AI Communications 23 (2010) 325–339
DOI 10.3233/AIC-2010-0477
IOS Press

Norm internalization in artificial societies

Giulia Andrighetto [a,*], Daniel Villatoro [b] and Rosaria Conte [a]
[a] *LABSS, Institute of Cognitive Science and Technologies, CNR, Rome, Italy*
[b] *Artificial Intelligence Research Institute (IIIA), Spanish National Research Council, Spain*

Abstract. Internalization is at study in social-behavioural sciences and moral philosophy since long; of late, the debate was revamped within the rationality approach to the study of cooperation and compliance since internalization is a less costly and more reliable enforcement system than social control. But how does it work? So far, poor attention was paid to the mental underpinnings of internalization. This paper advocates a rich cognitive model of different types, degrees and factors of internalization. In order to check the individual and social effect of internalization, we have adapted an existing agent architecture, EMIL-A, providing it with internalization capabilities, turning it into EMIL-I-A. Experiments have proven satisfactory results with respect to the maintenance of cooperation in a proof-of-concept simulation.

Keywords: Rich cognitive modeling, norms, internalization, agent architecture, autonomous agents, agent based social simulation

1. Introduction

The problem social scientists still revolve around is how autonomous systems, like living beings, perform positive behaviors toward one another and comply with existing norms, especially since self-regarding agents are much better-off than other-regarding agents at within-group competition. Since Durkheim, the key to solving the puzzle is found in the theory of internalization of norms [30,32,34,39,43]. One plausible explanation of voluntary non-self-interested compliance with social norms is that norms have been internalized.

Internalization occurs when

> a norm's maintenance has become independent of external outcomes – that is, to the extent that its reinforcing consequences are internally mediated, without the support of external events such as rewards or punishment [4, p. 18].

Agents conform to an internal norm because so doing is an *end* in itself, and not merely because of external sanctions, such as material rewards or punishment.

Norm internalization is one of the common themes running across all of the social-behavioral disciplines and there are not many. Not only sociologists, but also developmental, social and cognitive psychologists have perceived its crucial role in socialization [6,38, 44]. Drawing on the early work by Vygotzky (published in the US as late as 1978 [53]), Piaget [44] and Kohlberg [38], and several other psychologists showed that a parental attitude oriented to elicit norm internalization predicts children's later well being and even their inclination to other-regarding behavior [25].[1]

Despite these important contributions, however, the community's scientific definition and understanding of the process of norm internalization is still fragmentary and insufficient.

The main purpose of this paper is to argue for the necessity of a rich cognitive model of norm internalization in order to (a) provide a unifying view of the phenomenon, accounting for the features it shares with related phenomena (e.g., robust conformity as in automatic behavior) and the specific properties that keep it distinct from them (autonomy); (b) model the process of internalization, i.e., its proximate causes (as compared to the distal, evolutionary, ones; see [29,30]); (c) characterize it as a progressive process, occurring at various levels of depth and giving rise to more or less robust compliance (see [25] for a fine grained distinction of different degrees of norm internalization); and finally (d) allow for flexible conformity, enabling agents to retrieve full control [9] over those norms which have been converted into automatic behavioral responses [28].

Thanks to such a model of norm internalization, it will be possible to adapt existing agent architectures (see EMIL-A, [2,15]) and to design a simulation platform to test and answer a number of hypotheses and questions such as: which types of mental properties and ingredients ought individuals to possess in order to

*Corresponding author. E-mail: giulia.andrighetto@itc.cnr.it.

[1] See [40], for an overview of the research on Norm Internalization.

exhibit different forms of compliance? How sensitive each modality is to external sanctions? What are the most effective norm enforcement mechanisms? How many people have to internalize a norm in order for it to spread and remain stable? What are the different implications for society and governance of different modalities of norm compliance?

Throughout the paper, the process of norm internalization will be meant as a mental process that takes (social) norms as inputs and gives new terminal goals of the internalizing agent (from now on, the internalizer) as outputs.

Emotions, playing a significant role in this process, will not be investigated at this stage.

2. Related work

Norms frequently become internalized [46]. Norm internalization is mainly favoured by socialization institutions, such as family and school, working hard to make internalize a wide variety of norms, especially in young people, and by informal organizations of friends and peers.[2]

Contributions to explain internalization are sometimes based on *reinforcement learning* theory. Scott [46], for example, theorized that norm internalization leads to robust compliance, provided the external sanctioning system is *never* completely abandoned. Unfortunately, this explanation is incompatible not only with the view that *social norms can get internalized to the extent that they do not need social enforcement* [10, p. 2], but also with experimental evidence. For example, subjects playing ultimatum games are found to follow fairness considerations even when unobserved [11].

In the last few years, a strong renewal of interest around the notion of norm internalization appeared in the evolutionary game theoretic study of cooperation and prosocial behaviour. Gintis [29,30] argued that the increase in social complexity of early human society produced a rapidly changing environment, which in turn posed an adaptation problem to the genetic mechanisms for *altering* goals. Internalization of norms is adaptive because it *facilitates the transformation of drives, needs, desires and pleasures into forms that are more closely aligned with fitness maximization, while a purely genetic adaptive process would have taken orders of magnitude longer in time* [30, p. 62]. Though extremely interesting, this theory is focused on the distal (social and evolutionary) causes of norm internalization while leaving aside the proximate (internal and cognitive) causes.

Some authors (see [28]) conceive norm internalization as a process leading to a sort of *automatic* or *thoughtless conformity*. Epstein [28] writes: "when I'd my coffee this morning and I went upstairs to get dressed, I never considered being a nudist for the day". People, observes Epstein [28], blindly conform to the norm: they observe what the majority does and they act accordingly. Then the more they have done so in the past, the more they will redo it in the future. Agents learn not only which norms to conform to, but also how much they should think about them. In the author's view, internalization is learning not to think about norms.

Does this mean internalized norms are thoughtlessly complied with? What about the difference between an action done out of a 'sense of duty' and a habit? How can the internalizer gain control again over an automated action and refrain from applying a given routine? How can it move on when the traffic light is red but the policeman invites it to proceed? Even if another routine is activated by the new event (policeman invitation to proceed), how and why is one routine (stop) interrupted and fired the complementary one (move on)? How is the conflict solved?

The crucial necessity, here, is to provide a common ground, *an agent model that can exhibit different forms of norm internalizations, and, what is more difficult, can shift form one to the other*. We need to account for reversible routines, or, which is the same, for flexible conformity. In principle, a modular normative architecture nicely fits flexible automaticity and as it will be described in Section 6 EMIL-A seems a good candidate for this undertake.

3. Objectives

This work is aimed to propose agent based modeling, and in particular rich cognitive modeling, as a framework for casting a theory of the cognitive underpinnings of internalization and to characterize norm internalization as a progressive, multi-step process, leading from externally-enforced norms to norm-corresponding goals, and actions pursued for their own sake.

The normative agents we model are provided with a new measure, the *salience* of the norm, allowing them

[2] Gintis [29] claims that these formal and informal institutions favour vertical, oblique and horizontal transmission.

to detect the degree of activity of the social norms within their social setting. As we will show, this feature provides the agents with several advantages such as an additional criterion to better evaluate whether to observe the norm or not. The normative salience allows agents to leverage their normative social information when effectively addressing the norm internalization process.

We also provide a proof-of-concept simulation aimed to test the normative architectures discussed. The objective of the simulation is threefold:

1. Check the emergence and stability of cooperation in a simplified scenario. Agents perform dyadic interactions playing a classic Prisoner's Dilemma. Some agents are initialized with a *Cooperation* norm, and they will also defend this norm through the application of sanctions.
2. Observe run-time the different phases an agent goes through when the internalization process is taking place.
3. Analyze the interaction dynamics and their effect on different normative architectures. By allowing different types of agents to interact in a normative context, we can observe as system designer, the consequent effects of their behavior in the society. This knowledge would be helpful for policy-makers in open multi-agent societies.

