Information Sciences 245 (2013) 218–239

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Human-inspired model for norm compliance decision making

N. Criado [a,*], E. Argente [a], P. Noriega [b], V. Botti [a]

[a] University of Bolton, Deane Rd, Bolton, Greater Manchester BL3 5AB, United Kingdom
[b] Institut d'Investigació en Intel.ligència Artificial, Consejo Superior de Investigaciones Científicas, Campus de la UAB, Bellaterra, Catalonia, Spain

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 29 April 2013
Accepted 13 May 2013
Available online 20 May 2013

Keywords:
Norm compliance
Emotion
BDI agent

ABSTRACT

One of the main goals of the agent community is to provide a trustworthy technology that allows humans to delegate some specific tasks to software agents. Frequently, laws and social norms regulate these tasks. As a consequence agents need mechanisms for reasoning about these norms similarly to the user that has delegated the task to them. Specifically, agents should be able to balance these norms against their internal motivations before taking action. In this paper, we propose a human-inspired model for making decisions about norm compliance based on three different factors: self-interest, enforcement mechanisms and internalized emotions. Different agent personalities can be defined according to the importance given to each factor. These personalities have been experimentally compared and the results are shown in this article.

1. Introduction

One of the main goals of the agent community is to provide a trustworthy technology that allows humans to delegate some specific tasks to software agents. Such purpose requires that software agents consider the legislation, social norms, etc. that regulate the performance of the task that has been entrusted to them.

Humans do not always follow norms. Instead, deliberated and rational violation of norms is a conduct that can be observed in all human societies [9]. Thus, software agents must be able to make decisions about norm compliance similarly to the user that has entrusted the task. Otherwise, the results obtained by the agent would not make sense to the user who might refuse from delegating more tasks to software agents. For these reasons, we consider that developing procedures that allow software agents to make decisions about norm compliance is crucial.

The existing literature has not proposed procedures that allow users to configure their agents to make decisions about norm compliance as users would do. For example, in some works, such as [7,3], the decisions about norm compliance are based on rigid procedures defined off-line by the agent designer and hard-wired on agents. Static procedures assume that it is possible to define off-line which is the best decision in all circumstances. Other works, such as [5,22], propose mechanisms for making on-line decisions about norm compliance. Specifically, these mechanisms consider the effects of violating and obeying norms on the agent goals. However, there are works in the psychology scene [17] that claim that norm compliance is not only explained by rational motivations; i.e., the impact of norms and their enforcement procedures (sanctions and rewards) on the agent's goals. Besides that, there are emotional motivations, such as shame or pride, that sustain norm compliance in human societies. For this reason, we consider that it is necessary to endow software agents with mechanisms for making decisions about norm compliance by balancing between rational and emotional criteria, just as humans do.

With the aim of contributing towards the resolution of this open issue, this article proposes a set of functions that allow agents to determine their willingness to comply with norms according to rational and emotional factors. The way in which

* Corresponding author. Tel.: +44 01204903555.
E-mail addresses: n.criado@bolton.ac.uk (N. Criado), eargente@dsic.upv.es (E. Argente), pablo@iiia.csic.es (P. Noriega), vbotti@dsic.upv.es (V. Botti).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.017

agents take these motivations into account allows modelling different agent personalities. Moreover, we have carried out several experiments for illustrating the performance of these functions and the behaviour exhibited by the different agent personalities. This article is organized as follows: Section 2 describes how the existing literature has faced with the problem of making decisions about norm compliance in software agents; Section 3 introduces the running example used in this paper; Section 4 contains the basic definitions used in this paper; Section 5 describes the functions that allow agents to make decisions about norm compliance; Section 6 describes the experiments that have been carried out; and Section 7 contains conclusions and future works.

2. Related work

Conte et al. defined in [11] a *norm-autonomous agent* as an agent whose behaviour is influenced by norms that are *explicitly represented* inside its mind. The delegation of complex, dynamic and realistic tasks to software agents makes necessary the explicit representation of norms in agent minds. Moreover, agents with an explicit representation of norms are able to belong to different societies, to communicate norms and to reason about them [22]. Therefore, *norm-autonomous agents* should have capabilities for *acquiring* norms; i.e., agents should be capable of recognizing the norms that are in force in their environment [3]. Moreover, norm-autonomous agents may have motivations to *accept* these recognized norms [22]. Besides that, norm-autonomous agents should be endowed with capabilities for determining whether a norm concerns their case and it is *relevant* [21]. After the recognized norm has been accepted and considered as relevant, then this kind of agents must decide whether or not to conform to it. This decision about obeying or violating a norm is known as *norm compliance* decision. Next, we describe how the existing literature deals with this decision.

In [10], Castelfranchi et al. describe how an agent architecture can be extended with an explicit norm notion. This work only proposes the architecture for norm-autonomous agents (i.e., the logical connections among norms and mental states) and details which tasks and deliberations should be carried out by agents. However, the authors do not provide agents with concrete mechanisms for carrying out these tasks and deliberations. Specifically, the authors point out the need to create mechanisms for making decisions about norm compliance, but they do not specify how these mechanisms can be defined.

Dignum et al. have proposed in [15] an extension of the classic BDI architecture for considering norms. In this proposal, agents are capable of representing norms, determining when norms are active (i.e., relevant), and resolving conflicts among norms and existing intentions. Conflicts are solved by means of static preference orders. The authors claim that these orders can be defined by considering the social benefit of norms and the cost of fulfilling and violating obligations. However, they do not provide specific details about how these orders could be computed.

The work of Boella and Lesmo in [5] also proposes considering the consequences of obeying norms (i.e., the cost of norm fulfilment) and violating norms (i.e. the cost of sanctions) to make decisions about norm compliance. An important contribution of this work is that norm enforcers are considered as autonomous agents that have their own motivations and limited capabilities for detecting violations and applying sanctions. Thus, agents may have different motivations for violating norms: material impossibility, conflicts with other goals, the possibility of violating the norm without being detected, or the possibility of not being sanctioned. The decision about norm compliance is carried out by a utility function that takes care of all the above-mentioned factors. However, this paper does not provide any information about how the behaviour of the norm enforcer agent is modelled and how this information is used in the definition of the utility function.

The aforementioned proposals have made an important contribution by pointing out the main requirements for norm-autonomous agents. In addition, they provide intuitive ideas and recommendations to meet these requirements. Specifically, they identify the type of information that can be useful for making decisions about norm compliance. However, enough details about how this information is obtained, maintained and combined to implement algorithms that allow agents to make decisions about norm-compliance are not provided.

More recent works have also confronted with the development of norm-autonomous agents. These proposals are classified next into norm-oriented or goal-oriented according to the priority that agents give to norms with respect to their internal goals.

2.1. Norm-oriented agents

Norm-oriented agents have as main purpose the fulfilment of norms above the achievement of their internal goals.

In [21], Kollingbaum has presented the *noA* proposal. It is a practical agent architecture with an explicit notion of obligation and prohibition. In this proposal, obligations are the agents' motivations, whereas prohibitions restrict the choices of activities that agents can ideally employ. Thus, noA agents are norm-oriented agents that do not have internal motivations. Therefore, they will always try to fulfil all norms. Norm conflicts are the main cause of norm violation. Thus, this work does not consider the autonomous decision about norm compliance. In contrast, Kollingbaum's work is focused on the definition of algorithms and procedures for detecting and resolving norm conflicts.

Another example of norm-oriented agent are *Normative KGP* agents, which are described in [25]. This proposal consists in extending KGP *(Knowledge-Goal-Plan)* agents [20] with explicit normative notions such as obligations, prohibitions, and roles. KGP agents have internal motivations or goals. However, in case of a conflict between norms and goals, agents will always follow the behaviour specified by norms. Priority functions are used for solving possible conflicts among beliefs,

goals, intentions, and norms. However, this work has not proposed any mechanism for making decisions about obeying conflicting norms. Therefore, KGP are not autonomous to decide which norms the agent wants to comply with.

The architecture proposed by *Gaertner* in his thesis [19] is also an example of norm-oriented agent in which all norms are blindly followed. These norms are translated into intentions. These new intentions might be in conflict with the previous ones. As a solution to this problem, Gaertner proposes the use of an argumentation-based approach and a preference function.

2.2. Goal-oriented agents

In contrast to norm-oriented agents, goal-oriented agents have the main purpose of achieving their desires while trying to fulfil norms. Thus they have the capability of deciding about norm compliance.

In [7], Broersen et al. propose an extension of the BDI architecture with an explicit notion of obligation, known as *BOID*. This is one of the first proposals on norm-autonomous agents that describes how these agents can be designed in practise. Thus, BOID agents are formed by four *components* that are associated with Beliefs, Obligations, Intentions and Desires. Obligations are the external motivations of agents and their validity is taken for granted. In this proposal, agents can violate norms only due to a conflict. This type of conflicts is solved by means of a static ordering function that resolves conflicts between components and within components. According to the definition of these ordering functions, different types of agents can be defined. For example, agents in which the overruling order is B-O-I-D (i.e., beliefs over obligations, obligations over intentions and intentions over desires and intentions) give more priority to obligations than to their internal motivations (desires) and blindly obey norms without considering their intentions. Agents can be goal-oriented or norm-oriented according to the definition of the ordering function, but this ordering function is hard-wired on agents and cannot be modified.

One of the first proposals on goal-oriented agents that has explicitly considered the norm compliance dilemma is [22]. In this work, López y López et al. have proposed an agent architecture for developing norm-autonomous agents capable of interacting in norm-governed environments. Agents are autonomous for pursuing their own goals even if these goals violate the norms; i.e., agents are autonomous to come to a decision on norm compliance. For this reason, this work includes the notion of sanctions and rewards to persuade agents to follow the norms. In this work, López y López et al. have developed different strategies to allow agents to make decisions about norm compliance assuming that there is a material system of sanctions and rewards. Examples of these strategies are: *pressure*, norms with harmful sanctions are obeyed; *opportunistic*, only norms that are beneficial to the agent are respected; *fear*, all norms with sanctions are observed; *greedy*, norms whose fulfilment is rewarded are followed; and *rebellious*, no norm is respected. This work represents an important step towards the development of norm-autonomous agents capable of making flexible decisions about norm compliance. However, the deliberation about norm compliance is only based on the existence of an external mechanism of norm enforcement. Therefore, in absence of information about the enforcement mechanisms agents have no motivation to comply with norms. This proposal does not explain how agents comply with norms regardless of the existence of an enforcement system.

In all of the above-described proposals, norms are off-line programmed on agents [7] or agents are on-line informed by authorities about norms [22]. Therefore, agents are not capable of learning new norms on-line and adapting their behaviours according to these unforeseen norms. In relation with this feature, the *EMIL* proposal [2] has developed a framework for autonomous norm recognition. EMIL agents obey all recognized norms blindly without considering their own motivations. In a later work [3], the EMIL proposal has been extended for allowing agents to make decisions about norm compliance and to internalize norms. The decision about norm compliance is made considering the expected utility that agents should obtain if they fulfil or violate the norm.

In a more recent work, Fagundes et al. [18] presents a model for rational self-interested agents, which takes into account the possibility of violating norms. These agents are endowed with mechanisms for acquiring norms from a norm repository. The decisions about norm compliance are made by taking into account the expected utility of the states that would result if norms are violated and obeyed.

2.3. Discussion

Table 1 compares the main proposals on norm-autonomous agents described in this section. Specifically, this table illustrates performance of the proposed norm-autonomous agents with respect to their capabilities for: reasoning about norm compliance. The decision on complying with a norm implies that this normative goal has been selected between internal goals and other normative goals. Thus, the decision on norm compliance subsumes the resolution of conflicts among mental propositions and norms. However, we would like to make a difference between those works that consider conflicts as the only cause for norm violations and those ones that consider the fact that norms can be deliberately violated in the absence of a conflict with another mental attitude. Therefore, Table 1 has two different columns, labelled as *norm compliance* and *conflict resolution*, in order to point out how the norm compliance dilemma is considered.

This table makes a summary of the proposals reviewed by this section. Works of Castelfranchi et al. in [10], Dignum et al. in [15] and Boella and Lesmo' in [5] are more intuitive than formal. Specifically, these works specify which the main requirements for norm-autonomous agents are and which kind of information must be taken into account to meet these requirements. However, they do not propose any specific solution to meet these requirements. For example, Boella and Lesmo' in [5] have faced with the norm compliance problem by proposing the definition of a static utility function that consider the cost of

Table 1
Summary of proposals on norm-autonomous agents.

