Are Incentives Good Enough To Achieve (Info)Social Order?

Rosaria Conte Cristiano Castelfranchi

Division of AI, Cognitive & Interaction Modelling
IP-CNR, V.le Marx 15 - Rome - I-00137 - Italy
email: rosaria&pscs2.irmkant.rm.cnr.it

Abstract

In this paper, the role of incentives insocial order is questioned, based on a notion of incentive as additionalindividual utility, provided by an external entity, to actions achievingglobal utility.

Two notions of norms are compared: (1) inputs which modify agents'decisions through incentives (sanctions) and (2) prescriptions to executeobligatory action for intrinsic motivations. Two types of agents whichreason upon norms are also compared: (1) incentive based rational deciders, and (2) normative agents which are prescribed to execute norms for intrinsic reasons. The twotypes of agents are expected to have a different impact on norm compliance.Under suboptimal conditions of application of sanctions (uncertianpunishment), transgression is expectedto propagate more easily and rapidly among incentive-based agents thanamong normative agents. In particular, incentive-based agents are expectedto show a fast decline and even a collpase in compliance with the norms.Normative agents are expected to exhibit an oscillating behaviour, or at least a graceful degradation ofcompliance. Finally, the role of incentives is shown to have a lesserimpact on natural social agents than expected by a model of rationaldecision. What is worse, incentives have been shown to produce even negative effects on several aspects of social learningand norm compliance.

1 The Problem of Social Order in Agent MediatedInteraction

The problem of social order in naturalsocieties is a traditional concern of social philosophers and social scientists. With some right, rational action theory defines it asa dilemma, a yet unsolved (and insoluble?) incompatibility betweenindividual and global utility. If individual agents are rational, that is,if they act to maximise their individual utility, they will inevitably achieve a globally undesirable state ofaffairs (some will gain at the expense of others). Moreover, individualutility maximisation has long-term self-defeating effects, since all agents(supposedly rational), will exploitand be exploited at once. Not surprisingly, the traditional (analytical)solution to this dilemma proposed by rational action scientists is aforward-looking agent, which calculates the short and long-term effects ofaction. But with bounded rationality,a forward-looking agent cannot accomplish a thoroughly rational action.Hence, the necessity for means, social and institutional, designed toregulate societies and achieve a socially acceptable state of affairs.Rational agents' decisions must be modified through positive or negative incentives (sanctions). Indeed, sanctions andincentives provide *additional individual utility forself-interested agents to act according to some global utility*.

To achieve social order has become an urgent problem in agent mediatedinteraction, both in software agent-human agent, and in softwareagent-software agent interaction. This should not come as a surprise, sincehuman agents are self-interested and software agents are designed to act in the interest of and on behalfof their users (Rao, 1998; Crabtree, 1998). In infosocieties as well as innatural societies, local and global utility are often incompatible, andindividual utility maximisation is found to produce long-termselfdefeating effects (Crabtree, 1998). We will then speak of the problem of *infosocial* order as a new version of the old problem, which calls for much the samemeans and solutions already experienced in natural societies. Softwareagents scientists and designers are well aware of this necessity, asis documented by the recent studies in agent mediated electronic commerce(Dignum, 2000). Of late, the problem of infosocial order gave rise to a newfield of investigation, i.e. the field of electronic institutions (seeagain, Dignum, 2000). Given the impact of rational action and game theory on the multiagent systems field, onecould expect that means implemented to achieve infosocial order areinspired by the same principle mentioned above, that is, to provideadditional utility for software agents to act according to existing institutions. Hence, the efficiency of electronicinstitutions is expected to rely upon the efficiency of sanctions andincentives as inputs to software agents' rational decisions.

In this paper, we intend to investigate different ways of implementingagents which reason upon norms. As will be shown in the next section,norm-based reasoning is not necessary to obtain a norm-correspondingbehaviour. Other mechanisms have been implemented at the agent level, which do not allow for reasoning and decision based upon norms. These will be shortly examined in thenext section. The focus of this paper is on intelligent norm-driven action,that is, on action based upon a decision to comply with some norm. Inparticular, we intend to questionthe efficacy of sanctions and incentives in the achievement of socialorder. In natural societies, the efficacy of sanctions and incentives isfar from granted. Moreover, human agents do not always act upon rationaldecision, and normative action is not executed only when compliance is convenient in terms of individual utility.In natural societies, norms are even expected to be observed for intrinsicreasons, as ends in themselves. But what about infosocieties? Which impactcan incentives be expected to have on the achievement of infosocial order? After a short review of earlierwork on the implementation of social laws and conventions, we will comparetwo different views of agents reasoning about norms and otherinstitutions:

- Incentive-based rational deciders
- Normative agents, which are prescribed to be intrinsicallymotivated to comply with the norms.

In the following section, we will formulate specific consequences that canbe expected from either type of agents. Thereafter, some evidence fromnatural societies will be shown to match the expectations relative tonormative agents, rather than those relative to rational deciders. Some speculations about the relativedesirability of normative Vs. rational agents will conclude the paper.

