European University Institute

DEPARTMENT OF LAW

# EUI Working Papers

LAW 2012/27
DEPARTMENT OF LAW

INTENTIONAL COMPLIANCE WITH NORMATIVE SYSTEMS

Giovanni Sartor

**EUROPEAN UNIVERSITY INSTITUTE, FLORENCE**
**DEPARTMENT OF LAW**

# *Intentional Compliance with Normative Systems*

**GIOVANNI SARTOR**

EUI Working Paper **LAW** 2012/27

This text may be downloaded for personal research purposes only. Any additional reproduction for other purposes, whether in hard copy or electronically, requires the consent of the author(s), editor(s). If cited or quoted, reference should be made to the full name of the author(s), editor(s), the title, the working paper or other series, the year, and the publisher.

ISSN 1725-6739

© 2012 Giovanni Sartor

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
www.eui.eu
cadmus.eui.eu

## Abstract

I will address a challenge to mentalistic theories of norms, such as that developed by Cristiano Castelfranchi and Rosaria Conte, namely, the existence of large normative systems, which successfully direct people's thoughts and actions without being, in their entirety, mental contents of individual agents. I will argue that the cognitive attitudes and operations involved in compliance with normative systems are usually different from those involved in complying with isolated social norms. While isolated norms must be stored in the memory of the agents endorsing them, this does not happen with regard to large normative systems. In the latter case, the agent adopts a general policy-based intention to comply with the normative system as a whole, an intention that provides an abstract motivation for specific acts of compliance, once the agent has established that these acts are obligatory according the system. I will show how the endorsement of such a policy can be based on different individual attitudes, ranging from self-interest to altruistic, social or moral motivations. Finally, I will analyse how a normative system may both constrain powers and extend them, relying on this abstract motivation of its addressees.

## Keywords

Norms, Intentions, Normative Systems, Compliance, Obligations

# INTENTIONAL COMPLIANCE WITH NORMATIVE SYSTEMS

*Giovanni Sartor**

## 1. Introduction

The theory of norms is one of the (many) areas where Cristiano Castelfranchi has produced influential contributions, relevant to multiple disciplines (psychology, sociology, computing, legal theory, etc.). In his seminal book on "Cognitive and social action", co-authored with Rosaria Conte, an original perspective on normativity is developed, where norms are understood as twofold objects, having a mental as well as a societal side. Norms are viewed as complex mental objects, resulting from an architecture of goals and beliefs (Conte and Castelfranchi 1995, Ch. 5, 6, 7). In particular, Conte and Castelfranchi start with the idea that a normative belief consist in the belief that for everybody it is obligatory to accomplish a certain action. They argue that such a belief presupposes the belief that someone, the sovereign, wants the obligation to hold, and is accompanied by further beliefs about the sovereign, namely that the sovereign is disinterested and pursues legitimate goals. Finally, they argue that one's goal to perform an action is normative if it is relativised to the existence of a corresponding normative belief (the goal is pursued as long as the normative belief is held). Conte and Castelfranchi's account also includes the analysis of how one becomes a defender of a norm, rather than merely an addressee of it, and how normative attitudes can spread in society.

These ideas have been further developed in a number of contributions, where Castelfranchi and his col- leagues have broadened and deepened their analysis of normative attitudes and behaviour, and of the social dynamics related to the emergence of norms (Andrighetto et al. 2007). Moreover the analysis of normative aspects has been felicitously connected to other domains of inquiry, such as trust and conventions (see for instance among the recent contributions, Tummolini and Castelfranchi 2006, Tummolini et al. 2011).

I will here address a challenge to mentalistic theories of norms, i.e., the views that a norm's existence results from the norm itself being the content of appropriate mental states of the concerned agents (such as the shared belief that the norm is binding, and the goal or intention to comply with it). This challenge results from the fact that we follow not only shared social norms, but also complex normative systems: while shared social norms are represented in the mind of the concerned agents, large normative systems direct people's thoughts and actions without becoming, as a whole, mental objects for individuals.[1] We are often faced with systems of this kind in our daily life (the legal system, but also the prescriptions of an institutionalised religion, or the regulations of a company, a condominium, a regulated market, a teaching institution, a sociotechnical infrastructure such as an airport or a harbour, etc.). All norms of such a system cannot be stored in one's memory since they exceed human capacities (at least for the largest normative systems, such as a municipal law, containing many thousands, even millions, of rules) and moreover such norms persistently change as a consequence of intervening facts (such as the adoption of new regulations, new decisions interpreting, them, etc.). For instance, while each of us has some knowledge of a few rules of our legal system (the ones corresponding to shared moral rules, such as the prohibition of killing, or most frequently encountered, such as certain traffic rules, or governing one's particular activity, such as rules on software copyright for a computer programmer), generally the common citizen has a very

* To be published in In Paglieri, F., Tummolini, L., Falcone, R., and Miceli, M., editors, The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi. College Publications, London.

[1] The term agent is here used as in AI, to mean an entity endowed with cognitive capacities and capable of autonomous action; it is not used in the legal-economical sense of someone delegated to act on behalf of another.

vague idea of the content the law of his or her country, especially in technical domains such as tax law, land planning law, environmental law, etc.

When referring to a large normative system *N* an agent usually does not immediately find an answer to the question "What ought I to do?" (As it usually happens when applying a shared social norm). One rather needs asks oneself (or the appropriate expert) "What does *N* require from me?", i.e., "What ought I do to according to *N*?" The answer to this question ("I ought to do action *A* according to *N*") does not have, by itself, a motivating force for the agent. It is not a normative belief of the kind described in Conte and Castelfranchi (1995), but a belief about what is entailed by a normative system in combination with the relevant facts. The concerned agent may well refuse to take into account the system's requests (for instance one may ask oneself what a certain religion requires from oneself, without having the slightest intention to follow the prescriptions of that religion, whatever they may be).

I will suggest that the motivation to perform a particular action qualified as obligatory by a normative system results from a general intention to comply with the system as a whole. The latter attitude provides an abstract motivation for specific acts of compliance, once the addressee has established that certain actions are obligatory according to the system. I will show how the endorsement of such an intention can be based on different individual attitudes. Finally, I will analyse how a normative systems may both constrain social powers and extend them, relying on this abstract attitude of its addressees.

## 2. Preliminary Notions: Actions, Obligations, Norms

For analysing compliance, we need some basic notions. First, a way of expressing action and obligations is required. For actions I will use the simple *E* operator of Pörn (1977), though other action logics would be appropriate as well for this discussion of compliance (on the *E* operator see also Sergot 2001, for a different approach to action, see for instance, Horty 2001).

***Definition 1 (Actions)***

*Let proposition $E_jS$ describe agent j's positive action consisting in the production of state of affairs S, where "S" is any proposition. Thus $E_jS$ means "j brings it about that S". Similarly, let $\neg E_jS$ describe the negative action (the omission) consisting in not bringing about that S. Thus $\neg E_jS$ means "j omits to bring about that S" or "j does not bring it about that S". When the distinction between positive and negative action is not relevant, let us use $A_j$ to cover both. Let $\overline{A}_J$ denote the complement of Aj ($\overline{A}_J$ stands for $\neg E_jS$ if $A_j = E_jS$; it stands for $E_jS$ if $A_j = \neg E_jS$).*

For simplicity when an agent brings about its own action, I will not repeat the agent's name in the action's result. Thus, for expressing the idea that *John* smokes (*John* brings it about that he smokes, meaning that *John* does the action of smoking) rather than writing $E_{John}Smoke(John)$, I will write $E_{John}Smoke$

This notion of an action does not involve intentionality (an aspect which is involved in the notion of an action as a goal-directed behaviour in Conte and Castelfranchi 1995). I prefer to stick to this minimal understanding of agency since compliance with normative systems usually prescinds from an action's intentionality: holding the required behaviour is usually sufficient for compliance. Intentions may instead be relevant for the consequences of violations (where intention may be required, or negligence, for certain normative consequences to take place), an aspect that I am not considering here.

As an example of an action-proposition, consider the following

$$E_{John}Damaged(Tom) \tag{1}$$

which means "*John* brings it about that *Tom* is damaged", or more simply "*John* damages *Tom*" while the following

$$\neg E_{John}Damaged(Tom) \tag{2}$$

means "*John* does not bring it about is about that *Tom* is damaged", or more simply "*John* does not damage *Tom*". I do not need to discuss here the logic of *E*, which is a classical modal logic (if *A* and *B* are logically equivalent, then $E_xA \rightarrow E_xB$), including inference rule

$$\frac{A}{\neg E_x A} \tag{3}$$

and axiom schema

$$E_xS \rightarrow S \tag{4}$$

Inference rule (8) says that one cannot realise what is a logical theorem (a necessary truth). For instance since $A \vee \neg A$ is a necessary truth, being a theorem of propositional logic, *Tom* cannot be said to bring it about (it would hold independently of his action).

Axiom schema (4) says that that if the state of affairs *S* is realised though an action, then it is the case that *S*. For instance the fact that *Tom* makes it so that *Ann* suffers damage, obviously entails that *Ann* suffers damage:

$$E_{Tom}Damaged(Ann) \rightarrow Damaged(Ann) \tag{5}$$

***Definition 2 (Obligations and Prohibitions)***

*Let Ø denote obligation. $OE_jS$ means "it is obligatory that j brings it about that S". Similarly $O\neg E_jS$ means "it is obligatory that j does not bring about that S", or "it is forbidden that j brings about that S".*

For instance, the following means "it is obligatory that John makes it so that Tom is compensated", or more simply, "it is obligatory that *John* compensates *Tom*",

$$OE_{John}Compensated(Tom) \tag{6}$$

while the following means "it is obligatory that John does not makes it so that Tom is damaged", or more simply, "it is forbidden that John damages Tom".

$$O\neg E_{John}Damages(Tom) \tag{7}$$

I will not specify here a particular deontic logic, since the following considerations may apply to different deontic logics. The reader may assume, for instance, standard deontic logic, as characterised in Føllesdal and Hilpinen 1971, but my preference would to a simpler deontic logic, limited to the substitution of logically equivalent formulas inside the deontic operator, namely, the schema:

$$\frac{\mathcal{A}\leftrightarrow\mathcal{B}}{O\mathcal{A}\leftrightarrow O\mathcal{B}} \tag{8}$$

Permission can be defined as usually as the negation of a prohibition:

$$\mathcal{P}\boldsymbol{\mathcal{A}} \stackrel{\text{def}}{=} \neg O\bar{\mathcal{A}}$$

To keep the language as simple as possible, I shall not address how a deontic language can be enriched through Hohfeldian concepts (for a logical analysis, Sartor 2006), and how this this extension can be useful for addressing compliance (Siena et al. 2009). While I am making use of the E action logic, I consider that the ideas on compliance here developed are generally compatible also with approaches to deontic reasoning based on different logics for action.

