On Ex-Ante Law Enforcement in Norm-Governed Learning Agents*

Régis Riveret[1], Giuseppe Contissa[2,3], Antonino Rotolo[2], and Jeremy Pitt[1]

[1] Department of Electrical and Electronic Eng., Imperial College London, UK
[2] CIRSFID and Department of Legal Studies, University of Bologna, Bologna, Italy
[3] European University Institute, Florence, Italy

Abstract. We investigate law enforcement within a population of norm-governed learning agents using a probabilistic rule-based argumentation framework. We show that this formal framework can advantageously complete a traditional analysis based on expected utilities for the study of law enforcement systems when more realistic assumptions than hyper-rational agents or some behavourial phenomenon such as inertia are desired. This has significant implications for the design of systems of retributive justice for self-organising electronic institutions with endogenous resources, where the cost of monitoring and enforcement of laws and norms has to be taken into consideration.

1 Introduction

When norms are meant to control or guide autonomous agents, enforcement mechanisms are essential to back the compliance with normative systems. Enforcement refers to the promotion of compliance with the norms of the system by sanctioning agents: the sanctions are usually negative (punishments) while positive sanctions such as rewards can also be considered. Mechanisms of norm enforcement are fundamental to any normative system, and for this purpose many principles or techniques have been designed over time. Some enforcement systems are more oriented to the promotion of justice (in particular retributive justice) while others are biased towards some utilitarian measure such as efficiency.

In this paper, we make a preliminary investigation of law enforcement within a population of norm-governed learning agents using a probabilistic rule-based argumentation framework. In particular, the law enforcement agency is also represented by a learning agent, which can adapt the amount of surveillance according to the population profile of learning agents, who in turn can adapt their behaviour to comply, or not, with the norms. As a preliminary investigation, we focus on a simple scenario which presupposes that the violation of a rule or regulation is enforced by such a law enforcement agency. Crucially, we also take into consideration that any enforcement system has costs too.

* Part of this work has been carried out in the scope of the EC co-funded project SMART (FP7-287583).

Accordingly, this paper is organised as follows. The probabilistic rule-based argumentation framework is specified in Section 2, and its use by/within a population of norm-governed learning agents is presented in Section 3. The scenario and experimental results using such a system of law enforcement are described in Section 4, demonstrating that the internalisation of norms and the self-organisation of a system of law enforcement can yield the required compliant behaviour at 'acceptable' cost of enforcement. As a result, we show that using a traditional analysis in terms of expected utility can be misleading in the study of law enforcement systems, because learning agents tend to comply even though their surveillance is stopped. This has significant implications for the design of systems of retributive justice for self-organising electronic institutions with endogenous resources, where the cost of law enforcement has to be taken into consideration.

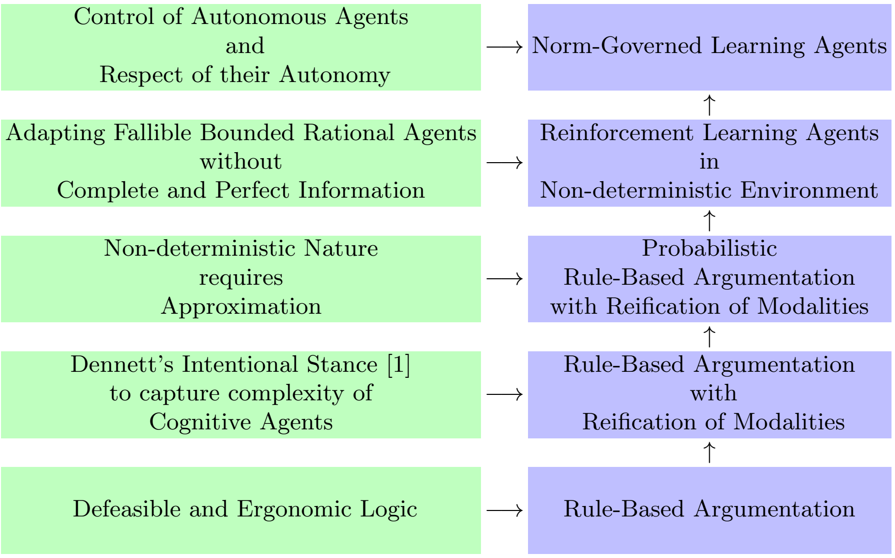
Control of Autonomous Agents
and
Respect of their Autonomy
Norm-Governed Learning Agents
Adapting Fallible Bounded Rational Agents
without
Complete and Perfect Information
Reinforcement Learning Agents
in
Non-deterministic Environment
Non-deterministic Nature
requires
Approximation
Probabilistic
Rule-Based Argumentation
with Reification of Modalities
Dennett's Intentional Stance [1]
to capture complexity of
Cognitive Agents
Rule-Based Argumentation
with
Reification of Modalities
Defeasible and Ergonomic Logic
Rule-Based Argumentation

Fig. 1. This diagram shows the layered architecture of our approach where each layer addresses some requirements by integrating techniques layer-by-layer.

2 Probabilistic Rule-based Argumentation

In this section, arguments and their conflict relationships are built from defeasible theories before a fixed-point semantics defines justified arguments. Finally, this argumentation framework is given a probabilistic interpretation.

Definition 1 (Language). *Let* Atoms *be a set of atomic formulas and* Lbl *a set of labels.*

Literals *The set of literals* $\mathrm{Lit} = \{\pm\psi | \psi \in \mathsf{Atoms}\}$ *consists of all atoms and their negations (we use* $\pm\phi$ *to cover the alternatives of affirmation and negation, i.e.* ϕ *and* $\neg\phi$*).*

Pure defeasible rules *have the form* $r : \phi_1, \ldots, \phi_n \Rightarrow \phi$ *where* $r \in \mathrm{Lbl}$, *and* $\phi_1, \ldots, \phi_n, \phi \in \mathrm{Lit}$. *Informally, this is a rule with identifier* r*, stating that if* $\phi_1, \ldots, \phi_n$ *hold then* ϕ *presumably holds. A rule with no antecedent, is written* $r : \quad \Rightarrow \phi$.

Preference ordering *Let* R *be a set of rules. Then* $\succ$ *is an antisymmetric partial order over* R*, i.e., if* $r \succ r'$ *then* $r' \not\succ r$*. Informally, a rule preference* $r_1 \succ r_2$ *states that rule* r_1 *prevails over rule* r_2.

Pure defeasible theories *A* pure defeasible theory *is a tuple* $\langle R, S\rangle$ *where* R *is a set of pure defeasible rules, and* S *is a set of preferences.*

Arguments are defined following [4], simplified to take into account that we just have one type of premises, namely, rules.

