SIXTH FRAMEWORK PROGRAMME

Project no. 033841

EMIL

Emergence in the loop: simulating the two way dynamics of norm innovation

Deliverable 1.2 EMIL-M

REPORT OF MODELS OF NORMS EMERGENCE, NORMS IMMERGENCE AND THE 2-WAY DYNAMIC

Due date of deliverable: 31.08.07 +45gg
Actual submission date: 05.10.07

Start date of project: 01.09.06 Duration: 36 months

Organisation name of lead contractor for this deliverable: CNR-ISTC

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

EMIL-M*: REPORT OF MODELS OF NORMS EMERGENCE, NORMS IMMERGENCE AND THE 2-WAY DYNAMICS

WORK PACKAGE 1.2 DELIVERABLE

1 * Parts of this deliverable have been published in **Andrighetto, G., Campennì, M, Conte, R., Paolucci, M.** 2007. On the Immergence of Norms: a Normative Agent Architecture. . In , *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington DC.* and **Conte, R., Andrighetto, G., Campennì, M, Paolucci, M.** 2007. Emergent and Immergent Effects in Complex Social Systems. In , *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington DC.*

TABLE OF CONTENTS

1 Introduction

Dealing with autonomous social **agents** (see Annex 3), **emergence** (see Annex 1) is in the loop between bottom-up and top-down processes: emergence of properties at the aggregate level cannot be effectively accomplished unless these feedback on the lower level through a complementary process of **immergence** (see Annex 3) into the mind and the behaviours of units (agents) at the lower level.

The objective of this work is to provide an analysis of emergence and immergence, and shed light on how they help account for **norm innovation** (see Annex 2).

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.), a **norm** (see Annex 1) emerges *as a norm* only when it is incorporated *into the **minds*** (see Annex 3) of the agents involved. In other words, it works as a norm only when the agents *recognise* it as such. In this sense, norm emergence implies its *immergence* into the agents' minds. When its normative, i.e. prescriptive, character is recognized by the agent, a norm gives rise to a full normative behaviour of that agent. With full normative behaviour we intend a norm-based one, i.e. the output of a decision whether to comply or not with the norm.

One still insufficiently explored (see Broersen et al. 2001) aspect of norms is the mental representations or mechanisms that allow them to affect the behaviours of autonomous intelligent agents or, to state it otherwise, that implement them. Norms not only regulate behaviour but also act on different aspects of the mind.

We will provide an analysis of the so called *inter agent* and *intra agent* processes involved in norm emergence. On the one hand, inter agent processes allow for norm transmission; on the other hand, intra agent properties and processes lead to norm immergence.

The deliverable includes:

- a model of intra-agent processes (par. 4); attention will be drawn on a normative architecture, EMIL-A, necessary for their accomplishment;
- a description of inter-agent processes (par. 5); special attention will be given to the mechanisms of emergence and diffusion of entities or properties at the aggregate level from interaction among agents, in order to point out how this allows a norm to be innovated;
- an account of EMIL-A in use (par. 6); the occurrence of interruptions, modifications and

deviations from the standard in vitro description provided in par. 4-5;

- an algorithm (par. 8) of the main abovementioned procedures and processes;
- a list of the variables (par. 9) the model refers to;
- a possible implementation of the Norm Recognizer (par. 10)

Three Annexes are included:

- Annex I: Emil Ontology, a common vocabulary of normative notions to work with, discussed and accepted by the EMIL partners.
- Annex II: Taxonomy of norm innovation, a document providing a first attempt to characterize and exemplify at least some fundamental types of norm innovation.
- Annex III: Glossary, containing a description of the technical notions entering the one or other aspect of the model presented.

2 *Norm-Innovation:* A Special Case of Immergence

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 spontaneous social norms. How do new social norms and conventions come into existence, and how are they abandoned? Lewis's (1969) theory of **conventions** (see Annex 1) does not account for the formation of shared reciprocal expectations of conformity. Of late, simulation studies about the selection of conventions have appeared, for example Epstein and colleagues' study of the emergence of social norms (2000, 2007) and Sen and Airiau's study of the emergence of a precedence rule in the traffic (2007) (for a review see the State of the Art. D 1.1). 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 (see Annex 2).

We will address only the case of norm instantiation: as in this type of norm innovation, the role of the norm recognizer (see par. 4.4) appears more crucial, illustrating how our normative architecture EMIL-A allows a new norm to be perceived and established as an instance of an existing norm.

3 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, in 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 (see also the State of the Art, D 1.1).

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. On the contrary, our claim is that only if the agent recognizes a norm she can decide whether to comply with it or not.

A further crucial feature rarely taken 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). Anticipation enhances the capacity of agents to face with complex social environments where they have to guide their attention to collect important (social) information for (inter-) action (Pezzulo 2007). As a module of a "general" cognitive architecture, EMIL-A should be provided with this special capacity.

As shown within the state of the art (cf. Del 1.1), we can find three main drawbacks in the simulation studies of norm-innovation. First, in most of them, norms are equalised to conventions, and the question of study is of a rather static sort, concerning the selection of one specific equilibrium. Instead, the present project points to an out-of-equilibrium theory of norm-innovation. Second, no attention is paid to immergence, and therefore to the role of mental mechanisms in norm-innovation. Instead, the role of the mind is crucial in our view of norm-innovation.

4 The Intra-agent Processes: EMIL-A

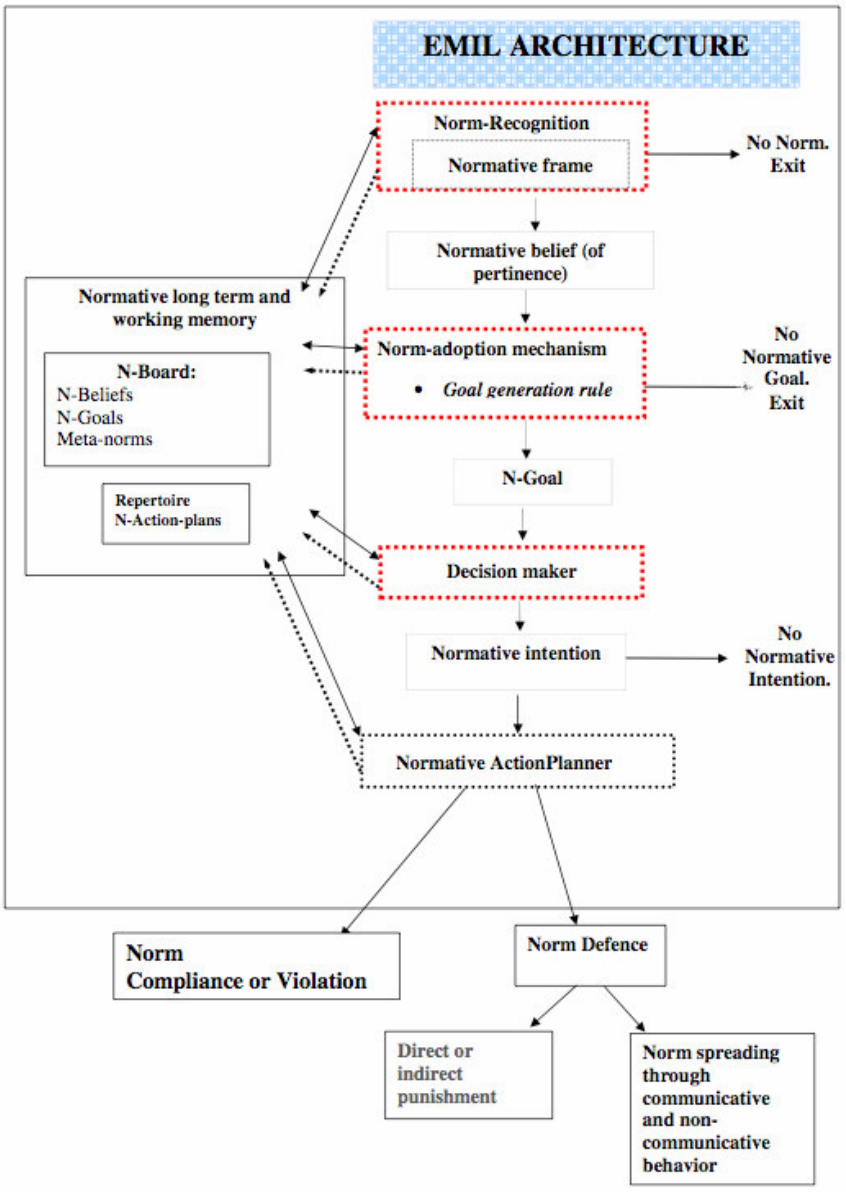


Figure 1: The main components of EMIL-A. It consists of four different procedures, indicated by the dotted boxes, four mental objects and a long term memory, indicated by the thick boxes. Dotted arrows indicate activation-search; bi-directional thick arrows stand for storing; one-directional thick arrow stands for information flow.

