions, typically using individual reinforcement learning or its extensions to collective tasks (see e.g. [12], [13]). Partalas et al. proposed in [14] 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. In all these systems, norms remain implicit whereas we are interested by constructing norms with explicit representations akin to logic formalisms.

Formal logics (typically deontic logics and argumentation, see e.g. [2]) are commonly investigated to represent and reason upon explicit norms, leading eventually to architecture for cognitive agents (see e.g. [4], [15]). 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 rewards). 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 Campenni at al. in [16] where BDI agents recognise norms by observing other agents. 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.

With regard to norm-synthesis, the problem was pioneered by the work of Shoham and Tennenholtz [17]. Fitoussi and Tennenholtz [18], 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 [19] included the implementation costs of norms and multiple design goals with different priorities. Christelis and Rovatsos [20] 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 [7] proposed a mechanism called IRON for the on-line synthesis of norms. IRON employs designated agents, often called 'institutional agents' [21], representing a norm-governed system/institution, and observing the interactions of the members of the system in order to synthesise conflict-free prescriptions without lapsing into over-regulation. The synthesised prescriptions are publicised to members of the system, but no sanction is associated to these prescriptions. Our work is fundamentally different: we target multi-agent systems and prescriptions are created for and by learning agents without designated agents receiving updates about the system interactions.

Exogenous or endogenous synthesised norm may involve their revision. Norm revision approaches are typically motivated by the following issues [22]: (a) Norms are conflicting and thus agents are unable to comply with them. (b) Some desirable state is rarely achieved, which indicates that there is insufficient guidance for agents. (c) Some undesirable state is frequently visited, which indicates that agents prefer to take the course of action leading to that state, the penalty notwithstanding. To address these issues, Corapi et al [22]

presented a framework for norm refinement using inductive learning. The learning mechanism is guided by the system designer who describes the desired properties of the framework through use cases, comprising (i) event traces that capture possible scenarios, and (ii) a state that describes the desired outcome. The main difference of our work and the approach of Corapi et al, therefore, concerns the fact that we do not rely on supervision, which is rarely available and requires substantial effort from the system designer.

Concerning norm change, studies in logic frameworks has attracted considerable attention since the Alchourrón, Gärdenfors, Makinson (AGM) framework of theory change [23]. For example, Boella and colleagues [24] have developed a formal framework for representing norm change. Their framework is produced by replacing the propositional formulas of AGM framework with pairs of propositional formulas—the latter representing norms—and adopting several principles from input/output logic. Governatori and colleagues [25], [26] have presented variants of a Temporal Defeasible Logic to reason about different aspects of norm change. Formal models of norm change, such as those mentioned above, are complementary to our work. We focus on a learning technique for self-governance that may be used along various norm change operations such as norm expansion, revision, contraction, annulment, etc, that satisfy properties including vacuity, recovery, and so on.

VI. Conclusion and future directions

We investigated the self-governance of learning agents, and more specifically the domain-independent (de)construction at run-time of prescriptive systems from scratch, for and by learning agents, without any agent having a complete information on the system.

The proposed solution is a development of the transfigurative mechanism coupled with a consensus system proposed in [8]. We allowed agents to repeal the prescriptions in force. So, the transfigurative mechanism enables learning agents to express learning policies (and thus behavioural patterns) into explicit prescription changes, and the consensus system allows agents to submit prescription changes for a vote and to vote eventually in favour or against a motion.

The simulations of a self-governed population of learning agents illustrated the benefits of our approach with regard to the convergence to desirable behaviours, they also indicated its scalability. Timeliness in run-time construction with learning agents appeared of the most importance. The results showed stability of the system since there was no back and forth enactments and repeals of prescriptions.

Future directions can be multiple. They include learning of joint actions and the transfiguration of these joint actions into complex prescriptive systems, distributed consensus systems (possibly in network) to avoid a central body collecting the votes. An interesting point regards a finer account of sanctions so that agents can self-govern with possible benefits of the principles of retributive justice.

Acknowledgment

Part of this work is supported by the Marie Curie Intra-European Fellowships PIEF-GA-2012-331472.

References

[1] R. Sutton and A. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998, vol. 116.

[2] G. Sartor, *Legal Reasoning: A Cognitive Approach to Law*. Springer, 2005.

[3] D. Gabbay, J. Horty, and X. Parent, Eds., *Handbook of Deontic Logic and Normative Systems*. College Publications, 2013.