Definition 2 (Argument). *An argument* A *constructed from a pure theory* $\langle R, \succ\rangle$ *has the form* $A_1, \ldots, A_n \Rightarrow_r \phi$*, where* $A_1, \ldots, A_n$ *are arguments built from* $\langle R, \succ\rangle$ *and* $r : \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \phi$ *is a rule in* R *such that:*

- $\mathsf{Conc}(A) = \phi$ *(the top-conclusion of* A*),*
- $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \ldots \cup \mathsf{Sub}(A_n) \cup \{A\}$ *(the sub-arguments of* A*),*
- $\mathsf{TopRule}(A) = r : \mathsf{Conc}(A_1), \ldots, \mathsf{Conc}(A_n) \Rightarrow \phi$ *(the top-tule of* A*),*
- $\mathsf{Rules}(A) = \mathsf{Rules}(A_1) \cup \ldots \cup \mathsf{Rules}(A_n) \cup \{\mathsf{TopRule}(A)\}$ *(the rules of* A*).*

Two kinds of argument-conflict are usually considered: rebuttal (clash of incompatible conclusions) and undercutting (attacks on inferences). For our purposes, we deal with rebuttals only. So, we assume a function $-$ over the set of literals, such that $-\psi = \neg\psi$ and $-\neg\psi = \psi$ where ψ is an atom. The semantics is Dung's grounded semantics [2].

Definition 3 (Argumentation framework and semantics).

Preference *An argument* A *is preferred over another argument* B*, denoted as* $A \succ B$*, iff* $\mathsf{TopRule}(A)$ *is preferred to* $\mathsf{TopRule}(B)$ $(\mathsf{TopRule}(A) \succ \mathsf{TopRule}(B))$.

Defeats *An argument* B *defeats an argument* A *iff,* $\exists A' \in \mathsf{Sub}(A)$*, such that* $\mathsf{Conc}(B) = -\mathsf{Conc}(A')$*, and* $A' \not\succ B$.

Argumentation framework *An argumentation framework is a pair* $\langle \mathcal{A}, \gg\rangle$ *where* $\mathcal{A}$ *is a set of arguments, and* $\gg \subseteq \mathcal{A} \times \mathcal{A}$ *is a binary relation of defeat. For any arguments* A *and* B*,* $B \gg A$ *iff* B *defeats* A.

Conflict-free set *A set* $\mathcal{S}$ *of arguments is said to be* conflict-free *iff there is no argument* A *and* B *in* $\mathcal{S}$ *such that* B *defeats* A.

Acceptable argument *An argument* A *is acceptable w.r.t. a set of arguments* $\mathcal{S}$ *iff any argument defeating* A *is defeated by an argument in* $\mathcal{S}$.

Characteristic function *The characteristic function, denoted* F_{AF}*, of an argumentation framework* $AF = \langle \mathcal{A}, \gg\rangle$*, is defined as* $F_{AF} : 2^{\mathcal{A}} \Rightarrow 2^{\mathcal{A}}$ *and* $F_{AF}(\mathcal{S}) = \{A|\ A \textit{ is acceptable w.r.t. } \mathcal{S} \subseteq \mathcal{A}\}$.

Admissible set *A conflict-free set* $\mathcal{S}$ *of arguments is admissible iff* $\mathcal{S} \subseteq F_{AF}(\mathcal{S})$*. If a set* $\mathcal{S}$ *is admissible then we write it* $\mathsf{adm}(\mathcal{S})$*. We denote* $\mathsf{Adms}(T)$ *the admissible sets of a framework* AF_T *built from a pure theory* T.

Grounded extension *A grounded extension* $GE(AF)$ *of a framework* AF *is the least fixed-point of* F_{AF}*. The grounded extension of a framework* AF_T *built from a pure theory* T *is also denoted as* $GE(T)$*. If an argument* A *in* $GE(T)$ *is such that* $\mathsf{Conc}(A) = \phi$*, then* T *entails* ϕ *and we write* $T \vdash_{GE} \phi$.

Justified argument and conclusion *An argument A and its conclusion is justified, $Just(A)$, with regard to a framework AF iff $A \in GE(AF)$.*

Example 1. Given a theory $T = (R, \succ)$ where $R = \{r_1 : \ \Rightarrow a; r_2 : \ \Rightarrow b; r_3 : a, b \Rightarrow c;\ r_4 : \ \Rightarrow d;\ r_5 : d \Rightarrow \neg c\}$ and $\succ = \{r_5 \succ_3\}$, we have the arguments:

$$\begin{array}{l|l|l} A_1 : \ \Rightarrow_{r_1} a & A_2 : \ \Rightarrow_{r_2} b & A_3 : A_1, A2 \Rightarrow_{r_3} c \\ A_4 : \ \Rightarrow_{r_4} d & A_5 : A_4 \Rightarrow_{r_5} \neg c & \end{array}$$

The grounded extension $GE(T)$ is $\{A_1, A_2, A_4, A_5\}$.

We present now the probabilistic argumentation framework given in [5], which will be used to describe agent's interaction. We first consider empirical probabilities to set an intuitive interpretation before moving to theoretical probabilities.

Empirical probabilities. Given a multiset $\Gamma = \{\langle R_1, S_1 \rangle, \ldots, \langle R_n, S_n \rangle\}$ of pure defeasible theories, we collect all rules and preferences in such theories into two sets $\mathsf{rul}(\Gamma) = \bigcup_{i=1}^{n} R_i$ and $\mathsf{sup}(\Gamma) = \bigcup_{i=1}^{n} S_i$. For simplicity, we fix the preference set: the preference set of each sample theory coincides with $\mathsf{sup}(\Gamma)$.

We can now proceed to assign probabilities to every rule in $\mathsf{rul}(\Gamma)$. The empirical marginal probability $\pi(r)$ that a rule r appears in a multiset Γ is: $\pi(r) = |\Gamma_r|/|\Gamma|$ where $\Gamma_r = \{T | T \in \Gamma, r \in \mathsf{rul}(T)\}$. Rules with probability 1 would appear in any theory whereas rules with probability 0 would appear in no theory.

A probabilistic defeasible rule has the form $\pi, r : \quad \phi_1, \ldots, \phi_n \Rightarrow \phi$ where π is a probability assignment, $r \in \mathrm{Lbl}$, and $\phi_1, \ldots, \phi_n, \phi \in \mathrm{Lit}$. A probabilistic defeasible theory is a tuple $\langle R, \succ \rangle$ of a set of probabilistic defeasible rules and a set of preferences over them.