	Conflict resolution	Norm compliance
Castelfranchi et al. [10]	√	
Dignum et al. [15]	√	
Boella and Lesmo [5]		√
NoA [21]	√	
Normative KGP [25]	√	
Gaertner [19]	√	
BOID [7]	√	
López y López et al. [22]		√
EMIL [3]		√
Fagundes et al. [18]		√

obeying a given norm and the possibility of being sanctioned. However, enough details concerning how this utility function can be defined are not provided. Later works have tried to close the gap between these intuitive ideas and more specific frameworks that allow the practical implementation of agents built upon these ideas. Some of them developed norm-oriented agents by omitting the agents' autonomy. Thus, they confront the problem of resolving conflicts among norms and other mental attitudes. In contrast, there are works that have faced with the agents' autonomy by using static mechanisms; e.g., the BOID architecture [7], which defines a static priority order among mental attitudes that is programmed on agents. These static mechanisms are suitable for controlled environments in which agents confront with foreseeable situations. However, complex scenarios in which agents should dynamically adapt require more flexible solutions to the norm compliance dilemma. As argued in [10] "if protocols that agents use to react to the environment are fixed, they have no ways to respond to unpredictable changes". Thus, they entail a limitation on the agent capacities for adapting to new societies or to the environmental changes. The work of López y López et al. [22], have explicitly proposed mechanisms for allowing agents to make a decision about obeying or violating a given norm at a specific moment by analysing the consequences of norms in terms of their economic cost. As being argued by López y López, in their proposal compliance with norms is only sustained by a material system of sanctions and rewards. Obviously, sanctions and rewards are one of the main motivations of agents when deciding to follow a norm. However, there are norms whose compliance is neither sanctioned nor rewarded. Therefore, there may be other explanations for norm compliance, such as emotions [17], that have not been considered by the existing works.

The present article represents a step towards the definition of flexible decision mechanisms for norm compliance. Our aim is to provide humans with reliable agents to which they can delegate tasks that are regulated by legal and social norms. Thus, humans may be liable for the agent's acts in front of the law. As a consequence, agents must be endowed with mechanisms that allow them to reason at execution time about the violation and fulfilment of norms similarly to their human user. Specifically, this paper details how rational and emotional reasons that explain compliance with norms in humans can be implemented in software agents. The decisions on norm compliance made by these software agents are expected to be more robust and reliable since norms are not only conducted by rational motivations but also by emotional motivations [3].

3. Running example

Over the course of this paper we will use an example to illustrate the human-inspired model for norm compliance decision making. This example consists of a software agent, called *assistant*, that draws up traffic routes according to the preferences that a human user has specified. These preferences may include time constraints, consumption requirements, avoidance of toll roads, and so on. Therefore, the routes suggested by the *assistant* agent indicate not only the particular ways or directions but also the speed at which the human should drive at each stretch for meeting the user requirements.

In order to calculate the most suitable route according to the user's preferences the *assistant* agent needs to know which are the norms that regulate the traffic in each region. If the suggested routes do not take into account norms, then the user that follows this route may be arrested and accused of a serious offence. Therefore, this scenario makes mandatory that software agents consider norms. These norms include both those formal norms that are defined explicitly in highway codes and those informal (i.e., social) norms that explain the attitude of the national population towards formal laws. There are some studies, such as the one made in [4], sustaining the hypothesis that social norms or national culture are more important than formal laws in the attitude and behaviour of the driver population. As a consequence, the same traffic norms do not have the same effect when they are applied in different countries.

The *assistant* agent cannot consider norms as hard-constraints that must be always respected since in some situations it may be not possible to find a traffic route that meets all the requirements and respects all norms. Even if this extreme situation does not occur, the user may be interested in taking advantage of violating some norms whenever possible (e.g., to arrive sooner). If the *assistant* agent considers norms as soft-constraints, then the user would have to program the *assistant* agent by checking all the norms and specifying its willingness to comply with each of them. Thus, the user would still make

the decisions about norm compliance. Moreover, the norms may change from region to region and along time. This entails that the user would need to re-program the *assistant* agent regularly. In this case, what is more desirable is to endow the *assistant* agent with mechanisms for finding an equilibrium point between the user requirements, which are the internal motivations; and the norms, which are the external motivations. Thus, the user could completely delegate the decisions about norm compliance and the selection of the route to the *assistant* agent. To perform this task successfully, the *assistant* agent must be able to make decisions about norm compliance similarly to its user; i.e., considering the impact of norms on the user goals and the emotional repercussions of norms. This case study will allow us to illustrate how our proposal allows software agents to consider and reason about norm compliance.

4. Preliminaries

The purpose of this paper is not to propose, compare or improve existing norm or agent models, but to make use of these models and propose a set of functions that can be used for making decisions about norm compliance based on rational and emotional criteria. The aim of this section is to provide the reader with the basic notions of norm, instance and normative agent that are used in this paper.

4.1. Norm definition

Several works on the existing literature on agents and norms (such as [12,22]) make a distinction between norms and instances. Following this distinction, *norms* define patterns of behaviours by means of *deontic modalities*: *obligations*, which define actions or goals that should be performed or satisfied by agents; *prohibitions*, which define actions or goals that should not be performed or achieved; and *permissions*, which define exceptions to the application of a more general norm of obligation or prohibition. Therefore, norms in our approach define a pattern of behaviour (or *norm condition* in our terminology) as obligatory, prohibited or permitted. In general, norms are not applied all time but include the notions of activation and expiration conditions. The *activation condition* defines when obligations, permissions and prohibitions must be instantiated and fulfilled by all agents under the influence of the norm. Instances remain active, even if the activation condition ceases to hold. The *expiration condition* defines the validity period or deadline of the instance. Finally, our norm notion also includes information about the enforcement mechanisms: *sanctions*, to punish agents which do not obey the norm and *rewards*, for rewarding norm fulfilment.

Definition 1 (*Norm*). A norm is defined as a tuple $\langle \Delta, C, A, E, S, R \rangle$, where:

- $\Delta \in \{\mathcal{O}, \mathcal{F}, \mathcal{P}\}$ is the deontic modality of the norm, determining if the norm is an obligation ($\mathcal{O}$), prohibition ($\mathcal{F}$) or permission ($\mathcal{P}$);
- C is a *wff* of $\mathcal{L}$[1] that represents the norm condition;
- A, E are *wff*s of $\mathcal{L}$ that describe the activation and expiration conditions, respectively;
- S, R are *wff*s of $\mathcal{L}$ that describe the sanction and reward, respectively.

In the proposed case study, it is very usual that there are traffic laws that try to prevent accidents. For example, a norm that obliges drivers to slow down when there is heavy rain in some area (A) is represented as follows:

$$\langle \mathcal{O}, slow(A), heavyRain(A), \neg heavyRain(A), fine(A), - \rangle \quad \text{(Heavy Rain Norm)}$$

Thus, the *assistant* agent knows that the speed must be reduced in routes that cross an area where there is heavy rain, or it may be fined.

4.2. Instance definition

An instance is an unconditional expression that binds a particular agent to an obligation, permission or prohibition. Any instance is created out of a norm once the activation condition holds according to the grounding of the activation condition as follows:

Definition 2 (*Instance*). Given a norm $n = \langle \Delta, C, A, E, S, R \rangle$ and a theory Γ of $\mathcal{L}$, an instance of n is the tuple $n_i = \langle \Delta, C', A', E', S', R' \rangle$, where:

- $\Gamma \vdash \sigma(A)$, where σ is a substitution of variables in A such that $\sigma(A)$, $\sigma(E)$, $\sigma(C)$, $\sigma(S)$, $\sigma(R)$ are grounded;
- $A' = \sigma(A)$, $E' = \sigma(E)$, $C' = \sigma(C)$, $S' = \sigma(S)$ and $R' = \sigma(R)$;

[1] $\mathcal{L}$ is a first-order predicate language whose alphabet includes: the logical connectives $\{\wedge, \vee, \neg, \rightarrow\}$; parentheses, brackets, and other punctuation symbols; and an infinite set of variables. Variables are universally quantified implicitly. Along this paper variables are written as any sequence of alphanumeric characters beginning with a capital letter. In addition, the alphabet contains non-logical predicate, constant and function symbols, which will be written as any sequence of alphanumeric characters beginning with a lower case.

For simplicity, we assume that once a norm is instantiated it is grounded. To ensure that all instances have not free variables, all variables that occur in E, C, S, R may be contained in A (i.e., $v_A \supseteq v_E \cup v_C \cup v_S \cup v_R$[2]).

Suppose that the *assistant* agent is informed by a meteorological server that there is heavy rain in a specific area $a1$. Thus, it knows that the Heavy Rain Norm has come into effect in area $a1$ and it is instantiated as follows:

$$\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), - \rangle \quad \text{(Heavy Rain Instance)}$$

4.3. Normative agent definition

A normative agent in this paper is defined as a practical reasoning agent [6] whose actions are directed towards its internal goals and the norms that regulate its environment. Specifically, this paper focuses on how a normative agent makes decisions about norm compliance on behalf of its user. To make such kind of decisions a normative agent considers the current circumstances of its user; i.e., the beliefs about the world in which its user is placed; and the user's requirements; i.e., the desires of its user. Besides that, we want that our normative agents can perform practical reasoning in a dynamic and uncertain world. For these reasons, the Graded BDI architecture, which has been described in detail in [8], is used in this paper. Specifically, this architecture has an explicit representation of graded mental attitudes, such as graded beliefs and desires, and fits perfectly the purpose of our paper.

We define a Normative BDI agent by extending the Graded BDI architecture with an explicit representation of norms and instances as follows:

Definition 3 (*Normative BDI agent*). A Normative BDI agent is defined as a tuple $\langle B, D, I, N, N_I \rangle$, where:

- B is the set of graded beliefs. This set is composed of (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0,1]$ represents the certainty degree associated to this proposition. The logical connective $\rightarrow$ is used to represent explanation and contradiction relationships between propositions. Thus, $(\alpha \rightarrow \beta, \rho)$ represents that the agent believes that α explains β, with a certainty degree ρ. Similarly, expressions such as $(\alpha \rightarrow \neg\beta, \rho)$ mean that the agent believes that proposition α contradicts proposition β, with a certainty degree ρ.
- D is the set of graded desires. This set is also composed of (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0,1]$ represents the desirability degree associated to this proposition. Negative desires are represented using the negation connective $\neg$ (i.e., $(\neg\gamma, \rho)$). Degrees of desires allow setting different levels of preference or rejection.
- I is formed by expressions such as (γ, ρ) expressions, where γ is a grounded formula of $\mathcal{L}$; and $\rho \in [0,1]$ is the intentionality degree of proposition γ. This intentionality degree is the consequence of finding a best feasible plan that achieves a state of the world where γ holds.
- N is the set of norms that affect the agent. It is composed of (n, ρ_s) expressions, where n is a norm, and $\rho_s \in [0,1]$ is a real value that assigns a salience degree to this norm. This salience represents the importance that the agent believes that the society gives to each norm.
- N_I is the set of instances that have been created out of N according to the agent beliefs B. N_I is composed of (n_i, ρ_r) expressions, where n_i is an instance, and $\rho_r \in [0,1]$ is a real value that defines the relevance of the instance. This relevance represents the degree in which the instance concerns the agent.

Thus, the sets B, D and I contain the cognitive elements, whereas the sets N and N_I contain the normative elements.

The calculation of the salience of norms and the relevance of instances is beyond the scope of this paper. However, we can assume that the salience of norms has been defined by the designer of the *normative system* [1], who determines the importance of norms in the norm hierarchy. In addition, we assume that each agent calculates the relevance of instances by considering the validity of the instance according to its current situation [12]; i.e., the degree to which the instance is pertinent to its current situation.