2 Related Work

Attempts to implement laws and conventions at the level of the agent goback to the early 90s. The necessity to achieve coordination in motion hasinspired the implementation of social laws in multiagent systems (Shohamand Tennenholz, 1992). Analogously,the necessity for robust performance in joint action and teamwork inspiredthe implementation of commitment (Cohen and Levesque, 1990a; Kinny &Georgeff, 1994) and conventions (Jennings and Mandami, 1992; Jennings,1995), and other norm-like mechanisms (such as responsibility, cf. Jennings, 1992). These models and thecorresponding systems present a twofold problem. On one hand, norms andlaws are implemented as action constraints, which ensure optimal efficiencyof the system, but grants no agent autonomy: agents cannot violate the norms. On the other, no innovation is allowedonline: agents are not enabled to acquire norms. These can modified andupdated by the programmer when the system is online.

Impulse to the implementation of norms as inputto agents' reasoning and decision-making comes from the rational action theory(Boman, 1999), which has hegemonial influence in multiagent systems (for acritique, see Castelfranchi and Conte, 1998). Based on the assumption thatdecisions are guided by agents' subjective preferences, norms are seen as external inputs to agents' decisions,because they modify the agents' preference order and therefore theirutility function through sanctions (if you violate the norm, you will get apenalty) or positive incentives (ifyou observe the norm, you will get a reward). This conceptualisation ofnorms does justice to the autonomy of agents, which are not simplyconstrained by norms, but decide whether to execute them or not on thegrounds of their criteria. At the same time, it allows norms to be updated and acquired online. Agents will receive(through communication) new normative inputs and the associated incentives,which they will take into account when acting. In this view, a norm-drivenaction is an autonomous action which implies norm-based reasoning and decision.

As will be shown in the next section, rationaldecision is one of the two major conceivable ways to implement intelligentand autonomous norm-driven action. Let us see the other and comparethem.

2 How to Implement Norm-Based Reasoning?

There are two main approaches to implement agents that reason and decideupon norms. These approaches depend upon two different notions of a norm,norms as prescriptions to execute an action based upon incentive (we willcall this incentive-based norms); norms as prescriptions to execute an action forintrinsic reasons.

In the case of incentive-based norms, a norm provides additional individualutility (usually through sanctions) for socially desirable action. Agentswhich are built upon this notion of norm will be here called rationaldeciders; rational decision is based upon a set of ordered preferences which gives rise to a subjective utilityfunction: given a choice, the agent will be rational if it performs theaction whichmaximises its utility. Sanctions and incentives must be provided in such adegree that the individual utility of socially desirable action is higher than the individual utility of socially undesirable action. Later on in thepaper we will examine a numberof specific effects that can be expected from this modality of normimplementation.

According to the second conceptualisation,norms are seen as prescriptions to execute actions based upon *intrinsic motivations* (we will call them motivating norms). Ithasbeen shown (cf. Conte & Castelfranchi, 1999) that one important aspect ofnormative prescriptions lies in the reasons why they ought be adopted. Inthis conceptualisation, a norm is more than a simple prescription about agiven action or decision. A norm prescribes not only *what* must (not) be done, butalso *why* it must (not) be done. Sanctions are not inherent to norms. They areconsequences of transgression, rather than reasons for obedience.

Normsprescribe that they be adopted just because they are norms, as intrinsic motivations. Indeed, this is the first and most importantcriterion for telling whether a given command is a norm or not. There aretwo orders of evidence that this is the case. First, agents (at leastnatural agents) are not (necessarily) informed about the entity of sanctions, nor will they feel entitled to askquestions about it, and still they can tell if the command is a norm ornot. When you get on a flight, you do not ask the staff members what is thesanction for smoking, although probably you have never known it. Agents which are built upon this notion ofnorm are here called normative agents.

The main difference between these two notions of norms is that sanctions(or incentives) are inherent to the former but not to the latter.Consequently, the main difference between rational deciders and normativeagents is that the former will execute thenorm only in presence of sanctions or incentives, while the latter arerequested to have an intrinsic motivation to obey the norm. This differenceneeds further consideration.

First, normative agents do often adoptthe norm for utilitarian reasons. But this is only a sub-ideal (in thesense logically defined by Jones and Pšrn, 1991) state of affairs. With rational deciders, the utilitariancalculation is not sub-ideal: rational deciders are not prescribed specificreasons for obedience.

Secondly, normative agents are expectedto decide whether to adopt a norm even if sanctions are not specified. Sucha condition, conversely, is undecidable for rational deciders: they will have no sufficient elements to decide.

Third, and consequently, there can always be a subset, however small, ofnormative agents which will adopt the norm for ideal reasons (intrinsicmotivations). But rational deciders cannot accept norms for intrinsicreasons, unless sanctions are also intended as internal. In such a case, of course, no significant difference holdsbetween normative agents and rational deciders. This leads us to precisewhat is here meant by sanctions and incentives.

2.1 Incentives and Sanctions

Here, we will provide an operational notion ofincentive. We speak about a positive incentive as an additional expectedbenefit of an action. More precisely, ws_i is anincentive for agent ag_i to perform a given action $a_i$when

- agentag_i does action a_i for any given goal g_i, and
- a_i brings about ws_i, a state of the world which achieves a further goal ofag_i, say, goal g_j,and
- ws_iincreases the value (or the utility) of $a_{i,}$so that ag_i is more likely toperform it as a means to achieve g_i.