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

- four procedures:
  - **Norm Recognition**, containing the **Normative Frame**
  - **Norm Adoption**, containing the **Goal Generation Rule**
  - **Decision Making**
  - **Normative Action Planning**[1].
- four types of representations:
  - **Normative Beliefs** (N-Beliefs**)**
  - **Normative Goals** (N-Goals)
  - **Normative Intentions** (N-Intentions)
  - **Meta-norms** defined as general rules telling agents how to reason, decide upon and apply specific norms (see Annex 3).
- one inventory containing the **Normative Board** and the **Repertoire of Normative Plans**, together with knowledge about the world and general normative knowledge.

The outputs of EMIL are two different kinds of normative **actions** (see Annex 3), compliance/violation and norm-defence (see par. 4.7).

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

- norm recognition module
- norm adoption module.

Then we will try to clarify how normative agents can be involved in inter-agent processes, focusing on norm instantiation as a case of norm-innovation. Finally, some words will be spent on normative routines and other shortcuts, speeding up and simplifying the application of this architecture.

[1] Regarding this module, we will not provide a description in this deliverable since we will refer to the description included in D3.1.

4.1 Normative mental objects

A brief description of the main components of the mental processing of norms will follow.

4.1.1 Normative belief

A belief that a given behaviour, in a given context, for a given set of agents, is forbidden, obligatory, permitted, etc. More precisely, the belief should be that "there is a Norm prohibiting, prescribing, permitting...". It implies the belief that a legitimate authority y has the **normative influencing goal** (see Annex 1) that the set of agents form a normative belief and goal in the interest of some superset of agents. The authority is included in this superset and may coincide with it. Indeed, norms are aimed at and issued for becoming such beliefs. In other words, norms must be acknowledged as such in order to properly work; this is their function. Normative beliefs, together with normative goals, are organized and arranged in the normative board (par. 4.5.1) according to the salience (see par. 4.5.1 and Annex 3) gained.

4.1.2 Normative belief of pertinence

Believing that a norm exists and concerns us requires at least a second group of beliefs: the beliefs of pertinence. The norm says what ought to be done by whom: (i) the obligation/permission/prohibition and (ii) the set of agents on which the imperative is impinging. For example, if I am addressed by a given norm (say, "be member of a professional order"), and the norm has to take effect on me, I must recognize this. The prescription is about a set or class of agents, and since I am an instance of that class, the norm applies to me.

4.1.3 Normative Goal

An internal goal is relativized to a normative belief. From a cognitive point of view, a goal is a wanted state of the world that might or not be verified and can be considered as a subset of the reasons for action. A normative belief gives rise to a normative goal for the subject to act in accordance with the norm itself. Two are the cognitive mechanisms fundamental for a normative goal to be formed:

- the **goal-adoption mechanism,** 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;

- the **goal generation rule**, an agent will have as a goal any state that implies that another of its goals will be achieved. Thus, thanks to the norm adoption mechanism, the normative belief of pertinence activates a *goal* that may generate a new normative goal.

As said before, normative goals are organized and arranged in the normative board (par. 4.5.1) according to the salience gained (see par. 4.5.1 and Annex 3).

4.2 Why bother with EMIL-A?

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

- Without a norm-recognition module, *how 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. **Sanction** (see Annex 1) is insufficient, since it also characterizes non-normative commands. The norm recognizer, endowed with the normative frame and enabled to have access to the normative board, shall attempt to answer this crucial question.
- Without norm-adoption, *how account for agents' autonomy*? Normative agents ought to be granted both **reasoning** (see Annex 3) and autonomy. Exposed to normative requests, they must decide whether to adopt them 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 cogency and of the existing beliefs.
- Without normative action of defence, *how account for social control*? This is decisive for spreading the norms through a population of autonomous agents. Probably based on a **normative equity principle** (see Annex 1), norm defence allows 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 deliverable, the whole normative architecture, together with the inter agent processes, are necessary to deal with norm innovation.

4.3 Normative Inputs

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

External normative inputs are **deontic commands** (see Annex I), prescribing that something is permitted, forbidden or obligatory, communicated either by the **legislator** (see Annex I) or by other members of the community. In our terms (Andrighetto et al. 2007; Andrighetto, Conte, Turrini, 2007), a norm appears as 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** (see Annex 1), which are formally empowered, but also by private citizens with regard to one another. In other words a norm is a deontic command, the power of which is inherent to the deontic itself [2].

As to *internal* inputs, EMIL-A may be activated by the decision-maker: sometimes the pursuit of some goal requires that a given norm be recognised, possible N-Goals be formed and potential conflicts among these and the former goal be solved.

[2] 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; persuasiveness and credibility of the source; compatibility with existing norms (norm conflicts often lead to violating one or the other); etc.

4.4 Mental Path of The Norm

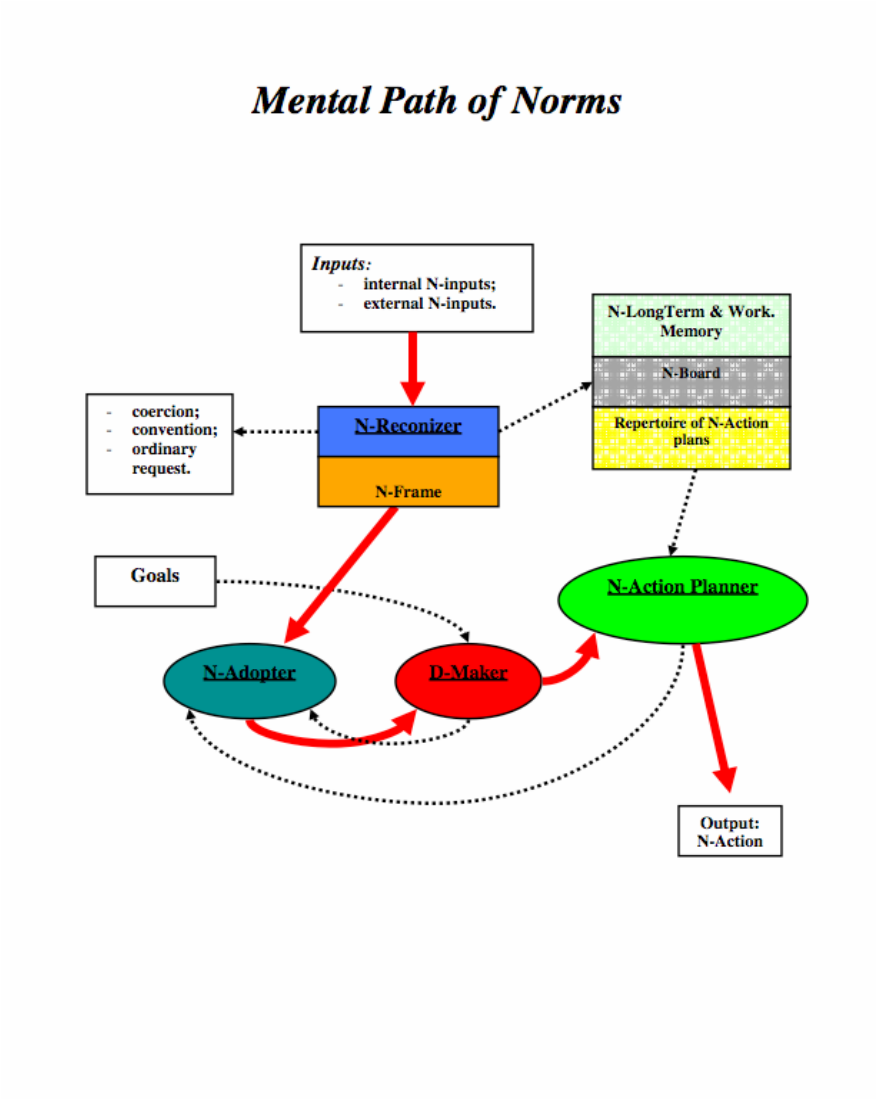


Figure 2: Tick red arrows represent the standard information flow. Dotted black arrows represent alternative directions of the information flow.

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 (see par.4.3)

After recognition, a norm becomes a *belief* in the mind of the agent stating that "there is a norm prohibiting, prescribing, permitting something". It is an **Observer**'s (see Annex I) N-Belief (Conte and Castelfranchi 1995; 2006). It may also be stored as a **normative belief of pertinence** (see Annex 1) stating that a norm exists and concerns us.

Norms work through social **goal-adoption** (see Annex 3): 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 *(pre-existent) goal* that may generate a new, normative goal thanks to a **goal generation rule** (see Annex 3). If no such a goal is generated, the norm will be violated.

An agent endowed with this particular kind of goal is allowed to compare it with any other goal (norm-decision maker) of hers and 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: direct or indirect punishment or norm spreading through communicative behaviours.