In the proposed example, the *assistant* agent is defined as a normative agent $\langle B, D, I, N, N_I \rangle$ where:

- B is formed by propositions that represent its beliefs about the world in which it is situated; i.e., the agent perceptions. For example, it has beliefs about the climatological conditions or the traffic congestion level. Moreover, the *assistant* agent knows explanation relationships between beliefs. These relationships allow the *assistant* agent to represent the potential consequences of actions or states. For example, the *assistant* agent knows that as the driving speed increases the braking distance also increases.
- D is formed by propositions that represent the user preferences; i.e., his/her intrinsic goals.
- In our proposal intentions are not considered as a basic attitude. Thus, the intentions of the *assistant* agent are generated on-line from the agent's beliefs and desires. See [8] for an explanation of the intention generation process.

[2] v_X is the set of variables occurring in any formula X.

- Suppose that the *assistant* agent is informed about the Heavy Rain Norm by a traffic authority. The traffic authority also has also informed the *assistant* agent that this norm has the highest priority in the highway code. As a consequence, the salience of this norm is set to 1 and the set N contains a proposition such as:

$$(\langle \mathcal{O}, slow(A), heavyRain(A), \neg heavyRain(A), fine(A), -\rangle, 1)$$

- As previously mentioned, a meteorological server informs the *assistant* agent that there is a heavy rain in a specific area $a1$ with a 75% of probability. For simplicity, we assume that the *assistant* agent calculates the relevance of the new instances as the certainty of the activation condition. Thus, the relevance of the instance is *0.75* and the set N_I contains a proposition such as:

$$(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, 0.75)$$

For the purpose of this paper it is only necessary to know that in Normative BDI agents the information flows from perception to action according to three main steps (see Fig. 1). Firstly, the agent *perceives* the environment and updates its beliefs, norms and instances (see Fig. 1a). Secondly, in the *deliberation* step, the desire set is revised (see Fig. 1b). For example, new desires may be created from the user preferences. Similarly, desires that have been achieved must be dropped. At this step the agent considers the set of instances that are relevant and makes a decision about which ones it wants to obey (this reasoning process is labelled as *norm reasoning* in Fig. 1b). As a result, new desires are created for fulfilling and/or violating norms. Finally, in the *decision making* step, desires help the agent to select the most suitable plan to be intended (see Fig. 1c).

In this paper we only focus on how a Normative BDI agent makes decisions about norm compliance. Other problems such as the reasoning about graded mental propositions [8], the acquisition of norms [3], or the instantiation of norms [12] have been addressed in other works.

5. Reasoning about norms

In the running example used in this article, the *assistant* agent must decide to what extent instances are observed or violated in the proposed routes (*deliberation* step). The mechanism responsible for this task should consider the salience of norms, the relevance of each concrete instance and the user preferences. This section proposes the rules and functions that allow Normative BDI agents to face with this complex issue.

The process by which agents extend their mental state according to their decisions about norm compliance (i.e., according to the instances that they want to follow or transgress) is what we name *norm reasoning*. As claimed in [3], usually normative elements (i.e., instances) generate new desires. Accordingly, we have considered the translation of instances into desires. For each instance in the N_I, the agent makes a decision about norm compliance (i.e., it calculates its willingness to comply with each instance) and updates its desire set accordingly. Thus, the two main problems tackled by our work are: the problem of making decisions about norm compliance; and determining the desirability of the desires created according to this decision. Depending on the desirability degree of these instance-based desires, they may generate new intentions to be executed or

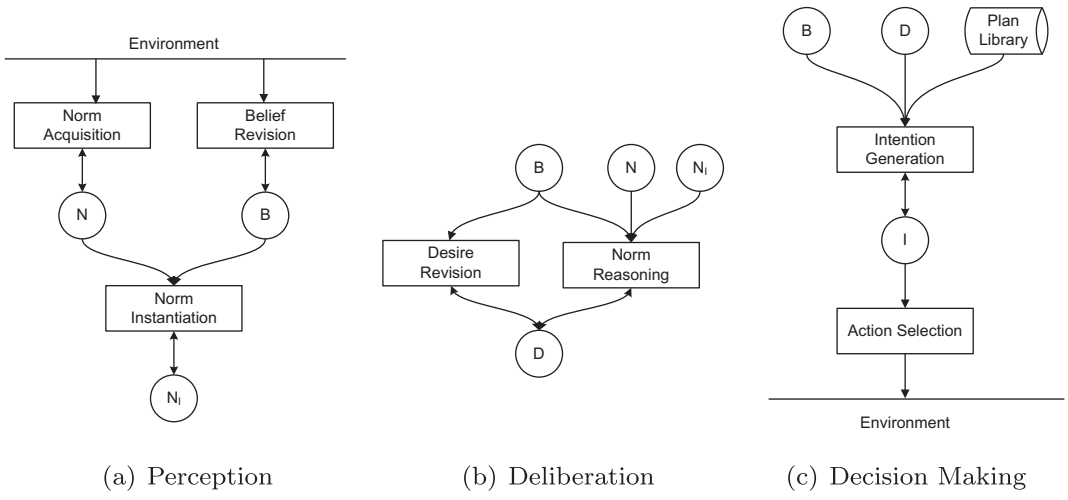


Fig. 1. Reasoning Phases in a Normative BDI Agent. The sets that contain the cognitive and normative elements (i.e., the sets B, D, I, N, N_I) are represented as circles. The reasoning processes are represented as boxes where: the input links represent the information used by the reasoning process, and the output links represent the information updated by the reasoning process.

Table 2
Operational rules of norm reasoning.

$\frac{(\langle \mathcal{O}, C', A', E', S', R' \rangle, \rho_r) \in N_I \wedge f_w(\langle \mathcal{O}, C', A', E', S', R' \rangle, B, D, N) > \theta}{\langle B, D, I, N, N_I \rangle \rightarrow \langle B, D^*, I, N, N_I \rangle}$ where $D^* = D \cup \{(C', f_i(\rho_r, \lvert f_w(\langle \mathcal{O}, C', A', E', S', R' \rangle, B, D, N)\rvert)\}$	(a)
$\frac{(\langle \mathcal{O}, C', A', E', S', R' \rangle, \rho_r) \in N_I \wedge f_w(\langle \mathcal{O}, C', A', E', S', R' \rangle, B, D, N) < -\theta}{\langle B, D, I, N, N_I \rangle \longrightarrow \langle B, D^*, I, N, N_I \rangle}$ where $D^* = D \cup \{(\neg C', f_i(\rho_r, \lvert f_w(\langle \mathcal{O}, C', A', E', S', R' \rangle, B, D, N)\rvert)\}$	(a*)
$\frac{(\langle \mathcal{F}, C', A', E', S', R' \rangle, \rho_r) \in N_I \wedge f_w(\langle \mathcal{F}, C', A', E', S', R' \rangle, B, D, N) > \theta}{\langle B, D, I, N, N_I \rangle \rightarrow \langle B, D^*, I, N, N_I \rangle}$ where $D^* = D \cup \{(\neg C', f_i(\rho_r, \lvert f_w(\langle \mathcal{F}, C', A', E', S', R' \rangle, B, D, N)\rvert)\}$	(b)
$\frac{(\langle \mathcal{F}, C', A', E', S', R' \rangle, \rho_r) \in N_I \wedge f_w(\langle \mathcal{F}, C', A', E', S', R' \rangle, B, D, N) < -\theta}{\langle B, D, I, N, N_I \rangle \rightarrow \langle B, D^*, I, N, N_I \rangle}$ where $D^* = D \cup \{(C', f_i(\rho_r, \lvert f_w(\langle \mathcal{F}, C', A', E', S', R' \rangle, B, D, N)\rvert)\}$	(b*)

they may be used to select the most suitable plan that achieves another goal that is more desired.[3] Norm reasoning rules depend on the deontic modality of the instance that is considered:

- *Obligations.* If the agent is willing to comply with an obligation, then a desire for reaching the goal imposed by the obligation is created (see Rule (a) in Table 2). Specifically, the f_w function calculates the agent willingness to comply with a given instance as a real value within the $[-1,1]$ interval. When it takes a value greater than a compliance threshold θ ($\theta \in [0,1]$), it means that the agent is willing to comply with the instance. If f_w takes a value lower than $-\theta$, then it means that the agent wants to violate the instance deliberately. An agent may decide to violate deliberately an obligation when it believes that this would entail good consequences (e.g., usually women drivers decide to violate the seat belt obligation when they are pregnant, since they believe that it is more safe for the baby). In this case, a new desire to violate the norm is created (see Rule (a*) in Table 2). The degree assigned to the new desires is calculated by the f_i function.
- *Prohibitions.* If the instance is a prohibition and the agent wants to obey it (i.e., when f_w takes a value higher than θ), then a negative desire is created to avoid the forbidden goal (see Rule (b) in Table 2). Similarly, Rule (b*) in Table 2 creates a desire for achieving the forbidden goal if the agent decides to violate the prohibition (i.e., when f_w takes a value lower than $-\theta$).
- *Permissions.* Finally, permissions do not infer positive or negative desires about the norm condition. In this proposal, we assume a closed world assumption, where everything is considered as permitted by default. Therefore, permissions define exceptions to the application of more general obligation and prohibition norms. As a consequence, they are only defined for creating a conflict with these more general norms. The resolution of conflicts among norms is beyond the scope of this paper and has been addressed by other works [21].

The rules contained in Table 2 explain how the new instances are considered for extending the agent mental state. Specifically, these rules create desires for fulfilling instances (i.e., when f_w takes a value higher than θ) and violating instances (i.e., when f_w takes a value lower than $-\theta$). When the value of f_w belongs to the $[-\theta, \theta]$ interval, then the agent is indifferent to the instance (i.e., the willingness is not enough to change the agent behaviour) and the agent decides to ignore the norm. Thus, no operational rule is executed and no desire is created in this case.

The relevance of instances changes along their life cycle (e.g., when a driver goes out of a tunnel the relevance of an instance of a norm that obliges to keep the car lights on decreases). As a consequence, instances must be reconsidered several times until they expire. Thus, replicated and/or contradictory desires about the norm condition with different desirability degrees may be inserted into the D set. Notice that Normative BDI agents carry out a desire revision process[4] that is responsible for maintaining and updating the desire set, taking control about replicated desires, inconsistent desires, etc.

As shown in the operational rules of Table 2, there are two key functions for the norm reasoning process: the *willingness* function (f_w) that calculates the willingness to comply with an instance, and the *internalization* function (f_i) that considers this willingness to assign a degree to the new instance-based desire. Next, the definition of f_i and f_w is explained.

5.1. Internalization function

The degree assigned to the instance-based desires is defined by the f_i function, which combines the relevance of the instance (ρ_r) and the motivation to comply with this instance (ρ_w) as a real value within the $[0,1]$ interval. Both conditions, the relevance of the instance and the motivation to comply with it, are required for creating a new desire. For example, if an instance has expired and it has no longer effect (i.e., ρ_r is 0), the desirability of any desire inferred from this instance must be *0* regardless of the agent willingness to comply with the instance. Moreover, the higher the relevance or the willingness is, the higher the desirability must be. Specifically, if both ρ_r and ρ_w take high values, then the new desire must have a degree

[3] The integration of the reasoning about desires with planning [28] has not been considered by our work.
[4] See [8] for more details about the desire revision process.

higher than ρ_r and ρ_w.[5] As a consequence, the combination among the uncertain values that cause the norm internalization is defined as a symmetric sum [16].

Definition 4 (*Internalization Function*). Let $\rho_r \in [0,1]$ be the relevance of an instance, and let $\rho_w \in [0,1]$ be the willingness to comply with this instance. The internalization function f_i: $[0,1] \times [0,1] \to [0,1]$ is given by:

$$f_i(\rho_r, \rho_w) = \frac{\rho_r * \rho_w}{1 - \rho_r - \rho_w + (2 * \rho_r * \rho_w)}$$

Remark. The internalization function f_i is a function such that: $f_i(1,1) = 1$ (i.e., when the instance is completely relevant and the agent is completely willing to comply with it, then the desirability of the instance-based desire is the highest); f_i has as null element 0 (i.e., when an instance is not relevant, then the desirability is *0*); f_i is increasing with respect to both arguments and continuous (i.e., the higher the relevance or the willingness is, the higher the desirability is). The internalization function is a variable aggregation operator (i.e., the behaviour of the operator depends on the values that are combined): when both $max(\rho_r, \rho_w) < 0.5$, the f_i function behaves as a conjunctive fusion operator (i.e., $f_i(\rho_r, \rho_w) \leqslant min(\rho_r, \rho_w)$); when both $min(\rho_r, \rho_w) > 0.5$, f_i behaves as a disjunctive operator (i.e., $f_i(\rho_r, \rho_w) \geqslant max(\rho_r, \rho_w)$); otherwise, f_i provides a combined result which is a compromise between ρ_r and ρ_w.

