On the Immergence of Norms: a Normative Agent Architecture

Giulia Andrighetto[1], Marco Campennì[1] [2], Rosaria Conte[1], Mario Paolucci[1]

1 LABSS - Istituto di Scienze e Tecnologie della Cognizione - CNR, via S. Martino della Battaglia 44, 00185 Rome, Italy
http://labss.istc.cnr.it
2 University of Modena and Reggio Emilia, Italy
giulia.andrighetto@istc.cnr.it

Abstract

The paper will describe a micro-social property, namely a module of an agent architecture, which is the immergent effect of norm-based regulation, and will illustrate how it allows for the recognition and conformity to existing norms, but also in at least one type of norm-innovation.

Introduction

Dealing with autonomous social agents, emergence is in the loop between bottom-up and top-down processes. Emergence of properties at aggregate level cannot be effectively accomplished unless properties feedback on the lower level through a complementary process of immergence into behaviours of units at the lower level. In complex social systems, where units at the lower level include intelligent agents, the process of immergence involves agents' minds, before and in order to become visible in their behaviours.

The aim of this paper is to provide an analysis of emergence and immergence processes, and shed light on how they help account for norm innovation.

In a view of norms as two-sided, external (social) and internal (mental) objects (Conte and Castelfranchi, 1995; Conte, 1998; Conte and Castelfranchi, 1999, etc.), norms emerge *as such* only when they emerge, not only *through* the minds of the agents involved, but also *into their minds*. In other words, they work as norms only when the agents *recognise* them, reason and take decisions upon them as norms. The emergence of norms implies their *immergence* in the agents' minds. Only when the normative, i.e. prescriptive, character of a command or other action is recognized by the agent, a norm gives rise to a normative behaviour of that agent. Immergence is a necessary correlate of emergence of at least a subset of macro-social phenomena, such as norms. In the social sciences, norms are usually conceived either as conventions – i.e., arbitrary solutions to problems of coordination that allow for multiple equivalent equilibria (Lewis, 1969) - or as solutions to problems of cooperation (Ullman-Margalit, 1977). In either case, both the exogenous and mandatory character of norms (see, Elster, 1983) are somewhat misperceived. Indeed, the conventionalistic view did progressively impose itself among social scientists, gradually replacing the latter. Furthermore, unlike the early treatment of conventions as proposed by Lewis, which heavily relied upon crucial cognitive notions, such as mutual expectations, the current view tends to underestimate the role of cognition in the emergence of conventions. A naïve conceptualization of "thoughtless conformity" (see for one example, Epstein, 2007) is spreading among social scientists, which cannot account for the complex mental dynamics leading agents to taking often implicit and not necessarily reflected upon decisions as to whether and when comply with a given convention. In my country, it will never occur to me to winder whether I should wear a chador before getting out in the morning. But if I am a visitor in a Muslim country, I'll probably spend a few minutes ruminating upon the opportunity of comply with such a convention, once detected it and the circumstances under which others observe it. More convincingly, I *automatically* stop at the traffic light turning red, except when I realize a policeman in the middle of a crossroad indicating me to drive on. What does thoughtless, automatic conformity really mean? Probably, it means that a number of decisions are de-activated under default conditions, and re-activated when current circumstances deviate from default. Far from a simplification, however, to account for these shortcuts requires further sophistication in the cognitive complexity of the agent than what our architecture is presently allowed.

Of course, our view of norms calls for a cognitive architecture of normative agents, which is not new to the field of agents and multiagent systems (think of the BOID architecture, for example). But preceding BDI approach to normative reasoning present some problems that the present approach aims to address, first of all norm-innovation. In previous work, norms are pre-established and built into the agents, which are enabled to reason and decide upon them. Instead, we endeavour to have agents finding out new norms and transmit them to one another. Whether they do effectively meet with a new norm or simply believe this to be the case, agents *de facto* make norms by finding them out in real matters, complying with them, and contaminating others with the same conjectures. Another still insufficiently explored (see Broersen et al., 2001) aspect of norms is the mental objects or mechanisms that allow them to affect the behaviours of autonomous intelligent agents, or to state it other wise, that implement them. Norms not only regulate behaviour but also act on different aspects of the mind.