The rest of this article is organized as follows. First, we will introduce a theory of mental dynamics of norms, in order to provide some basic concepts (see Section 4). Second, norm internalization will be analyzed, focusing on different types and levels of this process (see Section 4). We will provide some preliminary hypotheses about factors affecting internalization further on.

In Section 6, we will present a normative agent architecture, EMIL-I-A, able to account for different forms of norm internalization. In Section 7, we will present a simulation model aimed to test our normative architecture in a simplified scenario. In Section 7.4 results will be discussed. Finally, conclusions and future work will follow.

4. The cognitive dynamics of norms

In order to understand the process of norm internalization, some preliminary notions should be clarified. Building on Ullman-Margalit's definition of a norm [50] as *a prescribed guide for conduct which is generally complied with by the members of society*, we [2,15] define a norm as a behavior that spreads through a given society to the extent that the corresponding prescription spreads as well, giving rise to a shared set of *normative beliefs* and *goals*. A normative belief is a mental representation, held to be true in the world, that a given action is either obligatory, forbidden or permitted for a given set of individuals in a given context. On the other hand, a normative goal is an internal goal[3] *relativized* [4] to a normative belief: it is the will to perform an action *because* and *to the extent* that this is believed to be prescribed by a norm.

There are at least three main types of *normative beliefs*:

- The main normative belief, stating that: *there is a norm prohibiting, prescribing, permitting that . . .* [22,23,37,52].
- The normative belief of pertinence, indicating the set of agents on which the norm is impinging.
- The norm enforcement belief, indicating that a positive sanction is consequent to norm obedience and a negative sanction is consequent to norm violation.

In order to be compliant with the norm, the first two normative beliefs are necessary conditions: agents should recognize that there is a norm and that it applies to them. We claim that when individuals do not have them in mind, norms exert no effect on the behavior [2,15]. Furthermore, the more these normative mental representation are salient, the more they will elicit a normative behavior [11,16,55]. Norm compliance and norm salience are strongly intertwined: findings from psychology [7,17] and behavioral economics [12,54] have pointed out that drawing people's attention on a social norm and making it salient elicits an appropriate behavior. Making a norm salient typically means providing people with information about the behavior and beliefs of the other individuals [13, p. 4] (see Section 5), for a detailed description of the dynamics of salience.

[3] From a cognitive point of view, goals are internal representations triggering-and-guiding action at once: they represent the state of the world that agents want to reach by means of action and that they monitor while executing the action [20].

[4] A goal is relativized when it is held because and to the extent that a given world-state or event is held to be true or is expected [18]. An example is the following: tomorrow, I want to go sunbathing to the beach (relativized goal) because and to the extent that I believe tomorrow it will be sunny (expected event). The precise instant I cease to believe that tomorrow it will be sunny, I will drop any will to go to the beach.

Despite the main normative belief and the belief of pertinence, the norm enforcing belief is not a defining element of the norm, it simply enforces it. A *normative command* is a special command that is intended to be adopted by its addressees because it is normative and *norms must be obeyed* [52]. Of course this motive can be absent or weak in the minds of people, depending on the socialization and education process and the credit obtained by current institutions. *Sub-ideally*, norms are often complied with because they are enforced by a system of sanctions. But *ideally*, they are meant to be observed because are norms and should be complied with for their own sake.

However, a belief is not yet a decided action. Normative beliefs are necessary but insufficient conditions for norms to be complied with. What leads agents endowed with one or more normative beliefs to execute them, especially since, by definition, norms prescribe costly behaviours? How can norms generate goals?

Usually normative beliefs generate normative goals by reference to an external enforcement[5] (sanctions, approval, etc.). The agent calculates the costs and benefits of complying with or violating the norm and then decides how to behave. If no such a goal is generated, the norm will be violated. On the other hand, a norm is internalized when the norm addressee complies with it independent of external sanctions and rewards. In such a case, the normative goal is *no more relativized* to an expected sanction, but only to the main normative belief.

There are also other types and levels of internalization, so for example under some circumstances (see Section 5), the decision-making is avoided and the output is a conditioned action in the agent's repertoire fired by a perceived event. In the traffic light example, this consists of the sequence of movements necessary to activate the breaks of the car, a behavioral response so deeply internalized that one can hardly make it explicit.[6]

Norm compliance is expected to be more robust if norms are internalized then is the case when conducts are ruled only by external sanctions. If everybody internalizes a given norm, there is no incentive to defect and the norm remains stable [5, p. 1104]. Gintis argues that

> where people internalize a norm, the frequency of its occurrence in the population will be higher than if people follow the norm only instrumentally (i.e., when they perceive to be in their interest to do so) [29, p. 61].

Since it is very difficult that everybody in the population internalizes the norm, an interesting question for modeling, also suggested by Axelrod [5], is *how many people have to internalize a norm in order for it to remain stable*. This and other questions on norm internalization will be tested with a simulation model and some results will be shown in Section 7.

Moreover, in many circumstances, an agent that has internalized the norm will exercise a special form of social control, getting others to comply with the norm, reproaching transgressors and reminding would-be violators that they are doing something wrong. Norm defence is extremely important in the spreading and stabilization of norms over a population of autonomous agents [24]. As Axelrod [5] suggests, lowering the temptation to violate the norm might be not enough. Even in groups in which most people comply with the norm, if no one has an incentive to punish the remaining violators, the norm could still collapse. This means that internalizers are agents that not only comply with norms but also have an incentive to punish anyone else who does not comply with the norm (see [3], for a theoretical and simulative model of this phenomenon). If agents are driven to honour norms, they are likely to defend it, directly or indirectly.

5. Factors affecting internalization. Some preliminary hypotheses

Why do agents observe a norm irrespective of external enforcement? Far from providing a complete list of the factors favouring norm internalization, in the present work we will focus on some of the elements playing a key role in this process, such as:

- consistency;
- self-enhancing effect;
- urgency;
- calculation cost saving;
- norm salience.

Let us start with *consistency*. This mechanism operates at two stages: first by selecting which norm to internalize and later by enforcing it (self-enforcement)

[5] See [21,23] for a fine grained analysis of different reasons behind norm compliance.

[6] Interestingly, however, under the effect of other perceived events, conditioned actions may be blocked for the time interval required to process a disturbing or interfering event and restored later [9] in a semi-conscious fashion. Here, the behavioral response is automatic. For example, rather than stopping at the traffic light, go ahead to let an ambulance overtake.

and controlling that one's behavior corresponds to it (self-control). Consistency of new norms with one's beliefs, goals and previously internalized norms plays a crucial role in the selection process. Successful educational strategies favor internalization processes, often by linking new inputs with previously internalized norms. Analogous considerations apply to policymaking. Consider the antismoking legislation: the efficacy of antismoking campaigns based on frightening announcements and warning labels (e.g., sentences like 'Smoking kills' on cigarette packages, see [31]) is still controversial. One of the factors reducing the efficacy is the effect known as *hyperbolic discounting* ([14]; see also [45]), a psychological mechanism that leads to invest in goal-pursuit a measure of effort that is a hyperbolically decreasing function of the time–distance from goal-attainment and leads people to procrastinate energy-consuming work until the very last moment. Due to hyperbolic discounting, people, especially young people, are unable to act under the representation of delayed consequences of current actions. On the other hand, much more efficacious antismoking campaigns are those playing on previously emerged and diffuse set of social norms, such as the live-healthy precepts, highly consistent with the message they want to transmit.

