Cognitive dynamics of norm compliance. From norm adoption to flexible automated conformity

Giulia Andrighetto · Rosaria Conte

Abstract In this paper, an integrated, cognitive view of different mechanisms, reasons and pathways to norm compliance is presented. After a short introduction, theories of norm compliance are reviewed, and found to group in four main typologies: the rational choice model of norm compliance; theories based on conditional preferences to conformity, theories of thoughtless conformity, and theories of norm internalization. In the third section of the paper, the normative architecture EMIL-A is presented. Previous work discussed the epistemic module of this normative architecture, allowing for the generation of normative beliefs being formed. The fourth and fifth sections present the pragmatic modules of EMIL-A, i.e. norm adoption—leading to normative goals—and norm compliance—leading to their execution. Not only are several alternative reasons for norm adoption shown, but also several pathways to norm compliance are identified. Finally, a summary and ideas for future works conclude the paper.

Keywords Norm compliance · Norm internalization · Agent-based modelling · Cognitive modelling

1 Introduction

A common approach to norm compliance consists of assuming it as a one-step decision-making process, under the generic assumption that agents comply with norms when the costs of violation exceed the costs of compliance (Coleman 1989, 1990). From this prospective, norm compliance is explained in terms of rational

G. Andrighetto (✉) · R. Conte
Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy
e-mail: giulia.andrighetto@gmail.com; giulia.andrighetto@istc.cnr.it

R. Conte
e-mail: rosaria.conte@istc.cnr.it

choice, or as the strategy that an individual adopts in order to maximise his own utility. The act of obeying a social norm (for example, the norm that prescribes that one should give up one's seat on a bus to an old person) is, therefore, the rational choice that one must make if one wants to avoid punishments or to obtain rewards (Coleman 1989; Axelrod 1986). Within this approach, no general account of different modalities of norm compliance is usually proposed and the specific cognitive ingredients and pathways to norm corresponding behaviour are poorly analysed.

In this paper, a somewhat complementary view is taken. The process leading to norm compliance is shown to be interspersed with several checkpoints in which the norm addressee is asked to take complex decisions. Two premises are needed.

First, the model here presented builds on a definition of norm compliance as the result of a series of cognitive processes involving and operating upon norm-related mental representations, namely normative beliefs (NB) and normative goals (NG). Nonetheless, such a model is by no means intended to propose a fully deliberative view of norm compliance. The phenomena of norm internalization and flexible automated conformity (Conte 2009; Andrighetto et al. 2010b) are to some extent accounted for, showing the advantages of a broader cognitive perspective in which fully controlled and semi-automated processes of norm compliance can be integrated.

Second, an integrated view of norm compliance requires an agent based modelling approach, in which different ingredients and components are put together. In previous work, we presented the normative agent architecture EMIL-A[1] (Andrighetto et al. 2007, 2010a; Campennì et al. 2009), providing a picture of its epistemic component, responsible for the formation of normative beliefs. In this paper, the complementary issue is addressed and the pragmatic modules of EMIL-A, i.e. norm adoption—leading to normative goals—and norm compliance—leading to their execution, are presented and discussed.

2 Related work

According to an issue in the International Encyclopaedia of the Social Sciences, "No concept is invoked more often by social scientists in the explanations of human behaviour than 'norm'" (Encyclopaedia of the Social Sciences). Despite the importance of norms, there is still very little consensus on where they come from, how they work, and even why actors comply with them.

Current theories of norm compliance can be divided into four major groups. Theories belonging to the first group account for norm compliance in terms of *strategic reasoning*: on this view, agents comply with norms when the costs of violation exceed the costs of compliance (Becker 1968; Axelrod 1986; Coleman 1989). Agents calculate the expected utility of the alternative courses of action and

---

[1] This normative architecture has been developed within the EMIL project (EMergence In the Loop: simulating the two way dynamics of norm innovation), a FET-funded European project on the agent-based simulation of the two-way dynamics of norm innovation. See Lotzmann et al. (2012) for a description of EMIL-S, a tool box implementing EMIL-A.

choose the one that maximizes their expected utilities. In short, agents follow norms because it is individually rational.

Theories belonging to the second group account for norm compliance in terms of *social conformity* (Bicchieri 2006; Bicchieri and Xiao 2009; Bicchieri and Chavez 2010). Moving from the classic theory of convention proposed by Lewis (1969), Bicchieri and colleagues build on the notion of social expectation. In their view, individuals have conditional preferences for complying with norms. This means that their choice is dependent upon what the other individuals do (empirical expectations), and upon what the others expect should or ought to be done (normative expectation). This group of theories leaves a number of questions open, among which the most problematic, in our view, is to understand in which way the expectations—normative and empirical—are able to motivate individuals to comply with norms. It is not totally clear in which way norms are transformed into the will (i.e. the goal) to observe them (see Andrighetto and Castelfranchi, forthcoming).

The third group sees norm compliance not as a deliberate process but as form of *automatic, mindless, conformity*. According to Epstein (2007), social norms, once learned, are applied automatically rather than deliberated upon. For example, agents take no decision as to whether or not to get dressed in the morning or to use silverware while eating. Rather, they mindlessly execute the norms they have acquired while learning to get dressed or to eat. In this view, social norms are incorporated into social plans and executed thoughtlessly. The mindless conformity view has been questioned, under the consideration that norms are sometimes deliberately violated (for example, rather than stop at the traffic lights, to go ahead in order to let an ambulance overtake) (Andrighetto et al. 2010b). Does it mean deviance is as mindless as conformity, and that once learned to comply with or transgress against a norm, agents are bound to behave accordingly until they learn to do otherwise? Were this the case, the social world would be a much more predictable place than it actually is, but with far less flexible agents. In fact, the world is highly dynamic and uncertain and agents manage to adapt to it. Even if they learn how to behave under given circumstances and convert what they have learned into specified routines, they must be endowed with the capacity to defuse their routines when conditions change.

Within the fourth group of theories, internalization is suggested as the key to solving the puzzle of norm compliance (Scott 1971; Durkheim 1950; Gintis 2003). Internalization occurs when

> a norm's maintenance has become independent of external outcomes—that is, to the extent that its reinforcing consequences are internally mediated, without the support of external events such as rewards or punishment (Aronfreed 1968, p. 18).

In this case, compliance is seen as a product of internal sanctions that agents impose upon themselves. The norm internalization process has several advantages, such as increasing compliance and reducing the costs of norm adoption. Moreover, individuals who internalize norms are not only much better at complying with, but also at defending, them than are externally enforced individuals (Gintis 2003). Many questions remain to be answered, most of which revolve around the issue of

proximate causes (see Andrighetto et al. 2010b). How do people internalize norms, and what does this mean? What are internal sanctions and how can they replace external ones?

Curiously, though each of these four theories of norm compliance assumes specific psychological mechanisms, none of these theories has been systematically directly tested in regard to the cognitive processes they imply. What is still missing is an integrated socio-cognitive approach of different mechanisms, reasons and pathways to norm compliance, allowing experiments on the impact of different factors to be conducted and results on different measures of performance (e.g. behavioural conformity and stability) to be compared.

3 EMIL-A: foundational notions and architecture

3.1 Foundational notions

Building on Ullmann-Margalit's (1977) definition of a norm as a "prescribed guide for conduct which is generally complied with by the members of society", in the present work we refer to a norm as a behaviour that spreads through a given society to the extent that the corresponding prescription spreads as well, giving rise to a shared set of mental representations (namely, sets of beliefs and goals concerning the norm) (Conte et al. forthcoming; Conte and Castelfranchi 1995, 2006). In other words, we conceptualize norms as hybrid objects, consisting of a mental and a social side.

