From: AAAI Technical Report SS-93-05. Compilation copyright © 1993, AAAI (www.aaai.org). All rights reserved.

Norms as mental objects. From normative beliefs to normative goals.

Rosaria Conte* Cristiano Castelfranchi*

1. Introduction

The study of norms (Ns) is a recent but growing concern in the AI field (cf. [MOS92]; [SHO92]) as well as within other formal approaches to social action, such as Game Theory (where theories of Ns began to appear long before, cf.[LEW69]; [ULL77]). The normative is an essential dimension of social life. An explicit model of such dimension is crucial for any theory of social action. Notions such as "social commitment" (cf. [GAS91]; [COH90]; etc.) "joint intentions", "teamwork" [COH91], "negotiation" [DAV83], "social roles" [WER89] etc. will not escape a purely metaphorical use if their normative character is not accounted for.

Indeed, *one's committing oneself* before someone else causes the latter to be *entitled*, on the very basis of commitment, to expect, control, and exact the action one has committed oneself to. Social commitment is a customary, if not *the* most frequent, source of entitlements in everyday interaction: first of all, a promise creates an entitled *expectation* in the addressee, namely a belief that the promised course of action will be realized. As a consequence, the addressee feels entitled to *exact* the action expected, that is, to control its occurrence and *react* if it is not performed. In particular, the addressee is entitled to express a public *protest* and receive the consent and support of the community of witnesses before whom the action of commitment took place.

Furthermore, there is no true *negotiation* without *control for cheaters* and *norm of reciprocation* (cf. [GOU60]). When social agents negotiate an exchange of resources, they actually set up the value of the resources exchanged, that is, what each is *entitled* to expect from others in return, given the resources invested. Without a norm of reciprocation, such expectation would be easily disconfirmed when *social* exchange, as often happens, does not occur simultaneously. The utilitarian agent would spontaneously respect the agreement only if the partner is regarded as stronger. Otherwise, he would get away with no costs, in a word, he would cheat.

Analogously, joint intentions and teamwork arise from true negotiation and create *rights* based on commitments. Finally, roles are but sets of *obligations* and *rights*.

A *norm-abiding behavior* need not be based on the *cognitive processing of Ns* (it might be simply due to imitation), nor does a normative mind necessarily produce a behavior corresponding to Ns (Ns are operating in the mind even if the agent's final decision is to transgress against them). To model the normative reasoning does not imply to model a norm-abiding system. However, a large portion of autonomous agents' normative behavior is allowed by a cognitive processing of Ns. Therefore, Ns are not only an essentially *social* but also a *mentalistic* notion. Ns are indeed a typical *node* of the *micro-macro link*, of the link between individual agents and collective phenomena, minds and social structures. A view of normative systems as "emerging properties" of the social systems, improving coordination within the social systems themselves (cf. [MOS92]) is not sufficient.

On one hand, that of *coordination* is only *one* of the functionalities of Ns. Among others, consider the function of *identification*, (the good manners, for example, improve identification rather than coordination: on their grounds, *in-groups* are allowed to be *identified*, etc.); that of *protecting the single's interests beyond its own will* (as happens with the norms prescribing to fasten seat belts, to wear helmets when riding motorcycles

* SBSP - Social Behavior Simulation Project, Istituto di Psicologia, CNR, V.le Marx 15, I-00137 Roma, Italy.
Tel: (06) 8292626, Fax: (06) 824737, e-mail: PSCS@IRMKANT.BITNET

etc.) Furthermore, there are at least two types of norms of coordination: those which are aimed at *improving the agents' performances* (like the traffic norms), and those which are intended to reduce *displayed aggression* (like the norm of reciprocation).

On the other hand, the emergence paradigm gives an account of conventional norms and conforming behaviors (cf. [BIC90]), but leaves unexplained the *prescriptive character of norms*. Within the game-theoretic paradigm, a social norm is defined as an *equilibrium*, a combination of strategies such that "each maximizes his expected utility by conforming, *on the condition that* nearly everybody else conforms to the norm." (cf. [BIC90: 842]) Social norms are defined as behavioral regularities which emerge from the strategic agents' choices to conform. This is an account of the spreading of certain behaviors over a population of strategic agents, in a word, a model of social conventions. However, such a model does not account for a crucial aspect of the normative mechanism, which plays a role in the spreading of normative behaviors, namely the normative request. What is lacking in the game-theoretic definition aforementioned is a reference to the *agent's will that others conform to the norms*. A social norm, indeed, is such if it is associated at least with a general want that it be observed. Such a want is usually conveyed in many ways: from simple expectations (note that the term expectation is often used to refer to a hybrid mental object, namely a *goal/belief:* if you expect the weather tomorrow to be sunny and warm you both believe and wish it to be so), to explicit commands and requests; from implicit disapproval of transgressors, to explicit reproach etc..