The second factor playing a role in norm internalization is the *self-enhancing* effect of norm compliance: the norm addressee realizes that it achieves one of its goals by observing a given norm. Suppose I succeed in refraining from smoking and that after a few days, I realize an advantage that I had not perceived before: food starts to taste again. This discovery generates a goal (quit smoking to enjoy good food), not relativized to the norm but supporting it: I have converted the norm into an ordinary goal. Whether this goal will be strong enough to out-compete addiction is another matter.

Third, we focus on *urgency* (see, for example, [56]). In particular, one can argue that the more a given norm allows to answer problems frequently encountered under conditions of urgency, when time for decision-making is none or scanty, the more likely that norm will be internalized.

Fourth, we claim that agents are *parsimonious calculators*: under certain conditions, they internalize norms in order to save calculation and execution time. Imagine a driver's decision to stop at the traffic light when it turns red. Each time our driver approaches a red traffic light, it calculates the costs and benefits of complying with the norm: e.g., it predicts that if it does not stop, it will gain time, but that with a certain probability a fine will follow to its violation. It then chooses what is best for itself. After a certain amount of calcula, always giving the same output (e.g., the driver always decides to stop in order to avoid punishment), it will abstain from calculating: it will stop when the traffic light is red without thinking anymore on what to do. The agent will save calculus time, thus acting in a more effective way. Thinking declines when norms gain force and gradually stops once they are internalized.

This last point is strictly intertwined with another important factor favoring norm internalization: i.e., norm salience.

Norm salience (see [2,15]) is defined as the degree of activity and importance of a norm within a social group and a given context. The more salient the norm, the more likely it will be internalized as a conditioned action, a routine activated under specified conditions. Moreover, the higher the salience of the norm, the more deeply it will be internalized.[7]

The salience of a normative belief can vary depending on several social and individual factors. On one hand, the actions of others provide information about the importance of a norm within a group [11,17,27,33].

For example, the *surveillance rate* (frequency and intensity of punishment), the *quality of normative services* (e.g., if the road signs are well marked), the amount of *compliance* and the costs and efforts spared to *educate* the population to a certain norm are all cues signaling us the relevance of a norm within a group. On the other hand, norm salience is also affected by the individual sphere: it is dependent on how much that norm is entrenched with beliefs, goals, values and previously internalized norms of the agent [19,49].

6. Internalizer: The EMIL-I-A architecture

In order to account for the different forms, levels and processes of internalization, a rich cognitive platform, namely a BDI-type architecture is required and EMIL-A [2,15] seems a good candidate.

This normative architecture consists of mechanisms and mental representations allowing norms to affect the behavior of autonomous intelligent agents. As any BDI-type (Belief-Desire and Intention) architec-

[7] It has to be pointed out that the norm salience can also gradually decrease: for example, it happens when the agent realizes that norm violations do not receive any punishment or if the normative beliefs stay inactive for a certain amount of time, this meaning that the norm is not very active in the population anymore.

ture EMIL-A operates through modules for different sub-tasks (recognition, adoption, decision-making) and acts on mental representations for goals and beliefs in a non-rigid sequence.

For further references on how the norm recognition module works, we refer the reader to [15]. After recognition, a norm activates the three types of normative beliefs described in Section 4, that stored in the *normative board*.[8] Once generated or activated, normative beliefs will be inputed to the norm-adoption module: a normative goal – relativized to the expected enforcement – will be generated. In this condition the normative agent adopts the norm, because it wants to avoid punishment. The normative goal is then inputed to the decision-maker and compared with other goals possibly active in the system. The decision-maker will choose which one to execute and will convert it into a normative intention (i.e., an executable goal). Once executed, this normative goal will give rise to norm compliance and/or norm defense and/or norm transmission through communication. Otherwise, it will eventually be abandoned, solution that brings again to norm violation.

A crucial aspect of EMIL-A is the possibility to account for the occurrence of interruptions, modifications and deviations from the processes described so far: a norm can be internalized and even become a habit, a (semi) automatism, a routine behavior. In this work we have endowed EMIL-A with an internalization module, thus creating EMIL-I-A (EMIL Internalizer Agent).

EMIL-I-A internalizes a norm when two conditions apply: (a) the norm salience and (b) the cost-benefit computation time exceed a certain threshold. The internalizer is endowed with a *normative thermometer*, signaling him the social and individual salience of a certain norm. If the norm is highly salient, the agent will internalize it. The internalizer is also a computation costs optimizer. After having weighted for a certain number of times, the costs and benefits of complying or not with a certain norm (and having reached everytime exactly the same decision), the agent stops calculating and consider it the best choice. Once internalized, EMIL-I-A stops the normative deliberation and complies with the norm.

Norm salience is an important feature of our agents, improving their performance in several ways. It allows internalizers to observe norms in a *flexible* and *automatic* way. Salience enables the agents to *dynamically* monitor if the normative scene is changing and to adapt to it.[9] For example, in an unstable social environment, if the norm enforcement suddenly decreases, agents having highly salient norms are less inclined to violate them. A highly salient norm is a reason for which an agent continues to comply with it even in the absence of punishment. It guarantees a sort of inertia, making agents less prompt to change their strategy to a more favorable one. Vice versa, if a specific norm decays, internalizers are able to detect this change, ceasing to comply with it and adapting to the new state of affairs. Finally, if an agent faces an emergency or a normative conflict, norm salience allows him to decide which action to perform providing him with a criterion to compare the norms applicable to the context.

This is possible because our normative agents are as autonomous as socially responsive. They are autonomous in that they act on their own beliefs and goals (on the basis of their salience). However, they are also responsive to their environment and to the inputs they receive from it, especially to social inputs.

7. The model

In order to test the theory presented in this work we have provided our agent architecture with the necessary capabilities to internalize norms in a proof-of-concept multi-agent based simulation. Before explaining the simulation, we define the three types of agents whose performance we compare: strategic, normative and internalizers.

The decision-making process of a *normative* agent works in the following way. A *normative agent* (Fig. 1(b)) is a BDI agent which has (a) the normative belief of the existence of a norm (i.e., the main normative belief) and (b) the normative belief of a consequent sanction if the norm is not observed (i.e., the norm enforcement belief) and the goal of maximizing its utility. We hypothesize that *normative* agents observe norms because of the existence of this sanction (as it reduces its utility), i.e., if a norm is intensively defended through the application of sanctions, the normative agent will consequently observe it; on the other hand, when the norm is not defended, the normative agent will probably not observe the norm.

[8] The normative board is a portion of the long-term memory where normative beliefs are stored, ordered by salience.

[9] It is interesting to notice that this mechanism allows agents to record the social and normative information, without necessarily proactively exploring the world (e.g., with a trial and error procedure).

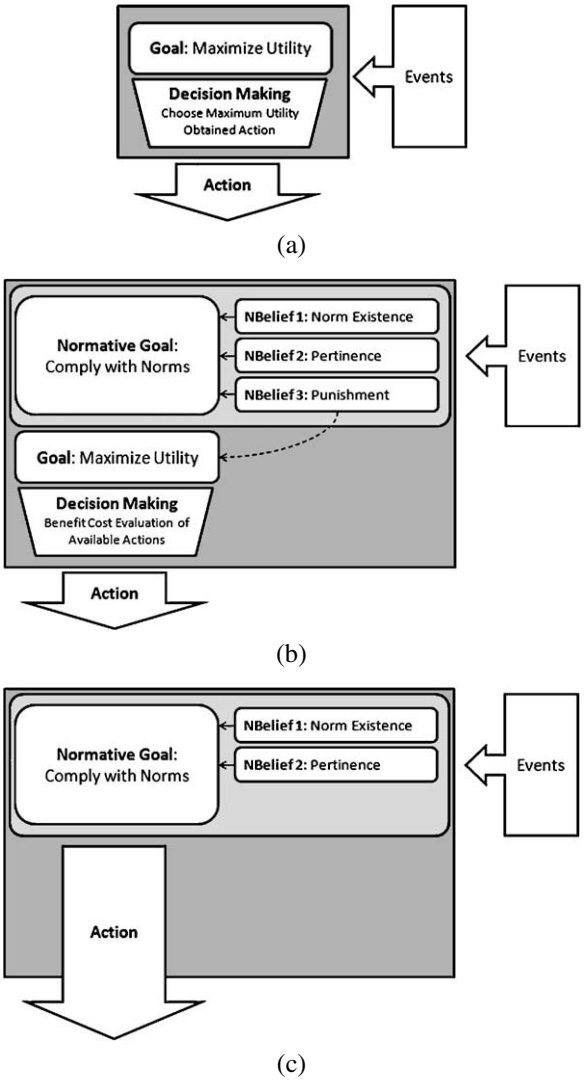

Fig. 1. Agent architectures. (a) Strategic, (b) normative, (c) internalizer.