With normative prescription, we indicate a request that a given action be done, or a world state be achieved, because it is obligatory. Beliefs are the cognitive way for any external input, including obligations, to access the mind. If an obligation is acquired anew, it will form the content of a new belief. In order for autonomous agents to undertake (or refrain from undertaking) a certain course of action, it is not sufficient that they know (i.e. they have the belief) that such a course of action is desired by someone. It is necessary for them also to have the goal of performing such an action.

Only via a new belief, can an obligation, as well as any request, affect the mind, and thanks to some (reasoning) rule give rise to the goal of fulfilling it. If x believes that there is an obligation and she has a previous goal that cannot be achieved (or is thwarted) if she does not fulfil the obligation, x will generate a new (normative) goal as a means for the previous one. More specifically, any agent x recognizing a given input as a norm forms at least the first of the following beliefs:

- *Main normative belief* (indicating the existence of the norm), which states that a given type of behaviour B, in a particular context C, for a given set of agents S, is forbidden, obligatory, permitted. More precisely, the belief states that "there is a norm N prohibiting, prescribing, permitting a". Beliefs supporting the creation of normative beliefs include:

    - The source of the prescription is a *formal* authority, held to issue (a specific set of) norms.
    - The source is not a formal authority, but the set of agents S, i.e. the source is a *distributed* one.

- • N is *impersonally addressed*, i.e. anyone belonging to S in circumstances C is required to comply with N.

- *Normative belief of pertinence* (indicating that the belief holder belongs to the set of agents on which the norm is impinging): x believes she belongs to S.

To these necessary normative beliefs, one further is often, but not necessarily, associated:

- *Norm enforcement belief*: the belief that normative compliance and violation are supported or enforced by positive or negative (informal) sanctions.

What is derived from these normative beliefs is a set of normative goals. We refer to *normative goals* as goals relativized to a normative belief. A goal is here meant in the very general sense derived from cybernetics, i.e. a wanted state of the world triggering and driving actions (Miller et al. 1960; Conte and Castelfranchi 1995; Conte 2009). A goal is relativized when it is held because and to the extent that a given world-state or event is hold to be true or is expected (Cohen and Levesque 1990). We distinguish between four types of normative goals:

- the *main* normative goal: the goal to comply with the norm;
- the normative *invocation* goal: the goal that the set of agents S on which the norm applies come to have the belief that "there is a norm N prohibiting, prescribing, permitting *a*";
- the norm *defence* goal: the goal that the action a prohibited or prescribed by the norm is realized;
- the norm *enforcement* goal: the goal that norm violations be punished.

These normative goals are generated on the basis of different (normative) beliefs. In this work, we will focus only on the main normative goal (for an analysis of the other three types of normative goals, see Andrighetto and Conte forthcoming).

For the main normative goal to be generated, the main normative belief and the belief of pertinence are necessary conditions, while the norm enforcement belief is not always required. In some circumstances, the agent decides to comply with the norm not to avoid punishment, but because she has the terminal goal that "norms be respected" (terminal adoption) (see Sects. 4, 5.3) or because she has internalized the main normative goal (see Sect. 7.1). The difference between terminal and internalized goals is subtle, but important. A terminal goal is a fully endogenous goal. When an instrumental goal is endogenised for whatever reason, it is internalised. Such a conceptual difference has profound implications, also at the psychological level (see, Conte et al. forthcoming). For example, a plausible hypothesis based on such a difference is that agents are more likely to be more committed to internalized than terminal goals, if only for reasons of cognitive balance. A repentant ex-smoker is more tenacious and systematic in condemning and discouraging smoking habits.

What was argued so far does not mean, however, that a believed obligation will always give rise to the goal of fulfilling it. However, the goal to fulfil the norm must be formed in order to check the conditions for its execution. Beliefs, including normative beliefs, cannot trigger but only guide goal execution and interact with

existing goals in the generation of new ones. A fortiori, a goal deriving from an obligation does not necessarily lead to a compliant action. As we will discuss later, a goal can be abandoned for a variety of reasons (for example because it turns out to be already achieved, or it is incompatible with some more important goals, etc.). Building on these ingredients, we can now introduce EMIL-A.

3.2 EMIL-A

EMIL-A has been presented at some length in several papers (for a complete overview, see Conte et al. forthcoming). It includes two main components, the *epistemic* and the *pragmatic* ones. The epistemic component is responsible for recognizing norms (it includes the *norm recognition* module). The pragmatic component is responsible for behaviour based on normative representations (it includes the *norm adoption* module and *norm compliance* module).

In Fig. 1 a sketch of the main components and mental dynamics of EMIL-A is provided. It includes:

1. Three types of representations:

   a. Normative Beliefs.
   b. Normative Goals.
   c. Normative Intentions: i.e. executable normative goals.

2. Three modules:

   a. Norm Recognition (epistemic component).
   b. Norm Adoption (pragmatic component).
   c. Norm Compliance (pragmatic component).

3. The norms' salience mechanism, which updates the salience of norms, according to external events (see Sect. 6.1)
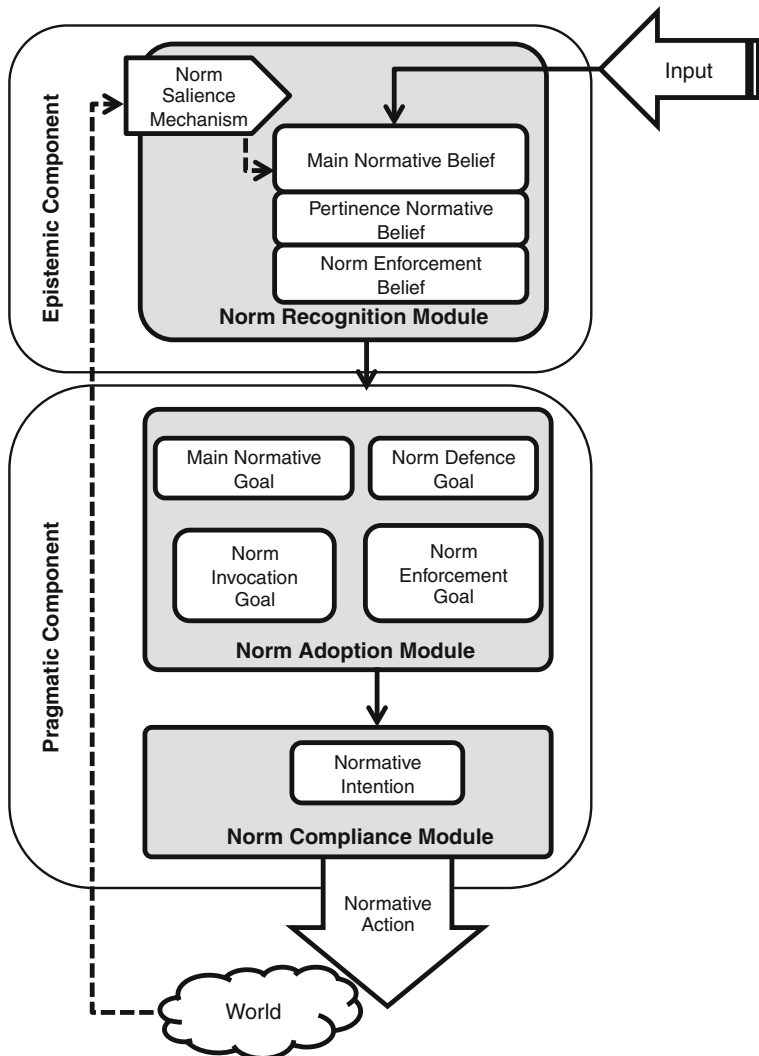
The *Norm Recognition Module* is the crucial component by means of which agents are able to infer that a certain norm is in force even when it is not already stored in their normative memory. It allows agents to form new normative beliefs processing the information received while interacting with or observing the other agents behaving in a common environment. The Norm Recognition Module detects whether or not the received social input refers to a normative belief already stored in the normative board. In the former case, it will update the salience of the corresponding norm accordingly. In the latter case, it will either form a new normative belief, or simply discard the input (for a detailed description of the norm recognition module, see Conte et al. forthcoming and Andrighetto et al. 2010a, b).