***Definition 3 (Norms)***

*I represent norms as defeasible conditionals*

$$[A \stackrel{n}{\Rightarrow} B] \tag{9}$$

*where A is a proposition and B is any kind of normative qualification, deontic or non deontic, and* $\stackrel{n}{\Rightarrow}$ *expresses normative conditionality, namely the link between an antecedent (possibly empty) and the normative consequent that is generated by that antecedent. A norm including variables stands for the set of all of its ground instances.*

I take normative conditionals to be non truth-functional, but to allow for (defeasible) modus ponens. Note that the conditional $A \stackrel{n}{\Rightarrow} B$ is not a statement of fact, but can rather be viewed as rule, according to which consequent *B* is produced (it holds, according to the normative system being considered) when the antecedent *A* holds. Here is an example of two deontic norms, the first stating that it is forbidden to cause damage to others, and the second that who causes a damage to another has the obligation to compensate the latter (in the following when obvious I drop the requirement $x \neq y$):

$$[x \neq y \stackrel{n}{\Rightarrow} O\neg E_xDamaged(y)] \tag{10}$$

$$[x \neq y \wedge E_xDamaged(y) \stackrel{n}{\Rightarrow} OE_xCompensated(y)]$$

The following is an example of a constitutive norm, saying that if we injure a person (make so that someone is injured), we cause damage to that person (injuring counts as damaging):

$$[E_xInjured(y) \stackrel{n}{\Rightarrow} E_xDamaged(y)] \tag{11}$$

Also concerning the normative conditional $\overset{n}{\Rightarrow}$ I will not provide a full logical account. I will just require that it enables defeasible detachment (*modus ponens*), i.e., that from *A* and norm $A \overset{n}{\Rightarrow} B$, the conclusion *B* can be inferred.

$$\{A, A \overset{n}{\Rightarrow} B\} \mid\sim B \tag{12}$$

I use the symbol $\mid\sim$ for non-monotonic derivability, assuming that normative conditionals are inherently defeasible, but in this paper I will not discuss defeasibility and its logical treatment (see Prakken and Sartor 2003, Sartor 2011).

Note that I do not distinguish deontic conditionals and constitutive or counts-as conditionals (Searle 1995, Jones and Sergot 1996, Grossi et al. 2008), assuming that the same inferences apply to both (on normative conditionality, see Sartor 2005; on the connection between deontic and constitutive conditionality, see Boella and van der Torre 2006). The following example shows how from a conditional and an instance of its antecedent we can defeasibly derive an instance of the conditional's consequent.

$$\{E_{Tom}Damaged(John), E_x Damaged(y) \overset{n}{\Rightarrow} OE_x Compensated(y)\} \mid\sim OE_{Tom}Compensated(John) \tag{13}$$

## 3. Relativised Normative Statements (in particular Obligations)

In addressing compliance we have to connect a normative system *N* (a set of norms) and (the propositions describing) the factual circumstances *C* relevant to *N*'s application. Here I am only interested in the obligations and the institutional facts that are generated by norms in *N*, when applied to facts in *C*. Thus we can assume that *C* contains (or entails) all factual literals (atomic sentences or negations of them) which are true in the real or hypothetical situation (the world) in which the norms have to be applied, without considering how the truth of such literals can be established. For simplicity's sake we can limit *C* to the factual literals that are relevant to the application of norms in *N*, matching literals in the antecedent of a norm in *C*. When the considered factual circumstances are those that hold in the real world (rather than in a merely possibly situation), i.e., they are the truths relevant to the application of *N* in the case at hand, I shall denote them through the expression *T (N).*

I will now introduce relativised normative statements, expressing that a proposition (in particular, an obligation) holds with regard to a normative system.

***Definition 4 (Relativised Normative Statements)***

*We say that any proposition B holds relatively to normative system N and circumstances C, and write* $[B]_{N,C}$ *iff* $N \cup C \mid\sim B$

$$[B]_{N,C} \stackrel{\text{def}}{=} N \cup C \mid\sim B \tag{14}$$

In particular when the proposition which is affirmed to hold is an obligation $O\mathcal{A}_x$, we abbreviate the corresponding normative statements $[O\mathcal{A}_x]_{N,C}$ *as* $\mathbb{O}_{N,C}\mathcal{A}_x$.

***Definition 5 (Relativised Obligation-Statements)***

*We say that it is obligatory relatively to N and C that x does A, and write* $\mathbb{O}_{N,C}\mathcal{A}_x$*, to express that* $N \cup C \mid\sim O\mathcal{A}_x$*:*

$$\mathbb{O}_{N,C}\mathcal{A}_x \stackrel{\text{def}}{=} N \cup C \mid\sim O\mathcal{A}_x \tag{15}$$

*According to Definition (5), a relativised obligation statement does not express a norm, but it expresses an assertion about the implications of norms (normative systems) and circumstances (in the terminology of Alchourrón 1969 and Alchourrón and Bulygin 1971 such assertions are called "normative propositions").*

*When we are referring to the true relevant circumstances of the real world—i.e., to the set of truths relevant to the application of N , denoted as T (N )—, rather than to circumstances of hypothetical situations, we simply write* $[B]_N$ *, or* $\mathbb{O}_N \mathcal{A}_x$*.*

$$[B]_N \stackrel{\text{def}}{=} N \cup T(N) \mid\sim B \tag{16}$$

$$\mathbb{O}_N \mathcal{A}_x \stackrel{\text{def}}{=} N \cup T(N) \mid\sim O\mathcal{A}_x$$

For instance, let us consider the following example, where $N_1$ includes a simplified version of the three norms above, and $C_1$ is limited to the fact that *John* injured *Tom*:

Example 1

$$C_1 = \quad \{E_{John}Injured(Tom)\}$$

$$N_1 = \{[E_x Injured(y) \overset{n}{\Rightarrow} E_x Damaged(y)];$$
$$[O\neg E_x Damaged(y)];$$
$$[E_x Damaged(y) \overset{n}{\Rightarrow} OE_x Compensated(y)]\} \quad (17)$$

It is easy to see that the following inferences holds on the basis of example (1):

$$(C_1 \cup N_1) \mid\sim E_{John} Damaged(Tom)$$
$$(C1 \cup N1) \mid\sim O\neg E_{John} Damaged(Tom)$$
$$(C1 \cup N1) \mid\sim OE_{John} Compensated(Tom) \quad (18)$$

Therefore, we can say that, relatively to $N_1$ and $C_1$, *John* has damaged *Tom*, it is obligatory that John does not damage Tom, and it is obligatory that *John* compensates *Tom*:

$$[E_{John} Damaged(Tom)]_{N1,C1} \wedge \mathbb{O}_{N1,C1} \neg E_{John} Damaged(Tom) \wedge$$
$$\mathbb{O}_{N1,C1} E_{John} Compensated(Tom) \quad (19)$$

If *John* has really injured *Tom* (and no other relevant circumstances obtain, such as exception excluding the application of the norms at issue), i.e., if $C_1 = T(N_1)$, we can simply say that according to $N_1$, *John* has damaged *Tom*, he ought not to damage him, and he ought to compensate him:

$$[E_{John} Damaged(Tom)]_{N1} \wedge \mathbb{O}_{N1} \neg E_{John} Damaged(Tom) \wedge$$
$$\mathbb{O}_{N1} E_{John} Compensated(Tom) \quad (20)$$

Here is another small example. The first norm in $N_2$ says that places open to the public are (count as) public places. The second says that if one is in a public place then one is forbidden to smoke.

Example 2

$$C_2 = \{OpenToPublic(LectureRoom); in(John, LectureRoom); E_{John} smoke\}$$
$$N_2 = \{[OpenToPublic(y) \overset{n}{\Rightarrow} PublicPlace(y)]; \quad (21)$$
$$[PublicPlace(y) \wedge in(x, y) \overset{n}{\Rightarrow} O\neg E_x Smoke]\}$$

We can say then say that according to $N_2$ given circumstances $C_2$ it is obligatory that *John* does not smoke ($\mathbb{O}_{N2,C2} \neg E_{Tom} Smoke$), and that *John* violates this obligation ($Violated_{N2,C2)} O\neg E_{John} smoke$)).

The extent of the set of action obligatory according to normative system *N* depends on the content of *N*, but also on the deontic logic we have adopted for *N*. For instance, if we adopt standard deontic logic for *N*, then if $N \mid\sim O\mathcal{A}$ and $\mathcal{A} \rightarrow \mathcal{B}$, then $N \mid\sim O\mathcal{B}$. This will not hold if instead we adopt the minimal deontic logic we described above (which requires that $\mathcal{A} \leftrightarrow \mathcal{B}$).

We can however recover the extent of obligations according to standard deontic logic, by defining a broader notion of a relativised obligation. For instance, following the idea of a logic of satisfaction, we could say that action A is weakly obligatory, relatively to a normative system *N*, if is entailed by actions that are obligatory relatively to the system.

The language of relativised obligation allows us to say that according to different normative systems different obligations hold. For instance, given that Canon law contains a universal norm prohibiting the use of contraception as well as a constitutive rule saying any action meant to make a sex act unfruitful counts as artificial contraception, and given that taking the pill in order to prevent pregnancy is meant to make subsequent sex acts unfruitful, we can conclude that according to the Canon law a woman, say Ann, is forbidden to take the pill in order to prevent pregnancy. Similarly, given that

Islamic law contains a norm that prohibits receiving interest on loans of money, we can say that according to Islamic law John is forbidden to receive interest on loans of money.

A notion of relativised permission can be provided that corresponds to the above analysis of an obligation.

***Definition 6 (Relativised Permission)***

*Let us say that it is permissible relatively to N and C that x does $\mathcal{A}$, and write $\mathbb{P}_{N,C}\mathcal{A}_x$ iff N and C entail $P\mathcal{A}_x$*:

$$\mathbb{P}_{N,C}\mathcal{A}_x \stackrel{\text{def}}{=} N \cup C \mid\sim P\mathcal{A}_x \qquad (22)$$

Note that according to this definition, saying that an action $E_xS$ is permissible relatively to normative system $N$ and circumstances $C$ ($\mathbb{P}_{N,C}\ E_xS$) does not amount to saying that it is not the case that $E_xS$ is forbidden relatively to the same system and circumstances ($\neg\mathbb{O}_{N,C}\ \neg\ E_xS$). Proposition $\mathbb{P}_{N,C}E_xS$ is not equivalent to $\neg\mathbb{O}_{N,C}\neg E_xS$, since the former holds when $N \cup C$ entails $PE_xS$, while the latter holds when $N \cup C$ does not entail $O\neg E_xS$ (see Alchourrón 1969, Alchourrón and Bulygin 1971).