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

Nevertheless, it is possible that the information concerning the norm is processed in other ways. The norm recognizer can produce two kind of normative belief: (a) a new normative belief stored in the normative board; (b) a normative belief that becomes input for the norm adopter. The norm recognizer can produce also a belief about a coercion, a behavioural regularity or about an ordinary request (in this case the output is not processed by the rest of the architecture).

The norm adopter works on inputs arriving from the norm recognizer or the normative action planner or the decision maker to produce an input for the decision maker.

The decision maker receives inputs from the norm adopter or from the non-normative architecture of the agent (i.e. its goals) to produce inputs for the norm adopter and the normative action planner.

The normative action planner can receive information from the decision maker but also directly from the normative board (i.e. in case of shortcuts) to produce an external output (i.e. a normative action) or an internal input for the norm adopter.

4.5 Norm Recognizer

The norm recognition module is the main entrance, so to speak, 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 other normative architectures did not render justice to the recognition procedure. On the contrary, as we will argue in the second part of the deliverable, this module is fundamental for norm innovation.

Simplifying, we can say that we recognize the existence of a given norm, or that a certain norm is in action if

- the norm is already stored in our (normative) memory;
- the norm is inferred or induced by the agent on the grounds of given indicators.

In the first case, the agent is facilitated by schemata, scripts, or other pragmatic structures (Wason and Johnson-Laird, 1972; Schank and Abelson, 1977; Fiske and Taylor, 1991; Barsalou, 1999; see Markus and Zajonc, 1985 for an overview) the norm is embedded in (see Bicchieri 2006, for a description). Once these are activated for any reason, the corresponding normative beliefs, expectations and behavioural rules are prompted.

The second option is followed when such scripts, and consequently the corresponding pattern matching operations, are not possible. The agent has no corresponding norm. This is why the norm recognizer module is needed. Indeed, the norm-recognizer that we are going to describe tries to answer the question as to how it is possible to recognize a norm even when this is still not stored in the agent's memory[3]. EMIL-A endeavours to model this second aspect of norm recognition, particularly crucial to account for norm innovation, as we will try to highlight later on in this document.

4.5.1 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 is that part of the long term memory where active norms are stored. It contains normative beliefs and normative goals, organized and arranged according to the *salience* gained. With *salience* we refer to the norm's degree of activation, which is a function of its permanence in the N-Board and the degree to which it is believed to be shared by members of the community: in a particular situation, a norm is more shared 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 own history.

[3] See par. 10 for a possible implementation of the Norm Recognizer.

Salience increases also depending on how often the norm is adopted, thus reaching the Decision Maker. If, for example, the norm is never adopted by the agent, then its salience rate begins to decrease, and sooner or later the N-belief will decay (see par. 6.1.1). On the contrary, if the norm is frequently processed by the N-Decision Maker, and turned into a normative action, its salience rate will increase.

4.5.2 The Normative Frame

If it is the case that the normative external input is an unknown norm the normative frame will be activated. This is a dynamic schema guiding recognition. It contains five slots for the properties defining a norm (Andrighetto, Conte, Turrini, 2007):

- **Deontic**
- **Source**
- **Role**
- **Enforcement mechanism**
- **Control** (see Annex 1).

Agents do not 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 as a norm, i.e. a command based upon a deontic.

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

Slots that cannot be filled in immediately are left in standby until further evidence is obtained.

A norm is defined as such by the set of variables described above. This is a static description of the result of normative agent action. We know that at *run time* the normative agent must process information becoming from world's observation or interaction with other agents to recognize norms or to produce new normative beliefs. A brief description of the normative agent in action is given later in the document in § 9.

4.6 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, that of norm-adoption. In the present model, norms are adopted *unless* agent has good reasons *not* to do so. In short, we believe agents have a weak disposition, a positive default, to

take normative requests into account and adopt them forming a corresponding N-goal, with a value corresponding to norm salience.

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: if one recognises an input as normative, one believes there is a good reason, however feeble, for accepting it. The value of a N-goal may be recomputed later on, while taking decision about whether to comply with the norm in question or not.

4.7 Decision maker

This procedure is at work every time a goal is activated, except when shortcuts are in place (see par. 6.1.3). It controls whether there are obstacles to the goal's pursuit, otherwise it puts it to execution, either checking for existing plans or routines or planning them anew.

There are some factors that can reinforce or weaken goals. In particular, the value of the active goal may be temporarily strengthened by the entity and probability of sanctions, and weakened by potential incompatible goals. 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, the value of the normative goal decreases with **cogency**, i.e. the perceived costs of norm violation. 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; he feels self-confident at driving and he finds the norm-compliance less cogent.

Consider also that if some norms are perceived to be violated, the value of the corresponding normative goal will decrease and observers will proportionately be discouraged from observing it.

4.8 Norm defence

Once an agent has been compliant with a norm, it is likely to become a Defender of the norm. In many circumstances, the compliant agent influences others to comply with the norm, warning, for example, someone else from transgressing against a norm, or reminding her she is doing something wrong. It can be due to a "norm-sharing" mechanism, according to which agents are likely to

respect the norms they believe to be useful, and want others to comply with these norms. This also accounts for the intuition that people do not want others in their own conditions to sustain lower costs, benefits being equal. This is a special application of the equity rule to the normative domain (cf. Conte and Castelfranchi, 1995, ch. 7). In order for own costs to be not higher than those of others, the compliant agent wants other norm addressees either to comply with the norm that the agent himself has already observed, or to be punished. Punishment can be either:

- Direct: imposition of sanctions, or
- Indirect: e.g., spreading of bad reputation.

This normative conduct might be viewed as strongly responsible for the spreading of norms over a population of autonomous agents. The larger the number of agents conforming to one given norm, the more they will be likely to urge others to conform with the same norm. Analogously, but on the other hand, as soon as some norms are perceived to be violated, the value of the corresponding normative goal will decrease and observers will proportionately be discouraged from observing it. Therefore, on one hand the agent is driven to defend, directly or indirectly, the norm she has complied with, on the other hand she is likely to abandon her normative goal and violate a norm that has been already violated by other agents subject to it.

4.9 Value added of EMIL-A

So far, the main problem in the study of norm emergence has been why agents comply with norms. In the present project, we propose to address a problem that is logically and pragmatically precedent to this, and this is how agents tell that something is a norm. Only afterwards, it makes sense to wonder why they comply with it. Indeed, these two problems point to two different ways to conceptualize autonomy: autonomy in goals (or, self-interest) and autonomy in beliefs. Agents not only decide which goals to pursue, but also which beliefs to hold, including the normative ones.

Hence we propose that normative recognition be seen as an inbuilt property of intelligent social agents, and an immergent effect of social regulation. In our view, this represents an important requirement of norm-innovation, as the process by means of which a given norm gains ground in a population is both gradual and complex: it is the result of both agents' interpretations of one another's behaviours, and of their transmitting such interpretations to one another. Norm-recognition plays a crucial role into this process.

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 violate norms, but have little problems in telling them; (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 psychological 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 shall stress later on, norms have a *motivational effect*. This claim is supported by evolutionary psychologists (see for example, Cosmides and Tooby, 1992), who refer to this motivation as *intrinsic* and granted by an innate normative module[4]. 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 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 on 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.

[4] To this view, we would like to object 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?

5 Inter-agents Processes[5]: a Scenario of Norm Innovation

Based on a gradient of novelty, we identified three main categories of norm innovation (for a description see ANNEX 2),

- *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.

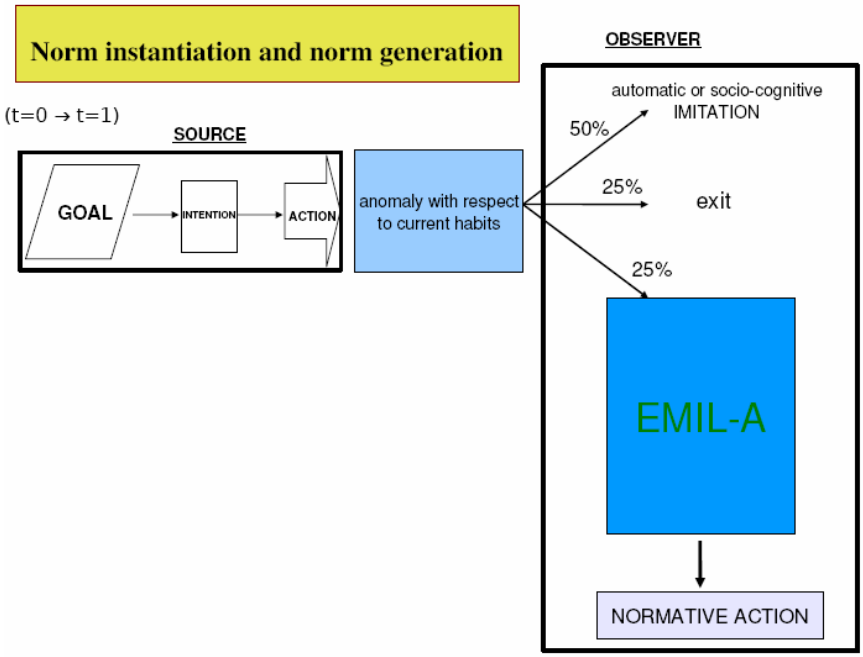


Figure 4

5 Is to be noticed that the link between intra and inter –agent processes is not well working yet. We will refine this connection as soon as the architecture will be really implemented, work now in progress (see par.8 and D3.1), but soon finalized.