The worldstate ws_i,then, is a positive side-effect of a_i, anincentive. A sanction is a negative side-effect. Agent ag_i may be informed about the side-effects of a_i,and still this action is not initialised by any of them, but by the agent'soriginal goal g_i. Nonetheless, actions' side-effects obviously interfere with agents'planning and decision-making. Suppose you want to get warm. Your planlibrary suggests several alternative plans: to turn on the heater, to wearwarm clothes, or to make a fire. Suppose that the wooden fire has a nice effluvium. Nexttime you want to get warm, you'll probably choose again the wooden fire,because the effluvium acted upon you as an incentive.

We will speak of a *socialincentive* (positive or negative), when an incentive iscontrolled (provided or not) by another entity ag_j, where ag_i - ag_j. Moreprecisely, agent ag_i has a socialincentive[1] to execute a_i, when

- ag_j has the power to bring about (or to obstacle)ws_i
- ag_j has the goal to influenceag_i to execute a_i,that is has the goal that ag_i decide to executea_i
- ag_j believes that goalg_i of agent ag_i isinsufficient for ag_i to put $a_i$to execution
- ag_j believes that ws_i will increasethe value of a_i for ag_i, and therefore the probability that ag_i willexecute it
- ag_j gets ag_i to know that ifag_i will perform a_i, ag_j will bring about (or prevent) ws_i.

A social incentive is therefore an additionalvalue or utility, provided by an external entity, which modifies theagent's decision. Such an external entity must have the power or capacity of bringing about a worldstaterelevant for a_i's goals. This will turn into asocial power of ag_j's: thanks to the power ofbringing about ws_i, ag_j has also power overag_i. Agents may control and influence otheragents also by providing incentives to them.

2.2 Incentive-Based Rational Deciders

Rational deciders calculate the subjectiveexpected value of actions according to their utility function. According toa classical strategy of choice, given an agent ag_i and a set of alternatives for action A_i =a_1, ..., a_n, the value of each alternative (taking into account its costs) per itsrelative probability of occurrence will be compared. That which yields themaximum utility (including the alternative "don't act") will be put toexecution.

How is it possible to have rational deciders to observe a norm? First, theymust be informed about the norm. Agents must be provided with criteria torecognise norms. For example, a norm may be a command imposed by a givenauthority and associated with given incentives (usually, negative). With norms, arational decider will perform the same calculation which is applied to anyother decision: the utility of norm compliance is computed in the usualway:

$v_c p_i + v_t(1- p_i)$

where v_c is thevalue of compliance; to simplify matters, we assume this value to be equalto the incentive (or sanction); p_i is theprobability of its occurrence, and v_t stands for the value of transgression, which isalways positive since a normative action is by default inconvenient. From this, it can be easily drawn the conclusion thatif an incentive is lower than the value of transgression, a rationaldecider will not comply with the norm, unless the probability of incentiveis lower than the complementary probability (to not receive incentive, or, which is the same, to undergosanctions).

2.3 Normative Agents

[1] >From now on, we will speak about social incentives, but will call themincentives for short.

Normative agents are cognitive agents which areprescribed to adopt norms as ends in themselves. Normative agents are hereseen as BDI-like agents, characterised by mental states, namely goals and beliefs, and the capacity toreason and act upon them. Normative agents form beliefs about a norm, maydecide to adopt it by forming a corresponding goal, and to achieve it, byexecuting a norm-driven action.

Ideally, norms not only prescribe a givenaction a_i, but also a given motivation forexecuting it, i.e. the goal to comply with the norm because it is a norm.

In the following two sub-sections, we willresume our model of norms presented elsewhere (cf. Conte & Castelfranchi,1999), and will show how such a model accounts for the motivationsprescribed by the norms.

2.3.1 Our formalism

The formalism used is a simplified version of Cohen and Levesque's (1990b)language for describing their theory of rational action. The languageappears as a first-order language with operators for mental attitudes andaction. Two modalities for beliefs and goals *(BEL x p)* and *(GOAL xp)* are defined according to the possible worlds semantics, andtherefore through accessibility relations. Two modalities for action *(HAPPENS e)* and*(DONE a)* express, respectively, events taking place in the world independent of theagents' actions and occurrence of actions. Finally, time is represented asan infinite sequence of events.

Beliefs and goals are given the usual possible world interpretation. As forconsistency, the Hintikka axioms for beliefs apply to this model (seeHalpern & Moses 1985). As for realism, goals are a subset of beliefs. (Theaccessibility relation ***G***, which defines the set of worlds in which goalsare achieved is a subset of the accessibility relation ***B***, which defines the set of worlds belief-accessible to a given agent.). Insuch a model, in fact, a goal is defined as a belief-compatible desire. (Inother words, agents cannot have goals which they believe to beunachievable.)

Many notions can be constructed on the groundsof these primitive modalities plus the operators ◊for "later", ; for "sequence" and ? for the procedure to testwhether a given proposition is true.