When a new normative belief is formed, the Norm Recognition Module will send information to the Norm Adoption module. This will use such information to decide whether or not to form the corresponding normative goal, based on the norm-adoption rule (see Conte and Castelfranchi 1995; Sect. 4).

Finally, the new normative goal will be imputed to the *Norm Compliance Module*. This consists in a decision-making procedure that takes a normative goal as an input and possibly puts it into execution, performing a normative action.

The procedure will put the goal to execution unless it is already realised or incompatible with more important goals. In the last two cases, the Normative Goal will be suspended until the conditions for its execution will be verified again.

The epistemic component of the EMIL-A architecture (the norm-recognition module) has been developed and described in previous work (Andrighetto et al. 2010a; Campennì et al. 2009). In the following sections, the pragmatic component is modelled with some detail (Fig. 1).

Fig. 1 Main components and mental dynamics of EMIL-A. It consists of different modules interacting with one another by means of input–output mechanisms. The Norm Recognition module plays a crucial role by informing both the Norm Adoption and the Norm Compliance modules. These two modules are responsible for the actions performed by the agent

4 Norm adoption: reasons and reasoning

We refer to the mechanism that leads from a normative belief to a normative goal as *norm adoption* (Conte and Castelfranchi 1995). An autonomous agent acts to achieve her own goals and must have reasons for choosing whether to act as she does. In particular, if an autonomous agent accepts another's (normative) request, she must have good reasons for doing so.

The general mechanism by which an autonomous agent adopts external requests, called adoption, has been described at some length in Conte and Castelfranchi (1995). Here, suffice it to say that an agent (the adopter) will adopt another agent's (i.e. the adoptee's) goal as hers, on condition that she, the adopter, comes to believe that the achievement of the adoptee's goal will increase the chances that she will in turn achieve one of her previous goals. For example, I will accept your request to lend you my laptop, if this is a means for me to borrow your fancy clothes tonight. When the external request is a prescription, a special application of this process occurs, i.e. norm adoption. I will adopt the norm if, say, I think that by doing so I avoid getting a fine, obtain others' approval, build a good reputation, etc. General adoption leads to form *social* goals (achieve somebody else's goals). Norm adoption leads to form *normative* goals. Norm adoption even when instrumental cannot be reduced to the consideration of normative beliefs, because it is the combined effect of goals + beliefs, which triggers the agent's activity, including its mental activity.

Norm adoption is a decision-based process, and not a deterministic one, in which the agent can decide to adopt a norm and form a normative goal for several reasons. In Conte and Castelfranchi (1999) and Conte (1998), two main types of norm adoption were identified: instrumental and terminal adoption.

Instrumental adoption: the subject adopts the normative goal if she believes she can get something in return (avoid punishment, obtain approval, praise, etc.). Normative goals can be formed for self-regarding reasons. This does not prevent the goal thus formed from being normative in the fullest sense. All that is needed for a goal to be normative is that it is based on norm-related representations. *Cooperative* adoption is a particular form of instrumental adoption. Norm-adoption is cooperative when it is *value-driven,* that is when the subject shares both the end of the norm and the belief that the norm achieves it. For example, an agent may decide to conform to the recycling norm because she believes that, by doing so, she helps to reduce our species' negative impact on the environment. In this case, the result of adoption is not only a normative, but also a moral goal, or, to be more precise, a normative goal based on a moral (value-driven) motivation. In cooperative norm adoption, a specific norm is adopted because and to the extent that it is believed to yield a positive, desirable state of the world, which is a value of its own. In such a case, the norm addressee will adopt only the norms that she shares.

Terminal adoption: the subject wants to observe the whole set of norms she is subject to as ends in themselves. She has the terminal goal or value that "the norms be respected" (Kantian morality). Terminal norm adoption implies that any norm deserves obedience while it exists. In such a case, a normative goal already exists and leads to the adoption of a norm as soon as one is found.

In the following subsections, some specific forms of reasoning belonging to these general categories of norm adoption are analysed. For each reasoning, the corresponding cognitive pattern is graphically represented according to some fundamental relationships between different mental representations or between them and the external world.

Here follows a brief glossary, introducing the main notions that will be use in the rest of the paper.

1. *Generate*: standing for a mental state generated anew. A belief may be generated by perceptions or by other beliefs. A goal is generated by at least one new belief and one previous goal.
2. *Instantiate*: indicating the relationship between a member of a class and the class it belongs to.
3. *Interfere*: referring to the relationship between an external event and a goal, such that the external event is about to compromise/realize the goal.
4. *Pursue*: pointing to the relationship between instrumental goals and their ends, or between a goal and the state of the world it is aimed at.
5. *Activate*: applying to the relationship between beliefs or perceptions and the goals currently inactive in the mind, which become active based on those perceptions or beliefs. A goal is active when it is operative in the mind, triggering decision-making, planning, or other mental operations.
6. *Interact*: concerning the relationship between a belief and the goal it activates. Such an interaction may generate a new goal via the goal-generation rule.

5 Types of norm adoption

In the present theory, a normative belief is necessary to generate a normative goal. But is it also sufficient? Our answer to the second question is no, a normative belief is not sufficient. At least another condition must be met, i.e. a goal for which the normative belief is relevant must already exist among the agent's goals.

A belief is relevant for a goal when it describes a state of the world that interferes positively or negatively with the goal's achievement or maintenance. When this is the case, the belief activates the goal in question, i.e. turns it on, at the same time pointing out (one of) the conditions that should be realized or removed in order to satisfy it. As a possible consequence, a new goal will be generated either as an instance of the activated goal, or as a means for achieving the condition interfering positively, and removing the condition interfering negatively with it.

In the following sub-sections, we will provide some examples of different types of norm adoption.

5.1 Apparent adoption

A norm may or may not be represented in the mind of agent x, but sometimes x's behaviour may involuntarily and inadvertently correspond to it. It may happen that x's behaviour *de facto* pursues a norm, independent of her mental states, when her

Fig. 2 Apparent adoption
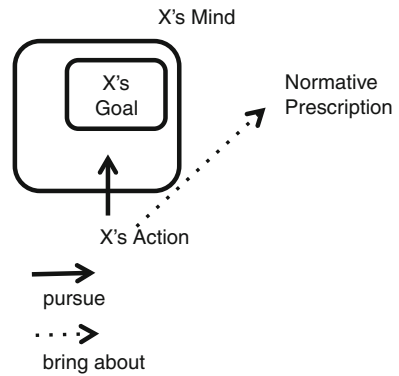
goal coincides with the world-state prescribed by the norm, although she does not aim to comply with the norm nor perceives this correspondence. In such a case, there is a mere norm-corresponding behaviour, based on an apparent form of adoption. The typical example is to study with a brilliant and good-looking classroom mate to enjoy his company, with the unlooked for result of fulfilling parents' or teachers', or even one's own, prescriptions (Fig. 2).