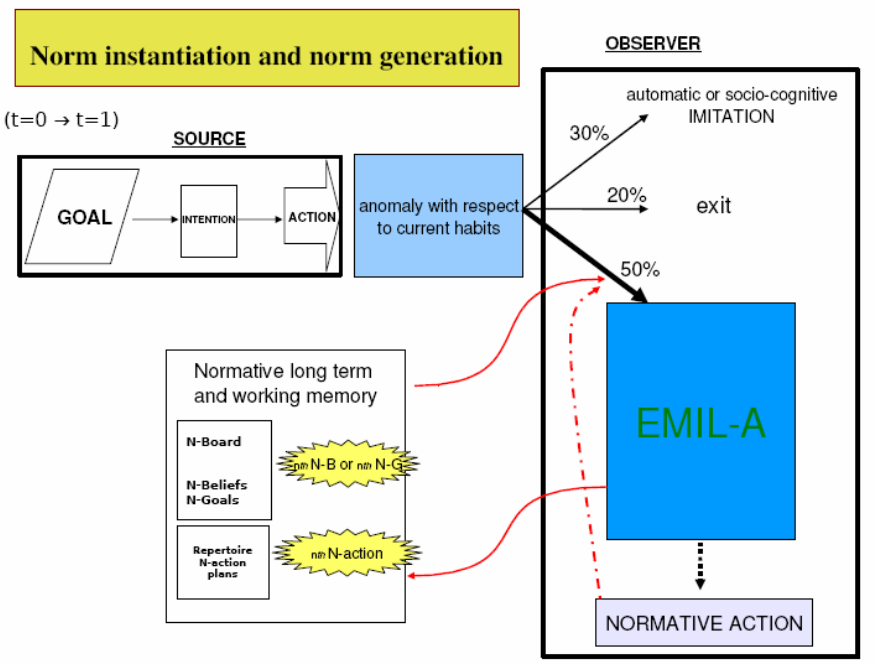


Figure 5

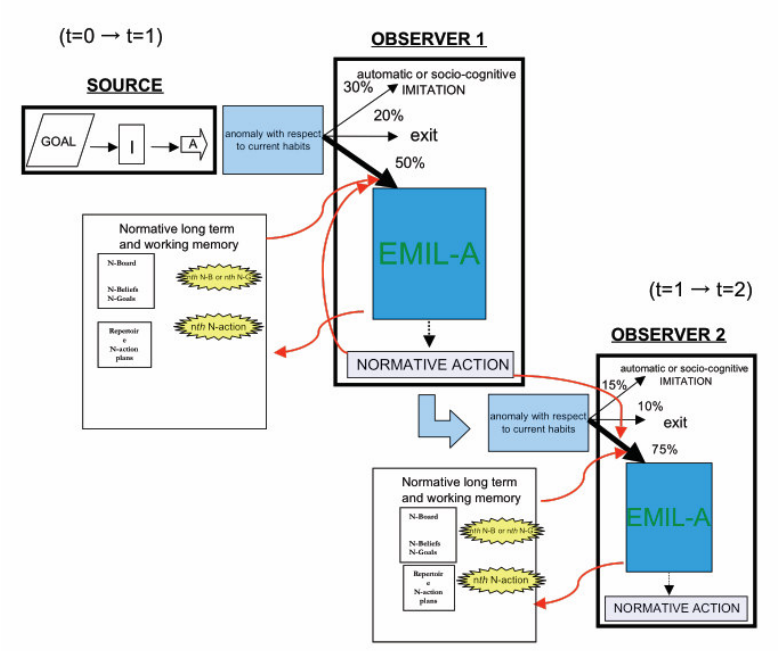


Figure 6

Figures 4-6: 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 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 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-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.

6 EMIL-A in use

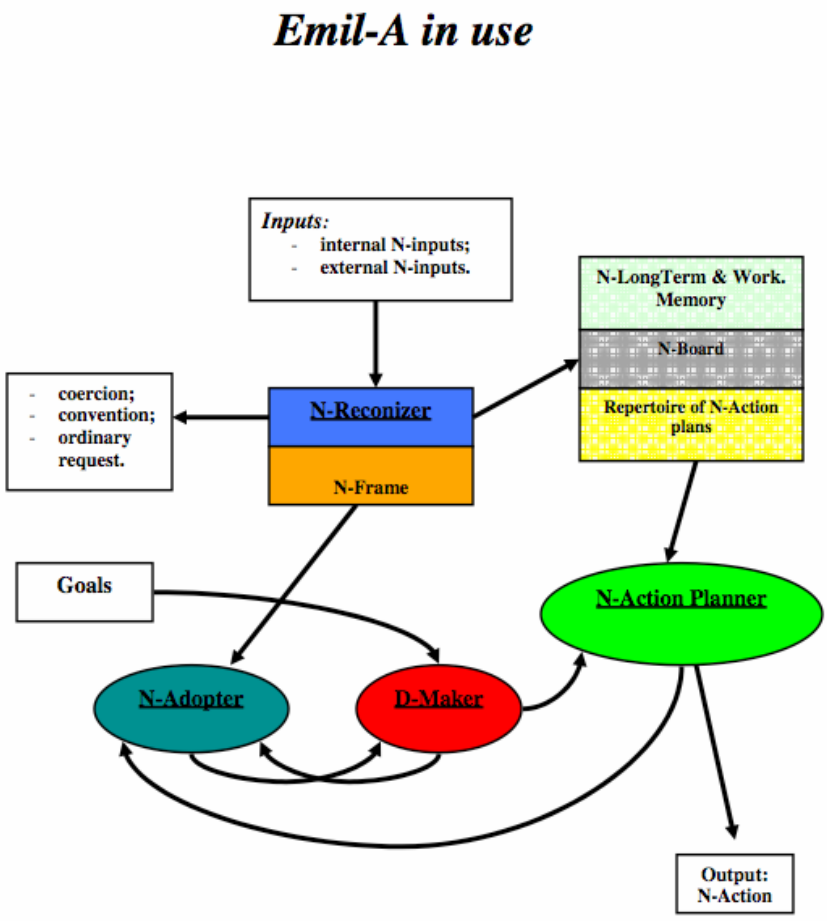


Figure 8: Tick black arrows represent all of the possible ways for flowing information

A crucial aspect of the agent architecture is the occurrence of interruptions, modifications and deviations from the processes described so far.

We will examine three specific phenomena:

- **decay**, i.e. the process by means of which a given norm decays leaving no trace in the agent,
- **internalization**, i.e. the process by means of which, once a goal has been generated by a N-belief, no track is maintained of its normative source,

- **shortcuts**, occurring when one or more decisions in the ordinary processing are bypassed; in particular, we will examine at some length reactive normative behaviours i.e. routines fired under given input conditions.

6.1.1 Decay

Are norms permanent objects in agents' minds? Certainly, not. As any other representation, norms may be acquired, modified and lost. Norms may disappear under the effect of cognitive and non-cognitive mechanisms.

In the former order of factors, we enumerate electro-chemical, physiological and traumatic phenomena: people affected by post-traumatic, post-surgical or chronic neurological disorders may exhibit behavioural anomalies and socio-pathologies consequent to a loss of social norms and conventions (Damasio, 1994; Anderson et al, 1999).

As to the multiple cognitive mechanisms responsible for the radical loss of norms, we would like to call the reader's attention at least to the following:

- norm-revision: reasons leading to a given input being recognised as a norm are found no more adequate. This may occur with
  - Perceived error in previous recognition: observer is led by the state of the world currently perceived (current behaviour of the source) to reconsider and revise previous interpretations.
  - Perceived change in the current state of the world: for example, observer perceives a modification in the source's behaviour.
  - Simple forgetting, generally associated to a gradually reduced salience of the norm, no longer fit to a changing environment.
- Norm-revocation: the reasons that led adopter to accept a norm are found no more adequate. Again, this includes
  - Revised (meta-)norms: decider finds the norm to be no more compatible with decision-based modifications of its normative board.
  - Modified personal goals: reasons for adopting the norm are insufficient as decider's goals have changed in the meantime. This may also be due to a non-cognitive cause, such as the normal course of life.