In our implementation, a *normative* agent knows that a norm exists and is able to predict that violators are punished. Both beliefs are presented as normative beliefs. However, agents do not know beforehand what the surveillance rates of the norm are. During the simulation, agents update (with their own direct experience and observed normative social information) the perceived probability of being punished. The decision-making of a normative agent is also sensible to a *risk tolerance* rate: when the perceived punishment probability is below the *risk tolerance* threshold, agents will decide to violate the norm; they will observe the norms otherwise. Although this process might provide agents with a maximum benefit, it yields the computational cost of evaluating each of the options at everytime step and a cost to the society as norm abiding agents will only behave normatively in the presence of punishment.

As we explained in Section 6, an *internalizer* agent is basically a normative agent who is able to internalize norms (same normative structure plus an internalization module). An *internalizer* (Fig. 1(c)) initially behaves as a normative agent. Meanwhile, the salience mechanism also works in the agents minds. This salience, as specified in Section 5, provides agents with a measure of importance of a norm within the population, and is constantly updated with the normative social and individual information available to agents. Norm observance, violations, norm defense, explicit deontic messages are some of the parameters that affect the norm salience.

From the technical point of view, an agent will *internalize* a norm when both following conditions are fulfilled:

1. The norm salience is above a certain threshold, indicating the norm is important within the society.
2. The agent rationale specifies that it has done the benefit–cost calculation for all the possible actions (as all normative agents do so during the decision-making process) for a certain number of times and that this cost of calculation times is now above its tolerance threshold.

Once a norm is internalized, *internalizers* do not make the benefit–cost calculation anymore and they observe the norm as an *automatism*. Nevertheless, the salience mechanism is still active and is still continuously being updated. In this way agents are able to unblock the automatism of a norm and return to the benefit–cost calculation stage.

The third types of agents are *strategic* ones. They do not know about the existence of the norms and will always choose the action that has provided them the maximum benefit in the past. This agents have been implemented as Q-learning agents as in [47,51].

7.1. Simulation model

In order to observe the dynamics of agents behavior, we have designed a multi-agent based simulation where the different types of agents can interact to perform the same task. The simulation is structured in the following way (see Fig. 2). Agents are connected with a social network which specifies their interaction

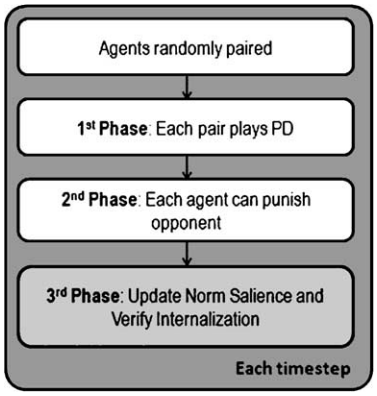

Fig. 2. Simulation process.

topology. Each timestep, they will repeatedly interact in randomly formed couples of neighboring agents and play a Prisoner's Dilemma (PD) (first stage). After the game, all agents can punish their opponents in case they defected (second stage). In order to provide a more realistic model of punishment than those present in [26], we include two different types of punishment: *strategic* punishment and *normative* punishment. Both of them yield a cost to the punisher, affecting also the utility of the punished defector and producing a deterrent effect. However, the normative punishment is also accompanied with a deontic message or a normative evaluation making explicit the norm existence. Both normative and internalizers inflict strategic punishment while the norm's salience of the agents is below a certain threshold (agents do not yet believe in the norm); once this threshold is reached, agents will continue punishing normatively, obtaining the same deterrent effect but also adding an educative perspective to the sanction. We consider that punishing normatively is also a meta-observance of the norm.

For the proposed agent architecture, agents need to process the normative social information available in their environment. This social information affects directly the norm salience mechanism which orchestrates the internalization process. In order to obtain this information we have added one more phase to the game: after all agents have chosen their first stage action (cooperate or defect) and their second stage action (punish or not), they can draw the following social normative information within their neighbours:[10]

- First stage cooperators: neighboring agents who chose *cooperate* in the PD.
- First stage defectors: neighboring agents who chose *defect* in the PD.
- Non-punished defectors: amount of neighboring agents that defected in the first stage and were not punished in the second stage by any other agent.
- Consistent strategic punishments observed: neighboring consistent agents[11] who have applied strategic punishment, or, neighboring agents who have received a strategic punishment from a consistent agent.
- Consistent normative punishments observed: neighboring consistent agents[10] who have applied a normative punishment, or, neighboring agents who have received a normative punishment from a consistent agent.
- Consistent educative messages observed: neighboring consistent agents[10] who have sent an educative message, or, neighboring agents who have received an educative message from a consistent agent.
- Consistent strategic punishments received: amount of strategic punishments a certain agent receives from a consistent agent.
- Consistent normative punishments received: amount of normative punishments a certain agent receives from a consistent agent.
- Consistent educative messages received:[12] amount of educative messages a certain agent receives from a consistent agent.

An aggregation of this social information provides agents with an estimate measure of the salience degree of a norm within the society and from their subjective point of view. The weights' values used in the aggregation calculation (interpreted from [17][13]) are shown in Table 1.

The resulting value from the aggregation functions is normalized between 0 and 1 and added accumulatively to the norm salience. We have not permitted salience to go above 1 or below 0. Normative agents update their probability of being punished (non-punished defectors divided by number of defectors)

[10] The amount of neighbours is defined by the structure of the social network in which agents interact.

[11] An agent is consistent if when it chooses to punish, has also cooperated in the PD.

[12] These normative messages are sent by educational agents. These types of agents are agents which communicate norms. The role of educative can be assumed by a normative agent or by an internalizer.

[13] The intuitive justification for the usage of these values is that of giving a higher weight to those social cues that are highly related to normative motivations and lower weights to those which would have selfish/utilitarian motivations.

Table 1
Normative social information weights for salience aggregation function

Social cue	Weight
Self first stage cooperation	0.99
Self first stage defection	−0.99
Self second stage strategic punishment	0
Self second stage normative punishment	0.99
Observed first stage cooperators	$0.33 \times n$
Non-punished defectors	$-0.66 \times n$
Consistent strategic punishments observed	$0.33 \times n$
Consistent normative punishments observed	$0.99 \times n$
Consistent educative messages observed	$0.99 \times n$
Consistent strategic punishments received	$0.33 \times n$
Consistent normative punishments received	$0.99 \times n$
Consistent educative messages received	$0.99 \times n$

each timestep with the same information by which they update their norm salience.

7.2. Experimental design

In order to study the behavior of the 3 types of agents, normatives, internalizers and strategics, some parameters are fixed for all the simulations performed. Agents are located in a fully connected network, allowing agents to potentially interact with all other agents present in the simulation.