In this paper we will provide an analysis of the so called *inter agents* and *intra agent* processes needed to deal with norm emergence. On the one hand, inter agents processes contribute to characterize the transmission of the norm; on the other hand, intra agent properties and processes define its immergence. As to the inter-agents processes, special attention will be paid to the mechanisms of emergence and diffusion of entities or properties at the aggregate level, from interaction among agents: this, aimed to point out how this allows a norm to be innovated.
As to the intra-agent processes, attention will be drawn on the normative architecture, EMIL-A necessary for its achievement.

Norm-Innovation: a Special Case of Immergence

Before illustrating EMIL-A, let us propose an operational taxonomy of norm-innovation. Norms are a highly adaptive artefact, emerging, evolving, and decaying. If it is relatively clear how legal norms are put into existence and then abrogated, it is much less obvious how the same process may concern social norms. How do new social norms and conventions come into existence, and how are they abandoned? Lewis's (1969) theory of conventions does not account for the formation of shared reciprocal expectations of conformity.

Indeed, some simulation studies about the selection of conventions have appeared of late, for example Epstein and colleagues' study of the emergence of social norms (2007), and Sen and Airiau's study of the emergence of a precedence rule in the traffic (2007). However, such studies investigate which one is chosen from a set of alternative equilibria.

A rather different sort of question concerns the innovation of social norms when no alternative equilibria are available for selection. We envisage at least three possible types of norm-innovation: 1. Norm-adaptation and extension: 2. Norm-instantiation 3. Norm-integration.

In the present paper, we will address only the case of norm instantiation, illustrating how our normative architecture EMIL-A allows a new norm to be perceived and established as an instance of an existing norm.

The Intra-agent Processes: EMIL-A

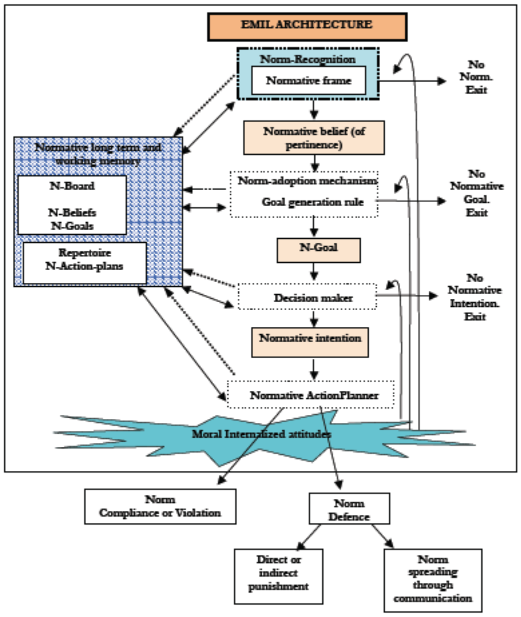
EMIL ARCHITECTURE
Norm-Recognition
Normative frame
No Norm. Exit
Normative long term and working memory
N-Board
N-Beliefs
N-Goals
Repertoire N-Action-plans
Normative belief (of pertinence)
Norm-adoption mechanism
Goal generation rule
No Normative Goal. Exit
N-Goal
Decision maker
No Normative Intention. Exit
Normative intention
Normative ActionPlanner
Moral Internalized attitudes
Norm Compliance or Violation
Norm Defence
Direct or indirect punishment
Norm spreading through communication

Figure 1: The main components of EMIL-A. It consists in four different procedures, indicated by the dotted boxes, three mental objects and a long term and working memory, indicated by the continuous boxes and a Moral Internalized Attitudes, indicated by the indented box. The dotted arrows indicate the act of asking, searching or activating and the two-directed continuous arrows the act of depositing.

Figure 1 illustrates the components of EMIL-A, consisting in:

Four different procedures:

- Norm Recognition, containing the Normative Frame;
- Norm Adoption, containing the Goal Generation Rule;
- Decision Making;
- Normative Action Planning.

Three different metal objects:

- Normative Beliefs;
- Normative Goals;
- Normative Intentions.

An inventory:

- A Normative Long Term Memory, containing a Normative Board and a Repertoire of Normative Action Plans.

Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A Moral Internalized Attitudes Module that, retroacting on the other procedures, can activate or deactivate them or directly impact on goals.

The outputs of EMIL are two different kinds of normative actions, compliance/violation and/or norm-defense.

In the rest of the paper, the main relevant features of EMIL-A will be introduced and discussed. In particular we will pay attention to:

- the functioning and role of the norm recognition module;
- the *positive default* bringing to the adoption of a normative goal;
- the motivational aspects of the normative mental representations;
- and the anticipatory and predictive nature of the whole system.