5.2 Instrumental adoption

Instrumental adoption is based on the goal-generating rule, i.e. any state of the world that is believed to lead to a wanted world-state will come to be wanted as well. Hence, x will adopt a norm instrumentally, when she believes that complying with it will lead her to achieve or maintain one of her previous goals (obtain reward, avoid punishment, etc.). In the following sections different types of instrumental adoption are identified.

5.2.1 Diligent adoption

Diligent adoption refers to technical norms, i.e., norms that are the effective means for the attainment of specific goals of the agent.

In this specific type of adoption, a normative goal is generated by a normative belief (e.g. the belief that you have to enter your PIN when using your credit card) interacting with an active goal of x's (e.g. the goal to get money from an Automated Teller Machine, ATM). As this is a case of instrumental goal-adoption, the normative goal is only a means for x to achieve her previous goal. To be noted, the link between the normative goal and the pre-existing goal is *natural*: the (technical) normative goal is a natural means for achieving x's goal. The normative belief refers to a technical norm (e.g. enter your PIN when using your credit card), and the (technical) normative goal results from the adoption of a technical norm (Fig. 3).

5.2.2 Artificial adoption

The main difference between artificial adoption and the preceding cases is that the normative goal is fully exogenous and the link between the normative goal and x's
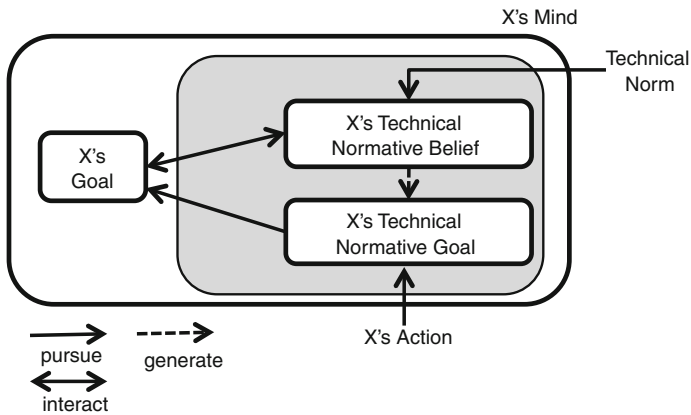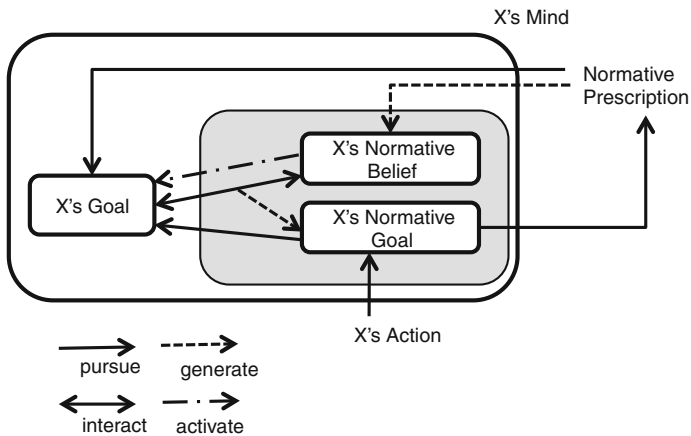
Fig. 3 Diligent adoption

Fig. 4 Artificial adoption

previous goal, from which the normative goal derives, is no longer a natural and direct link. It is an *artificial* or conventional link, in which x adopts the source's normative prescription in order to achieve her own goal.

To be noted, one further aspect of this configuration is the activation relationship holding between x's normative belief, generated by the source's prescription, and x's original goal: the normative belief must first activate x's previous goal and interact with it, in order to generate the new normative goal. For example, the relationship between doing your homework and getting out afterwards is by no means natural, but is established by the mother's normative will (Fig. 4).

5.2.3 Cooperative adoption

Cooperative adoption is a particular form of instrumental adoption, in which the subject adopts the normative goal to achieve not a personal but a common goal.

Norm-adoption is cooperative when it is *value-driven,* that is when the subject shares both the end of the norm and the belief that the norm achieves it. Cooperative adoption is mid-way between diligent and artificial adoption, and a shared norm is similar to a technical norm except that it represents a solution for a social, and not only individual, problem.

5.2.4 Adoption by trust

In adoption by trust, x adopts the norm because she considers it as a solution to overcome her bounded rationality on condition that it be adopted and performed by all group members. In terms of dependence theory (Conte and Sichman 1995; Conte and Sichman 2002; Castelfranchi et al. 1992), this means that x believes that members of group G, which she believes she belongs to, mutually depend (i.e. depend on one another) for realizing the solution in question (i.e. for satisfying Goal 1). Hence, x wants others to adopt the norm as well (i.e. Goal 2), and she adopts it to increase efficiency and to encourage others to do the same (Fig. 5).

5.2.5 Adoption by commitment

To understand this form of adoption and the corresponding cognitive configuration, the reader should perceive it in its unfolding over time. Suppose for example x's parents made her promise that she would graduate soon. At the onset, x only has Goal 1 (the goal to graduate soon). If x commits to this event (promise), she will generate a new obligation (a normative prescription) of herself, which she will perceive as a new normative belief. Such a normative belief will activate a pre-existing (moral) goal,
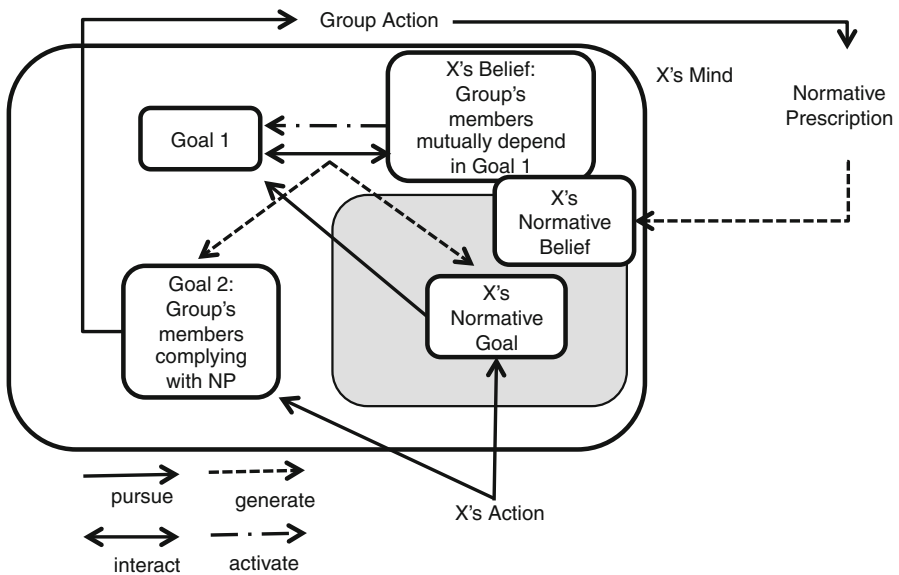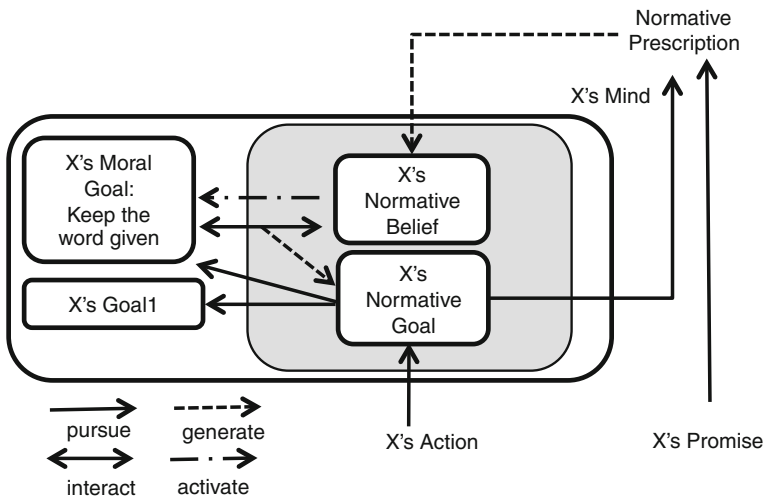