To implement this phenomenon requires more or less complex mechanisms. A simple solution is to include a forget rate, among the agent parameters, to be intended as the fact that if the salience rate of a norm remains very low for a certain period of time (to be specified), then the norm

decays from the agents' normative board mechanisms. However, more interesting solutions for norm revision and revocation exist. Indeed, norm-revision is already included in the norm-recognition process, as this is a non-linear process by means of which a given input (either a prescription or a regularly observed behaviour) is analysed.

The process of norm-revocation requires a dynamics of norms' salience, as a function of the frequency of norms' compliance and of the corresponding commands. It could also require a change in agents' personal goals: we would suggest that these may change only as to their motivational force, randomly varied over the population.

6.1.2 Internalization

This is the process by means of which a norm is transformed into an internal motivation. According to developmental psychologists and clinicians, this is often the case over the course of primary socialization, when agents (in this case, children) internalise their parents' commands, or those of other attachment figures.

Indeed, internalization is one of the most obscure aspects of the mind (see Meissner,1981; Wallis and Poulton, 2001), and we are not going to provide a conclusive treatment of this phenomenon here. Rather, we will present one way of interpreting and implementing it. Essentially, we will treat internalization as a twofold dynamics:

- gradual reinforcement of internal dispositions (i.e. increasing impact of dispositions on the motivational force of given ordinary goals) combined with
- gradually increased salience of the corresponding norm up to the point that it goes over the rank and it leaves the n-board, becoming a highly-valued goal.

This dynamics is triggered by the time of permanence of norms in the n-board: the earlier they are acquired, the more likely they are internalised.

6.1.3 Shortcuts. Normative routines

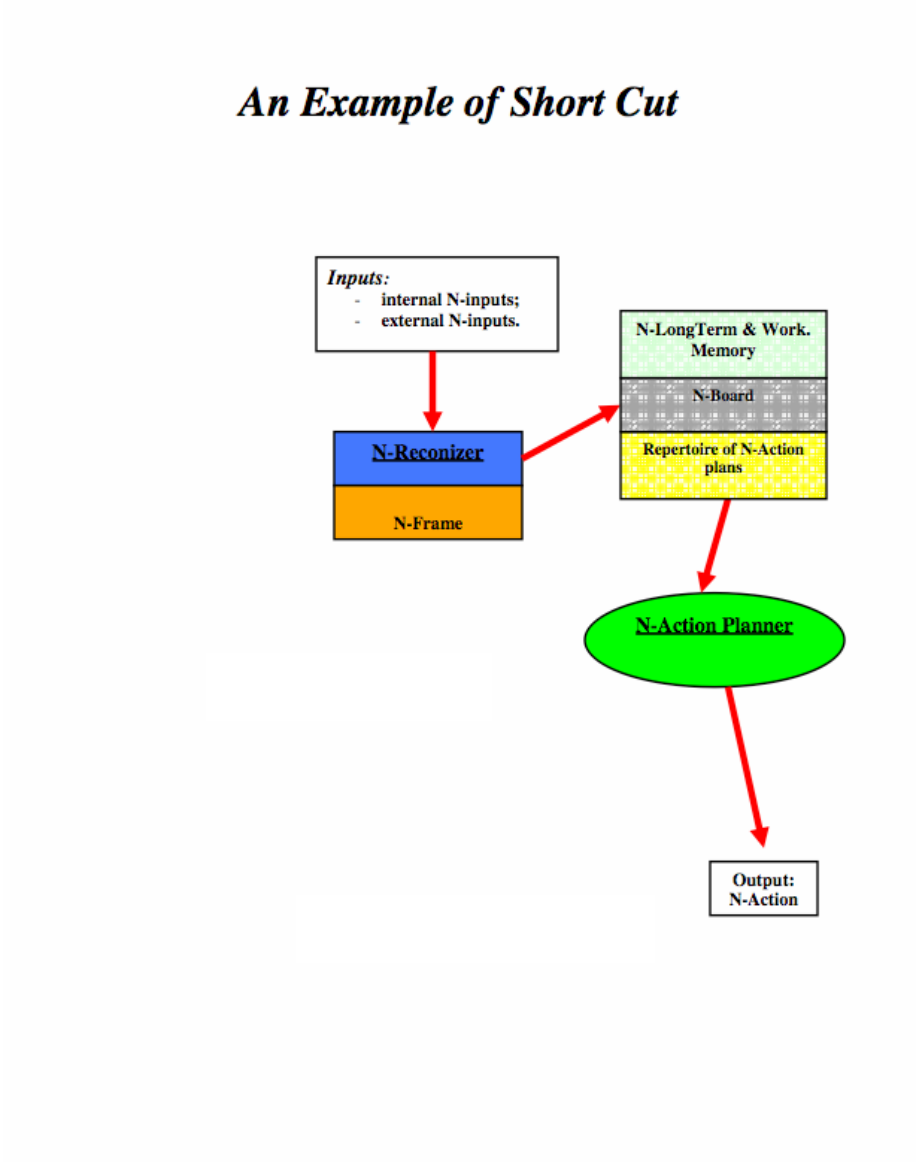


Figure 9: Tick red arrows represent standard information flow. Dotted black arrows represent alternative directions of information flow.

A phenomenon that shows similarities with the preceding one is the occurrence of shortcuts. However, in the preceding case, what is lost is the exogenous source of the norm, which is transformed into an ordinary goal.

Normative routines, on the contrary, are (semi)automatic executions of norms, fired whenever a given (subset of) input(s) is recognised. Automatisms are frequent in intelligent systems, including humans. What makes it possible is far from clear, whereas it is relatively easier to answer the question as to when it is likely to occur: the better and more frequent a given

behaviour, the more likely it is performed as an automatic, non-controlled (i.e. non-decided upon and inadvertent) routine (Bargh 1992). In the case of norms, the more frequent a certain normative action, the more likely it will become a routine: the normative action is performed thoughtlessly, on condition that some control mechanisms stay active. If for example normative conflicting inputs are detected, control mechanisms lock the shortcut, reactivating the EMIL-A standard path.

As to implementation, this can easily be implemented as a shortcut from certain conditioned obligations (e.g., traffic light turning red) to the corresponding action (stop), with no need for norm-adoption and decision-making. Of course, routines might interfere with other norms (as is the case when the traffic light is red but a policeman standing at the cross-road urges drivers to move on). How is it possible to solve conflicts in a non-controlled way? One answer to be explored might be that conflicting inputs render shortcuts inoperative.

7 Conclusions and Future Works

We have proposed a Normative Architecture of the Agent, EMIL-A, which is the immergent aspect of norm-based regulation, and we have illustrated how it allows for the recognition of, and conformity to, existing norms, and for the spreading of a new norm in at least one type of norm-innovation.

Anyway, the model sketched is still in progress: further investigation and a full implementation are needed. We are aware that:

- a deeper integration between EMIL-A components
- an analysis of moral and normative emotions (shame, remorse, or feeling of guilt) and dispositions and of the way both impact on different aspects of EMIL-A

would greatly enhance the scientific interest and plausibility of this architecture.

8 VARIABLES

Time:
- t

N-Board:
- N-Beliefs
- N-Goals

- Cogency:
  the weight of Goals to respect with context, indicates:
  - costs of violation
  - benefits of execution

- Salience:
  - objective: all norms have it in the same context (RANDOM_assigned);
  - subjective: produced by past experience & own history;

- AD_to_DM: how many times N is processed by N-Adopter and arrives at Decision Maker.

N-Frame (with its slots):
- Deontic:
  obligations
  forbearances
  permissions

- Source:
  personal
  impersonal

- Role:
  legislator
  addressees
  defenders
  observers

- Enforcement Mechanism:
  sanctions
  incentives

- Control:
  centralized
  decentralized

- Regularity Rate: adding strength of Norm Interpretation and confirms hypothesis of Norm.