5.2. Willingness function

As explained before, we are looking for designing and implementing agents much closer to actual human behaviours when deciding about norm compliance. Humans make decisions by balancing their internal motivations (i.e., their own desires) against other external motivations (e.g., social norms or laws). However, each person has his/her own personality; i.e., each person weights up these factors differently. Next, our human-inspired solution for making decisions about norm compliance is described.

The results calculated by the f_w function represent the agent willingness to comply with norms; i.e., it models the decisions about norm compliance. To calculate this willingness we have mainly considered the works of Elster [17] that analyse factors that sustain norms in human societies. In these works, Elster claims that compliance with norms can be explained by three factors: (i) *self-interest* motivations, which consider the influence of the fulfilment of norms on agent's goals; (ii) the *expectations* of being rewarded or sanctioned by others; and (iii) *emotional* factors that are related to internalized emotions such as honour (vs. shame) and hope (vs. fear). These three factors are modelled by three different functions that we denote as f'_w,f''_w and f'''_w, respectively. The agent's willingness to follow a concrete instance is calculated by the f_w function, which is defined as a weighted average.

Definition 5 (*Willingness Function*). Given an instance n_i, a set of beliefs B, a set of desires D and a set of norms N; the willingness function f_w is given by:

$$f_w(n_i, B, D, N) = \frac{w' \times f'_w(n_i, D) + w'' \times f''_w(n_i, D) + w''' \times f'''_w(n_i, B, D, N)}{w' + w'' + w'''}$$

where the weights w', w'' and w''' are defined within the [0,1] interval.

Remark. The value calculated by the willingness function f_w is a real value within the $[-1,1]$ interval that has been obtained combining the values of the three *willingness functions* (f'_w,f''_w and f'''_w) which are also defined within the $[-1,1]$ interval.

We assume that the weighted average is a suitable method to derive the central tendency of these three functions. The weights that each agent gives to these factors characterize the agent's personality and do not depend on the instance that is considered. In Section 6, we compare experimentally the main agent types that result from giving different values to w', w'' and w'''.

5.2.1. Self-interest (f'_w)

f'_w evaluates the agent interest from a utilitarian perspective; i.e., the utility is the good to be maximized. The utility of an instance is defined by considering the direct positive or negative consequence of the instance fulfilment. In case of an obligation, the direct consequence of the fulfilment of the obligation is the norm condition (i.e., C'). In case of a prohibition, obeying this prohibition implies that the norm condition will be avoided (i.e., $\neg C'$).

Definition 6 (*Self-interest*). Given an instance $\langle \Delta, C', A', E', S', R' \rangle$ and a set of desires D; the self-interest function f'_w is given by:

[5] Similarly, if there is a low relevance and willingness, then the new desire must have a degree lower than the relevance and willingness.

$$f'_w(\langle \Delta, C', A', E', S', R' \rangle, D) = \begin{cases} f_d(C', D) & \text{if } \Delta = \mathcal{O} \\ f_d(\neg C', D) & \text{if } \Delta = \mathcal{F} \end{cases}$$

where the function f_d (explained below in Definition 7) calculates the desirability of a proposition.

Remark. The desirability of a proposition β is a real value within the $[-1,1]$ interval such that: the -1 value means that the proposition β is absolutely rejected, a desirability value of 0 means that the agent is indifferent to β, and 1 means that the agent has maximum preference on β.

Definition 7 (*Desirability*). Given a proposition $\beta \in \mathcal{L}$ and a set of desires D; the desirability of β is defined as:

$$f_d(\beta, D) = f_{\oplus}(\beta, D) - f_{\ominus}(\beta, D)$$

where $f_{\oplus}(\beta,D)$ and $f_{\ominus}(\beta,D)$ are the positive and negative support of β, respectively.

Remark. The desirability of a proposition β is calculated by considering the positive and negative support of β; i.e., the degree of the desires hindered and favoured by β. Specifically, the positive support of a proposition β (i.e., $f_{\oplus}(\beta,D)$) is defined as the average among the desirability of all the desires that are favoured by β; i.e., desires $(\gamma, \rho_\gamma) \in D$ such that $\gamma \vdash \beta$. If there is not any desire that is favoured by β, then the positive support of β is *0*. Similarly, the negative support of a proposition β (i.e., $f_{\ominus}(\beta,D)$) is defined as the average among the desirability of all the desires that are hindered by β; i.e., desires $(\gamma, \rho_\gamma) \in D$ such that $\gamma \vdash \neg\beta$.

In the proposed case study, the *assistant* agent should make a decision about complying or not with the instance of the Heavy Rain Norm. Let us suppose that the human user has a new and fast car. He likes to show off the power of his new car and he has configured the *assistant* agent with this preference. Since area *a*1 is a crowded place, the human user has defined that he wants to pass across the area *a*1 as fast as possible. As a consequence, the *assistant* agent has a desire as the following $(\neg slow(a1), 0.9)$. Therefore, the interest on obeying this instance is the following:

$$f'_w(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), - \rangle, D) = f_d(slow(a1), D) = f_{\oplus}(slow(a1), D) - f_{\ominus}(slow(a1), D)$$
$$= 0 - \frac{0.9}{1} = -0.9$$

5.2.2. Expectations (f''_w)

f''_w models the impact of the external enforcement on agents. Specifically, the enforcement mechanism considered in this work consists in a material system of sanctions and rewards that modifies the utility that agents obtain when they violate or fulfil norms. According to this, f''_w considers how much the agent expects to lose from being penalized and how much it expects to gain from being rewarded. Since the fulfilment of any norm implies that the agent will be rewarded and not sanctioned, $f_{w''}$ is defined as the combination of the desirability of R' and $\neg S'$ (i.e. $f_d(R',D)$ and $f_d(\neg S', D)$). Specifically, we want that if both R' and $\neg S'$ are desirable to the agent, then the value of the expectation is higher than the desirability of R' and $\neg S'$.[6] Therefore, we have applied the MYCIN rules [26] for combining the desirability of the two consequences of norm fulfilment. Thus, the f''_w factor is defined as follows:

Definition 8 (*Expectation*). Given an instance $\langle \Delta, C', A', E', S', R' \rangle$ and a set of desires D; the expectation function f''_w is given by:

$$f''_w(\langle \Delta, C', A', E', S', R' \rangle, D) = \begin{cases} f_d(R', D) + f_d(\neg S', D) - (f_d(R', D) * f_d(\neg S', D)) & \text{if } f_d(R', D) \geqslant 0 \text{ and } f_d(\neg S', D) \geqslant 0 \\ f_d(R', D) + f_d(\neg S', D) + (f_d(R', D) * f_d(\neg S', D)) & \text{if } f_d(R', D) < 0 \text{ and } f_d(\neg S', D) < 0 \\ f_d(R', D) + f_d(\neg S', D) & \text{otherwise} \end{cases}$$

where f_d is defined as before.

Remark. The expectation function fw'' combines the desirability of the reward $f_d(R',D)$ and the undesirability of the sanction $f_d(\neg S', D)$ as follows: if both $f_d(R',D)$ and $f_d(\neg S', D)$ are positive, then f''_w provides a combined value that is higher than each individual desirability degree; if both $f_d(R',D)$ and $f_d(\neg S', D)$ are negative, then f''_w results in a value lower than each desirability degree; otherwise, f''_w results in a value that is a compromise among the two desirability degrees.

For simplicity, we assume that there is a perfect enforcement that always punishes offenders and rewards obedience. However, if agents are able to know the probability of being punished or rewarded, then the desirability of sanctions and rewards should be pondered with these probabilities.[7]

[6] Similarly, if both R' and $\neg S'$ are undesirable to the agent, then the value of the expectation is expected to be lower than the desirabilities of R' and $\neg S'$.

[7] For example, the probability of being rewarded when the obligation $\langle \mathcal{O}, C', A', E', S', R' \rangle$ is fulfilled can be represented as the certainty of the belief $(C' \rightarrow R', \rho)$.

In the proposed case study, let us suppose that fines in area $a1$ are not expensive. Therefore, user is not very worried about paying fines in this area. Thus, the *assistant* agent has a desire as the following $(\neg fine(a1), 0.25)$ and the enforcement of this norm is not very important to the agent:

$$f''_w(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, D) =$$
$$f_d(-, D) + f_d(\neg fine(a1), D) + (f_d(-, D) * f_d(\neg fine(a1), D)) = 0 + 0.25 - (0 * 0.25) = 0.25$$

5.2.3. Emotions (f'''_w)

f'''_w models the emotions triggered when the agent violates a given instance. Thus, f'''_w models the social repercussions of violating norms, whereas f''_w models the economic repercussions. We use the term emotion for representing the valued reaction of agents (i.e., the agent's cognitive interpretation) with respect to some aspect of the world (i.e., the reality) [23]. Thus, agents do not have an explicit representation and reasoning about emotions as in other proposals such as [14]. In fact, our proposal is not aimed at building emotional agents, but to develop norm-autonomous agents capable of understanding the most relevant emotions that are involved in the norm compliance decision. Specifically, agents are capable of anticipating, exhibiting and explaining those human emotions that are involved with the normative decisions. Thereby, the decisions about norm compliance are based on other criteria beyond utility.

As argued by Elster in [17], in human societies norms are sustained by the desire to avoid the disapproval of others. Following Elster's proposal, when the violation of norms is greeted with condemnation, *self-attribution* emotions (i.e., shame) are triggered on the offender. Moreover, the situations that are predicted to occur when norms are violated may cause *prospect* emotions (i.e., hope and fear) on the offender.

Self-attribution emotions (represented by the e_a parameter) calculate the disapproving of one's own censurable action and only sustain norm obedience. Thus, they are modelled as a real value within the $[0,1]$ interval that determines the evaluation (i.e., attribution) that the agent makes about itself when it violates a norm. *Prospect* emotions (represented by the e_p parameter) represent the fear and hope emotions triggered by the violation of an instance. They can sustain either the obedience or the violation of norms; e.g., in some conditions the violation of norms may entail desirable consequences. Thus, prospect emotions are modelled as a real value within the $[-1,1]$ interval that considers the possible outcomes of violating an instance. Positive values mean that the agent fears to violate the instance, since it believes that the violation may entail undesirable consequences. On the contrary, a negative value means that the agent considers norm violation as a hopeful possibility, since it would entail desirable consequences. The degrees of these two emotions (e_a and e_p) are also combined considering the MYCIN [26] rules. Thus, the f'''_w function calculates the agent emotional disposition to comply with an instance as a real number within the $[-1,1]$ interval as follows:

Definition 9 (*Emotions*). Given an instance n_i, a set of beliefs B, a set of desires D and a set of norms N; the emotions function f'''_w is given by:

$$f'''_w(n_i, B, D, N) = \begin{cases} e_a + e_p - (e_a * e_p) & \text{if } e_p > 0 \\ e_a + e_p & \text{otherwise} \end{cases}$$

where $e_a = f_a(n_i, N)$ is the value of the self-attribution emotions (see Definition 11) and $e_p = f_p(n_i, B, D)$ is the value of the prospect emotions (see Definition 13).

Remark. The emotions function fw''' combines self-attribution emotions e_a and prospect emotions e_p as follows: if both e_a and e_p are positive, then f'''_w provides a combined value that is higher than the emotion values; otherwise, f'''_w results in a value that is a compromise among the two emotion values.

For simplicity we assume that agents are equally prone to feel self-attribution and prospect emotions. If this is not the case, the value of these emotions should be weighted according to the agent propensity to feel these emotions.