## 4. Compliance

With the help of the notions introduced in the previous section, we can now address compliance. The issue of compliance can arise in very different context, as the following examples shows:

- Mary is appointed to a professorship. She signs a contract stating her commitment to comply with the University regulations.
- John enters a PhD program. He is directed to the booklet containing the regulations he has to comply with.
- Linda is appointed as a judge. She takes an oath to respect the Constitution and the laws of her country.
- Adolf Eichmann enters the SS. He takes an oath of obedience to death to Adolph Hitler and the superiors he has designated.
- Antony enters the Franciscan order. He promises to respect the body of regulations known as "The Rule of St. Francis" as well as the law of the Catholic Church.
- Mary, a shop-owner, receives a threats by gangsters belonging to a mafia organisation. She chooses to comply with all rules imposed by that organisation (monthly protection money, code of silence, etc.) to avoid problems with the bad guys.
- A digital agent enters and electronic marketplace. It commits to respect all rules of the marketplace.

In all these contexts the agent has taken the commitment (adopted the intention to) comply with a certain normative system. We can distinguish different notions of compliance. The first notion is behavioural compliance, which simply consist in behaving is such a way as to fulfil an obligation.[2]

***Definition 7 (Behavioural Compliance)***

*An agent x behaviourally complies with an obligation $O_N\mathcal{A}_x$ of a normative system N , iff the obligation holds according to N and x's behaviour counts as A according to N , i.e., iff*

$$\mathbb{O}_N\mathcal{A}_x \wedge [Ax]_N) \tag{23}$$

For instance, if a non-smoker does not smoke in a public office, ignoring that there is a prohibition to do so (she does not know about the prohibition in Example (2) above), she will still behaviourally comply with that prohibition. She will do that even if she is taking a siesta, and therefore is not aware that she is not smoking. On the basis of this notion of behavioural compliance we can develop the idea of conscious compliance, which consists in complying with a an obligation, while being aware that it is entailed by a certain normative system.

***Definition 8 (Conscious Compliance)***

*An agent x consciously complies with an obligation $OE_xS$ of a normative system N, iff x behaviourally complies with the obligation, believes that the normative system entails that obligation, and is aware of doing the required action, i.e., iff*

$$\mathbb{O}_N\mathcal{A}_x \wedge [Ax]_N \ \wedge Bel_x\,(\mathbb{O}_N\mathcal{A}_x) \wedge Bel_x\ ([Ax]_N\ ) \tag{24}$$

[2] As above, I will often omit to make explicit reference to the circumstances in which a normative set N is to be applied, assuming that an implicit reference is made to T (N ), the true circumstances relevant to the application of N .

*By assimilating knowledge and true belief we can say that x consciously complies with an obligation $O\mathcal{A}_x$ according to N iff x knows that, according to systems N, x is doing an action which is obligatory:*

$$Knows_x([\mathcal{A}_x]_N) \wedge Knows_x(\mathbb{O}_N \mathcal{A}_x) \quad (25)$$

Many instances of compliance with norms in a legal system are unconscious: the concerned agent is not aware that the law prescribes a certain behaviour, but behaves correspondingly, either motivated by moral or social norms or by any other factors (self interest, altruism, etc.).

## 5. Intentional Compliance

Here I am interested with acts of compliance motivated by the (belief) in the existence of an obligation relatively to a normative system. First of all the action considered must be intentional, namely motivated by the intention to perform it, and moreover such an intention must be motivated by the awareness of the obligation.

***Definition 9 (Intentional Compliance)***

*An agent x intentionally complies with an obligation according to N, when N entails that obligation ($\mathbb{O}_N \mathcal{A}_x$), and x's belief that this is the case ($Bel_x(\mathbb{O}_N \mathcal{A}_x)$) motivates x to intend to hold the prescribed behaviour ($Int_x \mathcal{A}_x$), which, in its turns motivates x to hold that behaviour ($\mathcal{A}_x$).*

$$\mathbb{O}_N \mathcal{A}_x \wedge (Bel_x(\mathbb{O}_N \mathcal{A}_x) \rhd^m Int_x \mathcal{A}_x) \wedge (Int_x \mathcal{A}_x \rhd^m [\mathcal{A}_x]_N) \quad (26)$$

where $\rhd^m$ denotes motivation, understood as mental causation.

This definition would require refinements, linked to the difficulties inherent to the notion of motivation, which I cannot address here. Let me just state that I take $M_1 \rhd^m M_2$ to be true, when both $M_1$ and $M_2$ are true and the fact that the agent instantiated $M_1$ was the reason why the agent subsequently instantiated $M_2$. With regard to the notion of an intention, I assume that the unconditioned intention to perform an action or omission consists in having the chosen goal to perform the action (for a discussion of the connection between goal and intentions, and for the proposal of a refined formalisation, see Castelfranchi and Paglieri 2007):

$$Int_x \mathcal{A} = CGoal_x A_x \quad (27)$$

For my purpose (and given that I do not need to distinguish actions and omissions) this simple notion of an intention will suffice.

Observe that one can perform an action wrongly believing that one is under an obligation. This is the situation where there is no obligation to behave in a certain way, but the agent believes that such an obligation exists.

$$\neg \mathbb{O}_N \mathcal{A}_x \wedge (Bel_x(\mathbb{O}_N \mathcal{A}_x) \rhd^m Int_x \mathcal{A}_x) \wedge (Int_x \mathcal{A}_x \rhd^m \mathcal{A}_x) \quad (28)$$

Let us now consider again example (1), and assume that $C_1 = T(N)$ ($C_1$ contains all true circumstances relevant to the application of N ). Given that

$$\mathbb{O}_N E_{John} Compensated(Tom) \quad (29)$$

we can say that John behaviourally complies with that obligation if *John* performs the obligatory action:

$$E_{John} Compensated(Tom) \quad (30)$$

We can say that *John* intentionally complies if he has the obligation, and his awareness of having the obligation leads him to intend to perform the obligatory action, which leads him to perform it:

$$Bel_{John}(O_N E_{John} Compensated(Tom)) \rhd^m Int_{John} E_{John} Compensated(Tom) \wedge$$
$$Int_{John} E_{John} Compensated(Tom) \rhd^m E_{John} Compensated(Tom) \quad (31)$$

## 6. Compliance with a Normative System

So far we have been considering compliance with a single obligation established by a normative systems. Now we need to consider compliance with a whole normative system, possibly including thousands of obligations (as any modern legal system).

***Definition 10 (Compliance with a Normative Systems)***

*An agent x complies with a normative system N, iff x complies with all obligations established by N. In other words, x complies with N, iff x performs every action $[\mathcal{A}_x]_i$ which is obligatory according to N:*

$$Complies_x(N) \stackrel{\text{def}}{=} [\mathcal{A}_x]_1 \wedge \ldots \wedge [\mathcal{A}_x]_n \quad (32)$$

where $[\mathcal{A}_x]_1 \wedge \ldots \wedge [\mathcal{A}_x]_n$ is the conjunction of every action or omission $[\mathcal{A}_x]_i$ such that $\mathbb{O}_N[\mathcal{A}_x]_i$, i.e., such that $N \cup T(N) \mid\sim O[\mathcal{A}_x]_i$.[3]

Complying with the whole of a normative system *N* (rather than with a single obligation) can be the object of a deliberation, on the basis of which an agent j adopts the corresponding goal, i.e., the goal of j's own compliance, which becomes j's intention to comply. Thus a consequentialist agent *j*, given that for him the utility of complying is higher than the utility of non-complying

$$u_j(Complies_j(N)) > u_j(\neg Complies_j(N)) \quad (33)$$

would assume that the utility of performing the action of complying (making so that he complies) is higher than the utility of omitting to do so

$$u_x(E_j Complies_j(N)) > u_j(\neg E_j Complies_j(N)) \quad (34)$$

Consequently, given the principle stated in definition (14) we can conclude that an agent j believing in Proposition (33) will adopt the intention achieve compliance.

$$Int(E_j Complies_j(N)) \quad (35)$$

However, it seems to me that the representation of the intention to comply in formula (35) above (namely, as an agent's intention to achieve a state of affairs where every obligation of that agent is fulfilled) fails to capture the usual state of mind of of an agent who has decided to comply with a normative system. In fact, an agent usually cannot have a precise mental representation of the state of affairs of full compliance, as specified in definition (10), since the agent ignores the norms in the system, and therefore cannot know what needs to be done to achieve full compliance. For instance, we all know that our country has a legal system, some of us know a few criteria for identifying the norms belonging to that system, but none of us knows all or most norms it contains. How can we intend to realise a state of affair of which we are not aware?

This objection be countered by conditionalising the actions to be performed to achieve compliance. Even if we cannot know what actions are obligatory, we can still intend to performs any action which happens to be obligatory. This is expressed by the following definition.

***Definition 11 (Compliance with a Normative Systems (Conditionalised Version))***

*An agent x complies with a normative system N, iff x complies with all obligations established by N. In other words, x complies with N, iff whenever an action or omission by x, denoted as $[\mathcal{A}_x]_1$, is obligatory according to N, x performs it:*

[3] To avoid infinite conjunction of redundant action propositions, we may add the requirement and for each such [Ax ]i there must exist an instance of a norm in N , whose conclusion is [OAx ]i .

$$Complies_x(N) \stackrel{\text{def}}{=} \bigwedge_{i \in [1 \ldots n]} (\mathbb{O}_N [\mathcal{A}_x]_i \rightarrow [\mathcal{A}_x]_i \quad (36)$$

*where* $\bigwedge_{i \in [1 \ldots n]}$ *(*$\mathbb{O}_N [\mathcal{A}_x]_i \rightarrow [\mathcal{A}_x]_i$*) stands for the conjunction of all formulas having the form* $\mathbb{O}_N [\mathcal{A}_x]_y \rightarrow [\mathcal{A}_x]_y$ *, one per each of x's action* $[\mathcal{A}_x]_i$ *prescribed by one of the norms of N.*

Also this representation, however, seems inadequate to me. Firstly, we do not know what antecedents of the conditionals included in the big conjunction will turn out to be true, and thus to what actions we are committing ourselves. Can we as rational agent intend, without qualifications, to bring about full realisation of an open set of demands whose content is unknown to us?

Secondly, even we could have a mental representation of the state of full compliance, we should know that this state of affairs is unlikely to happen: given the high number of obligations arising from the system, and the fact that we is not aware of many of them, we will most likely violate some of them, even though we are doing are best. How can one intend to realise a state of affairs being aware that most likely this state of affairs will not take place?

Thirdly, an agent committed to compliance should maintain its motivation even when the agent has failed to comply with one obligation, and even when when the agent deliberately chooses not to comply with one particular norm. But full compliance is an all or nothing state of affairs, which becomes impossible once one obligation is violated.