- Compliance

- Sanction

N is a norm **if and only if**

- Communicative act or behaviour has recognized as normative if Observer find that:

deontic is the reason of command (-> Observer can have)

9 NORM RECOGNIZER TOWARD IMPLEMENTATION

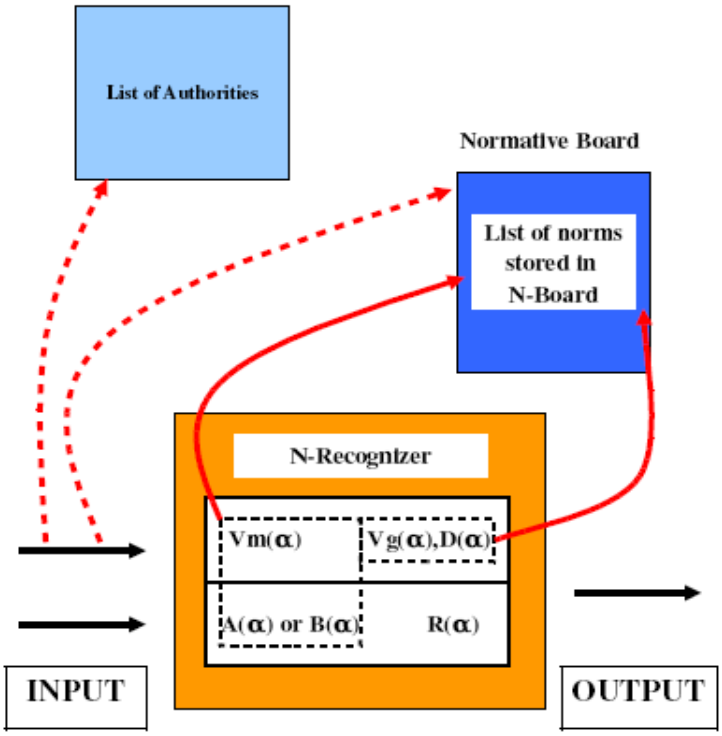


Figure 10: Black arrows indicate input and output flow; red arrows indicate the formation of a new normative belief, stored in the normative board (which can be a consequence of the third procedure, i.e. the n-frame analysis); dotted red arrows indicate the first procedure (search in normative board) and the second procedure (search in list of authorities).

Here we present a possible implementation of the norm recognizer module.

We assume that the norm recognizer receives input from other modules as a codified vector. Each input is a message that the norm recognizer receives from external (social) environment (i.e. another agent) or observing agents in interaction.

The norm recognizer may produce a normative belief as an output, which can be processed as an input by other modules (e.g. by the norm adopter or decision maker).

Legenda

- Beliefs and normative Beliefs (see Annex 3);
- Deontic (see Annex 1);
- Evaluation: there are two kinds of evaluations:
  - normative evaluations: sentences on what it's right or wrong, on what must or not must be

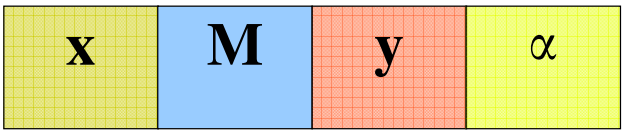
done (i.e. an example of moral evaluation is: *it is correct to respect the queue*) ;

- generic evaluations: the other evaluative sentences (i.e. an example of generic evaluation is: *your apartment is fine)*;

- Sanctions. We deal with normative sanctions as moral evaluations (negative): we assume that a sanction implies a (negative) evaluation of a norm violation (i.e. if *a* is a norm, *Vm*(□*a*) is a sanction because it's an evaluation in reply to an action transgressing against the norm) ;
- Assertions: we deal with assertions as generic sentences pointing to or describing states of the world;
- Behaviors: Behaviour refers to the actions or reactions of an agent, with regard to another agent or to the environment;
- Requests: requests of action made by another agent.

The Message

Each message is presented as an ordered vector consisting of four elements:



- the source (x);
- the modal through which the message is presented (M); there are six possible modals:
  - assertions (A)
  - behaviours (B)
  - requests (R)
  - deontics (D)
  - moral valuations (Vm)
  - generic valuations (Vg)
  - sanctions (S)
- the observer (y);
- the action transmitted (α).

Procedures

There are three procedures acting on module:

- search in the N-Board: it checks if the normative belief associated to the presented action already exists in the normative board; in this case, the message is not considered and the action is recognized as normative, otherwise it goes to next procedure;
- search in the list of authorities: it searches the authorities' list to check if the source of the message is a recognized authority (this is necessary but not sufficient because the authority must operate in a context of legitimacy); in this case (and if the action is presented as a deontic) the belief associated to the action becomes a normative belief and is inserted in the N-Board; otherwise it goes to next procedure;
- N- frame analysis: it analyzes slots characterizing the message and acts on its content (action).

Module in action

The module consists of two levels:

- Level_0: the access on this level is limited to actions presented by one of three modals A (assertions), B (behaviours), R (requests) only if the same action accessed in the past on Level_1 (i.e. the agent received in the past the same action as deontic (D), moral evaluation (Vm) or generic evaluation (Vg); this implies that the same action is present in the agent's architecture at level_1);
- Level_1: the access on this level is limited to actions presented by one of three modals D (deontics), Vm (moral valuations), Vg (generic valuations).

How a normative belief is formed

The normative agent interacts with the environment and with other agents. She receives information from both interactions: she can communicate with another agent in a lot of ways, but she can also observe the interaction of others and the related consequences.
The process is cumulative: it is necessary that the agent receives the same normative input in particular ways and repeatedly through the time; only in this case she can infer the existence of a certain norm from the received inputs.

As a consequence of this interaction (or observation), the normative agent can generate new normative beliefs in her mind only under certain conditions. There are many combinations of inputs that can generate the formation of a normative belief:

1) A message that contains a deontic (see Annex 1) is taken into account and stored in a higher level of her architecture (level_1); this will affect the agent's selective attention: further messages with the same content will be processed and stored at the same level, even if they are presented as (generic) evaluation (see Legenda above), whence the existence of a normative belief can be inferred.
2) The agent can perceive messages presented as moral evaluations (see Legenda above) while observing others' interaction. Even in this case, the agent becomes more attentive about new messages with the same content and even in this case she stores the relative belief in a high level of her architecture (level_1). She can infer the existence of a normative belief in two ways:

   i) when she receives other messages with the same content that are not presented as mere requests;

   ii) if she has already stored the same belief in her architecture (on level_0) as an assertion or behaviour (see Legenda above).
3) The agent receives messages about the same content (presented as different modals) from almost all of the other agents (only if at least one of the messages is presented as a deontic or a moral evaluation).

If the same actions appears in levels_0 as A or B and in level_1 as Vm, then the belief related to the action becomes normative belief and is inserted in the normative board (and deleted form levels 0 and 1); if the same action appears in level_1 as D and Vg, then the belief related to the action becomes normative belief and inserted in the normative board (and deleted from level 1). If an action presented as D comes from a recognized authority, then the belief related to the action becomes a normative belief and is directly inserted in the normative board; on the contrary, if the source is not a recognized authority, but the same action is later presented from a recognized authority, then the belief related to the action becomes a normative belief and is inserted in the normative board (and deleted from level 1).

Counter for the source

If an observer receives messages conveying the same action (at least one message must be a deontic or a moral valuation) from the most of the population (all – n, for $n < m\%(all)$), then the

belief related to the action becomes normative belief and is inserted in the normative board (and deleted from the levels in which it was).

Counter for the observer

If a source produces a message addressed to everyone (“erga omnes”, so the slot related to the observer is empty) as D or Vm, then the belief related to the action becomes normative and is inserted in the normative board.

Permanence in levels

The permanence of an action in a specific level is limited in time (while t < nTicks).

Local inconsistency

Inconsistencies are of two sorts:

1 the same modal refers both to one action and to its opposite,

2 the same action appears in the scope of two complementary modals.

Inconsistencies may occur within the same level. To solve such problem we use the counter for the sources: between action and not(action), the one presented by the major number of sources will be retained and proceed to higher level. We assume that all of the sources have the same credibility rate (i.e. it is not possible that an agent is more credible than another). This implies that all of the actions are dealt with in the same way.

10 Annex 1: EMIL ONTOLOGY[6]

Since norm innovation results from a complex collection of nested theoretical definitions, it is necessary to provide a shared ontology, or in other words, to forge a working vocabulary of interrelated notions. By ontology, we mean a conventional and operational tool, a set of theoretical notions that are defined one in relation to the others. Its goal is to make conceptual links explicit.

To have a first impression of the ontology, we will now have a brief overview of the main concepts.

Below, the reader is offered some guidelines for understanding the rationale of this ontology.

Rationale

The purpose of the present ontology obviously derives by the objective of the EMIL project, which is aimed at delivering a **simulation-based** theory of **norm-innovation**, where norm-innovation is defined as a **2-way dynamics**, inter-agent (emergence) and intra-agent (immergence).

Emergence is the process by means of which effects are generated by (inter)acting micro(-social) entities, and implemented upon, but *not* incorporated into, their rules (for a discussion see also Conte et al. 2007). With incorporation we consider the process by means the emergent effect gets represented in the producing system, and this representation contributes to replicate the effect.

Immergence, the process by means of which the emergent effect modifies the way of functioning of the generating system, affecting its generating rules or mechanisms in such a way that it is likelier to be reproduced (for a discussion see also Conte et al. 2007).

Hence, we need norm-related notions that are

dynamic to be compatible with a simulation-based investigation. Attention will be paid to modifications rather than typologies of norms and their functions.

[6] The references quoted were used only as a theoretical basis to define a shared and agreed language on normative notions among the EMIL Partners.