To estimate the value of these two emotions (e_a and e_p) an emotional model susceptible of being implemented in a software agent is required. Emotion mechanisms in the literature are often used in artificial agents as a method of improving action selection.[8] One of the emotional models that have made a deeper impact on the MAS field is the one developed by *Ortony, Clore and Collins* (OCC) in [23]. The representation of cognitive and normative elements in the Normative BDI architecture fits perfectly the cognitive factors considered by the OCC model. As a consequence, we use the OCC model for establishing the intensity of the emotions that are involved in the norm-reasoning process. The implementation of each one of these two emotional functions (i.e., f_a and f_p) is explained below:

- *Self-Attribution Emotions.* According to the OCC model, shame is a self-attribution emotion that is elicited by the evaluation of the actions that have been performed by the agent itself. Specifically, when humans evaluate their actions, this evaluation is usually made with respect to norms. In the same manner, actions of agents are self-evaluated as censurable

[8] See [24] for an analysis of emotion mechanisms for agents.

insofar as these actions contradict the norms. In particular, the shame that the agent will feel if it violates a given instance is defined by considering the importance (i.e., the salience degree) of those norms that are generalizations of this instance.[9] We formally define the set of norms that are generalizations of a given instance as follows:

Definition 10 (*Instance Generalization*). Given an instance $\langle \Delta, C', A', E', S', R' \rangle$ that has been created out of a norm contained in a set of norms N; the set of norms that are a generalization of this instance is calculated by the instance generalization function as follows:

$$f_g(\langle \Delta, C', A', E', S', R' \rangle, N) = \{(\langle \Delta_j, C_j, A_j, E_j, S_j, R_j \rangle, \rho_{a_j}) | (\langle \Delta_j, C_j, A_j, E_j, S_j, R_j \rangle, \rho_{a_j}) \in N, \Delta_j = \Delta, \text{ and exists a substitution} \sigma_j \text{ such that } C' \vdash \sigma_j(C_j), A' \vdash \sigma_j(A_j) \text{ and } E' \vdash \sigma_j(E_j)\}$$

Remark. The instance generalization function f_g determines that any norm $\langle \Delta, C, A, E, S, R \rangle$ can be seen as a generalization of a given instance $\langle \Delta', C', A', E', S', R' \rangle$ if the two normative propositions have the same deontic modality (i.e., $\Delta = \Delta'$) and there is a substitution σ such as the formulas $\sigma(C)$, $\sigma(A)$, $\sigma(E)$ can be derived from C', A', E', respectively. According to this definition, the self-attribution function f_a (used for calculating the value e_a) is defined as the maximum among the salience of these norms that are a generalization of a given instance.

Definition 11 (*Self-Attribution Emotions*). Given an instance n_i and a set of norms N; the intensity of the self-attribution emotions triggered by the violation of this instance is defined by the f_a function as follows:

$$f_a(n_i, N) = \rho_{s_{max}}$$

where $\rho_{s_{max}}$ is a real value within the [0,1] interval such that it exists $(n_{max}, \rho_{s_{max}})$ in $f_g(n_i, N)$ and for all (n_j, ρ_{s_j}) in $f_g(n_i, N)$ $\rho_{s_{max}} \geqslant \rho_{s_j}$.

For simplicity, we assume that agents have the same capabilities for achieving or avoiding the states represented in the norm conditions. Otherwise, the value of the attribution emotion should be calculated as the salience of the most important norm weighted by the agent capabilities to comply with this norm.[10]

- *Prospect Emotions.* According to the OCC model, the hope (vs. fear) emotion is triggered when a desirable (vs. undesirable) event is predicted. Therefore, the main factors on the intensity of hope (vs. fear) are the probability of the predicted event and the desirability (vs. undesirability) of this event. The fear and hope emotions that may be triggered if an instance is violated are defined by considering the desirability and probability of the consequences of violating an instance.

The consequences of violating an instance are a set of pairs (γ_j, ρ_j), where $\gamma_j \in \mathcal{L}$ represents a situation that is predicted to occur if the norm is violated; and $\rho_j \in [0,1]$ is the probability of this predicted situation. An obligation is violated when the norm condition is not achieved (i.e., $\neg C'$). Therefore, the consequences of violating the obligation are defined by considering those beliefs such as $(\alpha \rightarrow \gamma_j, \rho_j)$ and $\neg C' \vdash \alpha$.[11] In case of a prohibition instance, its violation entails the achievement of the norm condition (C'). Therefore, the consequences of violating a prohibition are calculated by considering those beliefs such as $(\alpha \rightarrow \gamma_j, \rho_j)$ and $C' \vdash \alpha$. We define formally the consequences of violating an instance as follows:

Definition 12 (*Consequences*). Given an instance $\langle \Delta, C', A', E', S', R' \rangle$, and a set of beliefs B; the predicted consequences of violating this instance are defined as follows:

$$f_c(\langle \Delta, C', A', E', S', R' \rangle, B) = \begin{cases} \{(\gamma_j, \rho_j) | (\alpha \rightarrow \gamma_j, \rho_j) \in B, \neg C' \vdash \alpha\} & \text{if} \Delta = \mathcal{O} \\ \{(\gamma_j, \rho_j) | (\alpha \rightarrow \gamma_j, \rho_j) \in B, C' \vdash \alpha\} & \text{if} \Delta = \mathcal{F} \end{cases}$$

The prospect emotions elicited by the violation of any instance are calculated by considering the desirability and probability of the consequences of violating this instance[12] as follows:

Definition 13 (*Prospect Emotions*). Given a set of beliefs B, a set of desires D, and an instance n_i; the prospect emotions triggered by the violation of this instance is defined by the f_p function as follows:

$$f_p(n_i, B, D) = \frac{-\sum_{\forall (\gamma_j, \rho_j) \in f_c(n_i, B)} \rho_j * f_d(\gamma_j, D)}{\sum_{\forall (\gamma_j, \rho_j) \in f_c(n_i, B)} \rho_j}$$

[9] Each instance is created out of a single norm. However, an instance can be seen as a particularization (i.e., instantiation) of more than one norm.

[10] For example, the capabilities to comply with the obligation $\langle \mathcal{O}, C', A', E', S', R' \rangle$ can be represented as the certainty of the belief (*canAchieve*(C'),ρ).

[11] It means that the negation of the norm condition ($\neg C'$) causes or explains γ_i with a probability ρ_j.

[12] Notice that the desirability and probability of the consequences of violating an instance are represented as beliefs and desires.

Table 3
Weight definition of the agent types compared in the experiments.

Agent type	w′ (Self-interest)	w″ (Expectation)	w‴ (Emotion)
Egoist	1	0	0
Cautious	0	1	0
Emotional	0	0	1
Egoist–Cautious	1	1	0
Egoist–Emotional	1	0	1
Cautious–Emotional	0	1	1
Egoist–Cautious–Emotional	1	1	1

where f_c and f_d are defined as before.

Remark. The prospect emotion function f_p calculates the prospect emotions elicited by the violation of an instance as a real value within the [−1,1] interval. In accordance with the previous definitions of the *willingness functions*, which define positive values as compliance sustaining, the f_p function has been defined as minus the weighted mean of the desirability of the effects of the violation. Therefore, the desirability of the consequences of the violation has been weighted by the probability of their occurrence (ρ_j). A positive value of f_p sustains norm compliance. Specifically, it means that the violation of the norm raises the agent's fears. A negative value of the f_p function sustains the violation of norms. It occurs when the agent hopes that the violation of the norm entails desirable consequences.

In the proposed case study, the value of the attribution emotion calculated by the *assistant* agent is 1, which is the salience degree of the Heavy Rain Norm. The *assistant* agent calculates the value of the prospect emotion by considering the consequences of not reducing the speed. Specifically, the *assistant* agent considers that not reducing the speed may cause an accident with a probability of 25% $-(\neg slow(a1) \rightarrow accident(a1), 0.25) \in B$. The human user does not want to cause an accident $-(\neg accident(a1), 1) \in D$. Therefore, $f_d(accident(a1), D)$ is -1 and the value of the prospect emotion is calculated as follows:

$$e_p = f_p(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, B, D) = \frac{-(0.25 * -1)}{0.25} = 1$$

and the value of the anticipated emotions is calculated as follows:

$$f'''_w(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, B, D, N) = e_a + e_p - (e_a * e_p) = 1 + 1 - (1 * 1) = 1$$

Let us assume that the human user has configured the *assistant* agent to consider the three willingness factors equally. Therefore willingness of the *assistant* agent to comply with the instance is calculated as follows:

$$f_w(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, B, D, N) = \frac{1 \times -0.9 + 1 \times 0.25 + 1 \times 1}{3} = 0.12$$

Assume that the compliance threshold (θ) is set to 0.1. Then, the norm reasoning rule for obligations (see Rule (a) in Table 2) is instantiated as follows:

$$\frac{(\langle \mathcal{O}, slow(a1), heavyRain(a1), \neg heavyRain(a1), fine(a1), -\rangle, 0.75) \in N_I \wedge 0.12 > 0.1}{\langle B, D, I, N, N_I\rangle \rightarrow \langle B, D*, I, N, N_I\rangle}$$

where $D* = D \cup \{(slow(a1), f_i(0.75, 0.12)\}$

Since 0.12 > 0.1, the *assistant* agent decides to comply with the norm and creates a new desire to achieve the norm condition. The degree of the new desire is calculated as follows:

$$f_i(0.75, 0.12) = \frac{0.75 * 0.12}{1 - 0.75 - 0.12 + (2 * 0.75 * 0.12)} = 0.29$$

and the set $D*$ contains now a proposition such as: $(slow(a1), 0.29)$.

6. Experimental results

This section compares the performance of the different agent types with respect to their decisions about norm compliance, which are modelled using the willingness function (f_w). As previously mentioned, the value of f_w is obtained combining the values of the three *willingness functions* (f'_w,f''_w and f'''_w) as a weighted average. The weights that each agent gives to these factors characterize each agent type. The experiments contained in this section are aimed at illustrating the performance of each willingness function individually and the effect of combining the different willingness functions. Given these three

Table 4
Parameters used in the experiments.

Parameter	Value
# Of agents	7
# Of non-grounded literals	10
# Of instantiations per non-grounded literal	10
# Of norms	[1,10]
Salience of norms	[0,1]
# of instances	[1,100]
Relevance of instances	[0,1]
# Of goals	10
# Of explanation relationships	[1,100]
Norm compliance threshold (θ)	0.1
# Of executions	1000

willingness functions, there are only 7 possible combinations[13] of them (excluding the empty set). As a consequence, we conducted experiments comparing the behaviour exhibited by these 7 main agent types that can be created with our model. Table 3 contains the characterization of the 7 agent types compared in this section.

To analyse how the willingness function works, to compare the main agent types and to validate our hypothesis about their behaviour, we have performed two different types of experiments. Firstly, we want to illustrate the performance of the functions proposed in this paper. With this aim we have performed a set of random experiments to illustrate the behaviour exhibited by the main agent types in a wide range of situations. Moreover, we analyse the effect of the different elements considered by the willingness functions on the norm-compliance decisions. These experiments are described in Section 6.1. In the second type of experiments we illustrate how the different types of agents make decisions about norm compliance in a concrete situation of the case study used in this paper. This experiment is described in Section 6.2. Finally, a discussion of the results of the experiments is contained in Section 6.3.

Since the experiments described in the following sections are aimed at analysing the performance of the willingness functions, other norm-reasoning problems faced by previous works [13] have been omitted.

6.1. Random experiments

6.1.1. Experiment description

We considered a scenario with the parameters that we sum up in Table 4. In each execution one agent of each type is created. These agents are affected by the same set of norms and instances. Moreover, all agents have the same desires and beliefs. Therefore, the only difference among agents is the way in which they make decisions about norm compliance; i.e., how they decide about which instances will be obeyed and which ones will be violated.

Environment Definition. In this simulator, we assume that the environment is described in terms of 10 different situations or states of affairs. Each one of these states of affairs is represented by a non-grounded literal (i.e., an atomic formula or its negation). Thus, we consider a set of 10 different non-grounded literals. We also assume that each one of the 10 non-grounded literals can be instantiated in 10 different ways. Thus, there are 100 grounded literals (these grounded literals correspond to the 10 ways in which each one of the 10 non-grounded literals can be instantiated).