$(HAPPENS\ a)$	anaction will happen next
$(DONE\ a)$	an action has just happened;
$(BEL\ x\ p)$	x has p as a belief
$(GOAL\ x\ p)$	x has p as a goal;
$(OUGHT\ p)$	there is an obligation whatsoever onproposition p;
$(AGT\ x\ e)$	x is the only agent of thesequence e;
$e_1 \leq e_2$	e_1 occurs beforee_2
$p?$	test action
$\Diamond p$	p will be true at somepoint in the future

A number of definitions, grounded upon theabove atomic predicates, are necessary to understand the formulae providedthroughout the paper. Most of them are drawn from Cohen and Levesque'smodel, and we present them here for the convenience of the reader unacquainted with that model. Some have beenintroduced by the authors and other collaborators in preceding works (Conteet al. 1991; Castelfranchi et al. 1992).

$$(DOES\ x\ a) \overset{def}{=} (HAPPENS\ a) \wedge (AGT\ x\ a) \tag{1}$$

This says that *x is the onlyagent of action a, which will happen next.* We need an analogous predicate for past actions,

$$(DONE - BY\ x\ a) \overset{def}{=} (DONE\ a) \wedge (AGT\ x\ a) \tag{2}$$

saying that, *x is the only agentof action a, which has just happened.*

Cohen and Levesque have also introduced thefollowing predicate to refer to sequences of world states,

$$(BEFORE\ q\ p) \overset{def}{=} \forall c(HAPPENS\ c; p?) \supset \exists a(a \leq c) \wedge (HAPPENS\ a; q?) \tag{3}$$

In words, *q comes before p when,for all events c after which p is true, there has been at least one event apreceding c, after which q was true*.

As for goals, Cohen and Levesque have introduced the notion ofachievement goal, which is defined as follows:

$$(A - GOAL\ x\ p) \overset{def}{=} (BEL\ x\ \neg p) \wedge (GOAL\ x\ \ \Diamond p) \tag{4}$$

that is, *x has an achievementgoal p if x believes that p is not true now but wants it to eventuallybecome true*. Throughout the paper, whenever the notion of goal is used, it will bemeant as an achievement goal in the above sense, unless otherwisespecified. Indeed, in our model (as well as in Cohen and Levesque's), anachievement goal is not yet an intention.

Cohen and Levesque's theory includes a notionof relativised goal:

$$\begin{aligned}(R - GOAL\ x\ p\ q) \overset{def}{=} {} & (A - GOAL\ x\ p) \wedge \\ & (BEFORE\ ((BEL\ x\ \neg q) \vee (BEL\ x\ p) \vee (BEL\ x \neg \Diamond p)) \\ & \quad \neg (A - GOAL\ x\ p))\end{aligned} \tag{5}$$

x has a goal p relativised to q, when x hasan achievement goal p, and before ceasing to have p as an achievement goal,x believes either that p is realised or unachievable or that the escapecondition q does not hold. Essentially, this means that x has p as long as and because hebelieves that q.

Our notion of a goal (Conte and Castelfranchi1995a) is slightly weaker than that allowed by Cohen and Levesque. Wepropose to treat goals as *realistic* desires,rather than *chosen* ones. In our terms, a goal is but a regulatory mental attitude which callsfor a series of operations, including some preliminaries, involved inplanned action. In other words, along the lines of classical AI planningsystems, we define a goal as a device which activates planning and *action*. In ourterms, a goal may be abandoned not only when it is believed to be fulfilledor unachievable, but also when it is found incompatible with another moreimportant goal.

2.3.2 Normative Beliefs and Goals

The pred *OUGHT* intuitively means that there is some sort of *obligation* on proposition p. For the time being, we take it as an atomic one-placepredicate, although it seems possible to further analyse it as some sort ofexternal reason which forces a given goal, namely the adoption of a givengoal. However, we will assume obligation as a primitive, which defines a set of worlds in which p followsfrom obligations. The relation of accessibility ***O*** is a subset of ***B***.

In our model, agents have normative beliefswhen they think there is an obligation on a given set of agents to do someaction.

In the following, x and y denote agent variables with *x -y* always implicitly stated, and a denotes an actionvariable.

We express the general form of a normativebelief as follows:

$$(N-BEL\ \ x\ y_i\ a) \overset{def}{=} (\Lambda_{i=1,n}(BEL\ \ x(OUGHT(DOES\ y_i\ a)))) \qquad (6)$$

in words, x *has a normativebelief about action a relative to a set of agents* y_i *if and only if* x *believes thatit is obligatory for* y_i *to do action a*. The predicate *OUGHT*here stands for an *obligation for a set of agentsy*$_i$ *to do action a*. A few words are needed to elucidate the semantics of our predicate*OUGHT*. This stands for an operator of obligation about any given state of theworld. However, it should be taken in a somewhat weaker sense than what isusually meant by obligation in traditional deontic logic. In fact, while intraditional deontic systems, p necessarily follows from obligation (that is to say, it is not possible thatat the same time p is false and obligatory), in other systems (Jones andPšrn 1991), two concepts need to be distinguished, one referring to deonticnecessity and the other to another type of obligation. The latter is defined as the circumstance in which a given propositionis both obligatory and possibly false in some sub-ideal world.

In order to express normative goals, anotherbelief is needed, namely a pertinence belief: for x to believe that he isaddressed by a given norm, he needs to believe that he is a member of the class of agents addressed bythat norm:

$$(P-\ N-BEL\ \ x\ a) \overset{def}{=} (\Lambda_{i=1,n}(N-BEL\ \ x\ y_i\ a)) \wedge (V_{k=1,n}(BEL\ \ x(x=y_k))) \qquad (7)$$

where P-N-BEL stands for normative belief ofpertinence; in words, *x has a normative belief of pertinencewhen he has a normative belief relative to a set* y_i *and an action a, and believes that he is included in* y_i.