Innovation-oriented, which is a special case of dynamics. By innovation, we mean a process designed or wanted by institutional or social agencies (if only an opinion movement). A merely conventionalist view of norms, a spontaneous and emergent dynamics, are insufficient to account for this process: rather than waiting for new regularities to emerge, agencies aim to *impose* new obligations or rights, new permissions or forbearances. In a word, new norms.

Hibrid, incorporated both in social and mental objects. In this perspective, the following views are deemed to be insufficient:

Epiphenomenal, maintaining that norms are but observable social patterns interpreted "as if" the patterns resulted from any normative force or process; on the contrary, we are interested into social patterns that result from the action of norms in society.

Behavioural, characterizing norms as observable regularities resulting from agents' squeezing each other into common practices. Conversely, we start from the assumption that it is important to look at what happens in the mind of agents in order to understand how norms operate.

Conventional, characterizing norms as conventions. Although necessary, this is still an insufficient view of norms, especially when we want to deal with innovation.

Command

A command is a coercive request of action, based upon (pretended) power over of the commander on the recipient.

Norm

A norm[7] is a social behaviour that spreads trough a population thanks to the spreading of a particular belief, i.e. the Normative-Belief, a belief that there is a command based on a deontic. 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. The normative request is enacted on the basis of the deontic.

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[8].

- **Meta-norms**, general rules telling agents how to reason, decide upon and apply specific norms.

- **Norm adoption**, the formation of a normative goal from a normative belief, thanks to some intervening rules (GGR and AR, see Annex 3).

- **Normative belief** A belief that a given behaviour, in a given context, for a given set of agents, is forbidden, obligatory, permitted, etc. More precisely, the belief should be that "there is a Norm prohibiting, prescribing, permitting...".

- **Normative belief of pertinence** Believing that a norm exists and concerns us requires at least a second group of beliefs: the beliefs of pertinence. The norm says what ought to be done by

[7] A lively debate on the concept of norm has been developed in several branches of Philosophy, Logic, Cognitive Science, Theory of Agents, Social Theory and Game Theory. Here are few references. **Social and Legal Philosophy,** Raz, 1975; Kelsen1991; **Logic of Action and Deontic Logic**, Horty 2001; Jones & Sergot 1993; Wright 1963; **Cognitive Science and Theory of Agents**, Conte & Castelfranchi 1995, 2006; Castelfranchi & Conte R 1999; **Social Theory and Game Theory,** Bicchieri 1990, 2006; Coleman 1990; Young 1998, 2006; Ullman-Margalit 1977; Therborn 2002; **Social Simulation**, Axelrod 1984,1986; Macy & Skvoretz 1998; Macy & Sato 2002; Sen & Airiau 2007.

[8] A social norm is a deontic command transmitted from one to another agent. It requires only addressees, although defenders and observers might be involved. It is located on the negative extremes along the dimensions of strength, explicitness. It is not issued by a personal authority, but it emerges gradually and spontaneously. It is incorporated into utterances and behaviours, and is not based upon explicit, defined and certain enforcement mechanisms .

whom: (i) the obligation/permission/prohibition and (ii) the set of agents on which the imperative is impinging. For example, if I am addressed by a given norm (say, "be member of a professional order"), and the norm has to take effect on me, I must recognize this. The prescription is about a set or class of agents, and since I am an instance of that class, the norm applies to me.

- **Normative equity principle**, agents want their normative costs to be no higher than those of other agents subject to the same norm.

- **Normative goal**, an internal goal relativized to a normative belief.

- **Normative influencing goal**, a goal to generate an obligation for a set of agents to do a given action.

- **Normative reasoning**, mental operation upon the internal representation of a given norm, which may lead to that norm being adopted, thereby forming a normative goal.

Convention

A convention (cf. Gilbert 1981, 1989; Lewis 1969; Sugden, 1986/2004, 1998; Young 1993, 1996) is a behavioural regularity, i.e. a practice or procedure widely observed by members of a given social network, based on

- the agent's goal of conforming to that behaviour in order to

act like the others,

- the mutual expectation that the others will conform to that behaviour as well.

More specifically conventions are a class of problems (arbitrarily selected from a potential of alternative candidates) classified as (pure) coordination games (viz. convention of keeping to the right (or left) when driving, pointed out by David Lewis, 1969), based on interdependency and mutual expectations.

The confine between conventions and norms is not clear-cut. Conventions may acquire a mandatory force over time, sometimes conventions get to be prescribed, and this is one factor leading to norm emergence. A good example is etiquette, which is halfway between a social norm (with obligations and possibly sanctions) and a convention. Greeting is a polite behaviour, and how to greet someone - whether by shaking hands or waving hallo - is ruled by conventions; on the other hand, when you receive greetings, it is mandatory to reply, probably due to the social norm of reciprocity.

Power in Commands and Power in Norms

According to the above given definition of command, this is "coercive request of action, based upon (pretended) power over (cf. Castelfranchi 1990; 2003) of the commander on the recipient."

According to the above given definition of Norm, Norms<Commands. That is a norm is a "coercive request of action, based upon (pretended) power over of the commander on the recipient.".

So we are going to talk about the ingredient Norm and Commands share, and the one that only norms have, making not true that Commands<Norms

Power over

Given the capability of an agent to bring about a set of world states and given the goals agents have in these objects we can define **power over** of a group of agents I towards a group of agents J as the possibility for I to realize/thwart a set of world states G wanted by J, such that J is not able without I to achieve G. This definition regards objective power over. We could analogously say that if J has power over I wrt G, I is **objectively dependent on** J for G. There exist epistemic variants of dependence. In fact we claim that J is dependent on I for G if J believes I has power over him wrt G.

Norm Frame

The frame is a set of features that characterize the norm and that define its crucial aspects, which are further decomposable. This means that when we specify a norm we need to say something about each of these aspects. A research question is thus whether these features are sufficient and necessary (thus, unique) conditions to construct a norm. We outline six aspects of the norm frame:

Deontic

A Deontic is basically a way of partitioning situations between good/acceptable ones and bad/unacceptable ones[9].

What is important for general recognition issues is that the authority from which the obligation emanates need to be recognized and accepted by the agents in order for the obligation to be dealt with and fulfilled or violated. As for validity based deontics, we can further distinguish them into:

- **obligations: i**t is obligatory to do so,
- **forbearances:** it is forbidden to do so,
- **permissions:** It is permitted to do so.

These constructions are clearly interdefinable. If the definitions, and even the intertranslations, are clear (Forbidden, Permitted, Obligatory) the mechanisms need to be investigated further. For mechanisms we mean: what happens when something is obligatory? And "when is something obligatory?".

Source

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

personal: "the locus from which the norm emanates" means "the nonempty set of agents that performed that action enabling the norm (and after which the norm existed)";

impersonal: "the locus from which the norm emanates" means "the community that enabled the norm". It is clear that considering the "community" equivalent to the "set of all agents" would make the two notions collapse. One of the aims of understanding impersonal source ought to be the understanding of this difference.

Normative Role

With normative role (Conte and Castelfranchi 1995) we mean the partition of the agents involved in a norm. We distinguish:

[9] On the concept of deontic see: Wright 1963, for validity based deontics; Meyer 1988, for deontic logic as a variant of dynamic logic; Alchourròn 1993, for recognition based deontics.

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 acquire beliefs about a norm, that is whether it is enforced, violated, emanated.

Enforcement mechanisms

These are the operations that attempt to modify agents' actions in order to make them compliant to a norm (cf. Axelrod 1984; Conte and Castelfranchi 2006; Conte and Paolucci 2001; Ulmann Margalit 1977).

We distinguish:

sanctions are enforcement mechanisms that inhibit agents' actions;

incentives are enforcement mechanisms that favour agents' actions.

The way actions can be favoured or inhibited follows precise paths in cognitive agents. In this sense enforcement mechanisms follow a path in agent minds, exploiting intra-agent processes. Moreover social artefacts can be used to sanction or to favour agents' actions. Reputation, for instance, can work as a normative sanction. But it can also be used as a normative incentive.

Control

Control is the way enforcement mechanisms are applied (Conte and Dignum 2001; Conte and Paolucci 2004). It implies both monitoring - that checks violation - and influence - that actively pushes cognitive agents' towards compliance. Normative influence will be analyzed further on.

They can be:

centralized: only one agent (individual or supraindividual) is entitled to sanction;

distributed: everybody is able to defend the norm.

Therefore, it has to be said that centralized control makes use of institutional rules for regulation, while distributed control does not presuppose any delegation.

11 Annex 2: TAXONOMY OF NORM INNOVATION

Types of norm innovation

Here below are listed examples of norm innovation, from poorly to strongly innovative.

Norm extension or adaptation

An existent norm is extended to new entities or social category, in such a way that its content is modified.