Norm Definition. In each execution a set of norms is randomly generated. The number of norms that are generated takes a random value within the [1,10] interval. Thus, we create one norm controlling each state of affairs at most. We assume that for each norm $\langle \Delta, C, A, E, S, R \rangle$ the elements C, A, E, S and R are randomly selected from the set of non-grounded literals. The salience of each norm gets a real random value within the interval [0,1].[14]

From each norm a set of instances is randomly created. Since each non-grounded literal can be instantiated in 10 ways, each norm can also be instantiated in 10 different ways. Therefore, the number of instances created out of each norm ranges randomly within the [1,10] interval and, as a consequence, the number of instances created in each execution ranges randomly within the [1,100] interval. For each instance $\langle \Delta, C', A', E', S', R' \rangle$ that has been created out of a norm $\langle \Delta, C, A, E, S, R \rangle$ the elements C', A', E', S' and R' are defined as one of the possible instantiations of C, A, E, S and R, respectively. The relevance of each instance takes a real random value within the [0,1] interval.[15]

Agent Definition. Agents pursue a set of *goals* that represent desired (vs. undesired) states of affairs; i.e., situations that the agent wants to achieve (vs. avoid). Thus, each goal is characterized by a goal condition that is a grounded literal representing the sate of affairs that is desired (vs. undesired); and a goal degree that represents the desirability (vs. undesirability) of the goal condition. Moreover, we assume that other states of affairs that are similar to the goal condition are also desired (vs.

[13] For simplicity we have only considered these agent types in which the w', w'', $w''' \in \{0,1\}$.
[14] Note that we describe later an experiment in which we analyse the impact of norm salience on the behaviour exhibited by the different agents.
[15] Note that the decisions about norm compliance are not affected by the relevance of instances, as explained in Section 5.2.

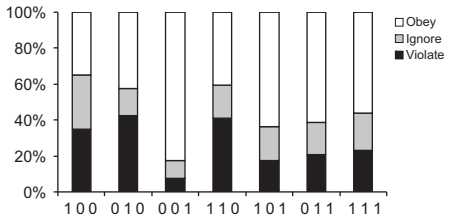


Fig. 2. Percentage of instances that are violated, ignored and obeyed on average when θ takes value 0.1.

undesired) with a lower degree. The more these states of affairs are similar[16] to the goal condition, the more desirable they are. Thus, each goal is also characterized by a goal distance that is a real value corresponding to the distance between the state specified by the goal and the most similar state that is not desired.[17]

In each execution, goals are randomly generated: the goal conditions are grounded literals that are randomly selected from the 100 grounded literal set; the goal degrees vary randomly within the $[-1,-0.75]\cup[0.75,1]$ interval—i.e., a positive (vs. negative) goal degree means a goal to achieve (vs. avoid) the goal condition—; and the goal distance varies randomly within the $[0,1]$ interval. For simplicity, we assume that only one instantiation of each non-grounded literal can be selected as a goal condition. Thus, in each simulation 10 goals are generated[18]; i.e., one per each non-grounded literal.

Besides the desirability of propositions, agents also use explanation relationships among propositions for making decisions about norm compliance. These explanation relationships are represented as graded beliefs such as $(\alpha\rightarrow\gamma,\rho)$, which means that α explains γ with a probability of β. Here, these relationships have been randomly generated. The antecedent (α) and consequence (γ) of an explanation relationship are randomly selected as grounded literals. The probability of these relationships (ρ) is a real randomly selected within the $[0,1]$ interval. In each execution, agents know a number of explanation relationships that ranges randomly within the $[1,100]$ interval[19]; i.e., it varies from 1 to the number of grounded literals considered in the simulations.

Agent Types. In the model proposed in this paper the decisions about norm compliance are made by considering three different factors: self-interest (f'_w), the enforcement mechanisms (f''_w) and the emotions triggered by the violation of norms (f'''_w). These three factors are combined in a single value (f_w) that is defined as a weighted average among the three willingness factors. Therefore, different agent personalities can be modelled according to the definition of the weights w', w'' and w''' (see Table 3). The three basic personalities are: *egoist*, *cautious* and *emotional*:

- *Egoist agents* ($w'=1$, $w''=0$ and $w'''=0$) only follow those norms that favour their goals or that avoid some undesirable state.
- *Cautious agents* ($w'=0$, $w''=1$ and $w'''=0$) comply with norms when they want to avoid the sanctions or when they are interested on the rewards.
- *Emotional agents* ($w'=0$, $w''=0$ and $w'''=1$) only consider the emotions that will be elicited if norms are violated. As explained in Section 5.2 agents are capable of anticipating both self-attribution and prospect emotions:
  - *Self-attribution emotion.* As explained before, the f_a function is defined as the maximum among the salience values of those norms that are a generalization of a given instance. In the experiments we have considered that an instance $\langle\Delta,C',A',E',S',R'\rangle$ can be seen as a generalization of a norm $\langle\Delta,C,A,E,S,R\rangle$ when the two have the same deontic modality and the expressions that are in C', A', E' are instantiations of the non-grounded literals in C, A, E, respectively.
  - *Prospect emotion.* In the experiments, the consequences of violating an instance are calculated considering the explanation relationships that have been randomly generated. For example, consequences of a prohibition instance $\langle\Delta,C',A',E',S',R'\rangle$ are calculated considering those explanation relationship that have as antecedent the non-grounded C'.

6.1.2. Results

Fig. 2 illustrates the performance of the different types of agents with respect to their decisions about norm compliance. This decision is modelled by the f_w parameter. Specifically, in Fig. 2 each agent type has been labelled according to the values given to the weights w', w'' and w'''. The values obtained by the f_w function have been classified again in three categories

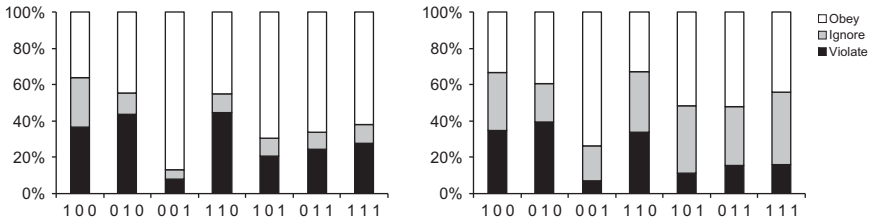
[16] We define a function $f_{distance}$ that determines the distance in terms of similarity between two grounded literals as a real number within the $[0,1]$ interval. A distance of 0 means "complete similarity" and a distance of 1 means "no similarity".

[17] Thus, the desirability of a grounded literal α with respect to a goal $\langle\beta,\gamma,\iota\rangle$, where β is the goal condition,γ is the goal desirability and ι is the goal distance; is calculated as follows:

$$\begin{cases}\gamma-\frac{f_{distance}(\alpha,\beta)*\gamma}{\iota} & \text{if } f_{distance}(\alpha,\beta)<\iota\\ 0 & \text{otherwise}\end{cases}$$

[18] Note that we describe later an experiment in which we analyse the impact of the number of goals on the behaviour exhibited by the different agents.

[19] Note that we describe later an experiment in which we change the number of explanatory relationships considered in the simulations.



(a) θ takes value 0.05 (b) θ takes value 0.2

Fig. 3. Percentage of instances that are violated, ignored and obeyed on average according to the value of θ.

according to the values of the *norm compliance threshold* (θ): deciding to *violate* (i.e., when f_w ranges within the $[-1,-\theta)$ interval); deciding to *ignore* (i.e., when f_w ranges within the $[-\theta,\theta]$ interval); and deciding to *obey* (i.e., when f_w ranges within the $(\theta,1]$ interval). Deciding to violate an instance means that the agent will try to behave contrary to the pattern of behaviour specified by the instance. Deciding to obey an instance means that the agent will try to follow the pattern of behaviour specified by the instance. Deciding to ignore an instance means that the agent will not change its behaviour due to the instance. Specifically, Fig. 2 shows the percentage of instances that belong to each one of the willing categories (i.e. *violate*, *ignore* and *obey*) when the compliance threshold (θ) is set to 0.1.[20] This experiment has been repeated 1000 executions to support the findings.

Regarding the three main agent personalities, it can be concluded that egoist agents (labelled as 1 0 0) are the most prone to ignore norms, since they only consider if the norm condition favours or hinders their goals. Cautious agents (labelled as 0 1 0) are not as prone to ignore norms, i.e., the percentage of ignored instances is lower. This is explained by the fact that cautious agents consider whether either the reward or the negation of the sanction favour their goals. Therefore, the percentage of instances that are indifferent in cautious agents is lower. In case of egoist and cautious agents there is a symmetric distribution of instances in the three willingness categories; i.e., egoist and cautious agents decide to obey as many norms as they decide to violate. This is explained by the fact that norms and desires are randomly generated and, as a consequence, norms favour or hinder the agent goals with the same probability. Finally, emotional agents (labelled as 0 0 1) are the most willing to obey norms; i.e., they are the most norm-oriented. This is explained by the fact that the attribution emotion (modelled by the f_a function) only sustains norm obedience. Moreover, the percentage of ignored norms in emotional agents is the lowest. This is explained by the combination between the prospect (modelled by the f_p function) and the attribution emotion. The prospect emotion considers the desirability of all the possible consequences of violating an instance. Thus, it is possible that the negative effects counteract the positive ones and the values obtained by the f_p function are near to 0. This value is combined with the value calculated by the f_a function, which is always positive, and f_w takes a value higher than θ.

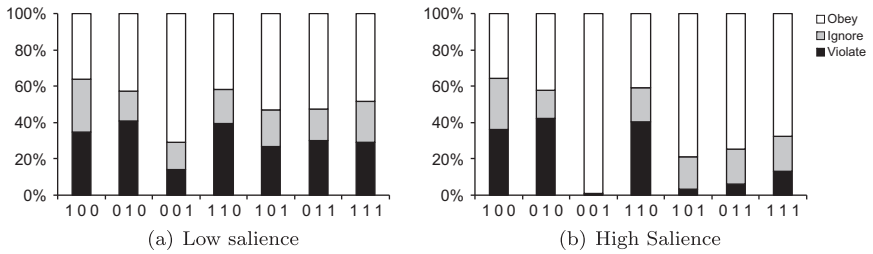
Other agent personalities can be defined from these three basic personalities by giving different values to the weights w', w'' and w'''. In this experiment, we have also analysed the behaviour of agents that use a mixed strategy for making decisions about norm compliance. Therefore, two or more willingness factors are considered in the calculation of the f_w parameter. As expected, all agents that consider emotions, (i.e., $w''' = 1$) have a tendency to decide to obey norms. Specifically, agents that consider the three willingness factors (i.e., $w' = 1$, $w'' = 1$ and $w''' = 1$) comply with less norms than the rest of agents that consider the emotional factor, since the influence of emotions is reduced by the other two factors. In case of agents that do not consider emotions (i.e., $w' = 1$, $w'' = 1$ and $w''' = 0$), the percentage of instances that are ignored is higher than in cautious agents. This is explained by the fact that the norm conditions, the sanctions, and the rewards are randomly generated; i.e., there is not any relationship between a norm and its enforcement. Therefore, it is possible that a norm favours one of the agent goals but the reward that the agent will receive hinders another goal. In this situation, the agent has motivations for violating the norm and motivations for following it. Thus, it decides to ignore the norm.

Compliance Threshold θ. In the previous experiment we have defined the compliance threshold (θ) as 0.1. To analyse the influence of this parameter on the decisions about norm compliance, the previous experiment has been repeated assigning different values to θ. Specifically, we have performed experiments in which θ is 0.05 and 0.2.

The compliance threshold represents how much a person is indifferent to norms; i.e., to what extent its behaviour is not influenced by norms. A high indifference means that the person is not affected by norms and that his/her activities are only directed by his/her own goals. As the indifference of a person decreases more norms become the motivations of his/her actions. This tendency does not depend on the concrete reasons why each person makes decisions about norm compliance. For example, an egoist person who is highly indifferent to norms decides to ignore more norms than another egoist person who is less indifferent to norms.

Fig. 3a and b shows the percentage of instances that belong to each one of the willingness categories when θ is 0.05 and 0.2, respectively. As we expect, the lower the compliance threshold is, the lower number of instances are ignored. On the contrary, when the compliance threshold takes higher values the percentage of ignored instances increases. Thus, the compliance threshold allows us to define different indifference degrees for each agent type.

[20] Note that we describe later an experiment in which we change the value of the compliance threshold.



(a) Low salience (b) High Salience

Fig. 4. Percentage of instances that are violated, ignored and obeyed on average according to salience of norms (ρ_s).