## 7. Policy-Based Intention to Comply

It seems to me that rather than committing itself to achieving full compliance, a reasonable agent could consider adopting a general compliance policy, namely the policy of intending to perform any action which is obligatory according to *N*. Thus the intention to comply will appear to be a policy-based intention, namely, an intention to act in a certain way under conditions characterised in a general way, so that they may be instantiated in different specific circumstances (on such policy-based intentions, see Bratman 1987, 87-92 and Bratman 1989,451 ff., for a formalisation in defeasible logic see Governatori et al. 2009, for some considerations, see also Sartor 2005, 31-40). According to this policy, the agent will comply whenever the conditions are met, giving a separate and independent relevance to each opportunity for compliance: the agent may fail to comply in one occasion (when the agent ignores that the conditions are met, or when overriding reasons exist defeating the application of the policy), but still keep a defeasible commitment to the policy and be governed by it in other occasions.

***Definition 12 (Policy-Based Intentions)***

*Let us represent policy-based intentions in the form:*

$$S \overset{i}{\Rightarrow} Int_j \mathcal{A}_j \tag{37}$$

*where* $\overset{i}{\Rightarrow}$ *is a non-truthfunctional connective (similar to* $\overset{n}{\Rightarrow}$ *for norms), meaning that the state of affairs S (the belief that it holds) triggers agent j's intention to do action* $\mathcal{A}_j$*.*

I assume that also that a modus ponens-like inference applies to $\overset{i}{\Rightarrow}$, so that:

$$\{S, S \overset{i}{\Rightarrow} Int_j \mathcal{A}_j \} \mid\sim Int_j (\mathcal{A}_j) \tag{38}$$

In fact, intentions often take a conditional form, which supports detachment. For instance, Tom, given that today is a working day and that he intends to work today if it is a working day, can conclude with the intention to work today.

$$\{workingDay(today), workingDay(today) \overset{i}{\Rightarrow} IntTomETomWork(today)\} \mid\sim$$
$$Int_{Tom}E_{Tom}Work(today) \tag{39}$$

Conditional intentions can have an abstract form, which enables multiple instantiations. For instance, agent Tom may have the following policy-based intention to work on any working day x:

$$workingDay(x) \overset{i}{\Rightarrow} Int_{Tom}E_{Tom}Work(x) \tag{40}$$

A general conditioned intention stands for the set of all of its ground instances, such as:

$$workingDay(Tomorrow) \overset{i}{\Rightarrow} Int_{Tom}E_{Tom}Work(Tomorrow) \tag{41}$$

so that from such a general intention, given a specific fact matching its antecedent, like

$$workingDay(Tomorrow) \tag{42}$$

Tom can infer the corresponding instance of the conclusion:

$$Int_{Tom}(E_{Tom}Work(Tomorrow)) \tag{43}$$

However, Tom does not need to store in his mind all of such ground instances (that he intends to work today if today is a working that, that he intends to work tomorrow if tomorrow is a working day, that he intends to work the day after tomorrow . . . ). he just needs the policy-based intention expressed in abstract terms, and can use it for specific inferences when needed.

Let us now consider how the commitment to comply with a normative system can be modelled as a policy- based intention.

***Definition 13 (Policy-Based Intention to Comply)***

*An agent j's commitment to comply with normative system N can be understood as the agent's j conditioned intention to do any action $\mathcal{A}_j$ that is obligatory according to N:*

$$\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j \tag{44}$$

Assume, for instance that *Tom*, while being in a place open to the public, is considering the implications of the normative system $N_2$ of example (2) (which says that places open to the public count as public spaces, and that it is forbidden to smoke in public places). Then *Tom* can establish that he is forbidden to smoke according to $N_2$:

$$\mathbb{O}_{N2} (\neg E_{Tom} Smoke) \tag{45}$$

Assume also that Tom has adopted the following policy-based intention to comply with N2:

$$\mathbb{O}_{N2} \mathcal{A}_{Tom} \overset{i}{\Rightarrow} Int_{Tom} \mathcal{A}_{Tom} \tag{46}$$

one of whose grounds instances is:

$$\mathbb{O}_{N2} (\neg E_{Tom} Smoke) \overset{i}{\Rightarrow} Int_{Tom} (\neg E_{Tom} Smoke) \tag{47}$$

From (45) and (47) Tom can conclude that he intends to abstain from smoking:

$$Int_{Tom} (\neg E_{Tom} Smoke) \tag{48}$$

As this example shows, the meaning of the policy-based intention to comply consists in its inferential role: it works in the agent's mind as defeasible rule, allowing the derivation of an instance of its conclusion given (the belief in) an instance of its antecedent. Its peculiarity in comparison to other inference policies is that its conclusion is an intention to be implemented, rather than a proposition to be believed. In conclusion, we have found two ways to understand the commitment (intention) to comply with a normative system *N* by an agent *j*:

- j's intention to realise the state of affairs where all obligations directed to j are satisfied through its action ($Int_j (E_j Complies_j (N)$)
- j's endorsement of the policy according to which j intends to comply with any N -obligation directed to itself ($\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j$)

It seems to me that there is only a one-way dependency between these two intentions. Adopting the latter policy-based intention is the most obvious way to realise (at least to some extent) the state of affairs of one's compliance. However, the converse does not hold: j may adopt the compliance policy, even when j does not intend to realise full compliance, knowing that it is not possible to achieve it. Moreover, such a policy may be limited by specific exceptions, whose detection would prevent the application of the policy (and would take *j* further away from full compliance), as I shall argue in section (9).

## 8. Compliance by Different Kinds of Agents

Compliance is neutral: the choice to comply with a normative system may result from the most different attitudes and goals. It is even doubtful whether in many cases a choice is involved in the adoption of the attitude to comply. When one lives in a certain community one tends to adopt the norms which are endorsed and followed in that community without the need of a specific act of choice. Correspondingly, when we know that our community has a normative system, but we don't know what rules belong to that system, we tend to adopt a general policy to comply with whatever rules will belong to that system, i.e., the policy-based intention above described. This happens in the communities in which we participate without an explicit choice (such as a country, a local community, a family, etc.), but also in those organisations that we enter by choice (a university, a company, a sport club, etc.), where a compliant attitude appears as a natural implication of one's choice to join a certain group or activity, rather than as a separate independent choice. Different explanations can be provided for the unreflected adoption of a determination to comply. For instance, it has been affirmed that humans are naturally endowed with the attitude of "docility", meant as "the propensity to behave in socially approved ways and to refrain from behaving in ways that are disapproved", and attitude that may have an evolutionary explanation since it "enhances human fitness tremendously by allowing children to enjoy a long period of dependence, and to acquire effective skills through learning" (Simon 1983, 64). So, it seems that humans living within a certain organisation or community would "naturally" desire to be included and approved, and consequently adopt the goal (the intention) to comply with the norms of that organisation or community.

This fact, however, does not exclude that one's intention to comply may be the result of a deliberate choice. Such a choice may provide the motivation for compliance even when one has no desire to be involved in a certain organisation or community. For instance a prisoner in concentration camp may choose to comply with the regulation of the camp, for fear of sanctions linked to non-compliance. He may also criticise those who do not comply (rather than approving of their courage), for fear of retaliation.

In other cases, a conscious deliberation to comply may support an existing insufficient commitment to do so. For instance a rebellious teenager may accept that he should comply with the school regulations (or with the law more generally) when convinced that non-compliance can easily get him into trouble.

Even people already having a certain propensity to comply may engage in a deliberation on whether to comply or not, when critically assessing whether they should or not maintain this attitude.

Different agents may have different ways of approaching the deliberation on whether they should comply with a norm or a normative system. For our purposes it is sufficient to focus on a broad category of agents, *consequentialist choosers*, namely, agents choosing their actions on the basis on an assessment of the consequences of such actions, an assessments determined by the expected utility (differential benefit) the agent expects as a result of the action. Here the notion of "result of an action" is understood is a very broad way, including the fact of adopting the action itself, as well as the further consequences of this fact (for a broad notion of consequentialism, see Pettit 1997).

I will distinguish two aspects involved in the assessment of a choice by an agent:

- the utility of action $\mathcal{A}_x$ according to agent *x*, denoted by $u_x\mathcal{A}_x$, i.e., the measure of the net desirability of that choice, according to *x*'s assessment,
- the impact of an action $\mathcal{A}_x$ on the well being of a subject *y* according to *x*, denoted by $w_y\mathcal{A}_x$, i.e., the measure of how much $\mathcal{A}_x$ advances or diminishes *y*'s well-being, according to *x*'s assessment.

Let us first characterise the general idea of a consequentialist chooser.

***Definition 14 (Consequentialist Chooser)***

*A consequentialist chooser x will intend to do an action $\mathcal{A}_x$ whenever x believes that the expected utility of doing that action is superior to the utility of not doing it:*

$$Bel_x(u_x(\mathcal{A}_x) > u_x(\overline{\mathcal{A}_x})) \rightarrow Int_x \mathcal{A}_x \quad (49)$$

Let us now distinguish different kinds of consequentialist choosers:

- Self-centred (egoistic). For a self-centred chooser *x*, the utility of a choice is equal to the choice's impact on *x*'s own well-being: $u_x(\mathcal{A}_x) = w_x(\mathcal{A}_x)$.
- Altruistic. For an altruistic chooser *x* the utility of a choice corresponds to its impact on the wellbeing of a set of agents, possibly including also (but not only) *x*: $u_x(\mathcal{A}_x) = w_{y\,1}(\mathcal{A}_x) + \ldots + w_{ym}(\mathcal{A}_x)$, where $y_1 \ldots y_m$ are the agents *x* considers relevant to its choice.
- Communitarian. For a communitarian chooser *x*, the utility of a choice corresponds to its impact on the wellbeing of *x*'s community: $u_x(\mathcal{A}_x) = w_g(\mathcal{A}_x)$, where *g* is the community *x* cares about.
- Utilitarian. For a utilitarian chooser *x*, the utility of a choice corresponds to the sum of its impacts on the wellbeing of each human being $u_x(\mathcal{A}_x) = w_{y1}(\mathcal{A}_x) + \ldots w_{y\,n}(\mathcal{A}_x)$ where $y_1 \ldots y_n$ are all human beings (by "utilitarianism", I mean the idea that the "standard of what is right in conduct, is not the agent's own happiness, but that of all concerned", Mill 1991, Ch. 2).