An empirical probabilistic defeasible theory is built from Γ: the set $\mathsf{probrul}(\Gamma)$ of the probabilistic rules from Γ, contains any rule in $\mathsf{rul}(\Gamma)$ expanded with the appropriate probability: $\mathsf{probrul}(\Gamma) = \{(\pi, r) | r \in \mathsf{rul}(\Gamma) \wedge \pi = \pi(r)\}$. The empirical probabilistic defeasible theory of a sample multiset Γ is the probabilistic defeasible theory $\langle R, S \rangle$ such that $R = \mathsf{probrul}(\Gamma)$ and $S = \mathsf{sup}(\Gamma)$.

Example 2. Let us have a sample multiset (with no preferences): $\Gamma = \{\langle \{r_1, r_2, r_4\}, \emptyset \rangle, \langle \{r_1, r_2, r_4\}, \emptyset \rangle, \langle \{r_2, r_3, r_4\}, \emptyset \rangle, \langle \{r_2, r_3, r_4\}, \emptyset \rangle\}$. Though this multiset is not statistically relevant, we use it to illustrate our concepts: $\mathsf{rul}(\Gamma) = \{r_1, r_2, r_3, r_4\}$, $\mathsf{sup}(\Gamma) = \emptyset$ and $\mathsf{probrul}(\Gamma) = \{(0.5, r_1), (1, r_2), (0.5, r_3), (1, r_4)\}$. The probabilistic theory of Γ is thus $\langle \mathsf{probrul}(\Gamma), \emptyset) \rangle$.

Finally, the empirical probability of the justification of an argument is: $P(Just(A)) = |\Gamma_A|/|\Gamma|$ with $\Gamma_A = \{T \,|\, T \in \Gamma_A, A \in GE(T)\}$.

Theoretical approach. We base it on Kolmorogov's framework. We assume that the sample space is Ω, an algebra on Ω is a set $F(\Omega)$ of all subsets of Ω (Ω belongs to $F(\Omega)$ and F is closed under union and complementation w.r.t. Ω). and, the following probability function P from $F(\Omega)$ to $[0, 1]$:

$$P(A) = \sum_{T \in A} P(T) \tag{1}$$

A sample space can be build from a multiset $\mathit{\Gamma}$ (but not necessarily) by gathering all rules and preferences in this multiset $\mathit{\Gamma}$. Let $\mathit{\Omega}_{\mathit{\Gamma}}$ denote all pure theories possibly constructed from $\mathit{\Gamma}$: $\mathit{\Omega}_{\mathit{\Gamma}} = \{\langle R, S\rangle \,| R \subseteq \mathsf{rul}(\mathit{\Gamma}) \wedge S = \mathsf{sup}(\mathit{\Gamma})\}$ In the remainder, we shall call these theories (possible) worlds or world theories.

In case of independent rules, the probability $P(T)$ of a pure theory T is:

$$P(T) = \prod_{r \in \mathsf{Rul}(T)} \pi(r). \prod_{r \in \mathsf{Rul}(\mathit{\Omega}) \setminus \mathsf{Rul}(T)} [1 - \pi(r)] \tag{2}$$

Unless otherwise specified, we do not assume that rules are stochastically independent: any set $\mathcal{T}$ of pure theories is attached with a potential $Q(\mathcal{T})$, and its probability is defined using an exponential model.

$$P(\mathcal{T}) = e^{Q(\mathcal{T})} / \sum_{\mathcal{T}} e^{Q(\mathcal{T})} \tag{3}$$

The *probability of justification of an argument* A is the sum of theories' probabilities in which A is justified: $P(\mathit{Just}(A)) = \sum_{T \in \mathit{\Omega}: A \in GE(T)} P(T)$.Similarly, the probability of a literal ϕ being justified is the probability of the set of worlds in which ϕ is justified: $P(\mathit{Just}(\phi)) = \sum_{T \in \mathit{\Omega}: T \vdash_{GE} \phi} P(T)$. So, the larger the proportion of world theories in $\mathit{\Omega}$ where ϕ is justified, the higher the probability that ϕ is justified.

Example 3. In this example we use atoms having the form $\mathsf{E}^t_i b$, indicating that agent i performs action b at time t. For instance, $\mathsf{E}^t_i left$ states that i drives on the left side of a road at time t. Let $\mathit{\Gamma}$ be a multiset of pure defeasible theories with the sample space $\mathit{\Omega}_{\mathit{\Gamma}}$: $\mathsf{sup}(\mathit{\Omega}_{\mathit{\Gamma}}) = \emptyset$ and $\mathsf{rul}(\mathit{\Omega}_{\mathit{\Gamma}}) = \{r^t_1, r^t_2, r^t_3, r^t_4\}$.

$$\begin{array}{l|l}
0.5, r^t_1 : \Rightarrow \mathsf{E}^t_{Tom} \mathit{left} & 1, r^t_3 : \quad \mathsf{E}^t_{Tom} \mathit{left} \quad \Rightarrow \neg \mathsf{E}^t_{Tom} \mathit{right} \\
0.5, r^t_2 : \Rightarrow \mathsf{E}^t_{Tom} \mathit{right} & 1, r^t_4 : \quad \mathsf{E}^t_{Tom} \mathit{right} \Rightarrow \neg \mathsf{E}^t_{Tom} \mathit{left}
\end{array}$$

Thus, rules r^t_1 and r^t_2 appear in half of the theories in this sample set. The sample space $\mathit{\Omega}_{\mathit{\Gamma}}$ can be represented by a table where each column is a world theory:

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}	T_{16}
r^t_1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
r^t_2	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
r^t_3	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
r^t_4	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
$\mathit{Just}(E^t_{Tom} \mathit{left})$	0	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0
$\mathit{Just}(E^t_{Tom} \mathit{right})$	0	1	0	0	1	1	0	0	0	1	0	0	1	1	0	0
$\emptyset$	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

We can compute, among others, the probability of the set $\mathcal{T}$ of world theories in which $E^t_{Tom} \mathit{left}$ is justified. Let's assume stochastic independent rules. $E_{Tom} \mathit{left}$ is justified in the set $\mathcal{T} = \{T_3, T_7, T_9, T_{11}, T_{13}, T_{15}\}$ of theories, thus $P(\mathit{Just}(E^t_{Tom} \mathit{left}) = P(\mathcal{T})$, and $P(\mathit{Just}(E^t_{Tom} \mathit{left}) = \pi(r_1)[1 - \pi(r_2)]$[4] or 1/4.