Then we will try to clarify how agents above defined can be involved in inter-agent processes; we will focus on the norm instantiation as a case of norm-innovation process.

Related Work

In evolutionary psychology, there are many efforts to define normative agents. Each definition focuses on one specific aspect of the problem. To make one good example, Sripada and Stich's model mechanisms (2006) of norm acquisition are proposed but no description how they work is given. Analogously, norm compliance is taken for granted without explaining how it works and to what extent it is compatible with agents' autonomy.

In AI, Broersen et al (2001) presents the so-called Belief-Obligations-Intentions-Desires or BOID architecture as a feedback loops mechanism, which considers all effects of actions before committing to them, and resolves conflicts between the outputs of its four components: each type of agent corresponds to a specific type of conflict resolution embedded in the BOID architecture. In all of them it is not contemplated that an agent can (or cannot) recognize an action as normative; *only if* the agent recognizes it as such she can decide whether to comply with it.

An other crucial feature rarely take into account is the anticipatory and predictive nature of normative agents. Anticipation (i.e. making decisions based on predictions, expectations, or beliefs about the future) is a vital component of autonomous cognitive agents living in social systems (Miceli and Castelfranchi, 2002). Agents incorporating reactive planning tend to be autonomous systems proactively pursuing at least one, and often many, goals. Anticipation enhanced performance of agents to face with complex social environments where they have to guide their attention to collect important (social) information to (inter-) act (Pezzulo, 2007).

EMIL-A architecture, beeing a part of "general" cognitive architecture, should be provided with this special capability.

Counterfactuals

The different parts of EMIL-A are necessary to deal with some tasks and theoretical questions, crucial to the normative domain:

- Without a norm-recognition module, we could not *rule out pure coercion,* or in other words discriminate between a norm and a mere coercion. Norms are more than mere commands of private agents obliging us to do or not to do something. The binding force is insufficient, since it also characterizes non-normative commands. The norm recognizer, endowed with the normative frame and able to have access to the base of knowledge stored in the normative board, shall attempt to give an answer to this crucial question.
- Without norm-adoption, *how account for agents' autonomy*? Normative agents ought to be granted reasoning and autonomy. Exposed to normative requests, they must decide whether to adopt it or not.
- Without normative decision making, *how account for norm violation, and conflict resolution*? A normative goal is not sufficient for agents to comply with norms. Several factors occurring within the process leading from normative goals to normative actions may cause agents to abandon the goal and violate the norm. One of such factors is conflicting (normative) goals. The decision maker helps decide whether and which (normative) goals to pursue on the grounds of their urgency and of the existing beliefs.
- Without normative action of defense, *how account for social control*? This is decisive for spreading the norms through a population of autonomous agents. Norm defense is based on a normative equity principle, which wants agents to sustain normative costs no higher than those sustained by other subjects to the same norm, benefits being equal.

As we will try to point out in the second part of this paper, the whole normative architecture, together with the inter agent processes, are necessary to deal with norm innovation.

Normative Inputs

EMIL-A receives and is activated by *internal* and *external* normative inputs.

External normative inputs are deontic commands, prescribing that something is permitted, forbidden or obligatory, communicated either by the legislator or from other members of the community. In our terms (Andrighetto et al. 2007; Andrighetto, Conte, Turrini, 2007), a norm is a prescription, or command, characterized by the use of deontics, which are reasons and bases for prescriptions. Deontics empower the command, substituting and rendering the exercise of personal power superfluous. This new kind of power may be exercised not only by institutional authorities, which are formally empowered, but also by private citizens with regard to one another. In other words a norm is a deontic command, whose power over is inherent to the deontic itself [1].

As to *internal* inputs, EMIL-A is activated:

- by the general architecture of the agent: sometimes the activation of goals in the agent's mind requires a sudden comparison with the existing N-Goals of the agent herself: this to test potential conflicts.
- by internalized attitudes, including social emotions and moral disposition, working as internal drives to norm compliance.

Mental Path of The Norm

To have an idea of how EMIL-A works, a sketch of an "ideal" and complete mental path of a norm will be provided. Probably, the standard path is rarely followed in its completeness and it is more plausible to consider that *shortcuts* take place during the computation of the normative input. For example, when you are driving and you see a red traffic light you will automatically stop. In this situation, it is not necessary that input follows the complete mental path.