Clearly, different kinds of consequentialist choosers will take different actions in the same situation. For instance when an action positively affects *x*'s welfare, but negatively affects relevant others to a larger extent, a self-centred agent will do it, but an altruistic (or utilitarian) agent will not. However all consequentialist choosers act with the purpose of increasing utility, and consequently, they should address the issue of endorsing the general policy-based intention of fulfilling any obligation established by a certain normative system, i.e., the intention to comply as expressed in by formula (44) above, in the following way. Assume that *j* believes that a higher utility will be obtained by adopting the policy to comply rather that by not having this policy (to express that a policy-based intention is considered as a whole in *j*'s reasoning about it, I enclose it in square brackets):

$$u_j([\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j]) > u_j(\neg[\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j]) \quad (50)$$

According to (50), making so that *j* has (acquires or maintains) the policy-based intention to comply is better than omitting to do that

$$uj(E_j[\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j]) > u_j(\neg E_j[\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \mathcal{A}_j]) \quad (51)$$

From (51), *j* can conclude that it intends to acquire (bring it about that it has) the intention-based policy to comply:

$$Int_j E_j[\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j(\mathcal{A}_j]) \quad (52)$$

Executing such an action, i.e., achieving that intention, would consist in adopting the policy-based intention to comply, namely, being ready to form the intention to perform an action $\mathcal{A}_j$ whenever (*j* believes that) this action is obligatory according to *N*.

For my purposes I do not need to engage in a discussion of the logic of meta-intention. It is sufficient to assume that a rational agent *x*, having the intention to perform the action consisting in adopting a (conditioned or unconditioned) intention $INT_x$ will perform such a mental action and acquire $INT_x$, according to the following schema:

$$Int_x E_x(INT_x) \rightarrow E_x INT_x \quad (53)$$

Given that actions are successful by formula (4) above, performing $E_x INT_x$ entails acquiring $INT_x$, i.e., in our example, adopting the policy-based intention to comply.

Various refinements and extensions of the consequentialist model of agency are indeed possible: intermediate positions could be distinguished (as when one is moderately altruistic, giving some importance to the well-being of others, but less importance than to one's own well being) or egalitarian-prioritarian elements may be introduced (so that the differential welfare or certain people is more significant than that of others). The bounds of rationality could also be considered, and the ways in which the social environment influences attitudes and choices. Finally, the analysis of compliance could also go beyond consequentialist reasoning, extending to cases where compliance follows from a deontological ethics (for a discussion of deontology and consequentialism, see Baron et al. 1997) or from a religious faith. All these refinements and extensions of the model here proposed are beyond the scope of this contribution, where I will limit my analysis to the simplistic typology of consequentialist reasoners just proposed.

## 9. Non-Compliance

An agent may also choose not to comply or to be indifferent to compliance. We can distinguish different ideas in this regard.

Firstly, the agent may be completely indifferent to compliance. In this case, for any obligation $O\mathcal{A}_j$, the fact that the obligation is prescribed by $N$ is no motivation for $j$ to perform. From $j$'s perspective, the $N$ -obligatoriness of an action is no reason to (intend to) do it (I write $A \not\Rightarrow B$ to mean that the conditional $A \Rightarrow B$ does not hold, is not applicable):

$$\mathbb{O}_N \mathcal{A}_j \not\overset{i}{\Rightarrow} Int_j \mathcal{A}_j \tag{54}$$

Secondly, the j may be diabolic, as far as N is concerned (in the sense of wanting to violate N 's obliga- tions just for the sake of doing it). For such a j, the very fact that an action Ax is obligatory according to N provides a motivation to violate N . In other terms, j has adopted the policy of doing the contrary of anything obligatory according to N :

$$\mathbb{O}_N \mathcal{A}_j \overset{i}{\Rightarrow} Int_j \overline{\mathcal{A}_j} \tag{55}$$

Thirdly, $j$'s commitment to compliance may be limited, since $j$ together with the compliance policy also adopts one or more exception-policies to it, namely, rules stating that the compliance policy does not hold under certain conditions (such rules would be undercutters, in the model of Pollock 1995, see also Prakken and Sartor 1997 and Prakken 2010). Different defeasible compliers may recognise different exceptions.

An opportunistic complier $j$ (the bad man, see Holmes 1897) makes an exception to the compliance policy whenever $j$ comes to believe that by violating an obligation it will get a higher personal advantage (well- being) than complying with it. Thus $j$ would adopt the following reasoning policy, which blocks the defeasible compliance policy of formula (44) above whenever the utility of non-compliance exceeds that of compliance: when the utility of doing $\mathcal{A}_j$ is inferior to the utility of not doing it, then the obligatoriness of $\mathcal{A}_j$ does not provide a (defeasibly sufficient) reason to have the intention to do it.

$$w_j \mathcal{A}_j < w_j \overline{\mathcal{A}_j} \overset{i}{\Rightarrow} (\mathbb{O}_N \mathcal{A}_j \not\overset{i}{\Rightarrow} Int_j \mathcal{A}_j) \tag{56}$$

Note that the opportunistic complier is not uncommitted toward compliance: $j$ still has the defeasible commitment to comply expressed by formula (44) above, but this commitment is overridden by the belief that non-compliance (in a particular case) would get $j$ a better outcome.

Effective sanctions could neutralise in many cases the opportunistic complier's exception, by making it so that that for any action $\mathcal{A}_j$, $j$'s expected utility of non-compliance (once that the risk of sanctions is also taken into account) is inferior to the utility of compliance. This however depends of the expected impact of the sanction on $j$, namely, on the amount of the punishment and its probability, which should outweigh the advantage that $\mathcal{A}_j$ would provide if there were no sanction.

Not all exceptions to the compliance policy are determined by self-interest. For instance, if Ann believes in some versions of natural law, or in some doctrine supporting civil disobedience, she would make an exception to her policy to comply with $N$, whenever she believes the $N$ requires her to do an action $\mathcal{A}_{Ann}$ which is (unbearably) unjust. Thus *Ann* would adopt the following policy, according to which when an action of her is unjust, then its obligatoriness according to $N$ is not a defeasibly sufficient reason for intending to do it:

$$Unjust(\mathcal{A}_{Ann}) \overset{i}{\Rightarrow} (\mathbb{O}_N \mathcal{A}_{Ann} \not\overset{i}{\Rightarrow} Int_{Ann} A_{Ann}) \tag{57}$$

Other kinds of exceptions could be distinguished. For instance an act utilitarian agent would make an exception to the compliance policy whenever it considers that complying causes more harm than good to humanity. Similarly a corruptible agent would make an exception to the compliance policy when by non-complying the agent would get a substantial differential personal advantage (the amount required for leaning toward non-compliance, being inversely proportional to the corruptibility).

Note that according to this construction of compliance, there is no direct clash (no-balancing) between one's conditioned intention to comply and the reason for holding a different behaviour. Rather the agent needs to consider whether such reasons instantiate an undercutter for the agent's intention to comply, i.e, whether they exclude the applicability of the compliance-policy to the situation under scrutiny. Such exception may also be introduced when an agent j is aware of its cognitive limitations.

## 10. Endorsement of Norms and Commitment to Comply

Research on social norms has recently addressed social processes through which norms are shared in a community, namely, the interlinked processes of the social emergence of norms and of their immergence in the mind of the concerned agents (Andrighetto et al. 2007). Besides considering the spontaneous emergence of shared customary rules, also the psycho-social process involved in compliance with authoritative orders has been studied (Conte and Castelfranchi 1999). However, I think that a further step is required to adequately capture the reasoning involved in the application of complex norm-systems.

Let us consider for instance a municipal tax law, such as the Italian one (which is a section of the larger Italian legal system). First of all, very few people have precise knowledge of a large set of rules from Italian tax law, and nobody's mind contains all of Italian tax law. It would be difficult to claim that such rules have "immerged" (and are stored) in the minds of Italian citizens since most the latter do not know (and have never known) most of those rules. What citizens share is only the ability to identify somehow the law in force in their country as distinguished from other laws (foreign or ancient laws) and a general commitment (in many case a very qualified one), to comply with this law and possibly some criteria to identify its main contents. Citizens also have some ideas on the implications of this law that are most important to them (e.g., that the law requires them to pay the income tax every year, that VAT has to be paid on purchases, etc.), but are unable to determine such implications with precision (on the distinction between identifying the law and determining its content, see Jori 2011).

Usually common citizens usually approach tax issues with the help of tax experts, who give them some indications of what obligations follow from tax law under specific real or hypothetical cases, what sanctions may follow from violating such obligations, what line of actions are most advantageous with regard to tax-law effects. On the basis of this fragmentary information, law-abiding people will determine how to comply with tax law. Let us try to analyse the reasoning process involved in applying this kind of normative information (and more generally all complex normative systems, such as advanced legal systems).

Let us assume that Tom has a general commitment to comply the normative systems *L* (the law), which includes many tax regulation (without knowing what it the precise content of *L*). In other words, he endorses the policy based intention to perform any action that is obligatory according to *L* (the law):

$$\mathbb{O}_L\, \mathcal{A}_{Tom} \overset{i}{\Rightarrow} Int_j\, \mathcal{A}_{Tom} \tag{58}$$

Tom is now wondering whether he should pay income tax on the capital gains he obtained by selling his house. Being committed to comply with the law, but not knowing what the law requires from him, Tom asks the tax expert Ann for advise. Assume that the Ann remembers that there is a rule in the tax code that establishes the requirement to pay income taxes on capital gains, but vaguely remembers that there are exceptions to it. This prompts Ann to look for exceptions, and she finds indeed one matching houses. This exception says (in a simplified form) that capital gains from the sale of houses purchased more than 5 year before the sale and inhabited by the seller are exempted from income tax. Assume that Ann's inquiry has let that to conclude that the legal system L contains certain norms:

$$\begin{aligned} L \supseteq \{ & SellsHouse(x) \overset{n}{\Rightarrow} OE_x PayIncomeTaxOnSale; \\ & BoughtMoreThan5YearsBefore(x) \wedge HasInhabitedHouse(x) \overset{n}{\Rightarrow} \\ & \neg(SellsHouse(x) \overset{n}{\Rightarrow} OE_x PayIncomeTaxOnSale)\} \end{aligned} \tag{59}$$

where the second norms in (59) says that under the indicated conditions the first one does not hold (is not applicable).

*Ann* then asks *Tom* whether at the time of the sale more that 5 years had elapsed from *Tom*'s purchase, and whether he has been living in the house. Assume that *Tom* replies positively to the first question and negatively to the second one. Then *Ann* says: "Dear, *Tom*, unfortunately you are legally bound to pay income tax on your gains". In fact, by combining the law *L* with these factual circumstances (let us assume these circumstances are the only relevant ones), *Ann* can see that the following inference holds:

$$L \cup \{\neg HasInhabitedHouse(Tom)\} \mid\sim OE_{Tom}PayIncomeT axOnSale \quad (60)$$

so that she can infer what she tells her client:

$$\mathbb{O}_L E_{Tom}PayI ncomeTaxOnSale \quad (61)$$

If *Tom* asks for an explanation, *Ann* would probably answer by saying that whenever one was has not lived in the house one sells, then according to the law one has the obligation to pay income tax:

$$SellsHouse(x) \wedge \neg HasInhabitedSoldHouse(x) \rightarrow \mathbb{O}_L E_{(x)}PayI ncomeTaxOnSale \quad (62)$$

Note that formula (62) does not express a norm of *L* (there is no norm in *L* which has exactly that content, see formula (59)). More generally (62) is no norm at all, but rather is a general conditional statement about *L*, namely the statement that in case that the seller has not inhabited the sold house, then *L* entails that the seller has to pay taxes on capital gains. Similarly, if *Ann* were contacted by *Tom* before making the sale, she would tell him: "Since you have not inhabited the house, you will have to pay income tax on your capital gain".