Now, *x*'s beliefs tellhim not only that there is an obligation to do action a, but also that theobligation concerns precisely himself.

We have not seen any normative goal yet. A normative goal is defined hereas a goal always associated with and generated by a normative belief. Letus express a normative goal as follows:

$$(N-GOAL\ \ x\ a) \overset{def}{=} (R-GOAL\ \ x(DOES\ \ x\ a)(P-N-BEL\ \ x\ \ a)) \qquad (8)$$

or, x *has a normative goalconcerning action a when he has the goal to do a relativised to hispertinence normative belief concerning a.* A normative goal of a given agent x about action a is therefore a goal thatx has as long as he has a pertinence normative belief about a:x has a normative goal in so far as he believesto be subject to a norm.

2.3.3 The Paradox of Normative Requests

What is the relationship between a normativebelief and a normative goal? This question should be examined from twodifferent perspectives.

From the point of view of the agent, anormative belief is *necessary* but not sufficientfor a normative goal to be formed, and *a fortiori*, a normative action to be executed. Elsewhere (Conte and Castelfranchi,1995a), we have examined several mechanisms of norm adoption, includinginstrumental and cooperative adoption. In other words, there may be several reasons for agents to adopt a norm: to avoidsanctions, to achieve positive side-effects (incentives), or even toachieve a goal which the norm is able to instore. In the latter case, theagents have one goal in common with the norms, or, better, with the system which has issued the norm.

From the point of view of the norm itself, anormative belief is not only *necessary* but oughtto be also *sufficient* for a normative goal to be formed. Agents must know that action isobligatory (N-belief) to have a normative goal concerning that action. Onthe other hand, if they have a normative belief, they ought to want toexecute it.

$$(N - BEL\ x\ y_i\ a) \supset (BEL\ x(OUGHT((P - N - BEL\ y_i\ a) \supset (N - GOAL y_i\ a)))) \qquad (9)$$

Sub-ideally, this may not be the case.*ought* to be the case; this is what the norm expects. Indeed, this is how a normcan be distinguished from other, coercive, requests or commands. All that anorm says is what must be done: provided the agent is dutifully informedabout it, it will have a normative will corresponding to it. Sanctions are consequent to action violations, and reasonably effects which agents learn to associate toit. In real matters, negative or positive incentives have a stronglymotivating role in norm compliance. But on the one hand, this is not alwaysand necessary the case: norms may and sometimes are observed for intrinsic reasons. On the other hand, this isa *sub-ideal*, however frequent, state of affairs (Jones and Porn, 1991), meaning thatonly in a subset of the worlds in which the norm is in force, a normativebelief is sufficient for a normative goal to arise and the corresponding action to happen. Thissubset is that of ideal worlds. In sub-ideal worlds, that is thecomplementary subset, a normative belief is only a necessary butinsufficient condition for a normative goal, and the latter is a necessary but insufficient condition for a normative action.

3 What Can Be Expected?

Which expectations can be made with regard tothe effects of the two architectures? Both types of agents can violate thenorm, since both types of agents are autonomous. Rational deciders will violate a norm when it isinconvenient for them to comply with it. Normative agents can violate anorm for a number of reasons, which include but are not reduced toutilitarian reasons, for example to solve a normative conflict. In case of a conflict between two norms, rational deciders areexpected to choose that which is most convenient, or least inconvenient tothem. On the contrary, normative agents are expected to apply the mostimportant one, irrespective of their own convenience. Furthermore, normative agents can violate a norm whichthey consider unfair.

More explicitly, we can formulate two generalexpectations:

- incentive-based deciders will comply with the norms to the extent that the(positive or negative) incentive is such that the utility of obedience ishigher than the utility of transgression (sanction is higher than theconvenience of transgression);
- normative agents will comply with a norm as long as eitherideal conditions apply (intrinsic motivations) or sub-ideal conditions apply (in this case they will behave asrational deciders) or ideal conditions apply and the norm is not unfair orcontrary to duty.

3.1 Rational Deciders' Impact

More specifically, incentive-based deciderswill violate a norm n_i as soon as one or moreof the following conditions applies:

- Sanctions are notimposed: an incentive-based decider will certainly violate a norm if nosanction is expected to follow from violation, since by definition inabsence of incentives normcompliance is individually irrational.
- Sanctions areexpected but are not specified: in such a condition a rational decider willeither infer the specification of sanctions, or will not take anydecision.
- Thesanction[2] for violating n_i is lower than the value of transgression withequal probability of application of the sanction (1/2).
- The sanction(negative incentive) for violating an incompatible norm n_j, where $(n_i - n_j) \wedge (n_i \supset \neg n_j)$ is higher. Thisaspect of norm-based decision-making is important especially in societies of growing complexity, where the set of normstends to increase, and conflicts among norms become more likely.
- The sanction(negative incentive) for violating the norm $n_i$is not or rarely applied: p_i tends to 0. Since the utility of norm compliance, as seen above, is equalto the value of incentive (or sanction) per its relative probability ofoccurrence (taking into account the utility of transgression), obviouslywith a probability proximate to zero, the utility of incentive is also nullified. Therefore, even with a moderately convenient value oftransgression, a rational decider is likely to violate the norm. Considerthat both the probability and entity of sanctions may be inferred byobserving others' behaviour: the more others violate, the less likely and/or severe the sanction is expected to be.This has further consequences which we will examine in the followingsection.