Fig. 5 Adoption by trust

Fig. 6 Adoption by commitment

i.e. keep the word given or fulfil the commitment. The interaction between the normative belief and the moral goal will generate a new normative goal, i.e. to graduate soon, relativized to the normative belief (i.e. the belief about the obligation deriving from commitment). To accomplish this goal will become a normative action that, while achieving the normative goal, will increase the probability that x's initial goal, Goal 1, and x's moral goal will be satisfied (Fig. 6).

5.2.6 Conditioned adoption

Conditioned adoption is the parallel of adoption by commitment. They are both based on some pre-existing moral principle or value, which in conditioned adoption is exemplified by reciprocity, or extinguishing one's debts. As in the preceding picture, the reader is asked to read the configuration reported above dynamically. At the onset, x has goal 1, which an action by y happens to achieve. Whether y acted accidentally or intentionally, does not matter: his action is sufficient to generate an obligation on x, i.e. return the favour. This will generate a new normative belief of x's that will activate her moral goal to extinguish debts or reciprocate adoption and benevolence. Consequently, the interaction between the normative belief and the moral goal will generate a new normative goal of x's, i.e. adopt one of y's goals (Fig. 7).

To be noted, had x acted to obtain y's favour in return, she would not have adopted a norm, but only a goal of y's. Interestingly, however, in such a case she would have had to strategically rely upon y's conditioned norm adoption. It would be a case of norm-based social reasoning, rather than norm adoption. When individuals realise how norm intelligence works and the kinds of behaviours it can result in, they can decide to rely on it to predict the conduct of others and influence their minds.

Conditioned adoption is obviously specular to a more general precept, *do unto others what they have done unto you*. Interestingly, this precept may be extended to
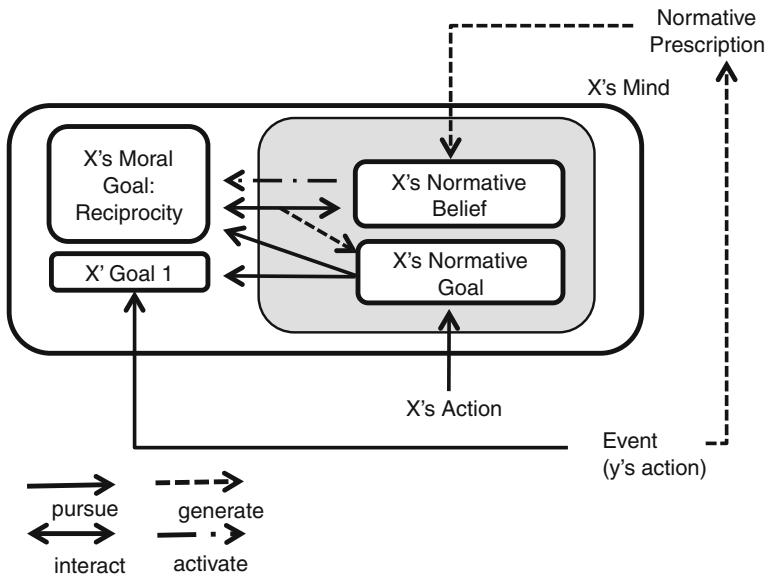
Fig. 7 Conditioned adoption

include a preventive variant by means of anticipation, i.e. "*Do(n't do) unto others what you (don't) want others to do unto you*": obtain reciprocity or at least prevent retaliation. To account for the preventive variant, one should simply read the picture above in a slightly different order, with x's action and goals preceding y's action.

A fairly special case of the preventive variant of conditioned adoption is the norm to fulfil expectations. Observers often attribute regularities to prescriptions. Regularities are perceived to derive their mandatory force from their consequences: once a behaviour *b* starts to spread over a population *P*, members of *P* expect it to be maintained. They anticipate future events based on such expectations, adapt their behaviour to it, and *count on* b (Conte and Paolucci 2002). Expected effects then turn into wanted effects. To disrupt regularities implies the betrayal of such expectations, which is the same as injuring others, compromising their goals.

5.3 Terminal adoption

Norm adoption is terminal, when the agent decides to comply with the norm because of the goal "norms must be obeyed". This type of adoption does not depend on the specific nature of the norm in question. Indeed, x may even believe the norm to be wrong, still she will adopt it as long as she believes such a norm to be in force (Fig. 8).

6 When are we most likely to form normative goals?

In this section, two aspects affecting the probability of generating a normative goal will be discussed, the norm salience and the nature of the pre-existing goal.
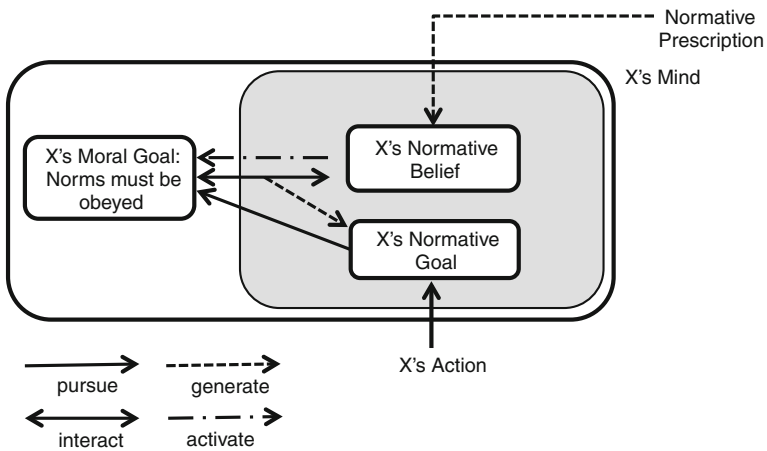
Fig. 8 Terminal adoption

6.1 Norm salience

A norm can be perceived as more or less salient. Norm salience is a measure that indicates how important and operative a given norm is perceived to be by group members (Andrighetto et al. 2010b; Bicchieri 2006; Cialdini et al. 1990; Xiao and Houser 2011). A norm still in force may be poorly operative. For example, capital punishment still exists in the military law of many European countries (in Italy, for example), but is never applied. Examples of poorly operative social norms are even more frequent: as manners prescribed by the etiquette become inoperative, they are soon adapted or replaced by new ones.