As the objective of this work is to observe the behavior of the internalizers, we will not focus on punishment dynamics (for further information we refer the reader to [3]); the cost of punishment will be fixed to 1 unit for punishers reducing violators utility in 4 units (1:4 punishment technology is used because it has been shown [42] to be more effective in promoting cooperation). The decision-making of both normative and internalizers is not affected by the cost of applying a punishment; however, the cost of being punished affects their first stage decision-making. All the simulations are populated with a fixed amount of 10 *educators*,[14] and a total population of 100 agents, varying the proportion of pure strategic and internalizers (e.g., in a population where there are the fixed 10 educators, and 20 internalizers, the other 70 agents are strategic agents). We remind the reader that an internalizer is a normative agent that will eventually internalize a behavior and that is the reason why we do not explicitly include normative agents in our experiments: by having internalizers, we already represent the dynamics of these normative agents.

Our hypotheses to be proven through experimentation are:

- Strategic and Normative agents do behave in a selfish efficient strategy, provoking a collapse of cooperation when punishment rates are low.
- Internalizer agents are able to maintain the social order imposed by social norms even if those are not defended (temporarily, allowing the system to recover from possible failures; permanently, salience will indicate the state of the norm, allowing internalizers to unblock the automatism generated by the norm and start the benefit–cost calculation process again).
- Internalizer agents are able to unblock normative automated actions when this norm disappears.

The results presented in this section are the average results of 25 simulations. All non-strategic agents are initialized with a constant propensity to violate norms if the perceived probability of being punished is equal to or lower than 30%. They are also given a constant exploration rate of 0.5% allowing them to take a first stage random action. To study the effect of punishment on norm internalization, we have designed several punishment probability distributions:

1. *Constant*: agents have a constant probability of being punished.
2. *Linearly increasing*: the probability continuously increases as the simulation runs.
3. *Linearly decreasing*: the probability continuously decreases as the simulation runs.
4. *Step down*: at a certain moment of the simulation, the probability of being punished drops from 1 to 0, i.e., from total punishments to no punishment.
5. *Step up*: at a certain moment of the simulation, the probability of being punished raises from 0 to 1, i.e., from no punishment to total punishments.

These different punishment probability distributions allow us to observe the dynamics of cooperation with the different types of agents. On the other hand, these punishment probabilities are completely unrelated with the second stage mixed strategy decision. Once agents decide whether to or not to punish a defector, then, this decision might be unachievable because of environmental conditions (simulated with these punishment probability distributions).

[14] Educators are agents hold the norm since their creation and also send educative messages.

7.3. Experimental results

This first experiment has the main objective of showing the norm internalizers dynamics when within a society formed by pure strategic agents and other internalizers. Information such as the salience of the norm and the number of internalizers in the *automatism phase* (when norms are internalized) are those that we need to observe carefully.

The experimental results in Figs 3 and 4 show the dynamics of the internalizers.[15] In this experiment, the punishment probability distribution decreases linearly. These results are obtained from simulations with a fixed amount of agents (=100), and changing the distribution of internalizers and strategic agents.

The x-axis specifies the timesteps of the simulation, the y-axis specifies the number of internalizers. In Fig. 3(a) the z-axis specifies the cooperation rates; the salience in Fig. 3(b) and the internalization rates in Fig. 3(c). We can observe (in Fig. 3(a)) that the amount of internalizers is directly proportional to the stability of the cooperation rates: the more the internalizers, the longer the cooperation. The explanation of the phenomenon is found in the dynamics of the internalizers: they start behaving as normative agents, and as the punishment probability is above their risk tolerance, they comply with norms. At a certain moment, those that are able to internalize, do so (as it can be seen in Fig. 3(c)). However, when the punishment rates decrease, strategic agents start defecting. As specified before, the fewer the internalizers in the population, the more the strategic agents. When the perceived punishment probability decreases, strategic agents start defecting. With a higher number of strategic agents, the amount of defections is also higher, affecting salience in a more radical way (as can be seen in Fig. 3(b)).

The conclusions drawn from this first experiment are confirmed by the following one (in Fig. 4). The simulation conditions are the same except for the punishment distribution, which in this case is a *Step down* distribution: agents will be able to punish up to timestep 600, after that, they will not be able to punish anymore.

In Figure 4(a), we can observe that in populations with a larger amount of internalizers, cooperation remains stable unlike in those with fewer internalizers. By observing the results in Fig. 4(c), we can analyze the dynamics of internalizers: internalizers take around 600 timesteps to internalize the norm; after punishment disappears (in timestep 600), internalizers remain in the internalization phase for a longer time when the number of them is higher.

In order to observe the dynamics of internalizers with respect to the amount of strategic and normative agents, we run an exhaustive experiment of the search space with all the punishment distributions presented. However, due to space constraints, we show the results related to a low amount (10) of internalizers (in Fig. 5). Obviously, by introducing a higher amount of internalizers, cooperation rates stabilize in a more robust way than with fewer internalizers.

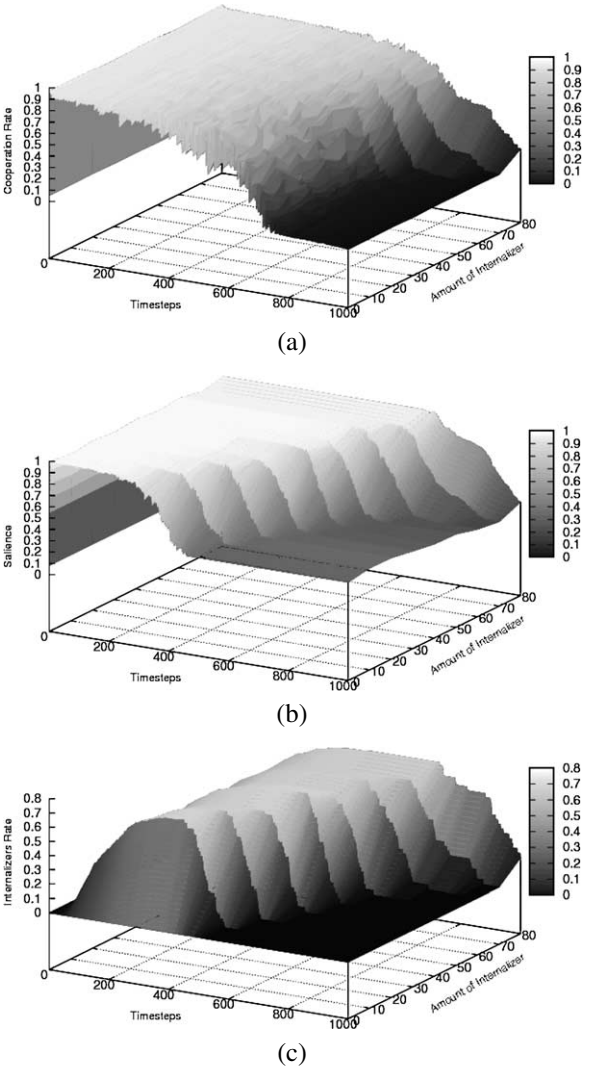

Fig. 3. Internalizer dynamics with a linearly decreasing punishment distribution. Internalizers are normative agents able to internalize. (a) Cooperation rates, (b) salience, (c) internalization rates.

[15] Internalizers are normative agents (that already hold the norm) and do have the capability to internalize the norms.