With a homogeneous society of incentive-baseddeciders, any of the above conditions is followed by a fast decline or even a collapse in compliance with a given norm. Theinconvenience of norm compliance will be detected sooner or later by allmembers of the society. Consequently, their decisions will rapidly convergeon norm violation, unless the externalsource of sanctions monitors the behavioural effects of agents' decisionsand takes efficient measures of norm enforcement, by either intensifyingthe application of sanctions or augmenting their entity.

[2] >From now on, we will speak of sanctions rather than incentives, becausenorms are enforced by sanctions more than by positive incentives. However,the formal reasoning can easily be extended to the other factor ofenforcement.

3.2 Normative Agents' Impact

On the other hand, normative agents are expected

- To comply with a normeven if sanctions are not imposed, or are not imposed explicitly.*A fortiori*, normative agents may comply with norms when they know that sanctions areimposed but their entity and probability of application is uncertain.
- To execute norms even though sanctions are such that the utility of normcompliance is lower than the utility of transgression. A heterogeneouspopulation of normative agents, where ideal and sub-ideal agents co-exist,ensures that even a small subset of agents will still apply the norm for intrinsic reasons.
- To comply with the norm n_i even when sanction is not or rarely applied. This is but a special case ofthe previous point. Of course, sub-ideal agents will converge on normtransgression. However, an even small number of stubborn agents will complywith a norm even when the sanctions are not or rarely applied.
- To comply with thenorms when others violate. A persistent execution of the norm in a smallshare of the population (ideal agents) is expected. This has interesting further effects at the globallevel: since sub-ideal agents, as well as rational deciders, are enable toinfer the entity and probability of incentives by observing others'behaviours, some persistence in norm execution will have the consequence to limit or counteract this inference.Some oscillatory effects can be expected: agents which perceive idealagents' behaviours will draw different conclusions on theentity/application of sanctions than others and will therefore be more likely to execute the norm. But as they perceive thebehaviours of other sub-ideal agents, who were not exposed to the influenceof ideal ones, they will go back to violation. Indeed, even ideal normativeagents may be affected by others'decisions. Frequent transgressions may be perceived as "demotivating": themore a given norm is violated, the more it is perceived as unfair orinadequate or ineffective. This perception may reduce an intrinsicmotivation to comply with that norm. However, no collapse in norm compliance is expected with normative agents butrather a "graceful" and non-linear degradation[3].
- To solve normconflicts even independent of the respective sanctions: with

  $(n_i - n\ _j) \wedge (n_i \supset \neg n_j)$

  normative agents are not necessarily expected to choose the norm whichgrants them the higher individual utility. Again, an even small number ofideal normative agents will still choose the norm which is more importantthan the other according to some plausible criteria (entity of the injury consequent to norm transgression,reparability of norm transgression, etc.).

In short, the general expectation that incentives are a good solution tothe problem of (info)social order should be reconsidered and mitigated.Incentives should be seen as useful means to enforce the norms, rather thanas sufficient mechanisms for modelling and implementing them. Social order cannot primarily

[3] Simulation studies should be carried out to confirm thisexpectation.

rely uponincentives and sanctions, unless sanctions are always severe and certain soas to lower the utility of transgression compared to the utility ofcompliance.

4 Evidence from Natural Societies

Things work much better if norms are executed for their own sake, that is,if at least a share of the whole society accepts and complies with thenorms for intrinsic motivations. But how is it possible that such type ofnorm exists at all? Or, better, howis it possible that autonomous agents have intrinsic reasons to comply witha norm? Does this type of agent really exist, or is it conceivable only ina morally ideal society?

A look at human societies shows some importantphenomena. First, real (social or legal) norms are *not* primarily defined as incentive-based prescriptions, but rather asprescriptions which ought to be accepted for their own sake. Secondly,incentives have a lower effect on norm compliance than should be expectedif a model of rational decision is accepted: natural agents take into account sanctions less than rational deciders areexpected to do. Third, incentives may bear negative consequences on normcompliance. Let us examine each phenomenon with some detail.

4.1 Incentives and the Concept of a Norm

What are real norms? Which roles do incentives play in their definition andrecognition? As said before, sanctions are neither necessary nor sufficientfor norms. People can tell and accept a prescription as a norm, even ifthey do not know and are not informed about the respective sanctions. Indeed, this is quite often thecase: agents take decisions in absence of "official" information about theentityand probability of sanctions. Moreover, agents may try to infer suchinformation, but they will neither expect that such information be providedby the source of sanctions, nor are they allowed to exact it. Indeed,agents may take it into account "privately". To calculate the entity and probability of sanctions is (considered) anaggravation of crime, because to observe the norm ought to be a sufficientmotivation. On the other hand, people may accept a command under threateven if they do not perceive it as a norm: agents may yield to intimidation even if they are perfectly awarethat it is illegal (people may surrender to an armed criminal but denounceher as soon as possible).