[4] $$\begin{aligned}
P(\mathcal{T}) &= P(T_3) + P(T_7) + P(T_9) + P(T_{11}) + P(T_{13}) + P(T_{15}) \\
&= \pi(r_1)\pi(r_3)\pi(r_4)[1 - \pi(r_2)] + \pi(r_1)\pi(r_4)[1 - \pi(r_2)][1 - \pi(r_3)] \\
&\quad + \pi(r_1)\pi(r_3)[1 - \pi(r_2)][1 - \pi(r_4)] + \pi(r_1)[1 - \pi(r_3)][1 - \pi(r_2)][1 - \pi(r_4)] \\
&\quad + \pi(r_1)\pi(r_2).\pi(r_3)[1 - \pi(r_4)] + \pi(r_1)\pi(r_2)[1 - \pi(r_3)][1 - \pi(r_4)] \\
&= \pi(r_1)[1 - \pi(r_2)]
\end{aligned}$$

Suppose now that some potentials are attached to sets of theories denoted by their justified actions: $Q(just(E_{Tom}left)) = Q(just(E^t_{Tom}right)) = 10$, $Q(\emptyset) = 0$. Using a Boltzmann distribution, we have: $P(just(E_{Tom}right)) = 6.e^{10}/(6.e^{10} + 6.e^{10}+5.e^0)$. So $P(just(E^t_{Tom}right)) = P(just(E^t_{Tom}left))$ (≈ 0.5) and $P(\emptyset) \approx 0$.

3 Learning agents

Let us informally introduce the language we use to define multi-agent systems. We assume that agents process external inputs from an objective environment into a subjective counterpart, and (successfully or not) act upon this objective environment. We formally cater for basic temporal epistemic features, agency and simple normative aspects by temporalised prefix operators for literals that we present with the following meaning:

$\mathsf{Hold}^t_i\phi$	It holds, from the viewpoint of an agent i at time t, that ϕ.
$\mathsf{Hold}^t_{obj}\phi$	It objectively holds at time t that ϕ.
$\mathsf{E}^t_i\phi$	The agent i is attempting at time t to bring it about ϕ.
$\mathsf{Obl}^t_{obj}\phi$	From an objective point of view, ϕ is obligatory at time t.
$\mathsf{Obl}^t_i\phi$	From the point of view of agent a, a ought to bring about ϕ at time t.

A formula indexed by *obj* indicates that this formula holds objectively, rather than being merely believed by an agent. We may say that *obj* embodies the objective point of view, that it only accepts what is true. Thus, any prefixed literal subscripted by *obj* is called an environmental literal, while those subscripted with i are called agents literals.

Definition 4 (Language for MAS). *Let* Atoms *be a set of atomic formulas,* Lbl *a set of labels, obj a parameter for literals,* $\mathrm{Ag} = \{i_1, i_2, \ldots\}$ *a finite set of agents,* $\mathsf{Times} = \{t_1, t_2, \ldots\}$ *a discrete totally ordered set of instants.*

Results *Let the set* $\mathsf{Results} = \{out_i(X) | i \in \mathrm{Ag} \cup \{obj\} \wedge X \in \mathbb{R}\}$ *where a proposition* $out_i(X)$ *(*$out_{obj}(X)$*) affirms that i (Nature) obtains utility* X*.*

Basic literals *Let* Lit *denote the set of basic literals consists of all atoms in* Atoms *and in* Results *and their negations.* $\mathrm{Lit} = \{\pm\psi | \psi \in \mathsf{Atoms} \ \vee \psi \in \mathsf{Results}\}$*.*

Action literals *have the form* $\pm\mathsf{E}^t_i\phi$ *where* $i \in \mathrm{Ag}$*,* $t \in \mathsf{Times}$*, and* $\phi \in \mathrm{Lit}$*. It affirms an agent attempts or not to realise the state of affairs* ϕ *at* t*.*

State literals *have the form* $\pm\mathsf{Hold}^t_i\phi$ *where* $i \in \mathrm{Ag} \cup \{obj\}$*,* $t \in \mathsf{Times}$*, and* $\phi \in \mathrm{Lit}$*. It affirms that according to an agent i or from an objective point of view, the state of affairs* ϕ *holds (does not hold) at time* t*.*

Obligation literals *have the form* $\pm\mathsf{Obl}^t_i\phi$ *where* $i \in \mathrm{Ag} \cup \{obj\}$*,* $t \in \mathsf{Times}$*, and* $\phi \in \mathrm{Lit}$*. It affirms an obligation to bring about* ϕ *at* t*.*

Rules *are probabilistic defeasible rules of the form* $\pi, r^t : \phi_1, \ldots, \phi_n \Rightarrow \phi$ *where* r *is a label (*$r \in \mathrm{Lbl}$*),* t *is a time (*$t \in \mathsf{Times}$*), and each* $\phi_1, \ldots, \phi_n, \phi$ *is a state literal, an action literal, or an obligation literal.*

We assume an incompatibility function $-$, which returns the set of modal literals which are incompatible for a given modal literal. Let t an instant of time ($t \in$

Times), let i denote an agent ($i \in \mathrm{Ag}$), let M denote a modality Hold^t_i, Hold^t_{obj}, E^t_i, Obl^t_i or Obl^t_{obj} and let ϕ be a literal. Each modal literal $\pm M\phi$ is incompatible both with its complement ($M\phi$ is incompatible with $\neg M\phi$ and vice versa) and with the formula obtained by substituting the embedded literal ϕ with its complement ($\pm M\phi$ is incompatible with $\pm M - \phi$ and vice versa). Let ψ denote an atom, we have:

$$\begin{array}{r|l} -M\psi = \{\neg M\psi, M\neg\psi\} & \quad -M\neg\psi = \{\neg M\neg\psi, M\psi\} \\ -\neg M\psi = \{M\psi, \neg M\neg\psi\} & \quad -\neg M\neg\psi = \{\neg M\psi, M\neg\psi\} \end{array}$$

A multi-agent system is encoded into a probabilistic theory whose rules are environmental rules or agent rules. The first holds in the environment, regardless of the beliefs of an individual agent, the second hold in the agent's mind. The fundamental difference between those rules is that the probabilities of environmental rules are assumed to be fixed, whereas probabilities of the agent rules may be changed by the agent itself in order to adapt to the environment. Thus, we partition the probabilistic theory encoding a multi-agent system into an environmental probabilistic theory (whose rules' probabilities are not meant to change), and probabilistic theories representing agents (whose rules' probabilities may change). So, each agent i is represented by a probabilistic theory $\mathcal{T}_i$ and the environment is represented by a probabilistic theory $\mathcal{T}_e$. A multi-agent probabilistic theory $\mathcal{T}$ is the union of agents' probabilistic theories and the probabilistic environment theory: $\mathcal{T} = \mathcal{T}_e \bigcup_i \mathcal{T}_i$.