After recognition, a norm becomes a *belief* in the mind of the agent stating that "there is a norm prohibiting, prescribing, permitting…". It is an Observer's N-Belief (Conte and Castelfranchi, 1995). It may also be stored as *a beliefs of pertinence* stating that a norm exists and concerns us.

Norms work through social goal-adoption, i.e. the fact that x believes that y wants p, is a reason for x to have (adopt) the goal that p, since and until y has it as a goal. Thanks to the norm adoption mechanism, the normative belief of pertinence activates a *(preexistent) goal* that, by a goal generation rule [2], may generate a new, normative goal. If no such a goal is generated, the norm is violated.

An agent endowed with this particular kind of goal is allowed to compare it with any other goal (norm-decision maker) of her and, to some extent, to choose which one will be transformed in N-Intentions, i.e. in executable goals.

The N-Goal can be transformed into a normative intention (and than into an action, i.e. a performed goal):

- of compliance;
- and/or defence: either of direct or indirect punishment or of norm spreading through communication;

or eventually be abandoned, solution that brings again to norm violation.

Norm Recognizer

The norm recognition module gives access to the EMIL-A architecture. Before an input is recognized as normative, the norm cannot immerge in the minds of agents and, as a consequence, cannot emerge in society. Agents need to be able to discriminate between norms and other social phenomena, such as coercion, ordinary requests, conventions, etc. Norm recognition explores the normative frame and the normative board, often resorting to the anticipatory and predictive capacity of the whole system (see the simulator described below).

Our claim is that existing normative architectures so far did not render justice to the recognition procedure. On the contrary, this module is fundamental for norm innovation, as we will argue in the second part of this paper.

The Normative Board. When EMIL-A has to deal with an external input, such as a NO SMOKING sign, the norm recognition module will explore the N-Board. Suppose a corresponding normative belief is found (DO NOT SMOKE WHEN PROHIBITED), a belief of pertinence is fired that will follow the path described previously.

The normative board contains normative beliefs and normative goals, organized and arranged according to *salience* that these normative objects have gained. With *salience* we refer to the norm's degree of activation, which is a function of its consistency with (shared) moral dispositions: in a particular situation a norm is more consistent than others, so its salience is higher. There are two types of salience:

- objective salience is valid for all agents in the same context;
- subjective salience originates from past experience and from own history.

The difference in salience between N-Beliefs and N-Goals has the effect that some of these normative mental objects will be more active than others and they will interfere more frequently and with more strength with the general cognitive processes of the agent.

The Normative Frame. If it is the case that the normative external input is an unknown norm the normative frame will be activated. It is a dynamic schema that allows us to recognize and categorize an external input as normative. It contains the properties defining a norm (Andrighetto, Conte, Turrini, 2007):

[1] Although necessary for the spreading of the prescribed behaviour, the normative command is insufficient: additional factors consist of the mandatory force (obligatoriness and enforcement) of the command; the persuasiveness and credibility of the source; compatibility with existing norms (norm conflicts often lead to violating one or the other); etc.

[2] The **goal-generation rule** (Conte and Castelfranchi, 1995):
If x has goal p and x believes that q is one of p's antecedents, then x will also have goal q.

Deontic: a way of partitioning situations between good/acceptable ones and bad/inacceptable ones. We distinguish deontics into:

- obligations;
- forbearances;
- permissions.

Source: the locus from which the norm emanates. We distinguish the source into:

- personal;
- impersonal.

Role: the partition of the agents involved in a norm. We distinguish:

- Legislators, the personal source;
- Addressees, those agents that are mentioned by the norm as allowed or not allowed to carry out a given action;
- Defenders, that is those agents that share and enforce the norm;
- Observers, those that acquires beliefs about a norm, that is whether it is enforced, violated, emanated.

Enforcement mechanism: operations that attempt to modify agents' actions in order to make them compliant to a norm. We distinguish:

- sanctions: enforcement mechanisms that inhibit agents' actions;
- incentives: enforcement mechanisms that favour agents' actions.

Control: the way enforcement mechanisms are applied. It implies both monitoring - that checks violation - and influence - that actively pushes cognitive agents' towards compliance. They can be:

- centralized: only one agent (individual or supraindividual) is entitled to sanction;
- distributed: everybody is able to defend the norm.

However, not all the slots need be filled in for a norm to be recognised. What is needed is that agents recognise

- the prescriptive and impersonal character of the N (Conte, and Castelfranchi, 2006);
- the entitled, legitimate/valid, in force and impersonal nature of the authority;
- the application of the norm *erga omnes*;
- the legitimate reactions or sanctions to transgressions.