I think that this example may suffice to show that norms included in large normative systems operate differently from social norms. When we learn social norms we permanently store them in our memory, as the content of appropriate normative beliefs and goals, so that they can directly govern our behaviour. On the contrary, we do not learn and store in our memory most norms included in a large normative systems. We rather possess some ideas about the existence of such a system and the ways to identify its content. When needed, we collect some fragmentary information about the system and combine this information with the relevant facts, both tasks being often delegated to experts. On the basis of this information we can conclude that the system requires us to perform certain actions. By combining such conclusions with our general commitment to comply with the system we adopt intentions to perform such actions.

## 11. Compliance by Officers

Certain normative systems have officers (typically judges) charged with ensuring compliance, in particular by sanctioning non-compliance.

For each obligation $OA_x$, let us denote with *Punished(x)*, the situation where *x* is punished. Let us assume for simplicity's sake that a single compliance officer, a judge named *Jud*, who is responsible for ascertaining and repressing all violations of *N*, and that the punishment is the same for all violations. In other words, let us assume that *N* contains a norm stating than whenever an obligation $OA_x$ is violated, the obligatory action not having been accomplished, it is *Jud*'s obligation to punish *x*:

$$OA_x \wedge \overline{\mathcal{A}_j} \overset{n}{\Rightarrow} OE_{Jud}Punished(x) \tag{63}$$

Note that for rule in formula (63) above to fire, $OA_x$ must be derivable from *N* itself in combination with the facts of the case (thus we do not need to substitute $OA_x$ with the metalevel normative proposition $\mathbb{O}\mathcal{A}_x$).

This representation is a simplification with regard to complex normative systems, where we have multiple interlocked rules determining who is in charge for each kind of violation, and we need to distinguish offi- cers charged with providing evidence of violations, officers charged with establishing whether a violation has taken place and order a sanction, officers charged with carrying out the sanction. In fact officers in a complex normative system have a shared task which is more complex that simple punishment, a task which may possibly be characterised as the development and maintenance of their normative system. To accomplish this task they need to coordinate, to some extent, their activities consisting in creating, modifying, interpreting and applying the norms in the system (see Shapiro 2002 who sees this activity as a shared cooperative activity in the sense of Bratman 1992). For the purposes of this paper, however, a simplistic analysis will suffice.

Thus compliance by *Jud* (in its role as law enforcer) could be expressed as follows:

$$CompliesWith_{Jud}(N) \overset{\text{def}}{=} \forall(x)(\mathbb{O}_N E_{Jud} Punished(x) \rightarrow E_{Jud} Punished(x)) \tag{64}$$

*Jud*'s commitment to a policy-based intention to comply could be expressed as the intention to punish anybody it has the obligation to punish (namely, anybody who violated a norms):

$$OE_{Jud}Punished(x) \overset{i}{\Rightarrow} Int_{Jud}E_{Jud}Punished(x) \tag{65}$$

Following the reasoning in Sections 8 and 9 above we may consider the various conditions under which different judges, having different concerns, could adopt policy (65): they could adopt it out self-interest (to advance their career, have a good reputation, etc.), altruism, communal interests, moral commitments, and any mixtures of these and other motivations. Moreover, they may subject this policy to various limitations. For instance, a corruptible judge will not apply the policy when there is great advantage to be gained through non-compliance.

## 12. Spreading Compliance

I will not examine here the social determinants of compliance and the social factors that encourage or discourage compliance, which would require me to address the many issues dealing with the theory of social norms (see, for instance, Conte and Castelfranchi 2006, Bicchieri 2011) and their connection to legal systems.

I will just observe that the expected utility of x complying with N often depends on how many other agents will comply with *N*, as officers or as private individuals. As the number of compliers of a normative systems *N* increases usually both the individual and social differential benefit of compliance (as compared to non-compliance) increases: in a context of compliance, legal and social sanctions for non-compliance are more likely to take place, compliance is more likely to have a socially beneficial effect, compliance can be viewed as an exercise in reciprocity. This explains why in a context of increased (decreased) compliance, individuals are usually more (less) motivated to comply: so both compliance and non-compliance tend to spread in the community. For most people there is a threshold of compliance-frequency that makes the utility of compliance positive, so that they would choose to comply when the threshold is overcome. However, this threshold may be different for different people (both officers or common fellows) who may be differently motivated (by the individual, social, or communal benefit of compliance).

Thus, compliance by agents who are sufficiently motivated only where there is a higher compliance frequency may depend on whether there is a sufficient number of other agents sufficiently motivated at lower compliance levels, who can bootstrap the process.

Clearly a more complex picture could be developed though a more accurate and diversified representation of motivations for compliance (see for instance Bénabou and Tirole 2006), which may indeed lead different agents to different choices. For instance one may comply with norms that others do not com- ply with, in order to better advertise one's commitment to the common good, or to get a confirmation of one's morality; a Kantian agent should be unmoved by the non-compliance by others to a norms the agent approves of; a "myopic" agent would only care about the compliance by the nearest neighbours, etc.

I cannot here address all the many issues concerning the spreading of normative attitudes, an issue to which Castelfranchi and his colleagues have dedicated a number of important contributions (see for instance Andrighetto et al. 2007). One general consideration, however, is that people will have normative expectations about the compliance by others, and this expectation will be strengthened insofar as other people as a matter of fact do comply. Here not only one's belief in the value of having a certain normative system, but also reciprocity is at issue, as well as the fact that people will make their own choices (and take risks) on the basis of the factual expectation that others will comply.

## 13. The Morality of Compliance

When a normative system *N* is generally complied with and enforced, there will be usually a general attitude of viewing compliance as morally obligatory. This may indeed support the adoption of the intention to comply.

Let us assume that an agent's morality *M* (the set of moral norms the agent endorses) contains a norm stating the obligatoriness of whatever is obligatory (for any agent *x*) relatively to a certain normative system *N*:

$$(\mathbb{O}_N \mathcal{A}_x \overset{n}{\Rightarrow} O\mathcal{A}_x) \in M \tag{66}$$

If an agent *j* believes in proposition (66), and that $\mathcal{A}_j$ is really obligatory according to *N* (i.e., that $\mathbb{O}_N \mathcal{A}_j$ ), *j* will conclude that the obligation to do $\mathcal{A}_j$ is entailed by morality:

$$M \cup T(M) \mid\sim O\,\mathcal{A}_x \tag{67}$$

Thus, *j* will view action $\mathcal{A}_j$ as morally obligatory (according to definition (5), i.e., *j* will believe that

$$\mathbb{O}_M \mathcal{A}_j \tag{68}$$

Rule (66), when applied to the norms governing a political organisation (typically a state) expresses the idea of the political obligation, namely, the moral obligation to obey the law.

The obligation to comply may be qualified by exceptions (e.g., one may argue that it is not morally obligatory to comply with norms enjoining a serious violation of human rights, or which are blatantly unjust or absurd) especially when non-compliance is done in public to convey a political message urging resistance or change, so that it may qualify as civil disobedience.

The idea that that there is a moral obligation to obey a normative system N can contribute to compliance with N , as long as the concerned agent j is committed to do what is required by morality (as identified by j itself), i.e., as long as j endorses the following policy:

$$\mathbb{O}_M \mathcal{A}_j \overset{i}{\Rightarrow} Int_x \mathcal{A}_j \tag{69}$$

Thus *j*, believing that it has the obligation to do action $\mathcal{A}_j$ according to *N* (i.e, $\mathbb{O}_N \mathcal{A}_j$ ) can use moral rule (66) to conclude that it has a moral obligation to do $\mathcal{A}_j$ (i.e., $\mathbb{O}_M \mathcal{A}_j$) and consequently use policy (69) to adopt the intention of doing $\mathcal{A}_j$. Those who endorse rule (66) will also tend to extend their moral condemnation to the violators of norms in *N*.

In conclusion, moral beliefs may ground or reinforce the endorsement of a policy to comply, but this is not always the case, since the adoption of such a policy may also follows from self-interest or other motivations, as shown above.

## 14. Compliance and Social Power

The model of compliance here proposed can be related to the theory of power and influence proposed in Castelfranchi (2003). The basic idea I will use is that an agent *j* influences another agent *k* when *j* makes it so that *k* adopts *j*'s goals, and that influence is a most important mechanism for social power.

Let us assume a state of generalised compliance, so that all (or most) addressees of normative system N have adopted a policy-based intention to comply with *N*, according to the model indicated in formula (44) above. I will argue that under these conditions a normative systems can be an efficient machine not only for limiting, but also for producing influence and power.

Obviously, normative systems can limit social influence. For instance, assume that John is physically stronger than Tom. If there were no legal system prohibiting the use of violence, John could influence Tom and induce him to (intend to) accomplish what John likes (working for John, paying John for protection, etc.), by threatening to use violence against Tom. However, this is no longer possible (or at least more difficult) when there is an effective legal system *N* which prohibits using violence against others. If John himself is rigorously committed to the policy to comply with *N*, then John will adopt the intention to abstain from prohibited actions, and therefore also from violence. In case John is not committed to compliance (or is only defeasibly committed to it, with his self-interest providing for an exception), the compliance of others (and in particular of the enforcement officers) will make it so that the criminal behaviour is prevented of at least made less attractive by the prospect of punishment. This should prevent the threat or make it not credible. Therefore John will not use the threat, or at least Tom will not be influenceable through it.

Let us now examine how normative systems, rather than limiting social influence, can extend it. We need to consider that what obligations are generated by *N* depends on two factors: the norms in *N* and the true relevant factual circumstances *T (N)*. This means that *N* can work as an input-output machine. The input consists in changes in *T (N)* (the creation of new relevant facts), and the output consists changes in the obligations entailed by *N*. The input can produce the output in two ways: *(a)* by providing (or removing) facts that produce obligations according to the norms in *N*, or *(b)* by changing the norms in *N*, these changes having an impact on the obligations derivable from *N*. In this section I will consider the first way of changing *N*'s obligations, and in the following I will address the latter.