Since sanctions are important in normative systems, we use the paradigm of reinforcement learning. At each time step t, any agent concurrently behave after sensing a grounded environmental randomly selected by Nature. For every agent, the selection of a behaviour, is simulated by a probability distribution over all possible agent's pure theories. The outcomes of agents' behaviours are then entailed via the grounded extension of agents' pure theories and the pure environmental theory. The outcomes obtained at t are then used to update a time $t+1$ every agent's probability distribution over behaviours.

Definition 5 (Grounded theory). *The grounded theory $T(GE(T'))$ built from a theory T' is the theory $\langle\{\quad \Rightarrow \phi \mid T' \vdash_{GE} \phi\}, \emptyset\rangle$.*

We partition an environment $\mathcal{T}_e$ into two sub-environments: a first sub-environmental theory $\mathcal{T}_{e1}$ is used to compute the environmental facts that agents can sense before behaving, and a second sub-environment theory $\mathcal{T}_{e2}$, called an outcome theory, is used to compute the outcomes of agents' acts. Accordingly, at each time t, we have $\mathcal{T}^t_e = \mathcal{T}^t_{e1} \cup \mathcal{T}^t_{e2}$, and agents behaves after sensing the grounded environment $T(GE(T^t_{e1}))$, denoted E^t_{e1} in the remainder.

Pure theory utilities and qualities. Once all the agents have behaved in a grounded environment E^t_{e1}, some outcomes are entailed by a pure outcome environment T^t_{e2}. Thus agents can evaluate the utility and quality of their behaviours, that is, the utility and quality of any of their pure theories T^t_j. Formally, the utility for the agent i of a pure theory T^t_j associated to a grounded environment

E^t_{e1}, a pure environmental outcome theory T^t_{e2} and other agents' theories T^t_l is:

$$u_i(T^t_j) = \sum_{\forall A_k \in GE(E^t_{e1} \cup T^t_{e2} \cup T^t_j \bigcup_l T^t_l) : \mathsf{conc}(A_k) = \mathsf{Hold}^t_i out_i(X_k)} X_k \tag{4}$$

Given the utility of a theory T^t_j selected at time t, its quality at time $t+1$ is computed using a weighted average over time:

$$Q(T^{t+1}_j) = Q(T^t_j) + \alpha.[u_i(T^t_j) - Q(T^t_j)] \tag{5}$$

where $\alpha \in [0,1]$. If T^t_j is not selected, then:

$$Q(T^{t+1}_j) = \beta.Q(T^t_j) \tag{6}$$

where $\beta \in [0,1]$. The parameter α is the weight of the latest utility in order to keep track of the non-stationary environment, the parameter β allows agents to forget unselected theories. At the initialisation of an agent i, say at time t_{init}, we assume an arbitrary value for any $Q(T^{t_{init}}_j)$.

Behaviours. If an agent probabilistic theory contains n rules, then the agent has to search what is the best theory amongst 2^n pure theories. To reduce the search space, we first assume that every agents' rules with a constant probability are independent. Thus, at each time t, a lottery is performed on the k fixed independent rules of agents' theory $\mathcal{T}^t_i$. Then, we make a partition of the set of the 2^{n-k} remaining theories using the notion of *behaviours* that regroups the set of pure theories which entail, on the basis of a grounded environment, the same (possibly empty) set of actions.

Definition 6 (Behaviour). *Let E^t_{e1} be a grounded environment at time t, let i denote an agent and let $\mathcal{T}_i$ be its probabilistic theory, let $\mathcal{E}^t_i$ be a (possibly empty) set of actions $\mathsf{E}^t_i\phi$ and let $\mathcal{T}'_i$ be the set of pure theories resulting from a lottery on $\mathcal{T}_i$ over every rule with a constant probability. We define the set of possible behaviours $\mathcal{B}_i(E^t_{e1})$ as a partition of $\mathcal{T}'_i$ such that a behaviour $B(\mathcal{E}^t_i, E^t_{e1}) \in \mathcal{B}_i(E^t_{e1})$ is the set of pure theories exactly entailing the set of actions $\mathcal{E}^t_i$:*

$$B(\mathcal{E}^t_i, E^t_{e1}) = \{T^t_i \mid T^t_i \in \mathcal{T}'_i \wedge \{\mathsf{E}^t_i\phi | E^t_{e1} \cup T^t_i \vdash_{GE} \mathsf{E}^t_i\phi\} = \mathcal{E}^t_i\}$$

As for notation, we shall write B^t_i or $B_i(E^t_{e1})$ as a shortcut for $B(\mathcal{E}^t_i, E^t_{e1})$.

Example 4. Running example. Tom is defined by a theory $\langle\{r^t_1, r^t_2, r^t_3, r^t_4\}, \emptyset\rangle$:

$$\begin{array}{lll|lll}
\_, r^t_1 : & & \Rightarrow \mathsf{E}^t_{Tom} \mathit{left} & 1, r^t_3 : & \mathsf{E}^t_{Tom} \mathit{left} & \Rightarrow \neg\mathsf{E}^t_{Tom} \mathit{right} \\
\_, r^t_2 : & & \Rightarrow \mathsf{E}^t_{Tom} \mathit{right} & 1, r^t_4 : & \mathsf{E}^t_{Tom} \mathit{right} & \Rightarrow \neg\mathsf{E}^t_{Tom} \mathit{left}
\end{array}$$

The rules prefixed with an underscore indicate that their probability can be changed. The rules r^t_3 and r^t_4 have a fixed probability, and thus, they are considered independent. A lottery is performed on these rules, and since their probability equals 1, the search space is reduced to $2^{4-2} = 4$ worlds (T_1, T_2, T_3, and T_4). So, at each instant t, Tom has the choice amongst the behaviours $B(\{\mathsf{E}^t_{Tom,\_}\mathit{left}\}, \emptyset) = \{T^t_3\}$, $B(\{\mathsf{E}^t_{Tom,\_}\mathit{right}\}, \emptyset) = \{T^t_2\}$, $B(\{\mathsf{E}^t_{Tom,\_}\mathit{right}, \mathsf{E}^t_{Tom,\_}\mathit{left}\}, \emptyset) = \emptyset$, and $B(\emptyset, \emptyset) = \{T^t_1, T^t_4\}$. The choice of Tom's behaviour will be based on a probability distribution over these behaviours.