As Fig. 3b shows, when θ is 0.2 the percentage of norms that are ignored increases more in agents that combine two or more compliance factors. This is explained by the fact that the willingness factors have been combined as a weighted mean where weights only take the values $\{0,1\}$. Thus, the weighted mean behaves as a compromise operator among the factors that are considered; i.e., the willingness factors whose weight is equal to 1. As a consequence, the combined result is higher or equal to the considered factor with the minimum value and lower or equal to the considered factor with the maximum value. Therefore, norms are obeyed (vs. violated) only when all the considered factors are higher (vs. lower) than θ (vs. $-\theta$). As mentioned above, norms and agents are randomly generated in an independent way, which makes difficult that all different factors take values higher (vs. lower) than θ (vs. $-\theta$).

Salience of Norms. With the aim of determining the effect of the salience of norms on the decisions about norm compliance, we also run out experiments varying the salience of norms. Specifically, we have performed experiments in which the salience of norms is low (i.e., lower than 0.5) and high (i.e., higher than 0.5).

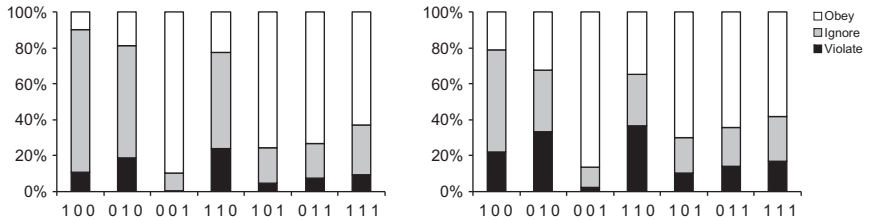
The salience of a norm represents the effect or influence of this norm on the society. The same norm may have different salience in different societies. For example, there are countries where the prohibition to run a red light is considered as a very important norm. Transgressions of that norm are infrequent and are sanctioned by the traffic authorities that impose fines and the society that condemns these acts. In other countries, the prohibition to run a red light is not considered as a very important norm. As a consequence, transgressions of that norm are more frequent even if the same fines are applied. Obviously, those persons that are more affected by the society are more sensitive to salience of norms. In contrast, those persons that do not care about what other people think are not affected by salience of norms.

Fig. 4a shows the results obtained when the salience of norms is low (i.e. $\rho_s \in [0,0.5]$). Fig. 4b shows the results obtained when the salience of norms is high (i.e. $\rho_s \in [0.5,1]$). As we expect, only those agents that consider the social repercussions of norms (i.e., emotions) are affected by the salience of norms. When the salience of norms is low (see Fig. 4a) the percentage of obeyed norms in agents that consider emotions decreases because norms are not very important to the society. As the salience of norms increases (see Fig. 4b) the percentage of obeyed norms in agents that consider emotions increases. This is explained by the fact that the society considers that norms are of high importance and a transgression of norms will be strongly criticized. As a consequence, all agents that care about what the society thinks follow the majority of the norms to avoid social reproaches. This is more noticeable in case of emotional agents ($w' = 0$, $w'' = 0$ and $w''' = 1$), since in this type of agents the influence of emotions is not reduced by other factors. Thus, in situations where the salience of all norms is very high it is preferable not using emotional agents, since they would behave as norm-oriented agents that follow almost all norms.

Agent Goals. In the previous experiments agents pursue 10 goals, which correspond to the number of non-grounded expressions used to generate the norms. Thus, agents are able to determine whether each instance is interesting to them (i.e., to what extent the norm condition, the sanction or the reward affects their goals). In this section we analyse what happens if agents pursue less goals and they are not always able to determine if an instance is interesting to them. Specifically, we have performed experiments in which agents pursue 3 and 6 goals.

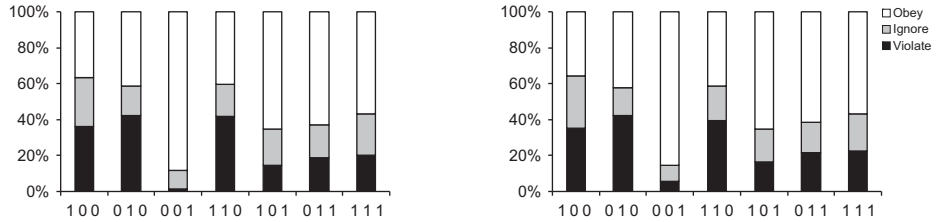
When an egoist person is not able to evaluate the effects of the norm fulfilment on his/her goals, then he/she decides not to change his/her behaviour and to ignore the norm. The same decision is taken when a cautious person is not able to determine whether the sanction or the reward of a norm hinders of favours this/her goals. Selfless persons not only consider the impact of norms and their enforcement on their goals and, as a consequence, they may decide to obey norms even if they do not know the effect of norms on their goals. Obviously, if a person has few goals (i.e., interests), then the probability that norms have an effect on his/her goals is low and more norms are ignored.

In this section, we describe the results obtained when we vary the number of goals that agents pursue. Fig. 5a and b shows the results obtained when the number of goals is 3 and 6, respectively. As we expected, agents that consider self-interest and expectation factors are the most affected by the number of goals. Specifically, when the number of goals is very low (see Fig. 5a) these kinds of agents have very few information to make decisions about norm compliance and the percentage of ignored norms increases. This is more noticeable in egoist agents, since they only consider if the norm condition hinders or favours their goals. In contrast, cautious agents consider the effect the sanction and reward have on their goals. As the number of goals increases (see Fig. 5b), the percentage of ignored norms decreases.



(a) The number of goals is 3 (b) The number of goals is 6

Fig. 5. Percentage of instances that are violated, ignored and obeyed on average according to the number of goals.



(a) The number of explanation relationships is 10 (b) The number of explanation relationships is 1000

Fig. 6. Percentage of instances that are violated, ignored and obeyed on average according to the number of explanation relationships.

As a consequence, when the number of goals that an agent pursues is low it is better not to use egoist agents, since they would not have enough information to determine the effect of norm fulfilment on its goals and the majority of instances would be ignored.

Explanation Relationships. This experiment consists in varying the number of explanation relationships that an agent knows. Specifically, we want to analyse what happens when the number of explanation relationships is very low (i.e., agents only know 10 explanation relationships) and when it is very high (i.e., agents know 1000 explanation relationships).

Explanation relationships model relationships between two states of affairs in which one causes the other. Thus, they represent the knowledge that a person has about the potential repercussions of his/her actions. Obviously, those persons that are more empathetic take into account the repercussions of norms when they made decisions about norm compliance. The more explanation relationships (i.e., potential repercussions) an empathetic person knows, the higher the probability of knowing a positive or negative repercussion of a given norm. In contrast, those persons that do not care about other people do not consider the potential repercussions of their actions.

Fig. 6a and b shows the results obtained when the number of explanation relationships is 10 and 1000, respectively. As we expect, only those agents that consider emotions are affected by the explanation relationships. When the number of explanation relationships is very low (see Fig. 6a) agents that consider emotions have very few information for deciding to fulfil/violate norms due to their good/bad consequences and the decisions about norm compliance are made according to the attribution emotion that only sustains norm compliance. As a consequence, the number of obeyed norms increases. This is more noticeable in agents that only consider emotions ($w' = 0$, $w'' = 0$ and $w''' = 1$), since the emotions are not combined with other factors. As the number of explanation relationships increases (see Fig. 6b), the number of obeyed norms decreases lightly. This is explained by the fact that agents have more information about the repercussion of norms and there is a higher probability of deciding to violate norms due to their bad consequences.

As a result of these experiments we can categorize different scenarios in which is more preferable not using a particular type of agent. For example, egoist agents are not suitable when the number of goals is very low, since the majority of norms are ignored. Emotional agents are not suitable when the salience of all norms is high, or when the number of explanation relationships is very low. In these situations emotional agents blindly follow all norms and it is preferable to combine emotions with other factors such as expectations or self-interest.

6.2. Case study

The previous experiments were aimed at providing a categorization of the behaviour exhibited by the main agent types in a general situation. For this reason, we generated the agent desires and norms randomly. However, the random generation of desires and norms may lead to unrealistic situations; e.g., a situation in which the behaviour of agents is determined by inconsistent desires and/or contradictory norms. In contrast, this section presents a more realistic situation in which the behaviour of agents is determined by consistent desires and rational norms. Specifically, we contextualise the experiments in the case study used in this paper and we describe how the different types of *assistant* agent make decisions about norm compliance. We have performed an experiment with the parameters contained in Table 5.

Table 5
Parameters used in the experiment for the traffic assistant case study.

Parameter	Value
# Of agents	7
# Of norms	1
Norm salience (ρ_s)	[0,1]
# Of areas	10
# Of instances	1
Instance relevance (ρ_r)	[0,1]
# Of goals	3
# Of explanatory relationships	10
Norm compliance threshold (θ)	0.1
# Of execution	1000

6.2.1. Experiment description

We assume that all assistant agents know the Heavy Rain Norm. Thus, the norm set (N) of the 7 agents contains a proposition as follows:

$$(\langle \mathcal{O}, slow(A), heavyRain(A), \neg heavyRain(A), fine(A), -\rangle, \rho_s) \in N$$

The norm salience degree (ρ_s) gets a random value within the interval [0,1].

We assume that the set of non-grounded literals considered by all *assistant* agents contains at least the literals $slow(A)$, $\neg fine(A)$ and $\neg accident(A)$,[21] which represent: decreasing the speed in area A, not paying a fine in area A and not having an accident in area A.

In each simulation, an instance of the Heavy Rain Norm is randomly created by selecting one concrete area in which there is heavy rain. We assume that the *assistant* agents can be used in 10 different areas ($\{a1, \ldots, a10\}$). Thus, the norm condition, the activation and expiration conditions and the sanction are instantiated with the area that is selected. For example, if area $a5$ is selected in a concrete simulation, then the following instance is added to the instance set (N_I) of all *assistant* agents:

$$(\langle \mathcal{O}, slow(a5), heavyRain(a5), \neg heavyRain(a5), fine(a5), -\rangle, \rho_r) \in N_I$$

The value of ρ_r is assigned a random value within the [0,1] interval. All agents know the same instance with the same relevance value.

All *assistant* agents pursue the same set of desirable states or *goals*. Specifically, we generate three goals (i.e., one per each non-grounded literal that is considered for making decisions about norm compliance). As previously mentioned, we assume that the user wants to show off the power of his new car and he does not want to slow down in crowded places. The area that is more crowded is randomly selected from the set $\{a1, \ldots, a10\}$. The desirability of slowing down in this area varies randomly within the $[-1, -0.75]$ interval. The desirability of slowing down in other areas depends on how much crowded they are. Similarly, the area in which more expensive fines are imposed is randomly selected from the set $\{a1, \ldots, a10\}$. The desirability of not being fined varies within the [0.75,1] interval. Finally, we assume that the user wants to avoid having an accident that involves other vehicles and their passengers. Thus, the area in which there is more traffic (e.g. the area in which there is a high traffic density) is also selected from the set $\{a1, \ldots, a10\}$. The desirability of not having an accident in this area varies randomly within the [0.75,1] interval. For example, Fig. 7 represents three goals that have been randomly generated. The X-axis represents the substitutions that can be applied to the non-grounded literals (i.e., to the literals $slow(A)$, $\neg fine(A)$ and $\neg accident(A)$) to give rise to the grounded literals, whereas the Y-axis shows the desirability degree of these grounded literals.

Moreover, the *assistant* agents know 10 explanatory relationships that represent the probability in which not slowing down in a specific area can cause an accident in this area. Specifically, it is represented as graded beliefs such as $(\neg slow(a) \rightarrow accident(a), \rho)$; where $a \in \{a1, \ldots, a10\}$ and the probability of this relationship (ρ) is a random real within the [0,1] interval.

Again we have performed 1000 executions to support the findings.

6.2.2. Results

Egoism is the concept of acting in one's own self-interest. Thus, we want that the decisions of egoist *assistants* are mainly influenced by the goals of their user. Since the user is only interested in showing off his car, an egoist *assistant* would decide to violate the Heavy Rain Norm in most cases. Cautious persons carry out prudent forethought to minimize risk. In this case, we want that cautious *assistants* decide to follow the Heavy Rain Norm to avoid sanctions (i.e., the economic repercussions). Finally, emotions provide the affective component to motivation. Thus, we expect that emotional *assistants* decide to comply with the Heavy Rain Norm to spare the user the embarrassment of performing reprehensible actions.

[21] Notice that these three literals are the ones that will be considered by the *assistant* agent when deciding about norm compliance.