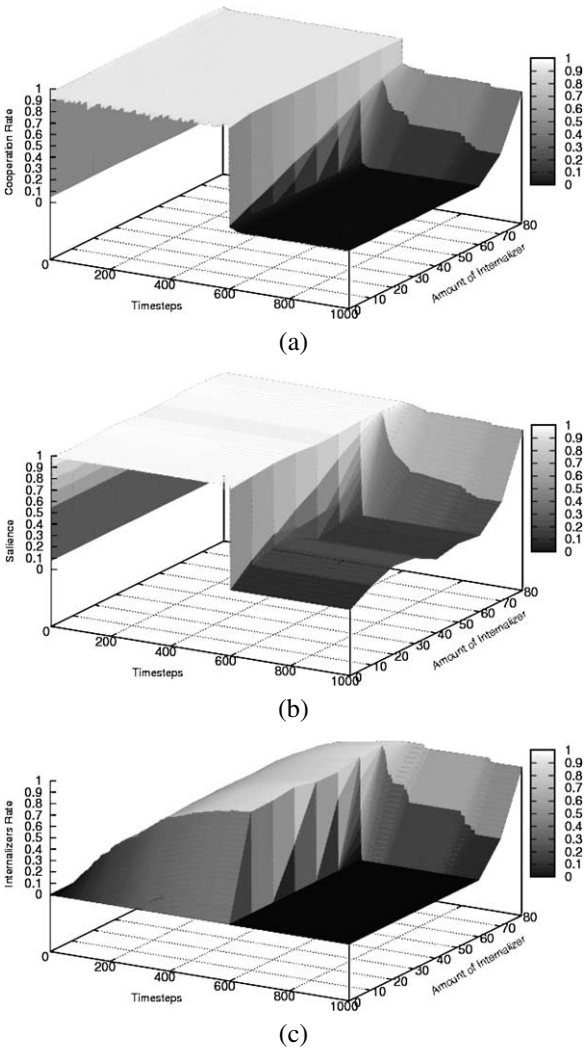

Fig. 4. Internalizer dynamics with punishment distribution step down at 600. Internalizers are normative agents able to internalize. (a) Cooperation rates, (b) salience, (c) internalization rates.

The experiment with a low amount of internalizers (Fig. 5) shows that cooperation rates are more unstable with a higher number of strategic agents. This phenomenon is caused by the natural propensity of strategic agents to exploration and utility maximization; when punishment rates decrease, they start defecting, thus leading to unblock the generated automatism. In the same figure, we can also observe how normative agents are sensitive to punishment probability distributions: when the perceived probability is below 0.3 (their risk tolerance rate), defection is the dominant strategy.

One important remark about the effects of internalization concerns the costs of punishments for society. In Fig. 6, we can observe how in populations with a higher amount of internalizers (and lower amount of strategic agents), the amount of punishments inflicted is lower than in those with fewer internalizers. A reduction of the amount of costly punishments would imply a significant reduction in agents' expenditures to maintain cooperation.

7.4. Discussion

The proof-of-concept simulation model has confirmed our initial hypotheses. Allowing different types of normative agents to interact provides system policymakers with a tool that can help them predict the dynamics of prosocial behavior.

Internalizers are endowed with a rich cognitive architecture allowing them to maintain high cooperation rates even when punishment rates are low. This phenomenon is not observed when dealing with populations of strategic agents, whose ultimate intention is to maximize their utility, leading to a general collapse of cooperation. We have observed an interesting phenomenon with normative agents, which will maintain cooperation when the punishment rates are above their risk tolerance threshold.

These results would lead us to think that a complete population of internalizers would be the best solution in terms of system performance; unfortunately, this is not true. Internalizers do need a certain amount of strategic or normative agents to unblock the normative automated actions when necessary. The norm salience mechanism allows the system to recover from a possible failure in the sanctioning structures. However, when the norm disappears, agents will eventually unblock the automatism generated by the internalization process and start the whole process of norm recognition and internalization again.

We have also observed that a significant number of internalizers is convenient for the society in general, as they keep the cost of punishment low.

8. Advantages and disadvantages of norm internalization

Hypotheses concerning the effect of internalization follow from the properties of internalization analyzed so far. To some extent the advantages of internalized norms are easily identifiable: norm compliance is expected to be more robust if norms are internalized than is the case when conducts are ruled only by external

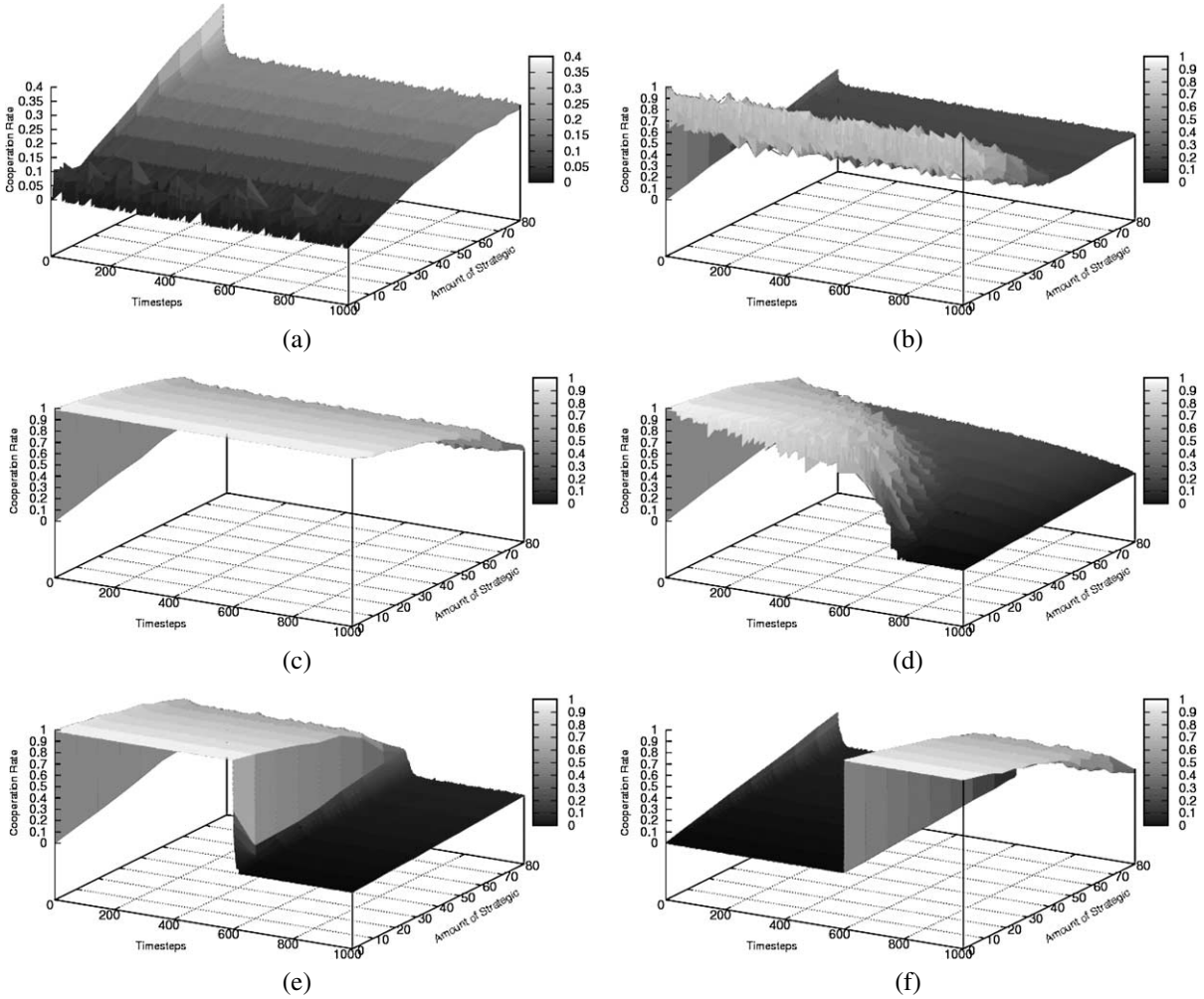

Fig. 5. Cooperation rate with different punishment probability distributions and with 10 internalizers. (a) Pr = 0.2, (b) Pr = 0.4, (c) Pr = 1, (d) linearly decreasing, (e) step down at 600, (f) step up at 600.