Behaviours' probability. Using equation (1), the probability of a behaviour B_i^t is the sum of the probabilities of its pure theories: $P(B_i^t) = \sum_{T_i^t \in B_i^t} P(T_i^t)$. Since the selection of a pure theory implies that other pure theories are not selected, theories do not necessarily have the same qualities over time. To reduce the search space, behaviours' theories are assumed to have the the same utilities and thus the same qualities. So, we use equations (5) and (6) to compute the quality of behaviours, and we assume the following exponential distribution:

$$P(B_i^t) = e^{Q(B_i^t)/\tau_i} / \sum_{B_i^t} e^{Q(B_i^t)/\tau_i} \tag{7}$$

where $Q(B_i^t)/\tau_i = [Q(T_i^t) + ln(|B_i^t|)]//\tau_i$, such that $T_i^t \in B_i^t$. When $Q(T_i^t) \gg ln(|B_i^t|)$, then $ln(|B_i^t|$ can be omitted to approximate behaviours' distribution. The learning parameter $\tau_i > 0$ modulates the "exploitation" of the pure behaviours that entail the highest outcomes and "exploration" of other behaviours.

Algorithm 1 Animation of a multi-agent system

```
- Initialise the system with a probabilistic theory $\mathcal{T}_e \bigcup_i \mathcal{T}_i$ with $\mathcal{T}_e = \mathcal{T}_{e1} \cup \mathcal{T}_{e2}$ ;
for $t = 0$ to $t_{end}$ do
  - Do a lottery on independent fixed probabilistic rules (so $\mathcal{T}_e^t$ describing the environment results in one pure defeasible theory $T_e^t$);
  - Compute the grounded environment $E_{e1}^t = T(GE(T_{e1}^t))$;
  for each agent $i$ do
    - Compute the set of possible behaviours $\mathcal{B}_i(E_{e1}^t)$;
    - Compute the distribution over the behaviours $\mathcal{B}_i(E_{e1}^t)$ using equation (7).
    - Do a lottery over the distribution over $\mathcal{B}_i(E_{e1}^t)$ resulting in one behaviour $B_i^t$;
  end for
  - Compute the grounded extension $GE(E_{e1}^t \cup T_{e2}^t \bigcup_i T_i^t)$;
  for each agent $i$ do
    for each behaviour $B_i^t$ do
      if $B_i^t$ was previously selected then
            $Q(B_i^{t+1}) = Q(B_i^t) + \alpha.[u_i(B_i^t) - Q(B_i^t)]$ where $\alpha \in [0, 1]$
      else
            $Q(B_i^{t+1}) = \beta.Q(B_i^t)$ where $\beta \in [0, 1]$.
      end if
    end for
  end for
end for
```

4 Law Enforcement Analysis

Common instruments used in legal systems to reduce and control dangerous activities are civil liability and direct regulation.

Civil liability is an “after-the-fact” (*ex post*) instrument: it places an obligation for one agent to pay compensation for damages once they have occurred. Harm or injury is therefore always required. Usually, also negligence is required, that is an injurer must pay compensation only when he acted with a level of care that was less than a standard of care appropriate for the given activity (“due care”). Liability does not deal directly with risk control, but with the damage that occurs once (and only if) a risk has materialized.

Direct regulation instead is a “before-the-fact” (*ex ante*) instrument: regulatory rules are aimed at setting standards for activities to reduce risks arising from such activities, so that every agent that intend to engage in a regulated activity is required to comply with the applicable standard and incur the related compliance cost. Regulatory rules are typically enforced through administrative or criminal sanctions for violations (injunctions, monetary compensations, fines or imprisonment).

Differently from liability, which always requires a harm, direct regulation prohibits a certain non-compliant behaviour, irrespective of any actual harm having been caused by the non-compliance. Another interesting difference is that under direct regulation, the amount of sanctions can be freely set by the authority. Often they are set at a level reflecting the social loss that would result from possible harms multiplied by the reciprocal of the chance of the wrongdoer's being caught. Under civil liability instead, the sanction always corresponds to the actual damage, unless punitive damages may be imposed (however, punitive damages are not allowed in many legal systems).

Many activities (car driving, transportation, industrial production, etc.) are controlled through an interplay of regulatory and liability rules, but achieving the optimal combination is usually a hard task. A dual system of regulation and liability may create the danger that the incentives created by each instrument will not be coordinated, and that agents and society will pay the cost of both systems and obtain the advantage of neither, i.e. the risk that the agent pays twice for its behaviour, while the society adds the cost of monitoring without a corresponding additional gain for the global wealth.

For our purposes, we assume a population of N agents having the possibility to perform an action with three levels of care: $\mathsf{E}_i^t\phi_1$, $\mathsf{E}_i^t\phi_2$ and $\mathsf{E}_i^t\phi_3$:

$$\begin{array}{r|l|l}
\_, r1_i^t : \ \Rightarrow \mathsf{E}_i^t\phi_1 & 1, nr_{12i}^t : \mathsf{E}_i^t\phi_1 \Rightarrow \neg\mathsf{E}_i^t\phi_2 & 1, nr_{13i}^t : \mathsf{E}_i^t\phi_1 \Rightarrow \neg\mathsf{E}_i^t\phi_3 \\
\_, r2_i^t : \ \Rightarrow \mathsf{E}_i^t\phi_2 & 1, nr_{21i}^t : \mathsf{E}_i^t\phi_2 \Rightarrow \neg\mathsf{E}_i^t\phi_1 & 1, nr_{23i}^t : \mathsf{E}_i^t\phi_2 \Rightarrow \neg\mathsf{E}_i^t\phi_3 \\
\_, r3_i^t : \ \Rightarrow \mathsf{E}_i^t\phi_3 & 1, nr_{31i}^t : \mathsf{E}_i^t\phi_3 \Rightarrow \neg\mathsf{E}_i^t\phi_1 & 1, nr_{23i}^t : \mathsf{E}_i^t\phi_3 \Rightarrow \neg\mathsf{E}_i^t\phi_2
\end{array}$$

When there is an obligation, an agent may internalise it:

$$\_, int_i^t : \mathsf{Obl}_{obj}^t\phi_1 \Rightarrow \mathsf{Obl}_i^t\phi_1 \big| \_, c_i^t : \mathsf{Obl}_i^t\phi_1 \Rightarrow \mathsf{E}_i^t\phi_1$$

The performance of each action is associated with an outcome (the higher the level of care, the lesser the payoff):

$$\begin{array}{l}
1, out1_i^t : \mathsf{E}_i^t \phi_1 \Rightarrow \mathsf{Hold}_i^t out_i(5) \\
1, out2_i^t : \mathsf{E}_i^t \phi_2 \Rightarrow \mathsf{Hold}_i^t out_i(10) \\
1, out3_i^t : \mathsf{E}_i^t \phi_3 \Rightarrow \mathsf{Hold}_i^t out_i(16)
\end{array}$$