In earlier works (Andrighetto et al. 2010a; Andrighetto and Villatoro 2011), we identified several factors contributing to norm salience, which provide many indicators for updating the related norm salience perceptions of group members. Among such factors, the following play a prominent role:

- *Explicitness of the normative request*. This is not only a measure of efficacy of the prescription, but also an indicator of the power and legitimacy of the normative source. The more explicit the prescriptive message, the higher the norm salience and as a consequence the more efficacious and effective the normative will is perceived to be (Cialdini et al. 1990).
- *Source legitimacy*. Norm salience is a function of the degree to which the prescriptive source is accepted and found legitimate (Faillo et al. 2010).
- *Transgression rate*. Norm salience is an inverse function of this variable. In turn, transgression is perceived either as an indicator of scarce importance of the norm, or as an indirect measure of inefficient control and enforcement and therefore of a weak or illegitimate source (Cialdini et al. 1990).
- *Norm enforcement typology*. If properly designed, the enforcement mechanisms play not only a coercive but also a *norm-signalling* role. We use *sanction* to refer to the mechanism that combines the coercive component with the

norm-signalling one (while we call *punishment* the enforcing mechanism that relies only on its coercive component) (Andrighetto and Villatoro 2011; Villatoro et al. 2011). Sanction draws people's attention on a number of explicit or inferred events: (a) the sanctioned conduct is perceived as violation of a norm; (b) this conduct is disapproved of; (c) a causal link is established between violation and sanction: "you are being sanctioned because you violated that norm"; etc. Sanctions convey a great deal of norm-relevant information that has the effect of increasing the salience of norms and promoting their spreading.

- *Norm's effect.* The more adequate the solution provided by the norm to a given social problem or the more equitable the social state of affairs it brings about, the more salient it is perceived to be. This indicator affects particularly technical norms, and yields the type of adoption that was named diligent earlier in the paper (Sect. 5.2.1). In general, when these types of norm are perceived as highly salient, they are adopted to avoid complex or time consuming reasoning and calculation.

Once recognized as norms, external inputs (gathered through observation or communication) are archived in a region of the norm recognition module named normative board (see Andrighetto et al. 2010a), and arranged according to their degree of salience. The more salient a norm, the stronger its activation power and the more likely it will be complied with (Cialdini et al. 1990).

If I hear people saying that ticket controls on the metro have been intensified, the salience of the norm to buy the ticket gets higher. Hence, my normative belief concerning paying the ticket (more specifically my norm enforcement belief) will activate the goal to avoid a loss (taking a fine), what will in the end generate the normative goal to purchase the ticket.

6.2 Nature of the goal

Goal activation is a function not only of norm's salience, but also of several quantitative and qualitative aspects of goals.

As to quantitative dimensions, the more *important* (or urgent) the goal, the more likely it will be activated by any interference. If it is lunchtime, the fresh smell wafting from the bakery at the corner will seriously challenge work on a paper. Instead, if it is only half an hour since I had breakfast, I will probably find it easier to tell myself that I recently decided to start a new diet, and that the paper is due tomorrow. Analogously, if I need to make myself accepted within the new social environment where I have recently bought an apartment, I may worry about my neighbours' intolerance to noise. In my previous, temporary, residencies, on the contrary, I used to pay little attention to the preferences of my fellow tenants.

Goal activation depends also on the *goal status* (Basso et al. 1993). As to status, based on prospect theory (Kahneman and Tversky 1979), a *maintenance* goal is more likely to be activated than an *achievement* goal, other things being equal. A maintenance goal is a goal that is already achieved and that holders try to maintain or restore whenever disturbing events threaten its realization. While an achievement goal is a not yet achieved goal that holders are either currently

pursuing or momentarily setting apart in wait for more favourable or feasible conditions of pursuit (for a formal difference between the two types of goal, see Cohen and Levesque 1990). As losses are suffered more than unachieved gains, interferences are likelier to activate maintenance goals more than achievement goals, again other things being equal. Turning to the normative context, would this imply that punishment works better than praise or approval? Not really. Rather, it implies that if one believes one already has approval and praise, one will feel less likely to put it at risk by behaving inappropriately or illegally, than would be the case if one has nothing to lose.

7 Norm compliance

Is norm adoption enough for the norm to be actually observed? Can we say that such a condition is sufficient for norm compliance? Unfortunately, not. The way to normative action is still quite long, and interspersed with checkpoints in which decisions might endanger the whole process. The main normative goal may be dropped at any point, along this complex itinerary. Worse, it may be the case the main normative goal is never dropped but the norm is not complied with. This is the case when interferences are beyond one's control. If the seat belt of my car breaks up while I am sitting in a traffic jam, there is little I can do but violate the norm of keeping seat belt fastened while sitting in the car. But there are other checkpoints.

First of all the new goal is checked against the current state of the world. The goal might turn to be already true in the world. A second check consists of evaluating the goal against other goals. If it is found incompatible with other more important ones (normative or non-normative), it will be dropped. Otherwise, it is fed into a planning module. This is not the forum for a detailed treatment of the planner's activity. We limit ourselves to addressing the question of when a normative goal, once formed, is most likely to be executed. In particular, two major factors are considered: a) the individual and social perceived effects of compliance/ violation, and b) educational role-playing.

On the one hand, agents executing normative goals are affected, among other factors, by possible unforeseen side effects of norm compliance. If agents perceive unpredicted positive effects of norm compliance on themselves, the normative goals will be reinforced and the likelihood they will be turned into actions will increase. This is similar to action feasibility: the normative goal, originally generated as a means for a specific goal (e.g. avoiding sanctions), might also prove useful in achieving other goals. For example, the goal's holder might realize that by respecting the defence to smoke in her room at work, she feels better at the end of the working day. The day after, she will probably have her norm abiding intention confirmed.

On the other hand, perceiving the antisocial effects of one's own or others' norm violation may reinforce one's normative goal and finally turn it into a normative intention. Suppose someone who usually practised pick pocketing could vividly perceive the effects of a good share of his successful undertakings. In most cases, this perception will be followed by a robber's reduced "activity", at least until the

memory of its effects fades away. In particular, norm violation is less likely to occur again if its perceived effects are accompanied by social emotions like empathy, shame, and guilt. Even professional killers need to "protect" themselves from social emotions and empathy, by keeping at a distance from their victims, learning as little as possible about them, and "acting" in cold blood, suddenly and cleanly, avoiding any sort of contact (Grossman 1995).

Finally, *norm invocation* is another factor increasing the chance that normative goals are executed. Others' norm invocation plays both a direct role, acting as a mechanism of norm enforcement and an indirect role, by affecting norm salience and therefore the generation of normative goals. Moreover, agents that happen to play a norm-defending role (for example, people blaming those that do not respect the queue) can be hypothesized to be less likely to violate the same norm, not only in the presence of the recipients of their former norm-defending messages, but also in their absence. This hypothesis is easily derived from both equity theory (Walster et al. 1978)—stating that people feel uneasy when the ratio between the benefits they receive compared to the costs sustained is both lower and higher than their fellows—and dissonance theory (Festinger 1957)—people feel uneasy when their behaviour and/or mental states are inconsistent.

But norm compliance is not always fully deliberative. In the following subsections, the process so far described is shown to shrink by avoiding some checkpoints and their related decisions. Two main shortcuts are analysed at some length, *norm internalization* and *thoughtless conformity*.

7.1 Norm internalization

Normative goals may be internalized. In previous work (Conte 2009; Andrighetto et al. 2010b), we propose that internalized normative goals are normative goals no more relativized to enforcement normative beliefs, but only to the main normative belief (see Sect. 3.1). In other words, the main normative goal is endogenised, i.e. it becomes an end in itself, needing no external enforcement to being complied with. When an internalized main normative goal is created, enforcement, if any, is self-administered (feeling of guilt, self-depreciation, loss of self-esteem, or other negative self-evaluations in case of violations, and pride, enhanced self-esteem, security, or other favourable self-evaluations in case of conformity) (Reykowski 1982).