For instance a normative system can make orders binding (for instance, the orders of a military commander to a soldier, or of an employer or manager to a worker), by making obligatory for the addressee of an order to comply with it. This idea could also be expressed by using the notion of institutional (norm-based) power (Jones and Sergot 1996, Gelati et al. 2002a, Sartor 2006, Hage 2011b, Hage 2011a, Tummolini and Castelfranchi 2006), but here a simpler representation will be provided, without expressly formalising the concept of institutional power. Assume the system *N* contains a rule according to which *Ann* has the obligation to do whatever action $\mathcal{A}_{Ann}$ is ordered by her manager *Tom* (for simplicity I do not consider the limitation of such an obligation in modern legal systems, where the order must pertain to the execution of the work, and respect the worker's rights and dignity):

$$E_{Tom}Order(\mathcal{A}_{Ann}) \overset{n}{\Rightarrow} \mathbb{O}\mathcal{A}_{Ann} \tag{70}$$

Assume that Tom does indeed order Ann to do something (for instance, to draft the minutes of a meeting):

$$E_{Tom}Order(E_{Ann}DraftMinutes) \tag{71}$$

so that this action-proposition becomes one the true relevant facts

$$(E_{Tom}Order(E_{Ann}DraftMinutes)) \in T\,(N\,)) \tag{72}$$

Given that *N* contains rule (70) and *T (N)* contains fact (71) the following holds:

$$N \cup T(N) \mid\sim OE_{Ann}DraftMinutes \tag{73}$$

so that we can say that according to *N* it is indeed obligatory that *Ann* drafts the minutes

$$\mathbb{O}_N E_{Ann}DraftMinutes \tag{74}$$

Assume that Ann has adopted the general compliance policy of formula (44) above relatively to normative system *N*, so that she intends to do whatever action of her is obligatory according to *N*:

$$\mathbb{O}_N \mathcal{A}_{Ann} \overset{i}{\Rightarrow} Int_{Ann}\, \mathcal{A}_{Ann} \tag{75}$$

Policy-based intention (75) and normative proposition (74) entail that *Ann* will adopt the intention to draft the minutes

$$Int_{Ann}E_{Ann}DraftMinutes \tag{76}$$

Thus, given that *Ann* is committed to comply with *N*, *Tom* can influence her. By ordering any action, he modifies *T (N)* and makes it so that $N \cup T(N)$ entails the obligatoriness of that action, which makes it so that *Ann* adopts the intention of doing that action. Note that this power by *Tom* does not depend on his personal qualities (*Ann* may dislike *Tom* or believe that he an incapable idiot), it only depends on the content of the normative system, on the relevant facts, and on *Ann*'s commitment to policy-based intention (75).

A normative system N can also provide individuals with the possibility of binding themselves, i.e., of undertaking obligations according to *N*, or more generally of creating any normative positions concerning themselves. For this purpose it is sufficient that *N* contains the following rule:

$$[E_x\, Promise\, (\mathcal{A}_x) \overset{n}{\Rightarrow} O\mathcal{A}_x] \tag{77}$$

meaning that whenever an *x* promises to do $\mathcal{A}$ then *x* has the obligation to do $\mathcal{A}$.

A complied with (and protected through sanctions or other means of social pressure) normative system containing the rule in (77) enables agents to create credible commitment for themselves (given the costs of non-compliance), on the basis of which others can act (e.g., I promise to give 1,000 euros to the person who will bring back to me my lost dog), or can be induced to take similar commitments, as in contracts (on a more general approach to contract, which views them as means to create not just obligations but any kind of normative positions see Gelati et al. 2002b, Sartor 2006, Hage 2011b). For example, assume the following: 1) system *N* contains the rule in (77), 2) I promised that I will give 1000 euros to the best law student of this year; 3) Ann is this year's best law student. It follows that according to N , I have the obligation to give 1000 euros to Ann.

## 15. The Machine of the Law

Let us now consider how an agent (a legislator) can have the ability to introduce new norms in *N*. For this purpose, we need to assume that *N* is a dynamic normative system (Kelsen 1967), including meta-rules determining what new norms will belong to *N*. For simplicity I shall leave temporal aspects implicit even though they are essential in an adequate account of normative dynamics (see Governatori et al. 2007). So, let us assume that *N* includes a meta-norm saying that whatever norm φ is issued by the legislator *Leg* is included in *N* (φ is a variable ranging over norms):

$$[E_{Leg}\ Issued\ (\varphi) \overset{n}{\Rightarrow} \varphi \in N\ ] \tag{78}$$

I cannot here develop the analysis of the dynamics of normative systems, which would require a discussion on how to model defeasibility and time (see for instance Governatori et al. 2006). Thus, for our purposes it is sufficient to characterise N as the minimal set satisfying the following equality:

$$N = \{[E_{Leg}\ Issued\ (\varphi) \overset{n}{\Rightarrow} \varphi \in N\ ]\} \cup \{\psi : N \cup T\ (N\ ) \mid\sim \psi \in N\ \} \tag{79}$$

According to equation (79), N is defined as containing the meta-norm of (78) (which would work as the "constitution" in a logical sense of *N*, following Kelsen 1967) plus every other norm that is qualified as being in N according to N itself. i.e., any norm $\psi$ ($\psi$ is a variable ranging over norms such that *N* entails the proposition that $\psi$ is contained in *N* (for a presentation of this idea, see Sartor 2009, on modelling legal systems through metanorms, se also Yoshino 1995, Yoshino 1997 and Hernandez Marín and Sartor 1999).

Alternatively we could assume that that the content of equality (79) is rephrased by a fundamental norm, which is not does not belong to *N*, but constitutes the ultimate ground for membership to *N* (as a Kelsenian Grundnorm, or as a Hartian rule of recognition, see Hart 1994).

$$([E_{Leg}\ Issued\ (\varphi) \overset{n}{\Rightarrow} \varphi \in N\ ] \in N\ ) \wedge ((N \cup T\ (N) \mid\sim \psi \in N\ ) \overset{n}{\Rightarrow} \psi \in N\ ) \tag{80}$$

The two-pronged norm in (80), let us call it *Fundamental*, states the norm empowering the legislator is in *N*, and that all norms are in *N*, whose membership to *N* is entailed by *N* itself.[4] Then *N* can be defined as the minimal set of the norms whose legality is entailed by *Fundamental*, together with the relevant facts.

$$N = \{\varphi : (T\ (N) \cup Fundamental) \mid\sim \varphi \in N\ \} \tag{81}$$

Given this background (i.e., either equation (79) or (81)), let us assume that legislator accomplishes the action of issuing a new norm, for instance, a norm prohibiting any agent x to smoke:

$$E_{Leg}\ Issued(O\neg E_x Smoke) \tag{82}$$

The accomplishment of the action described in this formula is a fact, which is added to the true factual circumstance *T (N)*. With this addition, the following holds according to the rule of formula (78) above (when useful for clarity, I bracket norms included in meta-linguistic expression):

$$N \cup T\ (N) \mid\sim [O\neg E_x Smoke] \in N) \tag{83}$$

Consequently *N* contains norm $O\neg E_x Smoke$, according to formula (79):

$$[O\neg E_x Smoke] \in N \tag{84}$$

Since it now holds that

[4] The rule in (80) can also be rephrased as having a single conclusion (using variables in a very liberal way):

$$n\ ((\varphi = [E_{Leg}\ Issued(\varphi) \overset{n}{\Rightarrow} \varphi \in N\ ]) \vee (N \cup T\ (N\ ) \mid\sim \varphi \in N\ )) \overset{n}{\Rightarrow} \varphi \in N$$

$$N \cup T(N) \mid\sim O\neg E_{Tom}Smoke \tag{85}$$

so that we can say that now smoking is forbidden to Tom according to N :

$$\mathbb{O}_N \neg E_{Tom}Smoke \tag{86}$$

The legislator can use the power provided by formula (78) above to put a judge in charge of punishing violators. To achieve this result, the legislator just has to perform the action of issuing a norm to that effect, namely a norm saying that the judge J ud should punish any agent who violates a norm in N , i.e., any agent who does the opposite of what is obligatory for that agent:

$$E_{Leg}Issued(O\mathcal{A}_x \wedge \overline{\mathcal{A}}_x \overset{n}{\Rightarrow} OE_{Jud}\,Punished\,(x)) \tag{87}$$

As a consequence of this legislative action, *N* now contains the issued norm

$$[O\mathcal{A}_x \wedge \overline{\mathcal{A}}_x \overset{n}{\Rightarrow} OE_{Jud}\,Punished\,(x)] \in N \tag{88}$$

with means that J ud has, according to N , the obligation to punish any violator.

Assume now that both *Ann* and *Jud* have the policy-based intention to comply with *N*, and that *Ann* views non-compliance as immoral. Then, on the basis of the statement of the legislator, *Ann* will adopt the intention not to smoke, *Jud* will adopt the intention to punish smokers in public places, and *Ann* would believe that anyone who smokes in a public place behaves immorally.

The legislator can also confer to another agent, the administrator *Admin*, the ability to insert new norms in *N* (delegated legislation) by enacting such norms (while respecting certain legal constraints on *Admin*'s legislative action):

$$[E_{Leg}Issued(E_{Admin}Issued(\varphi) \wedge E_{Admin}RespectConstraints(\varphi) \overset{n}{\Rightarrow} \varphi \in N\,] \tag{89}$$

As a consequence of the action described in formula (89) and the characterisation of *N* in (79), the norm empowering *Admin* is now contained in *N*:

$$[E_{Admin}Issued(\varphi) \wedge E_{Admin}RespectC\ onstraints(\varphi) \overset{n}{\Rightarrow} \varphi \in N\,] \in N \tag{90}$$

Consequently whatever new norm $\varphi$ is issued by *Admin*, respecting the relative constraints (concerning the content of φ or the procedure for its creation), that norm will be inputted in *N*. In this way, the legislator transfers to *Admin* the legislator's ability to influence people's behaviour, by exploiting their commitment to compliance.

Not only the generalised commitment to comply with *N* provides the legislator (and its delegatees) with the possibility to influence the behaviour of compliers and judges. It also provides those who are able to influence the legislator with the ability to influence the behaviour of all others. Assume for instance that *Tom* is the leader of the party having the majority in the legislative assembly. Then *Tom* can make it so that the legislator adopts the intention to introduce (or repeal) a norm $B \overset{n}{\Rightarrow} \mathcal{A}$, to make it so that the population intends to do (and does) action $\mathcal{A}$ under circumstances *B*.

A normative system supported by a generally endorsed policy-based intention to comply can thus work as an input-output machine, empowering those who can control its input: by providing appropriate normative and factual inputs, they can obtain corresponding intentions and actions and so implement their aims. As Karl Olivecrona put it "[t]he purpose of the lawgivers is to influence the actions of men, but this can only be done through influencing their minds" (Olivecrona 1971, 21-2, Spaak 2009). Thus legislators (and those able to influence them) can use the "machinery of the law" for reaching their social, political (and sometimes personal) purposes (see Pattaro 2009, Pattaro 2005). Normative systems, in a way, precede certain social powers, and provide for their foundation. The extent of norm-systems based powers may indeed be very large, which explains why developed legal systems contain constitutional limitations and controls over the exercise of such powers (such as democratic procedures for electing the legislative body, judicial review over legislation and administration, more generally, an institutional system of "checks and balances").