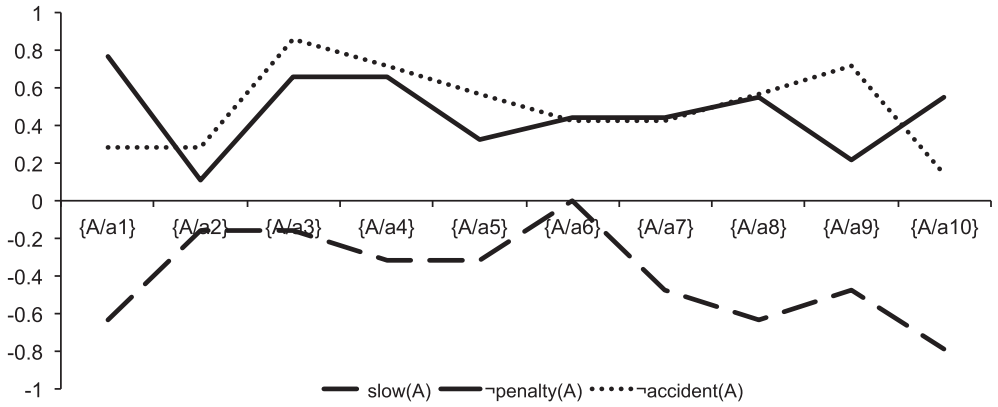


Fig. 7. Example of the goals for the assistant agent. The *Y*-axis shows the desirability and the *X*-axis shows the substitutions that can be applied to the non-grounded literals to create the grounded literals.

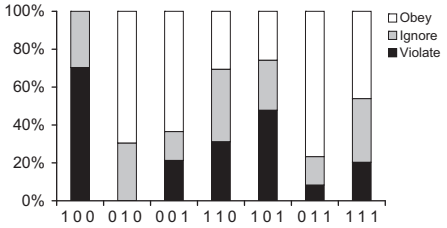


Fig. 8. Percentage of instances that are violated, ignored and obeyed on average per each type of assistant agent.

Fig. 8 illustrates the performance of the different types of *assistant* agents with respect to their decisions about norm compliance. As we expect, an egoist *assistant* never decides to comply with the Heavy Rain Norm (i.e., never decides to slow down). In most cases, it decides to violate the norm and behaves against the norm by deciding to accelerate instead of decelerate. Only when there is heavy rain in an area that is not crowded, it decides to ignore the norm and it maintains its behaviour (since there are not much people to whom show how fast is his/her car). A cautious *assistant* decides to comply with the obligation in most cases and, as a consequence, it decides to decrease the speed. When there is a heavy rain in an area in which speeding fines are cheap, the *assistant* agent decides to ignore the norm and it maintains the speed. Finally, an emotional *assistant* decides to reduce the speed in almost all cases. Only when the norm is not salient and the *assistant* does not fear of having an accident (i.e., when there is heavy rain in a low traffic density area in which there is a low probability of having an accident) the norm is ignored. The behaviour exhibited by the three main agent personalities is completely different; i.e., egoist agents are lawbreakers since they never decide to obey norms, cautious and emotional agents are norm-abiding since they never decide to violate norms (they only decide to obey or ignore the norm). In general, agents defined as a combination of the basic personalities have a smoother behaviour and they decide to violate, ignore and obey norms. However, agents that combine the expectation factor with the emotional factor (i.e., agents labelled as 011) are the most norm-abiding. This is explained by the fact that these agents are a combination of the two norm-abiding personalities, which entails that almost always norms are obeyed.

6.3. Discussion

As shown in the results provided in this section, the human-inspired model proposed in this paper allows agents to make decisions about norm compliance autonomously. However, the behaviour of an agent depends on the willingness factors that it considers and, as shown in the experimental results, it is predictable to some degree. Thereby, humans would be able to delegate tasks controlled by norms to software agents. For example, we have illustrated that each kind of *assistant* agent behaves as we expected; i.e., the decisions made by the *assistant* agent are sensible according to the willingness factors that it considers. Thus, the human-inspired model allows different types of *assistant* agents to be implemented. As a consequence, this model can be used by different kinds of users that want to customize the routes proposed by the *assistant* agent according to their personality and preferences.

It should be noted that improving the agent capabilities for making decisions about norm compliance obviously comes at an additional temporal cost. Specifically, Normative BDI agents must evaluate each instance against its desire set for calculating the self-interest and the expectation factors. To calculate the prospect emotion, agents must evaluate each instance against its desire and belief sets to determine the desirability of the repercussions of instances. In the calculation of the attri-

bution emotion multiple substitutions are applied to determine the norms that are a generalization of each instance. This step may be computationally expensive if the number of instances, norms and substitutions is high. However, this problem can be easily avoided if instances are annotated with the norm that has created the instance and the attribution emotion is simply calculated as the salience degree of this norm.[22]

Finally, we would like to remark that our proposal is agnostic with respect to the rationality of agents [27]; i.e., our model can be used by both perfect rational agents and bounded rational agents. Usually, the theories of bounded rationality introduce uncertainty on the knowledge of the agent. Our model has been designed to be used by agents that have graded beliefs. Thus, our model can be used by agents with perfect knowledge (in this case the certainty of beliefs takes values 0 and 1), but also by agents with uncertain knowledge (in this case the certainty of beliefs takes real values within the [0,1] interval). Another way in which rationality can be bounded is by assuming that agents have incomplete information about alternatives and incomplete information about consequences. Our model has been also designed to be used by agents that have graded desires, intentions and explanatory beliefs. Thus, our model can be used by agents with perfect information about alternatives (in this case the certainty of explanatory beliefs, desires and intentions takes values 0 and 1), but also by agents with uncertain knowledge (in this case the certainty of explanatory beliefs, desires and intentions takes real values within the [0,1] interval). Therefore, our model can be used with either certain or uncertain knowledge and with either complete or incomplete information about alternatives and consequences.

7. Conclusions

This paper answers a main question related to the possibility of developing software agents that consider norms as humans would do. In response to this issue, this paper proposes a human-inspired model that allows software agents to consider both their preferences and the norm repercussions when they determine their willingness to comply with norms. The repercussion of norms is not only defined in terms of the utility of norms and the economic cost (vs. benefit) of the sanctions (vs. rewards), but also in terms of the social repercussions of norms (i.e., emotional factors). In particular, the human-inspired model anticipates the emotions that will be elicited if the norms are transgressed. As far as we are concerned, this is the first model of norm-autonomous agent that considers emotions as a motivation for norm compliance. We believe that this complex behaviour is required in applications such as: social simulation scenarios, environments in which humans and agents interact in a realistic way, scenarios in which humans delegate tasks to personal software agents, and so on.

The way in which the model combines rational and emotional factors allows different agent personalities to be implemented. To illustrate the behaviour exhibited by the different agent personalities we carried out a set of experiments. The results explained in this paper demonstrate that the decisions about norm compliance are predictable according to the agent personality. Thus, the human-inspired model can be adjusted according to the personality traits of the user. Moreover, these experiments analyse the influence of the different elements that are considered by our model. Finally, we provide the reader with guidelines to adjust the human inspired model (i.e., selecting the most suitable personality for agents).

As future work, we plan to evaluate to what extent the behaviour exhibited by the human-inspired model is in accordance with the user preferences and personality by testing our model in experiments in which humans will take part.

Acknowledgments

This paper was partially funded by the Spanish government under Grants CONSOLIDER-INGENIO 2010 CSD2007-00022, TIN2009-13839-C03-01, TIN2008-06701-C03–03, TIN2008-04446 and by the FPU Grant AP-2007-01256 awarded to N. Criado. This research has also been partially funded by the Generalitat de Catalunya under the Grant 2009-SGR-1434 and Valencian Prometeo Project 2008/051.

References

[1] C. Alchourrón, E. Bulygin, Normative Systems, Springer-Verlag, WienNew York, 1971.
[2] G. Andrighetto, M. Campenni, R. Conte, M. Paolucci, On the immergence of norms: a normative agent architecture, in: Proc. of AAAI Symposium, Social and Organizational Aspects of Intelligence, 2007.
[3] G. Andrighetto, D. Villatoro, R. Conte, Norm internalization in artificial societies, AI Communications 23 (4) (2010) 325–339.
[4] C. Bicchieri, The Grammar of Society: The Nature and Dynamics of Social Norms, Cambridge Univ Pr, 2006.
[5] G. Boella, L. Lesmo, Deliberate normative agents, in: Social Order in MAS, Kluwer, 2001.
[6] P. Bourdieu, Practical Reason: On the Theory of Action, Stanford Univ Pr, 1998.
[7] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, L. van der Torre, The boid architecture – conflicts between beliefs, obligations, intentions and desires, in: Proc. of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), ACM Press, 2001, pp. 9–16.
[8] A. Casali, L. Godo, C. Sierra, A graded bdi agent model to represent and reason about preferences, Artificial Intelligence 175 (7-8) (2011) 1468–1478.
[9] C. Castelfranchi, Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations, Journal of Applied Logic 1 (1) (2003) 47–92.
[10] C. Castelfranchi, F. Dignum, C. Jonker, J. Treur, Deliberative normative agents: principles and architecture, Intelligent Agents VI. Agent Theories Architectures, and Languages (2000) 364–378.
[11] R. Conte, C. Castelfranchi, F. Dignum, Autonomous norm acceptance, LNCS 1555 (1999) 99–112.

[22] This simplification does not take into account that an instance can be seen as a particularization of more than one norm. Thus, it assumes that norms that generate similar instances have similar salience degrees.

[12] N. Criado, E. Argente, V. Botti, Normative deliberation in graded BDI agents, in: Eighth German Conference on Multi-Agent System Technologies (MATES-10), LNAI, vol. 6251, Springer, 2010, pp. 52–63.
[13] N. Criado, E. Argente, V. Botti, Open issues for normative multi-agent systems, AI Communications 24 (3) (2011) 233–264.
[14] M. Dastani, J. Meyer, Programming agents with emotions, in: Proc. of the European Conference on Artificial Intelligence (ECAI), IOS Press, 2006, pp. 215–219.
[15] F. Dignum, D. Morley, E. Sonenberg, L. Cavedon, Towards socially sophisticated BDI agents, in: Proc. of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2000, pp. 111–118.
[16] D. Dubois, H. Prade, A review of fuzzy set aggregation connectives, Information sciences 36 (1-2) (1985) 85–121.
[17] J. Elster, Social norms and economic theory, Journal of Economic Perspectives 3 (4) (1989) 99–117.
[18] M. Fagundes, H. Billhardt, S. Ossowski, Reasoning about norm compliance with rational agents, in: Proc. of the European Conference on Artificial Intelligence (ECAI), IOS Press, 2010, pp. 1027–1028.
[19] D. Gaertner, Argumentation and Normative Reasoning, PhD thesis, University of London, 2008.
[20] A. Kakas, P. Mancarella, F. Sadri, K. Stathis, F. Toni, The KGP model of agency, in: Proc. of the European Conference on Artificial Intelligence (ECAI), 2004, p. 33.
[21] M. Kollingbaum, Norm-Governed Practical Reasoning Agents, PhD thesis, University of Aberdeen, 2005.
[22] F. López y López, M. Luck, M. dInverno, A normative framework for agent-based systems, Computational & Mathematical Organization Theory 12 (2) (2006) 227–250.
[23] A. Ortony, G. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge Univ Pr, 1990.
[24] T. Rumbell, J. Barnden, S. Denham, T. Wennekers, Emotions in autonomous agents: comparative analysis of mechanisms and functions, Autonomous Agents and Multi-Agent Systems (2011) 1–45.
[25] F. Sadri, K. Stathis, F. Toni, Normative KGP agents, Computational & Mathematical Organization Theory 12 (2) (2006) 101–126.
[26] E. Shortliffe, B. Buchanan, A model of inexact reasoning in medicine, Mathematical Biosciences 23 (3-4) (1975) 351–379.
[27] H.A. Simon, Models of bounded rationality, Emperically Grounded Economic Reason, vol. 3, MIT press, 1997.
[28] R.H. Thomason, Desires and defaults: a framework for planning with inferred goals, in: Proc. of the International Conference on Principles of Knowledge Representation and Reasoning (KR), Morgan Kaufmann, 2000, pp. 702–713.