Agents don't need to understand nor agree about the specific function of a norm. They must respect it because it is a norm (or, sub-ideally, because of surveillance and sanctions), but in any case, they need first to recognize it is a norm.

The properties defining a norm are variables that can assume values within a defined range.

The normative frame works as a simulator, which helps interpret or forecast the nature of the inputs agents come across with. If, for instance, one of the properties of the input matches with one of the variables listed above, a normative interpretation is put forward. Successive experience will either confirm or dismantle this hypothesis.

In the meantime, the other slots will be left in standby until the hypothesis is either verified or rejected. If further checks are confirmatory, the agent will form a new normative belief. The Normative Frame, thanks to our modelling, allows to recognize a norm even if detecting only few (salient) normative properties; thanks to them the other properties will be deduced by default.

To simulate the results of a hypothesis, the normative frame must be interfaced with the agent's knowledge of the world. Simulations helps the agent to recognize unknown norms, disambiguate opaque interpretations and forecast the effects of different decisions.

Rather than conjecturing an innate and universal disposition to norm compliance, a hypothesis that evolutionary psychologists would probably subscribe too (but see also Horne 2007), we propose that normative recognition be seen as an inbuilt property of intelligent social agents, and an immergent effect of social regulation.

Unlike moral dispositions, it is poorly sensible to subjective variability, and rather robust. It allows us to (a) account for the universal appearance of norms in human and primate societies; (b) render justice to the intuition that humans may violate norms, but have little problems in telling norms; (c) account for the evolutionary psychological evidence (see Cosmides and Tooby, 1992) showing that agents easily apply counterfactual reasoning to find out social rules, but find it difficult to do so with logical ones; finally, (d) explain why, as pointed out by developmental psychology data, norm acquisition follows a stable ontogenetic pattern starting quite early in childhood (Nucci, 2001; Cummins, 1996; Piaget, 1965; Kohlberg, and Turiel, 1971).

In short, the intuition behind our normative architecture is twofold: on one hand, dealing with norms is based upon a universal capacity to tell norms, on the other this capacity is supported by a norm frame, an internal "model of a norm" that agents use as a frame of reference. As we will stress later, norms also have a *motivational effect*. This claim is again supported by evolutionary psychologists (see for example, Cosmides and Tooby, 1992), who refer to this type of motivation as *intrinsic* and granted by an innate normative module. We would like to object to this view that the existence of a norm module is either too strong or insufficient: it is too strong because it leaves no room to autonomy and norm violation. It is insufficient because little is said about how it effectively works: what are norms? How are they learned? What is their internal processing, the path they follow in the mind? The motivational nature of the norm can be understood only if we explode such complex mental representations in their components, N-Beliefs, N-Goals and N-Intentions, and pay attention to the mental path they follow and to the procedures and rules that assure their elaboration. The emphasis laid on the innate and universal features of EMIL-A should not be mistaken, leading to think that no space is left to subjective variability. If norm recognition is a must, equally accomplished by a vast majority of agents, moral attitudes - i.e. the results of normative and moral experience accumulated during lifetime that affect different normative procedures - are not. They are definitely subjective.

Furthermore, the reinforcement effects that occur on different EMIL-A procedures vary among agents.
Personal experience, for example, impacts norm salience. Analogously, the normative frame, being in constant interaction with the social environment and the other procedures, is liable to their influence. In these terms, a normative architecture is allowed to elegantly ignore the culture/nurture controversy.

Towards Norm Adoption

Imputed by normative requests, the agent will generate a normative goal thanks to the norm adoption procedure. This does not imply, by the way, that the request will certainly be complied with. Our claim is rather that, whenever an input gives rise to a normative belief of pertinence, a new process starts: the agent wonders *why not* to adopt it. In short, we believe agents have a weak disposition, a positive default, to take normative requests into account and adopt them unless there are good reasons *not* to do so. Unlike ordinary adoption, in which agents must have positive reasons for adopting others' requests, in norm adoption a baseline reason is provided, as pointed out before, by the deontic itself: either one does not recognise it under a command, or one believes there is a good reason, however feeble, for accepting the command. The normative goal may be extremely weak, and its value may be re-evaluated later on, when the decision making procedure will compare the effects of violation with the costs of compliance. Suppose that at night, while approaching a crossroad, I see the traffic light turning red. It is late in the night and neither cars nor pedestrians are visible. It is also most unlikely that any policeman is observing me. In this situation, together with the costs of violation, also the value of the normative goal decreases. We define *cogency* the criterion for choosing whether to execute a (normative) goal or not, and compares the costs of executing it with the effects of dropping it; while *salience* is the ranking order of norms, and is related to other norms. Getting back to our example, we can have two different agents:

– agent 1 is a recently qualified driver; probably for her the norm *stop-if-traffic_light-is-red* is urgent because she feels uncertain in driving;
– agent 2 is an expert driver; she feels self-confident at driving and she finds the norm-compliance less cogent.