[4] G. Governatori and A. Rotolo, "BIO logical agents: Norms, beliefs, intentions in defeasible logic," *Autonomous Agents and Multi-Agent Systems*, vol. 17, no. 1, pp. 36–69, 2008.

[5] S. Sen and S. Airiau, "Emergence of norms through social learning," in *IJCAI*, Morgan Kaufmann Publishers Inc., 2007, pp. 1507–1512.

[6] R. Conte, G. Andrighetto, and M. Campenni, Eds., *Minding Norms: Mechanisms and dynamics of social order in agent societies*. Oxford Scholarship, 2013.

[7] J. Morales, M. López-Sánchez, J. A. Rodríguez-Aguilar, M. Wooldridge, and W. Vasconcelos, "Automated synthesis of normative systems," in *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 483–490.

[8] R. Riveret, A. Artikis, D. Busquets, and J. Pitt, "Self-governance by transfiguration: From learning to prescriptions," in *DEON*. Springer, 2014, pp. 177–191.

[9] A. M. Polinsky and S. Shavell, Eds., *Handbook of Law and Economics*, 1st ed. Elsevier, 2007, vol. 1.

[10] R. Riveret, A. Rotolo, and G. Sartor, "Probabilistic rule-based argumentation for norm-governed learning agents," *Artificial Intelligence and Law*, vol. 20, no. 4, pp. 383–420, 2012.

[11] Y. Shoham and M. Tennenholtz, "On the emergence of social conventions: modeling, analysis, and simulations," *Artificial Intelligence*, vol. 94, pp. 139–166, 1997.

[12] D. Villatoro, J. Sabater-Mir, and S. Sen, "Social instruments for robust convention emergence," in *IJCAI*, AAAI Press, 2011, pp. 420–425.

[13] C. Yu, M. Zhang, F. Ren, and X. Luo, "Emergence of social norms through collective learning in networked agent societies," in *AAMAS*, International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 475–482.

[14] I. Partalas, I. Feneris, and I. P. Vlahavas, "Multi-agent reinforcement learning using strategies and voting," in *ICTAI*. IEEE Computer Society, pp. 318–324.

[15] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre, "The BOID architecture: Conflicts between beliefs, obligations, intentions and desires," in *AGENTS*. ACM, 2001, pp. 9–16.

[16] M. Campenni, G. Andrighetto, F. Cecconi, and R. Conte, "Normal = normative? the role of intelligent agents in norm innovation," *Mind and Society: Cognitive Studies in Economics and Social Sciences*, vol. 8, no. 2, pp. 153–172, 2009.

[17] Y. Shoham and M. Tennenholtz, "On social laws for artificial agent societies: off-line design," *Artificial Intelligence*, vol. 73, no. 1-2, pp. 231–252, 1995.

[18] D. Fitoussi and M. Tennenholtz, "Choosing social laws for multi-agent systems: minimality and simplicity," *Artificial Intelligence*, vol. 119, no. 1-2, pp. 61–101, 2000.

[19] T. Ågotnes and M. Wooldridge, "Optimal social laws," in *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 667–674.

[20] G. Christelis, M. Rovatsos, and R. P. A. Petrick, "Exploiting domain knowledge to improve norm synthesis," in *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 831–838.

[21] M. Esteva, J. Rodriguez-Aguilar, J. Arcos, C. Sierra, and P. Garcia, "Institutionalising open multi-agent systems," in *ICMAS*. IEEE Press, 2000, pp. 381–382.

[22] D. Corapi, A. Russo, M. De Vos, J. A. Padget, and K. Satoh, "Normative design using inductive learning," *Theory and Practice of Logic Programming*, vol. 11, no. 4-5, pp. 783–799, 2011.

[23] C. Alchourrón, P. Gärdenfors, and D. Makinson, "On the logic of theory change: Partial meet contraction and revision functions," *Journal of Symbolic Logic*, vol. 50, no. 2, pp. 510–530, 1985.

[24] G. Boella, G. Pigozzi, and L. van der Torre, "Normative framework for normative system change," in *AAMAS*. ACM Press, 2009, pp. 169–176.
[25] G. Governatori and A. Rotolo, "Changing legal systems: Abrogation and annulment. Part I: Revision of defeasible theories," in *DEON*. Springer, 2008, pp. 3–18.
[26] ——, "Changing legal systems: Abrogation and annulment. Part II: Temporalised defeasible logic," in *NORMAS*, G. Boella, G. Pigozzi, M. Singh, and H. Verhagen, Eds., 2008, pp. 112–127.