The norm internalization process has several advantages, such as increasing compliance and reducing the costs of norm adoption. Moreover, individuals who internalize norms are not only much better at complying with, but also at defending, them than are externally-enforced individuals (Gintis 2003). An effect of the latter prediction is that this form of norm internalization is decisive for distributed social control. Internalization is not only a mechanism of private compliance, but also a key factor of social enforcement. Individuals who have internalized the norm, will comply with it with no need for external enforcement, and in many circumstances also persuade others to observe the norm by reproaching transgressors and reminding would-be violators that they are about to do something wrong (Andrighetto et al. 2010b).

Why do agents observe a norm irrespective of external enforcement? Factors affecting norm internalization should be investigated cross-methodologically,

confronting cognitive and psychological models with empirical evidence. Here we provide some preliminary hypotheses.

We suggest that a norm's *consistency* with one's beliefs, goals, values, emotions and previously internalized norms plays a crucial role in favouring its internalization. Successful educational strategies favour internalization processes, often by linking new inputs with previously internalized norms or values. Values give people reasons for new actions: "If something is good, it should be pursued" (Miceli and Castelfranchi 1989, p. 181). Conversely, when a norm is inconsistent with a person's beliefs and values, this may find herself in a state of cognitive dissonance (Festinger 1957). Cognitive dissonance, being psychologically uncomfortable, motivates the person to reduce it and leads to avoid information that is likely to increase it (Harmon-Jones and Mills 1999).

As argued by Deci et al. (1994), supporting *self-determination* is another factor promoting the internalization of values, attitudes and regulatory structures, such as norms. According to Deci and Ryan (1985, 1987), the central parameter mediating the effects of external events on intrinsic motivation is whether an individual perceives contexts as *supportive* of her autonomy (i.e. encouraging the individual to make her own choices) or as *controlling* (i.e. pressuring the individual toward a specific activity or toward particular outcomes). It has been shown that when supervisors or teachers behave in an autonomy-supportive manner (i.e. provide subordinates with options, acknowledge the subordinate's perspective, and provide a meaningful rationale for the request) rather than a controlling manner (i.e. use threats and deadlines and use rewards to control behaviours), subordinates display high levels of intrinsic motivation (Deci et al. 1981; Pittman et al. 1982; Ryan 1982).

Another important factor in favouring the process of norm internalization are *emotions* (Gintis et al. 2005; Haidt 2003; Tangney et al. 2007; Widegren 1998; Schwartz 1977; Schwartz and Howard 1981). As pointed out by Miceli et al. (2006), "agents may perform (or avoid performing) an action in order (not) to feel a certain emotion: I may give you a present to feel the joy of making you happy, or do my own duty not to feel guilty" (pp. 852–853). Thus, a given norm can be complied with not only on the grounds of the agent's expectations about its outcome and evaluations of its costs and side effects, but also in order to feel (or not to feel) the associated emotions. For example, the anticipation of pride or guilt (or in the presence of other agents, the anticipation of shame[2]) is one factor making people comply with or abstain from violating the norm.

Norm's salience is a fourth factor contributing to the internalization of norms. As said in Sect. 6.1, the more salient the norm is believed to be, the higher its impact on the goal to comply with it. The salience of the norm can increase to the point that the norm becomes internalized, i.e. converted into an internalized (normative) goal or even an automated action (see, Sect. 7.2). The belief that a norm is salient affects its compliance in several ways.

Once agents have internalized the norm is less likely to drop it and its compliance is more stable and frequent.

---

[2] The feeling of discomfort at having done something wrong not only by one's own norms but also in the eyes of those whose opinions matter to you.

7.2 Thoughtless conformity

Agents may sometimes turn normative actions into automated behaviours, as happens when a given input (for example, the red traffic light) leads the norm-addressee (in our example, the car driver) to fire a norm corresponding behaviour (stop at the crossroads) in a non-deliberative but semi-automated way. In our theory (Conte 2009; Andrighetto et al. 2010b), a norm is conformed to automatically, or thoughtlessly (Epstein 2007), when a normative action is performed as a conditioned action, independent of norm-based decisions and representations. An agent stopping at a red light does not need to recognize the input as normative. All she needs to do is perceive a given input, which will automatically activate the corresponding conditioned action, consisting in the sequence of movements necessary to activate the car's break, a behavioural response so deeply internalized that one can hardly make it explicit. The normative beliefs may still be present in the agent's mind, but they are not the reason why she executes that behaviour.

However, automated responses are rarely needed one hundred percent of the time. Indeed, a completely automatized answer may become counterproductive or even dangerous. Agents need to be flexible enough to be able to block a given routine when given conditions activate inconsistent prescriptions. For example, a car driver stopping at the red traffic light might see a policeman asking her to move on. In such a case, the car driver needs to be able to retrieve control of her action, block the automatism, and decide which input should be given priority.

To account for flexible automatism requires a slightly innovative way to model and implement conditioned actions, which has not been implemented so far. In flexible routines, alternative courses of action are attached to conditions, but only one of them is allowed under normal conditions, i.e. in absence of inconsistent inputs. The others are inoperative but still represented. When inconsistent inputs are met, the agent is required to retrieve alternative actions and choose which one to apply under the new conditions.

The main factors leading to respond to norms in an automatic way are *norm salience*, *explicitness* and *operationality*. The more frequent and operational the normative action, and the more standardized the context in which it is executed, the more likely the agent will automatize it, executing it in a partially conscious manner, unless external events block execution. The reason is rather obvious: the normative system often needs to rely on robust repeated performances, almost unfailingly executed with no waste of time (when the light is red or turns yellow start pressing the break until the car stops). In such cases, the more automatic the response, the better. Agents will waste no time in taking decisions, choosing whether or when to comply with the norm or not, possibly coming up with a less efficient action.

8 Conclusions

In this paper, the pragmatic component of the normative architecture EMIL-A has been modelled with some detail. Whilst previous descriptions of EMIL-A and the

simulations so far run (Campennì et al. 2009; Andrighetto et al. 2010a, b) provided a complete picture of the epistemic component, responsible for the formation of normative beliefs, in the present work the complementary issue is addressed: how do agents form and execute normative goals.

The pragmatic component includes two main modules, *norm adoption*, possibly leading to normative goals, and *norm decision-making*, possibly leading to normative intentions. Different types of norm adoption and the corresponding mental configurations have been analysed. Finally, the main factors allowing normative goals to be turned into effective actions have been examined. In particular, the process of norm internalization has been illustrated with some detail, as distinct from automated normative actions. The latter, usually made to coincide with thoughtless conformity, is seen here as a flexible automatism, allowing agents to profit from the advantages of automated behaviours but at he same time holding the capacity to retrieve control of the deliberative power when required by the circumstances. Future work is expected to implement this model on our agent platform, in order to check its validity and its effects on different simulated scenarios.

References

Andrighetto G, Conte R (forthcoming) Norms' Dynamics as a Complex Loop. In: Conte R, Andrighetto G, Campennì M (eds) Minding Norms. Oxford Series on Cognitive Models and Architectures, Oxford University Press

Andrighetto G, Castelfranchi C (forthcoming) Norm compliance: the prescriptive power of normative actions. Paradigmi.