In sum, after a normative request a normative goal gets formed in the mind of the addressee. The agent takes it into account, and tends to accept it unless there are good reasons for not doing so. This way of reasoning is inherent to the normative dimension, and diverges from a general modality of reasoning and social reasoning.

Inter-agents Processes: a Scenario of Norm Innovation

Let us better define the norm-innovation taxonomy above presented. Based on a gradient of novelty, we identified three main categories of norm innovation,

- *norm extension or adaptation*: an existent norm is extended to new entities or social category, in such a way that its content is modified;
- *norm instantiation*: a new norm is perceived and established as an instance of an existing norm;
- *norm integration*: a norm is determined by the integration of conflicting norms.

To better understand this process, it is necessary to examine at least one type of norm innovation, namely norm instantiation.

Usually, there are at least two agents involved in an episode of norm innovation: an agent source executing a given (normative) action and an agent observing it.

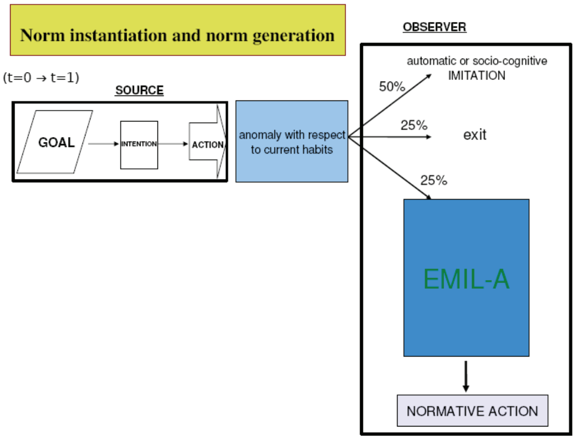
Norm instantiation and norm generation
(t=0 → t=1)
SOURCE
GOAL
INTENTION
ACTION
anomaly with respect to current habits
OBSERVER
automatic or socio-cognitive IMITATION
50%
25%
exit
25%
EMIL-A
NORMATIVE ACTION

Figure 2

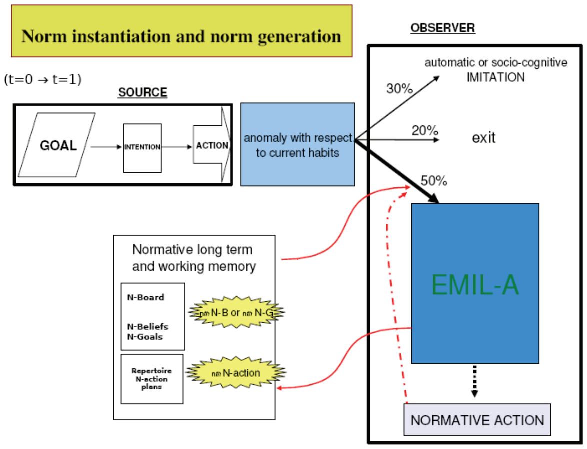
Norm instantiation and norm generation
(t=0 → t=1)
SOURCE
GOAL
INTENTION
ACTION
anomaly with respect to current habits
OBSERVER
automatic or socio-cognitive IMITATION
30%
20%
exit
50%
EMIL-A
NORMATIVE ACTION
Normative long term and working memory
N-Board
N-Beliefs
N-Goals
nth N-B or nth N-G
Repertoire N-action plans
nth N-action

Figure 3.