4.2 Incentives in Norm Enforcement

Incentives do not enforce compliance as much as expected. Humans are ratherheterogeneous with regard to normative decisions, although their decisionsare often perceived as utilitarian. Statistics about crimes do not confirmthe expectations allowed by the model of incentive-based decision. First, the average application of sanctions for certain crimes (burglary androbbery) is very moderate, and in some countries is close to 1%.Consequently, the utility of compliance should be close to 0, andcompliance should collapse. Nonetheless, the majorityof humans has never committed this type of crimes. Secondly, and moreover,the entity and probability of sanctions are not equivalent indecision-making: it is well-known that, with equal probability ofapplication, compliance does not increase with the severity of sanctions.

Third, frequent transgressions certainly contribute toencourage transgression. But this is not only because the perception offrequent transgressions affects the computation of the utility of normcompliance. Other mental processes occur: either the formation of a normative belief is obstacled by theassumption that a disregarded norm is bad or unfair or inadequate and thelegislator is weak and ineffective; or the normative goal is abandoned,because the control system is ineffectiveand unfair, and does not deserve obedience.

4.3 Bad Effects of Incentiveswith Human Agents

Good experimental evidence indicates that incentives may render a badservice to norms. Not only positive incentives have been found to reduce orinhibit intrinsic motivation (what is called overjustification; for a reentwork, cf. Lepper, forthcoming): when agents receive a reward for an activity which they were intrinsicallymotivated to accomplish, their intrinsic motivation will decrease. What isworse, negative incentives may reduce the unpleasantness of transgression (Greene et al.,1976): the lower the sanction, and the more the agents which comply withthe norm will be attracted to violation. Social psychologists explain thesefindings in terms of self-perception (Bem, 1972): the less my action (compliance) is justified by someexternal factor, the more I need to find an internal reason for it. I willtherefore be led to develop some good feeling or positive attitude withregard to it. If I complied with a norm which is not enforced by severe sanctions, I must have had a good reasonto do so. The norm must be an important one, or else, I may start to thinkthat to comply with that norm is good for its own sake. I develop anintrinsic motivation towards that norm, or towards the norms in general.

But why are incentives applied, then? We all know that they are appliedrather frequently. Rewards are used in education and learning with goodresults. The same is true for sanctions: parents keep punishing children when they do something wrong. Delinquents are imprisoned, although lessoften than desirable. Fraud and deception are castigated by the community.Social psychologists suggest some answers to this question. First, thesmaller the incentives the better (Greene et al., 1976). Secondly, they work much better in improving the qualityof performance than in motivating action (Tang and Hall, forthcoming),which is why rewards work better in physical and mental learning than inmoral and social education. Thirdly, they work when no intrinsic motivation has developed yet (Tang and Hall,forthcoming). Once the desirable behaviour has appeared, incentives ceaseto be useful and may even demolish the good job done. Fourth, and moreover,they work at their best *if agents perceive them as side-benefits, or additionalmotivations*, rather than as unique or primary reasons foraction (Hennessey and Zbikowski, 1993).

5 Final Remarks. Normative Agents Vs RationalDeciders: Which One would You Prefer to Deal with?

In this paper, the role of incentives in socialorder has been questioned, based on a notion of incentive as additionalindividual utility, provided by an external entity, to actions achievingglobal utility.

Two notions of norms have been defined and compared: (1) inputs whichmodify agents' decisions through incentives (sanctions) and (2)prescriptions to execute obligatory action for intrinsic motivations. Twotypes of agents which reason upon norms were also compared: (1) incentive based rational deciders, and (2) normative agents which are prescribed to execute norms for intrinsicreasons. Expectations about the effects of these two types of agents onnorm compliance have been formulated. With relatively inefficientapplication of sanctions (punishment), transgression propagates more easily and rapidly among incentive-based agents thanamong normative agents. Under suboptimal conditions of application ofsanctions (uncertain punishment), normative agents are expected to exhibitan oscillatory or at least a graceful degradation of compliance, while incentive-based agents are expected toshow a fast decline and even a collapse. Finally, the role of incentives innatural societies has been discussed. This role is shown to have lesserimpact on natural social agents than expected by a model of rational decision. What is worse, incentives havebeen shown to produce even negative effects on several aspects of sociallearning and norm compliance.