There are probabilistic rules defining the case where an accident may occur: the higher the level of care, the lesser the probability that an accident occurs. Notice that an accident does not affect agents' payoff.

$$\begin{array}{rll}
0.01, ko.1^t : & \mathsf{E}_i^t \phi_1 & \Rightarrow \mathsf{Hold}_{obj}^t accident(i) \\
0.05, ko.2^t : & \mathsf{E}_i^t \phi_2 & \Rightarrow \mathsf{Hold}_{obj}^t accident(i) \\
0.1, ko.3^t : & \mathsf{E}_i^t \phi_3 & \Rightarrow \mathsf{Hold}_{obj}^t accident(i) \\
1, out4_i^t : & \mathsf{Hold}_{obj}^t accident(i) & \Rightarrow \mathsf{Hold}_{obj}^t out_{obj}(-200)
\end{array}$$

We assume that at time 100, an obligation to act with care enters in force:

$$1, obl^t : t > 100 \Rightarrow \mathsf{Obl}_{obj}^t \phi_1$$

We settle an agent representing the enforcement agency (here a police) who has the possibility to sense and punish the violation of other agents:

$$\begin{array}{rll}
\_, mon^t : & t > 100 & \Rightarrow \mathsf{E}_{police}^t monitor, \\
1, viol1_i^t : & \mathsf{E}_{police}^t monitor, \mathsf{E}_i^t \phi_2, \mathsf{Obl}_{obj}^t \phi_1 & \Rightarrow \mathsf{Hold}_{police}^t viol(i), \\
1, viol2_i^t : & \mathsf{E}_{police}^t monitor, \mathsf{E}_i^t \phi_3, \mathsf{Obl}_{obj}^t \phi_1 & \Rightarrow \mathsf{Hold}_{police}^t viol(i),
\end{array}$$

with the following outcome rule:

$$\begin{array}{rll}
1, sanc_i^t : & \mathsf{Hold}_{police}^t viol(i) & \Rightarrow \mathsf{Hold}_i^t out_i(out^{fine}), \\
1, cost^t : & \mathsf{E}_{police}^t monitoring & \Rightarrow \mathsf{Hold}_{police}^t out_{police}(out^{mon}.N), \\
1, inp_i^t : & \mathsf{Hold}_{police}^t viol(i) & \Rightarrow \mathsf{Hold}_{police}^t out_{police}(-out^{fine}).
\end{array}$$

We run the simulations with $out^{fine} = -30$ and $out^{mon} = -4$. The rule *sanc* indicates that the amount of the fine of a careless agent. The rule *cost* expresses that the enforcing agency is monitoring all the N agents, and that the cost of monitoring each agent is -4.

Next, we will proceed as follows: firstly, we use traditional calculus to investigate the expected utilities of enforcement regimes. Then, we investigate enforcement within a population of learning agents.

4.1 Traditional calculus using expected utilities

Let's first make a traditional analysis using expected utilities. The overall expected utility of an agent i is $EU_i = EU_i(\emptyset) + EU_i(\phi_1) + EU_i(\phi_2) + EU_i(\phi_3)$:

$$\begin{array}{ll|l}
EU_i(\emptyset) & = 0 & EU_i(\phi_2) = 10.P(just(E_i\phi_2)) \\
EU_i(\phi_1) & = 5.P(just(E_i\phi_1)) & EU_i(\phi_3) = 16.P(just(E_i\phi_3))
\end{array}$$

The associated expected global wealth including the cost of potential accidents for a population of N agents are:

$$\begin{array}{lll|ll}
EW(\emptyset) & = 0 & & EW(\phi_2) = N.[10 - 0.05 \times 200] & (0) \\
EW(\phi_1) & = N.[5 - 0.01 \times 200] & (3.N) & EW(\phi_3) = N.[16 - 0.1 \times 200] & (-4.N)
\end{array}$$

When no obligation to act with care is enforced, then $EU_i(\emptyset) < EU_i(\phi_1) < EU_i(\phi_2) < EU_i(\phi_3)$, and a rational agent shall act with negligence. However, the expected global wealth EW would be negative ($EW(\phi_3) = -4.N$).

As a remedy of this tragic situation, an obligation can be introduced to maximise the global wealth, that is, when agents act with care. If $EU_i^{mon}(\phi_3) < EU_i^{mon}(\phi_1)$ so that agents act with care, then the system has to satisfy $16 + P(just(\mathsf{Hold}^t_{police}(\text{viol}(i)))).out^{fine} < 5$. We arbitrarily set $out^{fine} = -30$. Thus, a rational agent will act with care if the probability of being fined is superior to 11/30, formally $P(just(\mathsf{Hold}^t_{police}(\text{viol}(i)) > 11/30$. In this case, the expected global wealth will be positive: $EW(\phi_1) = 3.N$.

However, this calculation does not take into account the cost of enforcement. We assume that the cost of monitoring N agents is fixed: $N.out^{mon}_{ante}$ with $out^{mon}_{ante} < 0$. Thus, the expected global wealth is now:

$$EW_{ante}(\phi_1) = EW(\phi_1) + N.out^{mon}_{ante} \qquad (N.[3 + out^{mon}_{ante}])$$

Hence, if $3{+}out^{mon}_{ante} > 0$ then it is worth monitoring. In case the cost of monitoring is prohibitive ($out^{mon}_{ante} < -3$)then we return to a tragic scenario where the global wealth shall be negative.

Ex post surveillance offers an alternative. This alternative would meet up the doctrine of "cost internalisation": roughly speaking an agent has to pay for his damages but can continue to damage as long as he pays for it. However, the solution of cost internalisation, besides the moral counter-reasons, has a cost too. If $out^{mon}_{post} < -3$, then we return to a tragic situation.

Next, we show that, on the assumption that the law enforcement agency can adapt to agents' behaviour, the analysis of the global wealth in terms of expected utility is misleading: indeed a law enforcement can stop monitoring for some periods since learning agents internalise norms.

4.2 Simulation of learning agents

Let's now move to the study of the scenario with learning agents. Some simulation results of typical runs are presented in Figures 2 where, for any agent, the 'learning temperature' τ has been set to 1, the discount factor α to 0.1 and the forget factor β to 0.9. For any experiment, in a first phase, a majority of agents learn to act with negligence because this behaviour has the highest quality, but, in a second phase, the evolutions of behaviours and global wealth vary in function to the probability of surveillance:

- For a fixed probability of enforcement, $P(just(\mathsf{Hold}^t_{obj}(\text{viol}(i)) = 0.2$, agents behave with even more negligence since the level of care ϕ_2 is not worthy, and thus the global wealth continues to decrease.
- For $P(just(\mathsf{Hold}^t_{obj}(\text{viol}(i)) = 0.34 \quad (< 11/30)$, though the agents shall not behave with care according to the calculus of expected utility, the simulation shows that the value 0.34 seems a limit value where a majority of agents may get advantage of behaving with care on the long run. Notice that the sudden loss of the probability of negligent behaviour is caused by some temporal concentration of monitoring that may randomly occur.