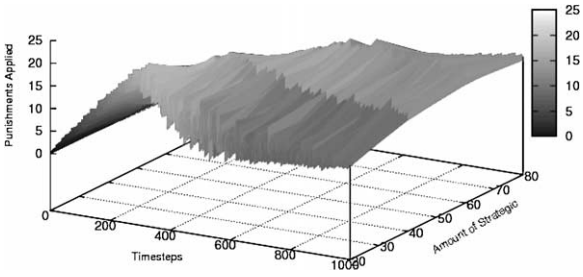

Fig. 6. Punishments applied in a linearly increasing probability distribution.

sanctions because they emancipate the norm addresses from external, sanctioning entities.

However, what should we expect from the comparison between internalized and fully endogenous mental states? Internalized goals are here hypothesized to be more persistent and lead to a more vigorous goal-attainment [8] than originally inner goals. The argument is based on prospect theory [36] (for a recent contribution, see [1]), which assumes loss aversion, i.e., people tendency to strongly prefer avoiding losses to acquiring gains, as a prominent feature of human beings. Internalized goals are already formed in the mind: unlike fully endogenous goals, internalized ones are selected among goals initially acquired under the effect of external influence. The more effort is invested in the attainment of these goals, the lesser they will be abandoned later, the more vigorously will they be attained.

A specific hypothesis is based on the *confirmation bias* (according to which people are likely to accept

inputs that confirm their beliefs and to reject disconfirming ones [41]; for a recent contribution, see [48]). Based on it, the internalizer is expected to show higher intolerance with regard to norm violation than both those who follow the norm under external enforcement and those who are spontaneously motivated to behave accordingly. Violation is a disturbing factor for the internalizer, which might lead it to weaken and even revise the commitment made. Hence, norm internalizers are expected to be more consistent and compliant than externally-enforced norm observers and endogenously motivated agents. A further consequence of the theory is that agents are much better at defending the internalized norms than externally-enforced observers. A consequence of the latter prediction is that norm internalization is decisive, if not indispensable, for distributed social control. Internalization is probably not only one mechanism of private compliance, but also a factor of social enforcement.

In short, internalization is a good predictor of compliance and second order cooperation (i.e., urging others to comply with the norms [35]). But what are, if any, the disadvantages of internalization? Again, the theory leads to formulate some hypotheses. First, internalization takes long and it is not necessarily successful: self-training may be too hard and it often requires several trials before getting through. People almost never quit bad habits and antisocial conducts on the very first try. Second, failures may have counter-effects: after a number of unsuccessful attempts, loss of self-esteem and feelings of helplessness may render too weak and fragile future private commitments, and jeopardize internalization. The question of course is what are the factors that may favor its success on the first try. Third, internalization emphasizes selection of inputs, autonomy. Moral autonomy is often encouraged at least in western societies, but it may have counter-effects as well. For example, it may lead to excessive variance in compliance. Fourth, how does internalization perform in societies characterized by a high degree of norms, possibly in sharp conflict with one another? One might expect that internalization is incompatible with perceived norm and value conflicts. Could it be that the future of societies is characterized by fragile and variable internalization? Another question for investigation.

9. Conclusions and future works

When Vygotsky first formulated his theory of internalization, he noted that only "the barest outline of this process is known" [53, p. 57]. We do not know, yet, how people manage to internalize beliefs and precepts with a reasonably adequate success, partly because we still do not agree about what to investigate or what we mean by this notion. Consequently, no useful notion and model is available for applications, despite its wide and profound implications. Questions such as how norm internalization unfolds, which factors elicit it, which are its effects, obstacles and counter-indications, are issues of concern for all of the behavioral sciences. The internalization of social inputs is indispensable for the study and management of a broad spectrum of phenomena, from the development of a robust moral autonomy to the investigation and enforcement of distributed social control; from the solution to the puzzle of cooperation, to fostering security and fighting criminality, etc. After a cognitive definition of the subject matter, the paper presents and discusses the building blocks of a rich cognitive model of internalization as a multi-step process, including several types and degrees of internalization. Next, factors favoring different types of internalization are discussed. The modular character of BDI architectures, like EMIL-A is shown to fit the approach advocated in the paper.

In this work we have also implemented and presented the results of the internalization module that has been incorporated into the existing platform, creating the new EMIL-I-A (EMIL Internalizer Agent). This new implementation allowed us to perform experiments and observe the behavior of internalizers in societies with different types of agents. The results obtained from the experiments allowed us to observe how EMIL-I-A indeed goes through all the phases of internalization when a norm is salient and returns to its *normative* behavior when the norm is no more salient. In order to achieve this result, a certain amount of strategic agents is needed within the society.

What is the value added of a rich cognitive model of internalization, as compared to simpler ones (e.g., reinforcement learning)? There are several competitive advantages. First, reinforcement learning does not account for the main intuition shared by different authors, i.e., the idea that internalization makes compliance independent of external enforcement. Second, a rich cognitive model, namely a BDI architecture with its specific modules, accounts for different types and degrees of internalization, bridging the gap between self-enforcement and automatic responses. Third, a BDI architecture accounting for different levels of internalization allows flexibility to be combined with automatism, as well as thoughtless conformity with auton-

omy. A BDI system can host automatism, but a simpler agent architecture does not allow for flexible, innovative and autonomous action.

In our agenda for future work there are several points to cover. As an immediate work to perform, and after observing the satisfactory results of our proof-of-concept simulation, we plan to apply our EMIL-I-A architecture to a more realistic scenario. We believe that the abstract simulations performed in this work have a significant meaning as they represent a widespread problem (Prisoner's Dilemma). However, bringing our theories to a more realistic scenario would allow the reader to better understand the advantages provided by the internalization process, as well as observe their important utility within the society.

The application of our architecture to a more realistic scenario will also help us to address the second problem we deal with. What happens when an agent has two or more independent norms conflicting with one another? Bringing back the example posed along the paper, what would an agent do when facing a red traffic light but listening to an ambulance siren? Should it wait or move on? Salience is a decisive element for us to solve these kind of problems.

Acknowledgements

This work was supported by the Spanish Education and Science Ministry [Engineering Self-*Virtually-Embedded Systems (EVE) project, TIN2009-14702-C02-01]; Proyecto Intramural de Frontera MacNorms [PIFCOO-08-00017] and the Generalitat de Catalunya [2009-SGR-1434]; the European Science Foundation EUROCORES Programme TECT, funded by the Italian National Research Council (CNR); the EC Sixth Framework Programme and European project COST Action IC0801 Agreement Technologies. Daniel Villatoro is supported by a CSIC predoctoral fellowship under JAE program. We also thank the CESGA and Rede Galega de Bioinformatica for the technical support.

References

[1] M. Abdellaoui, H. Bleichrodt and C. Paraschiv, Loss aversion under prospect theory: A parameter-free measurement, *Management Science* **53**(10) (2007), 1659–1674.

[2] G. Andrighetto, M. Campennì, R. Conte and M. Paolucci, On the immergence of norms: a normative agent architecture, in: *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence*, Washington, DC, 2007.

[3] G. Andrighetto, D. Villatoro, R. Conte and J. Sabater-Mir, Simulating the relation effects of punishment and sanction in the achievement of cooperation, in: *Proceedings of the Eight European Workshop on Multi-Agent Systems*, Paris, 2010.

[4] J.M. Aronfreed, *Conduct and Conscience; The Socialization of Internalized Control Over Behavior*, Academic Press, New York, 1968.

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

[6] A. Bandura, *Social Learning Theory*, Prentice Hall, New York, 1976.