However, which lesson can be drawn fromobservation of natural societies andextended to infosocieties? Is the observation of natural societies anyrelevant for software agent engineering? Our answer is, yes, ifapplications to agent-mediated interaction are considered. In this context,agent scientists and designers face an important pragmatic task: to design systems which can interact with one anotheror with humans in a useful, reliable, and trustworthy way *from the point of view of the human user*. The good question then is, with whom does a human agent prefer to interactwith? More specifically, when it comes to execution of norms, which one ispreferable, a rational decider or a normative agent. Here, it is necessaryto distinguish the two main roles that a software agent is expected to playin agent mediated interaction: that of user *representative* and that of *partner*. In e-commerce, for example, a system represents a given user in findinggood partners for bargain, giving assistance in negotiation, etc.. Butinteresting applications under development see software agents as partners of negotiation (cf.,). As to the role of representative, a rationaldecider which is benevolent to its user, has her preferences as itsultimate goals and applies strategies to maximise her utility is probablythe best choice. But as to the second role, that of *partner*, it is not so clear what should be preferred. Ultimately, one prefers todeal with trustworthy agents. But are incentive-based rational deciderstrustworthy partners? Is it preferable to deal with a system which respectsthe norms onlyin the interests of its own user (and therefore to the extent that this isconvenient to her), or with a system which, so to speak, takes normsseriously and respects them for their own sake? More specifically, whichcondition is more encouraging from thehuman agent point of view, an efficient and severe sanctioning system, or asociety of trustworthy partners? If the system is not efficient enough, itis certainly preferable to have a chance to meet agents which respect thenorms independent of sanctions. But even if the sanctioning system were efficient enough, wouldn't it bemore appealing to have at least a chance to deal with "good guys", meetnice partners? Isn't it better from the human point of view to know thatyour partner behaved correctly not because it was more convenient to do so, but because of its good will? Atthis stage, these questions do not allow for a conclusive answer. But wethink that we should be prepared to provide a pondering answer in the nearfuture.

References

Bem, D.J. 1972. Self-perceptiontheory. In L. Berkovitz (ed) *Advances in Experimental SocialPsychology*, New York, Academic Press.

Boman, M. 1999. . In Conte, R., Falcone, R., e Sartor, G. s.i. ofArtificial Intelligence and Law on "Agents and Norms" 7, 1999.

Castelfranchi, C. e Conte, R. Limits of economic rationality for agents andMA systems. *Robotics and Autonomous Systems*,Special issue on Multi-Agent Rationality, Elsevier Editor

Castelfranchi, C., Miceli, M., Cesta, A., 1992, *DependenceRelations among Autonomous Agents.* In Y. Demazeau, E. Werner(eds),*Decentralized AI* - 3,215-31. Amsterdam: Elsevier.

Cohen, P. R. & Levesque, H. J.1990b. Persistence, Intention, and Commitment. In *Intentionsin Communication*, ed. by P.R Cohen, J. Morgan & M.A. Pollack,33-71. Cambridge, MA: The MIT Press.

Cohen, Ph. & Levesque,H. (1990a). Intention is choice with commitment. *ArtificialIntelligence*, 42(3), 213-261.

Conte, R. e Castelfranchi, C.>From conventions to prescritions. Towards an integrated view of norms.Artificial Intelligence and Law 7: 323-340, 1999.

Conte, R. & Castelfranchi, C.1995a. *Cognitive and social action*. London: UCLPress.

Conte,R., Miceli,M.,Castelfranchi,C., 1991, *Limits and Levels of Cooperation.Disentangling Various Types of Prosocial Interaction.* InDemazeau,J.P. Mueller (eds), *Decentralized AI-2*, Y. , 147-157. Armsterdam: Elsevier.

Crabtree, B. What Chance SoftwareAgents, *The Knowledge Engineering Review*, 13,1998, 131-137.

Greene, D., Sternberg, B., and Lepper, M.R., 1976. Overjustification in aToken Economy. *Journal of Personality and SocialPsychology*, 57: 41-54.

Halpern,J.Y., Moses,Y.O.,1985, *A Guide to the Modal Logics of Knowledge andbelief. Proceedings of the Ninth IntenationalJoint Conference on Artificial Intelligence*, 480-91. Los Altos, CA: Kaufmann.

Hennessey , B.A. and Zbikowski, S.M. 1993.Immunizing children against thenegative effects of reward: A further examination of intrinsic motivationfocus sessions, Creativity Research Journal, 6:297-307.

Jennings N. (1995).Commitment and Conventions: the foundation of coordination in multi-agentsystems. *The Knowledge Engineering Review* , 8,.

Jennings, N. (1992). On being responsible, in Y. Demazeau & E. Werner (eds)*Decentralized Artificial Intelligence 3*,Amsterdam: Elsevier Science Publisher, 93-102.

Jennings, N. R. &Mandami, E. H. (1992). Using joint responsibility to coordinatecollaborative problem solving in dynamic environments. In *Proceedings of the 10th National Conference on ArtificialIntelligence*, San Mateo, CA: Kaufmann, 269-275.

Jones A.J.I , Pšrn,I.,1991, *On the Logic of Deontic Conditionals.* InJ.J.C. Meyer, R.J. Wieringa (eds), *First International Workshop on Deontic Logic in ComputerScience*, 232-47.

Kinny, D. & Georgeff,M. (1994). Commitment and effectiveness of situated agents. In*Proceedings of the Thirteenth International Joint Conferenceon Artificial Intelligence*, IJCAI-93, Sydney, 82-88.

Lepper, M.R. (forthcoming) Theoryby numbers?Some concerns about meta-analysis, *AppliedCognitive Psychology*

Rao, A.S. A Report on Expert Assistants at the AutonomousAgents Conference.*The Knowledge Engineering Review*, 13, 1998, 175-1179.

Shoham, Y. &Tennenholtz M. (1992). On the synthesis of useful social laws in artificialsocieties. *Proceedings of the 10th National Conference onArtificial Intelligence*, San Mateo, CA: Kaufmann, 276-282.

Tang, S. and Hall, V.C.(forthcoming) The overjustification effect: A meta-analysis, *Applied Cognitive Psychology*