Examples:

1. Etiquette changes:

    a. Switch from the greeting form "How do you do" to "How are you?". This change is due to the necessity of moving to a more informal way of relating with others.
    b. *Toitoyer*: In countries like Italy and France the formal "Lei/Vous" is more and more substituted by the informal "Tu", to address a conversation partner.
    c. Some courtesy formulas, having once an established and formal form, are becoming nowadays more and more informal.

2. Extension of old rights to new social categories:

    a. Extension of the right to vote to foreigner people of recently acquired citizenship.
    b. Extension of the right to get married to gay and lesbian people.

3. Extension and modification of copyright laws at the domain of open source.

These modifications are due to cultural changes. A subcase of norm adaptation is "contrary-to-duty" obligations: these are norms that are not applicable in particular contexts and hence are modified. One example is traffic regulation: traffic jams and illegal car parking. When a norm prescribing an action collides with the material impossibility to follow the prescription, the regulatory function of the (legal) norm is compromised and a new (social) norm arises in turn, provided that it is more flexible than the previous one. In Italian cities, for example, there are areas with insufficient parking spaces: in these areas can happen that people park their car in illegal ways,

provided that they either leave their telephone number or provide a means to move their car if it is blocking someone else's car. This is a social norm working when the legal norm is impossible to comply.

Norm instantiation

A new norm is perceived and established as an instance of an existing norm.

Examples:

1. Net-etiquette: norm that did not exist before the introduction of electronic mail system. The existing norm here is rules governing socially acceptable behaviour. The Net-etiquette will result in a series of manners that are either new or taken from the general etiquette.
2. Non-sexist language. In the scientific community the practice of a non-sexist language has caught on. For example, instead of the old-fashioned "He", now the use of the gender-neutral pronoun "S/he" has been encouraged. The existing norm here is a prescription on how to talk about a social group in order not to discriminate it.
3. Etiquette in multicultural society. In a multicultural society people must decide whether to allow every groups symbol and expressions or to limit them. Here the existing norm is the freedom of creed. The instantiation consists in the respect for each group values and the general will that every social group have the same amount of freedom of expression: every group will be allowed to exhibit religious symbols, or else no group will.
4. Etiquette in mass society: the big amount of people that travel all around the world ask for the introduction of new rules of conduct, such as:

    a. stopping when someone is taking a picture. Here the existing norm can be considered the will of not invading someone else's privacy, avoiding to appear in their pictures.
    b. speaking in an international, well known language, generally in English, when in a conversation foreigners are present, so that they can participate and nobody feels left out. Here the existing norm can be the will of including everybody in communication and not emarginating anybody.

Norm integration

A norm is determined by the integration of conflicting norms.

Examples:

- Symbolic infibulation: the recently proposed "soft" infibulation, a practice consisting in a merely symbolic cut of the female genital, resulting in a pinprick to cause only a drop of blood. This practice will save the ritual, without damages or pain. Here we have a so called cultural conflict, instantiated in a specific contrast between two opposite norms implying different values (to legalize infibulation as a necessary practice, to forbid infibulation as a harmful and useless practice; and the consequent vision of human being and human/community rights).

12 Annex 3: GLOSSARY

Actions, both physical and social, including actions on others' minds (social influence), i.e. a performed goal.

Adoption rule (AR), a corollary of the GGR; if an agent believes that adopting a goal of a given agent is a means for his obtaining one of his own goals, he will adopt that goal.

Autonomous agent, endowed with the capacity to generate and pursue its own goals.

Beliefs, states of the world as it is (to a variable degree of certainty). These may be:

- non social;
- social, about others, including their mental states.

Belief of pertinence, an agent has a normative belief relative to a set of agents, and believes that he is included in the set.

Cogency, perceived cost of violation.

Goals, wanted states of the world that might or not be verified.

Goal-adoption, an agent adopts another agent's goal when he wants the latter to achieve that goal.

Goal generation rule (GGR), an agent will have as a goal any state that implies that another of its goals will be achieved.

Intentions, executable goals.

Mind, a cognitive regulatory apparatus consisting of symbolic representations of goals and beliefs, and the capacity to manipulate (confront, modify, etc.) them.

Salience, norm's degree of activation, which is a function of its permanence in the N-Board and the degree to which it is believed to be shared by members of the community.

13 References

- Alchourrón, C. 1993. Philosophical Foundations of Deontic Logic and the Logic of Defeasible Conditionals. In Meyer e Wieringa , Deontic Logic in Computer Science, Chichester, Wiley, pp.43-84.

- Anderson SW.; Bechara A.; Damasio H.; Tranel D.; Damasio AR. 1999. Impairment of social and moral behaviour related to early damage in human prefrontal cortex. Nat Neurosci. 2:1032–1037.

- 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.

- Axelrod, R. 1984. The Evolution of Cooperation, New York, Basic Books.

- Bargh, J. A. 1992. The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects.
- The American Journal of Psychology 105, (2): 181-199.

- Barsalou, L. W. 1999. Perceptual symbol systems. Behavioral and Brain Sciences 22: 577-660.

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

- Bicchieri, C. 1990. Norms of cooperation, Ethics100: 838-861.

- 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press, New York.

- 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.

- Castelfranchi, C. 1990. Social power: A point missed in multi-agent DAI and HCI. Decentralized A.I pp. 49–62.

- 2003. The micro-macro constitution of power. Protosociology pp.18-19.

- Coleman, J. S. 1990. Foundations of Social Theory. Harvard University Press, Cambridge MA.

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

- Conte, R., Andrighetto, G., Campennì, M, Paolucci, M. 2007. Emergent and Immergent Effects in Complex Social Systems. In Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington DC.

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

- 1999. From conventions to prescriptions. Towards a unified theory of norms. AI&Law 7: 323-340.

- 2006. The Mental Path of Norms. Ratio Juris 19 (4): 501 – 517.

- Conte R., and Dignum F. 2001. From Social Monitoring to Normative Influence. JASSS 4 (2)

- Conte R., and Paolucci M. 2001. Intelligent Social Learning. JASSS 4 (1).

- 2004. Responsibility for Societies of Agents. JASSS 7 (4).

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

- Damasio, AR. 1994. Descartes' error: Emotion, rationality and the human brain. New York: Putnam.
- Epstein, J. 2000 Learning to be thoughtless: social norms and individual computing.
- Center on Social and Economic Dynamics Working Paper, No. 6

- 2006. Generative Social Science. Studies in Agent-Based Computational Modeling. Princeton-New York: Princeton University Press.

- Fiske, S. T.; and Taylor, S. E. 1991. Social cognition (2nd edn.). New York: McGraw Hill.

- 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.
- Gilbert, M. 1981. Game Theory and Convention. Synthese 46: 41–93.

- 1989. On Social Facts. Princeton: Princeton University Press.

- Horty, J., F. 2001. Agency and Deontic Logic. Oxford University Press, Oxford.

- Jones, A. J. I. and Sergot, M. J. 1993. On the Characterisation of Law and Computer Systems: The Normative Systems Perspective. In Meyer & Wieringa (eds) Deontic Logic in Computer Science: Normative Systems Specification. John Wiley & Sons.

- Kelsen, H. 1991. General Theory of Norms. Oxford, Clarendon. (1st ed. 1979).

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

- Markus, H., and Zajonc, R. B. 1985. The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology, pp. 137-229, 3rd Edition. New York: Random House.

- Meissner, W. W. 1981. Internalization in Psychoanalysis, International Universities Press, New York.

- Meyer, J.-J. 1988, A Different Approach to Deontic Logic: Deontic Logic Viewed as a Variant of Dynamic
- Logic. Notre Dame J. of Formal logic 29 (1): 109-136

- 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.

- Raz, J. 1975. Practical reason and norms. Oxford University, Oxford.

- Schank, R. C., and R. P. Abelson. 1977. Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.

- 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.

- Sugden, R. 1986/2004. The Economics of Rights, Co-operation, and Welfare, 2nd ed. New York: Palgrave Macmillan.

- 1998. The Role of Inductive Reasoning in the Evolution of Conventions. Law and Philosophy 17: 377–410.

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

- Wallis, K. C.; and J. L. Poulton 2001. Internalization: The Origins and Construction of Internal Reality. Open University Press, Buckingham and Philadelphia.

- Wason, P.; and Johnson-Laird, P. 1972. Psychology of Reasoning: Structure and Content. Harvard University Press, Cambridge, MA.

- Wright, G. H. von. 1963. Norm and Action. A Logical Inquiry. Routledge and Kegan Paul, London.

- Young, H. P. 1993. The evolution of conventions. Econometrica 61: 57-84.

- 1998. Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, Princeton, NJ.

- 2006. Social Norms, Prepared for the New Palgrave Dictionary of Economics Second Edition, Steven N. Durlauf and Lawrence E. Blume (eds), Macmillan, London.