Andrighetto G, Villatoro D (2011) Beyond the carrot and stick approach to enforcement: an agent-based model. In: Kokinov B, Karmiloff-Smith A, Nersessian NJ (eds) European conference on cognitive science. New Bulgarian University Press, Sofia

Andrighetto G, Campennì M, Conte R, Paolucci M (2007) On the immergence of norms: a normative agent architecture. In: Trajkovski GP, Collins SG (eds) Emergent agents and socialities: social and organizational aspects of intelligence, Papers from the AAAI Fall Symposium, Technical Report FS-07-04. Menlo Park, CA: The AAAI Press

Andrighetto G, Campennì M, Cecconi F, Conte R (2010a) The complex loop of norm emergence: a simulation model. In: Deguchi H et al (eds) Simulating interacting agents and social phenomena. Agent-Based Social Systems, Springer, pp 19–35

Andrighetto G, Villatoro D, Conte R (2010b) Norm internalization in artificial societies. AI Commun 23:325–339

Aronfreed JM (1968) Conduct and conscience. The socialization of internalized control over behavior. Academic Press, New York

Axelrod R (1986) An evolutionary approach to norms. Am Politi Sci Rev 80(4):1095–1111

Basso A, Mondada F, Castelfranchi C (1993) Reactive goal activation in intelligent autonomous agent architecture. In Proceedings of AIA'93—First International Round-Table on "Abstract Intelligent Agent", ENEA, Roma, January 25–27

Becker GS (1968) Crime and punishment: an economic approach. J Politi Econ 76(2):169–217

Bicchieri C (2006) The grammar of society. The nature and dynamics of social norms. Cambridge University Press, Cambridge

Bicchieri C, Chavez A (2010) Behaving as expected: public information and fairness norms. J Behav Decis Mak 23(2):161–178

Bicchieri C, Xiao E (2009) Do the right thing: but only if others do so. J Behav Decis Mak 22:191–208

Campennì M, Andrighetto G, Cecconi F, Conte R (2009) Normal = normative? The role of intelligent agents in norm innovation. Mind Soc 8:153–172

Castelfranchi C, Miceli M, Cesta A (1992) Dependence relations among autonomous agents. In: Werner E, Demazeau Y (eds) Decentralized A. I. 3. Elsevier Science Publishers, Amsterdam, pp 215–227

Cialdini RB, Reno RR, Kallgren CA (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 58(6):1015–1026

Cohen PR, Levesque HJ (1990) Persistence, intention, and commitment. In: Cohen PR, Morgan J, Pollack MA (eds) Intentions in communication. MIT Press, Cambridge, pp 33–71

Coleman J (1989) The rational choice approach to legal rules. Chic Kent Law Rev 65:177–191

Coleman J (1990) Foundations of social theory. Belknap, Cambridge

Conte R (1998) L'Obbedienza intelligente Laterza, Bari

Conte R (2009) Rational, goal governed agents. In: Meyers RA (ed) Springer encyclopaedia of complexity and system science. Springer, Berlin

Conte R, Castelfranchi C (1995) Cognitive and social action. UCL Press, London

Conte R, Castelfranchi C (1999) From conventions to prescriptions. Towards an integrated view of norms. Artif Intell Law 7:119–125

Conte R, Castelfranchi C (2006) The mental path of norms. Ratio Juris 19(4):501–517

Conte R, Paolucci M (2002) Reputation in artificial societies. Social beliefs for social order. Kluwer, New York

Conte R, Sichman JS (1995) DEPNET: how to benefit from social dependence. J Math Sociol 20(2–3):161–177

Conte R, Sichman JS (2002) Dependence graphs: dependence within and between groups. Comput Math Organ Theory 8(2):87–112

Conte R, Andrighetto G, Campennì M (eds) (forthcoming) Minding Norms. Oxford Series on Cognitive Models and Architectures, Oxford University Press

Deci EL, Ryan RM (1985) Intrinsic motivation and self-determination in human behaviour. Plenum Publishing Co, New York

Deci EL, Ryan RM (1987) The support of autonomy and the control of behaviour. J Pers Soc Psychol 53:1024–1037

Deci EL, Nezlek J, Sheinman L (1981) Characteristics of the rewarder and intrinsic motivation of the rewardee. J Pers Soc Psychol 40:1–10

Deci EL, Eghrari H, Patrick BC, Leone D (1994) Facilitating internalization: the self-determination theory perspective. J Pers 62:119–142

Durkheim E (1950) The rules of sociological method. The Free Press, Princeton-New York

Epstein JM (2007) Generative social science. Studies in agent-based computational modeling. Princeton University Press, Princeton-New York

Faillo M, Grieco D, Zarri L (2010) Legitimate Punishment, Feedback, and the Enforcement of Cooperation. Working Paper 16/2010. Department of Economics University of Verona

Festinger L (1957) A theory of cognitive dissonance. Row Peterson, Evanston

Gintis H (2003) The hitchhiker's guide to altruism: gene-culture coevolution and the internalization of norms. J Theor Biol 220(4):407–418

Gintis H, Bowles S, Boyd RT, Fehr E (eds) (2005) Moral sentiments and material interests: the foundations of cooperation in economic life. MIT Press, Cambridge

Grossman D (1995) On killing: the psychological cost of learning to kill in war and society. Little, Brown and Company, Boston

Haidt J (2003) Elevation and the positive psychology of morality. In: Keyes CLM, Haidt J (eds) Flourishing: positive psychology and the life well-lived. American Psychological Association, Washington DC, pp 275–289

Harmon-Jones E, Mills J (eds) (1999) Cognitive dissonance: progress on a pivotal theory in social psychology. American Psychological Association, Washington, DC

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. Econometrica 47(2):263–291

Lewis D (1969) Convention. A philosophical study. Harvard University Press, Cambridge

Lotzmann U, Möhring M, Troitzsch KG (2012) Simulating the emergence of norms in different scenarios. Artif Intell Law. doi:10.1007/s10506-012-9130-y

Miceli M, Castelfranchi C (1989) A cognitive approach to values. J Theory Soc Behav 19(2):169–193

Miceli M, de Rosis F, Poggi I (2006) Emotional and non-emotional persuasion. Appl Artif Intell Int J 20(10):849–879

Miller GA, Galanter E, Pribram KH (1960) Plans and the structure of behaviour. Henry Holt, New York

Pittman TS, Emery L, Boggiano AK (1982) Intrinsic and extrinsic motivational orientations: reward-induced changes in preference for complexity. J Pers Soc Psychol 42:789–797

Reykowski J (1982) Development of prosocial motivation: a dialectic process. In: Eisenberg N (ed) The development of prosocial behaviour. New York Academic Press, New York, pp 377–394

Ryan RM (1982) Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J Pers Soc Psychol 43:450–461

Schwartz SH (1977) Normative influences on altruism. In: Berkowitz L (ed) Advances in experimental social psychology, vol 10. Academic Press, New York, pp 221–279

Schwartz SH, Howard JA (1981) A normative decision-making model of altruism. In: Rushton JP, Sorrentino RM (eds) Altruism and helping behaviour. Lawerence Erlbaum, Hillsdale, pp 89–211

Scott J (1971) Internalization of norms: a sociological theory of moral commitment. Prentice-Hall, Englewoods Cliffs

Tangney JP, Stuewig J, Mashek DJ (2007) Moral emotions and moral behavior. Ann Rev Psychol 58:345–372

Ullmann-Margalit E (1977) The emergence of norms. Oxford University Press, Oxford

Villatoro D, Andrighetto G, Conte R, Sabater-Mir J (2011) Dynamic sanctioning for robust and cost efficient norm compliance. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp 414–419

Walster E, Walster GW, Bershcheid E (1978) Equity: theory and research. Allyn and Bacon, Inc., Boston

Widegren O (1998) The new environmental paradigm and personal norms. Environ Behav 30(1):75–100

Xiao E, Houser D (2011) Punish in public. J Public Econ 95:1006–1017