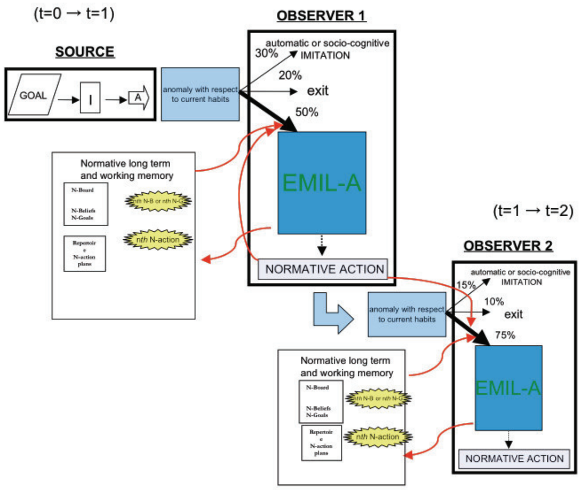
(t=0 → t=1)
SOURCE
GOAL
I
A
anomaly with respect to current habits
OBSERVER 1
automatic or socio-cognitive IMITATION
30%
20%
exit
50%
EMIL-A
NORMATIVE ACTION
Normative long term and working memory
N-Board
N-Beliefs
N-Goals
nth N-B or nth N-G
Repertoire N-action plans
nth N-action
(t=1 → t=2)
OBSERVER 2
automatic or socio-cognitive IMITATION
15%
10%
exit
75%
EMIL-A
NORMATIVE ACTION
anomaly with respect to current habits
Normative long term and working memory
N-Board
N-Beliefs
N-Goals
nth N-B or nth N-G
Repertoire N-action plans
nth N-action

Figure 4.

Figures 2-4: Arrows between source and observer stand for possible choices; the thicker the arrow, the more likely the corresponding choice. Continuous boxes represent agents; arrows pointing to boxes stand for causal processes. Curved continuous arrows represent reinforcing effects: when EMIL-A agent consults its normative board, it may receive an answer that reinforces the normative choice. Dotted curved arrows represent the same process at time t+1.

Let us analyse the situation step by step, over time.

- At time zero (t = 0) Source has a Goal that produces an Intention, and finally an Action.
- At time t + 1 (i.e. t = 1), Observer perceives that action. If Observer finds an anomaly with respect to current habits (anomaly plays a decisive role because it elicits interpretation), there are two possible scenarios:
  - *equiprobability*: in principle in absence of elements that can help us to decide, we have the same probability *p* for each option (*p* = 1 / *number of choices*): Observer can imitate Source thoughtlessly (25%); or else she may choose Source because she perceives her as successful (intelligent social learning, 25%); Observer may decide to ignore the anomaly (25%) and behave as usual; finally, Observer may decide to consider the action as normative (25%).
  - *salience:* if there is a high state of activation of a given n*th* N-*Belief* and N-*Goal* in the normative board and if the anomaly observed is consistent with it, the probability of interpreting the action observed as normative becomes higher than any other option; Observer will execute a n*th* N-*action* in the repertoire of normative plans.
- At time t + 2, in any case, once a N-action has been taken path-dependence reinforces the normative "path": if Observer finds a "similar" situation, it will be induced to choose the same path. In this case, Observer reacts to the anomaly as in the previous case, but the probability of each reaction is not static, predetermined; there is a dynamic process: if Observer chooses to interpret the action as a normative one, this choice is more reinforced than others, its probability increases each time and as a consequence that of others decreases.

Each time, the normative choice gains more and more weight; the synergy of the normative board and the repertoire of normative action plans impact on the probability of normative interpretation. The Observer's normative action reinforces her choice and that of other observers (if any). Hence, this mechanism is crucial for norm innovation.

What happens if there is more than one observer? Presumably, the self-reinforcing choice mechanism works both within the agent (for each observer) and between them: the level of activation in the normative board reinforces the interpretation of each observer and her normative action reinforces both her normative choice and that of the next one. This dynamic process involves N-*Beliefs* and N-G*oals* in the normative board and N-*actions* in the repertoire of N-action plans: each time a n*th* normative belief is active (i.e. it is salient), the normative interpretation is reinforced. Moreover, the consequent normative action reinforces the normative interpretation.

In the case of more-than-two observers, we find an avalanche effect of normative interpretation reinforcement: each observer reinforces not only her own normative interpretation in two steps (normative board and normative action), but also that of the next one with her normative action. Norm innovation is an inherently inter-agent process: only observers can innovate norms, since they need to perceive implicit commands, or adopted commands in a Source's behaviour. The new action may be even produced accidentally; nonetheless, if it fits salient norms, it may easily be interpreted as an instance of it. Once this interpretation has been done, the job is done: the higher the number of observers at subsequent times, the more likely and fast a new norm will establish.