[7] A. Bandura, Social cognitive theory of moral thought and action, in: *Handbook of Moral Behavior and Development*, Vol. 1, W.M. Kurtines and J.L. Gewirtz, eds, Lawrence Erlbauml, Hillsdale, NJ, 1991, pp. 45–103.

[8] J.A. Bargh, P.M. Gollwitzer, A. Lee-Chai, K. Barndollar and R. Troetschel, The automated will: unconscious activation and pursuit of behavioral goals, *Journal of Personality and Social Psychology* **81** (2001), 1004–1027.

[9] J.A. Bargh, P.M. Gollwitzer, A. Lee-Chai, K. Barndollar and R. Trotschel, The automated will: Nonconscious activation and pursuit of behavioral goals, *Journal of Personality and Social Psychology* **81**(6) (2001), 1014–1027.

[10] K. Basu, Social norms and the law, in: *The New Palgrave Dictionary of Economics and Law*, P. Newman, ed., 1998, available at SSRN: http://ssrn.com/abstract=42840.

[11] C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge Univ. Press, New York, 2006.

[12] C. Bicchieri and A. Chavez, Behaving as expected: Public information and fairness norms, *Journal of Behavioral Decision Making* **23**(2) (2010), 161–178.

[13] C. Bicchieri and X. Erte, Do the right thing: But only if others do so, MPRA Paper 4609, University Library of Munich, Germany, 2007.

[14] W.K. Bickel and M.W. Johnson, *Time and Decision*, Chapter Delay discounting: a fundamental behavioral process of drug dependence, Russell Sage Foundation, New York, 2003.

[15] M. Campennì, G. Andrighetto, F. Cecconi and R. Conte, Normal = normative? The role of intelligent agents in norm innovation, *Mind and Society* **8** (2009), 153–172.

[16] R. Cialdini and N. Goldstein, Social influence: Compliance and conformity, *Annual Review of Psychology* **55** (2004), 591–621.

[17] R.B. Cialdini, R.R. Reno and C.A. Kallgren, A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places, *Journal of Personality and Social Psychology* **58**(6) (1990), 1015–1026.

[18] P.R. Cohen and H.J. Levesque, Intention is choice with commitment, *Artificial Intelligence* **42**(2,3) (1990), 213–261.

[19] A.M. Collins and M.R. Quillian, Retrieval time from semantic memory, *Journal of Verbal Learning and Verbal Behavior* **8** (1969), 240–247.

[20] R. Conte, Rational, goal-oriented agents, in: *Encyclopedia of Complexity and Systems Science*, R.A. Meyers, ed., Springer, 2009, pp. 7533–7548.

[21] R. Conte and C. Castelfranchi, *Cognitive and Social Action*, University College of London Press, London, 1995.

[22] R. Conte and C. Castelfranchi, From conventions to prescriptions. towards a unified theory of norms, *AI and Law* **7** (1999), 323–340.

[23] R. Conte and C. Castelfranchi, The mental path of norms, *Ratio Juris* **19**(4) (2006), 501–517.

[24] R. Conte and F. Dignum, From social monitoring to normative influence, *Journal of Artificial Societies and Social Simulation* **4**(2) (2001).

[25] E.L. Deci and R.M. Ryan, The "what" and "why" of goal pursuits: Human needs and the self-determination of behaviour, *Psychological Inquiry* **11**(4) (2000), 227–268.

[26] A. Dreber, D. Rand, D. Fudenberg and M. Nowak, Winners don't punish, *Nature* **452** (2008), 348–351.

[27] N. Epley and T. Gilovich, Just going along: Nonconscious priming and conformity to social pressure, *Journal of Experimental Social Psychology* **35** (1999), 578–589.

[28] J. Epstein, *Generative Social Science. Studies in Agent-Based Computational Modeling*, Princeton Univ. Press, Princeton–New York, 2006.

[29] H. Gintis, The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms, *Journal of Theoretical Biology* **220**(4) (2003), 407–418.

[30] H. Gintis, The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions, *Journal of Economic Behavior and Organization* **53** (2004), 57–67.

[31] C.E. Goodall, Modifying smoking behavior through public service announcements and cigarette package warning labels: A comparison of Canada and the United States, PhD thesis, Ohio State University, OH, 2005.

[32] J.E. Grusec and L. Kuczynski, *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*, Wiley, New York, 1997.

[33] M.D. Harvey and M.E. Enzle, A cognitive model of social norms for understanding the transgression-helping effect, *Journal of Personality and Social Psychology* **41** (1981), 866–875.

[34] C. Horne, The internal enforcement of norms, *European Sociological Review* **19**(4) (2003), 335–343.

[35] C. Horne, Explaining norm enforcement, *Rationality and Society* **19**(2) (2007), 139–170.

[36] D. Kahneman and A. Tversky, Prospect theory: an analysis of decision under risk, *Econometrica* **47** (1979), 263–291.

[37] H. Kelsen, *General Theory of Norms*, Oxford Univ. Press, USA, 1979.

[38] L. Kohlberg, *The Psychology of Moral Development: The Nature and Validity of Moral Stages*, 1st edn, Harper & Row, San Francisco, 1984.

[39] M. Mead, *Cultural Patterns and Technical Change*, The New American Library, New York, 1963.

[40] M. Neumann, Norm internalisation in human and artificial intelligence, *Journal of Artificial Societies and Social Simulation* **13**(1) (2010), 12.

[41] R.S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, *Review of General Psychology* **2** (1998), 175–220.

[42] N. Nikiforakis and H.-T. Normann, A comparative statics analysis of punishment in public-good experiments, *Experimental Economics* **11**(4) (2008), 358–369.

[43] T. Parsons, *The Structure of Social Action. A Study in Social Theory with Special Reference to a Group of Recent European Writers*, Free Press, New York, London, 1937.

[44] J. Piaget, *Moral Judgement of the Child*, Free Press, New York, 1965.

[45] H. Rachlin, *The Science of Self-Control*, Harvard Univ. Press, Cambridge, London, 2000.

[46] J. Scott, *Internalization of Norms: A Sociological Theory of Moral Commitment*, Prentice-Hall, Englewoods Cliffs, NJ, 1971.

[47] S. Sen and S. Airiau, Emergence of norms through social learning, in: *Proceedings of IJCAI-07*, 2007, pp. 1507–1512.

[48] R.J. Sternberg, *Critical Thinking in Psychology*, Chapter Critical thinking in psychology: It really is critical, Cambridge Univ. Press, Cambridge, 2007.

[49] Y. Sun and B. Wu, Agent hybrid architecture and its decision processes, in: *International Conference on Machine Learning and Cybernetics*, 2006, pp. 641–644.

[50] E. Ullman-Margalit, *The Emergence of Norms*, Clarendon, Oxford, 1977.

[51] D. Villatoro, S. Sen and J. Sabater, Topology and memory effect on convention emergence, in: *Proceedings of the International Conference of Intelligent Agent Technology*, IEEE Press, 2009.

[52] G.H. von Wright, *Norm and Action. A Logical Inquiry*, Routledge & Kegan Paul, London, 1963.

[53] L.S. Vygotskii and M. Cole, *Mind in Society: The Development of Higher Psychological Processes*, L.S. Vygotsky, M. Cole et al., eds, Harvard Univ. Press, Cambridge, 1978.

[54] E. Xiao and D. Hauser, Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange, *Journal of Economic Psychology* **30**(3) (2009), 393–404.

[55] E. Xiao and D. Houser, Emotion expression in human punishment behavior, *Proc. Natl. Acad. Sci. USA* **102**(20) (2005), 7398–7401.

[56] H. Zhang and S.Y. Huang, Dynamic control of intention priorities of human-like agents, in: *Proceeding of the 2006 Conference on ECAI 2006*, Amsterdam, The Netherlands, 2006, IOS Press, The Netherlands, pp. 310–314.