## 16. Conclusion

I have first considered how obligations can be relative to a particular normative systems, and I have pro- vided a meta-logical representation of this idea. Then I have analysed the intention to comply with a normative system, affirming that that the commitment to comply must be understood as a policy-based intention. I have then considered why consequential choosers may come to this determination, as simple addressees or enforcement officers. Finally I have developed some considerations on how compliance can spread and how it can both restrain and provide power.

The study of compliance with normative systems involves various aspects I could not address here. First of all there is the issue of interpretation, i.e., of determining the content of the normative system to be complied with, on the basis of the available materials (texts, cases, practices, values, etc.), a problem that legal theorists have been discussing for centuries, and on whose epistemological-methodological nature the debate is still on-going. Other important issues concern modelling contrary to duty obligations (and other technical aspects of deontic logic), taking into account cooperation between the involved agents and dependencies and trust relationships between them, addressing negotiation and argumentation regarding how to comply and the consequences of violations, considering how a shared awareness of each one's intention to comply and a shared belief in a duty to comply can contribute to compliance.

Finally, the model here presented provides a minimal understanding of the internal point of view towards a normative system (i.e., the point of view of an agent that has chosen to use that system as a guide to its own behaviour). The analysis of such a point of view can be developed by adding further requirements, which may or may not apply with regard to particular normative systems or addressees of them: a social or conventional dimension (expected compliance by others contributes to motivate one's compliance), a shared dimension (there is a common awareness of each one's intentions to comply, or a common intention to comply), a cooperative dimension (the intention to comply concerns participation in a common project), a hierarchical-authoritative dimension (there an individual or collective agent having certain properties, such as those described in Conte and Castelfranchi 1995, 84ff, who issues and implements the norms), a believed moral dimension (compliance appears to the concerned agent as the content of a moral obligation), a claimed moral dimension (those producing and enforcing the system claim that there is a moral obligation to comply with it), a moral dimension tout court (there is a moral obligation to apply the system or comply with it), etc.

I think however that such aspects, are complementary but independent of the model developed here, which only assumes that the addressees of a normative system adopt a policy-based intention to comply with it, regardless of the reasons supporting this intention and the ways in which the system's content is identified.

While this work is still very preliminary, I hope it can provide some clues on how the project of identifying the cognitive basis of normative behaviour, a project to which Castelfranchi and his collaborators have given so many important contributions, can be extended to complex systems of norms, rather than being limited to social norms or specific orders of particular authorities.

## References

Alchourrón, C. E. (1969). Logic of norms and logic of normative propositions. Logique et analyse 12, 242–68.

Alchourrón, C. E. and E. Bulygin (1971). Normative Systems. Vienna: Springer.

Andrighetto, G., M. Campenni, R. Conte, and M. Paolucci (2007). On the immergence of norms: A normative agent architecture. In Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence.

Baron, M., P. Pettit, and M. Slote (1997). Three Methods of Ethics: A Debate. London: Blackwell.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior,". American Economic Review 96, 1652–78.

Bicchieri, C. (2011). Social norms. In Stanford Encyclopedia of Philosophy. Stanford University.

Boella, G. and L. van der Torre (2006). A logical architecture of a normative system. In Deontic Logic and Artificial Normative Systems, pp. 24–35. Springer.

Bratman, M. (1987). Intentions, Plans and Practical Reasoning. Cambridge, Mass.: Harvard University Press.

Bratman, M. E. (1989). Intention and personal policies. Philosophyical Perspectives 3, 443–469.

Bratman, M. E. (1992). Shared cooperative activity. Philosophical Review 101, 327–41.

Castelfranchi, C. (2003). The micro-macro constitution of power. ProtoSociology, 18-19, 208–65.

Castelfranchi, C. and F. Paglieri (2007). The role of beliefs in goal dynamics: Prolegomena to a construc- tive theory of intention. Synthese 155, 237–63.

Conte, R. and C. Castelfranchi (1995). Cognitive and Social Action. London: University College of London Press.

Conte, R. and C. Castelfranchi (1999). From conventions to prescriptions. towards a unified theory of norms. Artificial intelligence and Law 7, 323–40.

Conte, R. and C. Castelfranchi (2006). The mental path of norms. Ratio Juris 19, 501–17.

Føllesdal, D. and R. Hilpinen (1971). Deontic logic: An introduction. In R. Hilpinen (Ed.), Deontic Logic: Introductory and Systematic Reading. Dordrecht: Reidel.

Gelati, J., G. Governatori, A. Rotolo, and G. Sartor (2002a). Actions, institutions, powers: Preliminary notes. In G. Lindemann, D. Moldt, M. Paolucci, and B. Yu (Eds.), International Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA'02), pp. 131–47. Hamburg: Fachbereich Informatik, Universita¨t Hamburg.

Gelati, J., G. Governatori, A. Rotolo, and G. Sartor (2002b). Declarative power, representation, and mandate: A formal analysis. In Proceedings of the Fifteenth Annual Conference on Legal Knowledge and Information Systems (JURIX), pp. 41–52. Amsterdam: IOS.

Governatori, G., V. Padmanabhan, A. Rotolo, and A. Sattar (2009). A defeasible logic for modelling policy-based intentions and motivational attitudes. Logic Journal of IGPL 17, 36–69.

Governatori, G., M. Palmirani, A. Rotolo, R. Riveret, and G. Sartor (2006). Norm modifications in defeasible logic. In Proceedings of Jurix 2006, pp. 13–22. Amsterdam: IOS.

Governatori, G., A. Rotolo, R. Riveret, M. Palmirani, and G. Sartor (2007). Variants of temporal defeasible logics for modelling norm modifications. In Proceedings of Eleventh International Conference on Artificial Intelligence and Law, pp. 155–9. New York, N. Y.: ACM.

Grossi, D., J.-J. C. Meyer, and F. Dignum (2008). The many faces of counts-as: A formal analysis of constitutive rules. Journal of Applied Logic 6, 192–217.

Hage, J. C. (2011a). A model of juridical acts: Part 1: The world of law. Artificial intelligence and Law 19, 23–48.

Hage, J. C. (2011b). A model of juridical acts: Part 2: The operation of juridical acts. Artificial Intelligence and Law 19, 49–73.

Hart, H. L. A. (1994). The Concept of Law (2nd ed.). Oxford: Oxford University Press.

Hernandez Mar´ın, R. and G. Sartor (1999). Time and norms: A formalisation in the event-calculus. In Proceedings of the Seventh International Conference on Artificial Intelligence and Law (ICAIL), pp. 90–100. New York, N. Y.: ACM.

Holmes, O. W. (1897). The path of the law. Harvard Law Review 10, 457–78. Horty, J. F. (2001). Agency and Deontic Logic. Oxford University Press.

Jones, A. J. and M. J. Sergot (1996). A formal characterisation of institutionalised power. Journal of the IGPL 4, 429–45.

Jori, M. (2011). Del diritto inesistente. Pisa: ETS.

Kelsen, H. (1967). The Pure Theory of Law. Berkeley, Cal.: University of California Press.

Mill, J. S. (1991). Utilitarianism. In J. Gray (Ed.), On Liberty and Other Essays, pp. 131–201. Oxford: Oxford University Press. (1st ed. 1861.).

Olivecrona, K. (1971). Law as Fact (2nd ed.). London: Stevens.

Pattaro, E. (2005). The Law and the Right, a Reappraisal of the Reality that Ought to be, Volume 1 of Treatise of legal Philosophy and General Jurisprudence. Berlin: Springer. Pattaro, E. (2009). From ha¨gerstom to ross and hart. Ratio Juris 22, 532–48.

Pettit, P. (1997). The consequentialist perspective. In Three Methods of Ethics: A Debate, pp. 92–174. London: Blackwell.

Pollock, J. L. (1995). Cognitive Carpentry: A Blueprint for How to Build a Person. New York, N. Y.: MIT.

Pörn, I. (1977). Action Theory and Social Science: Some Formal Models. Dordrecht: Reidel.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. Argument and Computation 1, 93–124.

Prakken, H. and G. Sartor (1997). Argument-based extended logic programming with defeasible priorities.Journal of Applied Non-classical Logics 7, 25–75.

Prakken, H. and G. Sartor (2003). The three faces of defeasibility in the law. Ratio Juris 17, 118–39. Sartor, G. (2005). Legal Reasoning: A Cognitive Approach to the Law, Volume 5 of Treatise on Legal Philosophy and General Jurisprudence. Berlin: Springer.

Sartor, G. (2006). Fundamental legal concepts: A formal and teleological characterisation. Artificial Intelligence and Law 21, 101–42.

Sartor, G. (2009). Legality policies and theories of legality: From Bananas to Radbruch's formula. Ratio Juris 22, 218–43.

Sartor, G. (2011). Defeasibility in legal reasoning. In J. Ferrer (Ed.), Essays in Legal Defeasibility. Oxford: Oxford University Press. (Forthcoming.).

Searle, J. R. (1995). The Construction of Social Reality. New York, N. Y.: Free.

Sergot, M. J. (2001). A computational theory of normative positions. ACM Transactions on Computational Logic 2, 581–662.

Shapiro, S. J. (2002). Law, plans and practical reasoning. Legal Theory 8, 387–441.

Siena, A., J. Mylopoulos, P. A., and A. Susi (2009). Designing law-compliant software requirements, In Proceeding of the 28th International Conference on Conceptual Modeling (ER'09), pp. 472–86. New York, N.Y.: Spencer.

Simon, H. A. (1983). Reason in Human Affairs. Stanford, Cal.: Stanford University Press.

Spaak, T. (2009). Naturalism in Scandinavian and American realism: Similarities and differences. In Dahlberg (Ed.), De Lege, Uppsala-Minnesota Colloquium: Law, Culture and Values, pp. 33.83. Upp- sala: Iustus.

Tummolini, L., G. Andrighetto, C. Castelfranchi, and R. Conte (2011). A convention of (tacit) agreement betwixt us. Synthese.

Tummolini, L. and C. Castelfranchi (2006). The cognitive and behavioral mediation of institutions: To- wards an account of institutional actions. Cognitive Systems Research 7, 307–32.

Yoshino, H. (1995). The systematization of legal metainference. In Proceedings of the Fifth International Conference of Artificial Intelligence and Law (ICAIL). New York, N. Y.: ACM.

Yoshino, H. (1997). On the logical foundations of compound predicate formulae for legal knowledge representation. Artificial Intelligence and Law 5, 77–96.


European
University
Institute
European University Press