Conclusions and Future Works

In this paper we have proposed a Normative Architecture of the Agent, EMIL-A, which is the immergent effect of norm-based regulation, and we have illustrated how it allows for the recognition and conformity to existing norms, but also in at least one type of norm-innovation.

As pointed out, this point of view implies a dynamical, anticipatory and interactive attitude with respect to the social environment. Anyway, the model sketched in this paper is still in progress: further investigation and a full implementation are needed. In particular, we are aware of the risk of an excess of consistency in EMIL-A, derived by the reinforcement effects inside the model, that could be avoided only by introducing some sort of variation.

We are also aware that:

- a deeper integration between EMIL-A components
- and the modelling of a certain number of shortcuts taking place during the computation of the normative input

would guarantee the architecture more plausibility.

The achievement of these tasks is a starting point for future works.

Acknowledgments

This work was supported by the EMIL project (IST-033841), funded by the Future and Emerging Technologies programme of the European Commission, in the framework of the initiative "Simulating Emergent Properties in Complex Systems". The EMIL-A normative architecture owes its denomination to the above mentioned EMIL Project, since it is developed within its frame.
We would like to thank EMIL partners and in particular Rainer Hegselmann and Klaus G. Troitzsch for helpful comments on previous versions of this paper.
We would also thank Federica Mattei for the linguistic revision of the paper.

References

Andrighetto, G.; Conte, R.; Turrini, P.; and Paolucci, M. 2007. Emergence In the Loop: Simulating the two way dynamics of norm innovation. In *Proceedings 07122 of the Dagstuhl Seminar on Normative Multi-agent Systems,* Dagstuhl, Germany.

Andrighetto, G.; Conte, R.; and Turrini, P. 2007. EMIL Ontology, Technical Report, 00307, LABSS-ISTC/CNR.

Barkow, J.; Cosmides, L.; and Tooby, J. 1992. *The Adapted Mind: Evolutionary psychology and the generation of culture.* NY: Oxford University Press.

Broersen, J; Dastani, M.; Hulstijn, J.; Huang, Z.; and van der Torre, L. 2001. The BOID Architecture. Conflicts Between Beliefs, Obligations, Intentions and Desires, In Proceedings of the fifth international conference on Autonomous agents, Montreal, Quebec, Canada. 9 – 16.

Conte, R., and Castelfranchi, C. 1995. *Cognitive and social action*, London: London University College of London Press.

Conte, R. 1998. *L'obbedienza intelligente*. Bari: Laterza.

Conte, R., and Castelfranchi, C. 1999. From conventions to prescriptions. Towards a unified
theory of norms. *AI&Law* 7: 323-340.

Conte, R., and Castelfranchi, C. 2006. The Mental Path of Norms. *Ratio Juris* 19 (4): 501 – 517.

Cummins, D. D. 1996. Evidence for deontic reasoning in 3- and 4-year olds. *Memory and Cognition* 24(6): 823-829.

Epstein, J. M. 2006. *Generative Social Science. Studies in Agent-Based Computational Modeling*. Princeton-New York: Princeton University Press.

Horne, C. 2007. Explaining Norm Enforcement. *Rationality and Society* 19(3): forthcoming.

Kohlberg, L., and Turiel, E. 1971. Moral development and moral education. In G. Lesser, ed. Psychology and educational practice. Scott Foresman.

Lewis, D. K. 1969. *Convention: A Philosophical Study.* Cambridge Mass.: Harvard University Press.

Miceli, M., and Castelfranchi, C. 2002. The mind and the future: The (negative) power of expectations. *Theory & Psychology*, 12, 335-366.

Nucci, L. P. 2001. *Education in the Moral Domain*. Cambridge University Press.

Pezzulo, G. 2007. Anticipation and Future-Oriented Capabilities in Natural and Artificial Cognition. In *Proceedings of the 50th Anniversary of Artificial Intelligence*, 67-70. Springer LNAI.

Piaget, J. 1965. *The moral judgment of the child.* The Free Press: New York.

Sen, S., and Airiau, S. 2007. Emergence of norms through social learning. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence.* Forthcoming.

Sripada, C., and Stich, S. 2006. A Framework for the Psychology of Norms. In P. Carruthers, S. Laurence and S. Stich, eds., *The Innate Mind: Culture and Cognition*, 280-301, Oxford University Press.

Ullman Margalit, E. 1977. *The Emergence of Norms.* Oxford: Clarendon Press.