– When $P(just(\mathsf{Hold}^t_{obj}(\mathrm{viol}(i))) = 0.4$, then agents clearly comply to avoid fines (as foreseen by the approach with expected utilities), and the decrease of global wealth is slowly stopped to finally increase at a steady step. Though not being drawn in the figures, the option of fixed monitoring with high frequencies results quite inefficient due to its prohibitive cost.

The previous simulation settled a surveillance with a fixed probability of monitoring. We move now to the case of a learning enforcement agency which can adapt the amount of surveillance by taking into consideration negligent agents and the occurrence of accidents. To do so we replace the previous rule $out4^t_i$ with this one: $1, out4^t_i : \mathsf{Hold}^t_{obj} accident(i) \Rightarrow \mathsf{Hold}^t_{police} out_{police}(-200)$. Once the law enforcement agency enters into action, then agents start behaving more carefully. When the number of negligent agents in combination with the occurrences of accidents is low enough to undermine the utility of surveillance, the enforcement is dramatically reduced. However, due to the inertia of learning, most of the agents continue to behave with care even though the surveillance has become infrequent. With learning agents, the experimental probability of surveillance

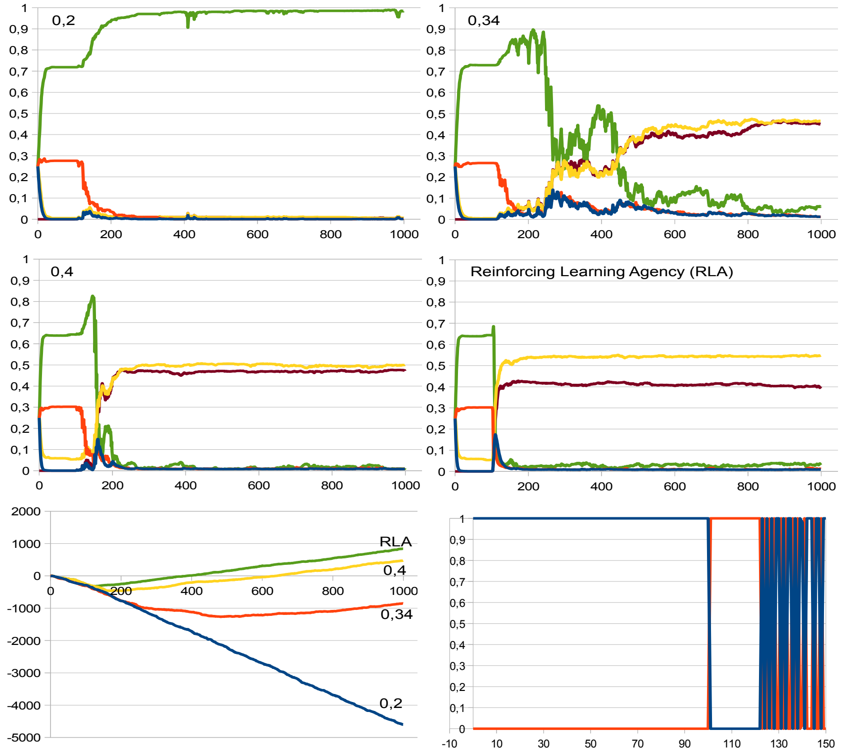

Fig. 2. Top Graphs: Probability of citizen behaviours vs. time. Green: ϕ_3 (negligence), Orange: ϕ_2, Yellow: ϕ_1 (careful act), Brown: Compliance via Internalisation, Blue: Inaction. **Bottom-Left**: Global Wealth (per agent) vs. time. **Bottom-Right**: Probability of the behaviours of a learning enforcement agency vs. time. Blue: Inaction, Orange: Monitoring.

is now about 0.33 over the last hundreds steps. This is relevant when compared to the simulation with the fixed probability of 0.34. Notice that when the monitoring is fixed at 0.4, the global wealth is increasing almost as good as in the scenario with a learning enforcement agency. Another advantage of a learning agency is at the introduction of the obligation, because agents start acting with care more quickly. An advantage of a monitoring fixed at 0.4, is that almost all agents act with care and there are few transgressions.

5 Conclusion and related work

In this preliminary work, we investigated law enforcement within a population of learning agents. We hope that this approach will allow us to shed new lights on the analysis of law enforcement, in particular with respect to more traditional calculus using expected utilities.

Many researches in fields such as Law, Philosophy of Law or Law & Economics inspect issues and solutions of enforcement. Many amongst them settle an utilitarian framework to design enforcement regimes (see e.g [8,7,6]) while we are interested to get further insights on the assumptions of norm-governed learning agents expressed in probabilistic rule-based argumentation. Our approach is thus closer to other logic-based simulation. For example, [3] showed that there are some situations where the cost of enforcement is sufficiently prohibitive that a certain level of non-compliance can be supported, supposing that the compliers were willing to indulge the 'enfant terribles' for the sake of the collective and not everyone was behaving that way. As more and more compliers became non-compliers then 'the system' had to start doing the monitoring and imposing the enforcement, and paying the costs of doing so. So the system not only had to have graduated sanctions, but also had to have a mechanism to customise the system of graduated sanctions according to the environment, which included the distribution of compliance/non-compliance tendencies in the population.

References

1. D. C. Dennett. *The Intentional Stance.* The MIT Press, Cambridge, MA, 1987.
2. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.
3. J. Pitt and J. Schaumeier. Provision and appropriation of common-pool resources without full disclosure. In *PRIMA*. Springer, 2012.
4. H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2011.
5. R. Riveret, A. Rotolo, and G. Sartor. Norms and learning in probabilistic logic-based agents. In *DEON 2012*. Springer, 2012.
6. S. Shavell. Liability for harm versus regulation of safety. *The Journal of Legal Studies*, 13(2):357–374, 1984.
7. S. Shavell. A model of the optimal use of liability and safety regulation. *The Rand Journal of Economics*, 15(2):271–280, 1984.
8. D. Wittman. Prior regulation versus post liability: The choice between input and output monitoring. *The Journal of Legal Studies*, 6(1):193–211, 1977.