Therefore, here it is intended to draw the attention on both the following aspects:

- the **prescriptive** character of Ns, that is, their role in controlling and regulating the behaviors of agents subject to them;

- a need for a **mentalistic** notion of Ns allowing cognitive agents to become normative agents as well. Ns are not yet sufficiently characterized as mental objects (cf [SHO92]). Here, some crucial problems concerning the mental nature of Ns will be raised and some initial solutions proposed. The problems are the following:

  - How are Ns represented in the agents' minds? Should they be seen as a specific mental object, and if so, which one?
  - Which relation do they bear with beliefs and goals? How can they regulate the agents' behaviors?
  - How is their prescriptive (and not simply conventional) character (that is, a more or less explicit request and the corresponding duty) expressed?
  - Why does an autonomous agent comply with norms, thus fulfilling others' expectations?

2. The normative coin.

What is a norm in an agent's mind? This question leaves aside another fundamental question, namely what is a N tout court. Although prioritary, we will not face this question here (however, see [CAS91]). Therefore, we will not try to answer the question what the "external" side of Ns is, namely, how Ns are encountered in social life, which functionalities they have, which features define *de facto* (independent of agents' beliefs) a normative source, etc..

Here, we will focus exclusively on the "internal" side of Ns, that is: Which cognitive role do Ns play in the agents' minds and in which format are they represented? What kind of mental object is a N? What is a normative source in the agents' beliefs?

There are at least two distinct ways in which *norms can be implemented* on a computer system: as *built-in functioning rules* and constraints (like production rules) or as explicit and *specific mental objects* (i.e. obligations, duties, etc.) distinct from, say, goals and beliefs. Of course, there might be *intermediate solutions*: for example, and in line with the game-theoretic view, one might think of noms as a way of describing the *behaviors of strategic agents in interaction*: no specific normative object is implemented in those agents' minds although cognitive action is allowed. A further alternative consists of implementing Ns just as goals, namely *final ends*: In this case, cognitive agents would be allowed to choose among their (competing) goals (instead of simply applying procedures and routines) but they would treat Ns as any other goal of theirs. At least two undesirable

consequences seem to derive from the latter eventuality: the observance of Ns would come to depend exclusively upon the agents' subjective preferences, and there would be no social control and influencing. To implement norms as specific mental objects, although a costly alternative, is required in a *cognitive modelling* approach to social agenthood, in which the present work is framed. Moreover, the cognitive regulatory endowment of the human species has undoubtedly profited by the explicit representation of norms. A crucial but difficult task, beyond the scope of the present work, would then be to explore these advantages and confront them with those of all alternatives considered.

Our language draws on [COH90]'s model for describing rational interaction. For the readability of the paper, let us provide the semantics of the main predicates used and the axiom (A1) here called the axiom of goal-generation (for further analysis, see [COH90], [CON91a], [CON91b]):

a:	action;
p and *q*:	states of the world;
v:	a worldstate positively valued;
x ,*y*, and z	single agents;
(BEL x p):	*x* has *p* as a belief
(GOAL x p):	*x* wants that *p* is true at some point in the future.
(OUGHT p):	an obligation concerning any given proposition *p*;
(DONE x a) = def (DONE a) ∧ (AGT x a)	action *a* has been done by agent *x*;
◊p	*p* is true at some point in the future.
((GOAL x p) ∧ (BEL x (q ⊃ ◊p))) ⊃ ◊(GOAL x q) (A1)	if *x* wants *p* and believes that if *q* than *p* will follow, then *x* will want that *q* as well.
(OBTAIN x p)=def (GOAL x p) ∧◊(p ∧ (BEL x p))	*x* obtains *p* iff *p* is a goal of *x*'s and later it will be true and *x* will believe so.
(GOAL-CONFL x y p q)=def (GOAL x p) ∧ (GOAL y q) ∧ (p ∨ q)	goals which consist of incompatible propositions are in conflict.

2.1. The normative belief.

Ns are generally considered as prescriptions, directives, commands. Even when expressed in any other type of speech act they are meant to "direct" the future behavior of the agents subject to them, their addressees (As).

To this end, they ought to give rise to some new goal in an A's mind (cfr. [CON91b]): for autonomous agents to undertake (or abstain from) a course of action, it is not sufficient that they know that this course of action is wanted (by someone else) (cf. [ROS88]; [GAL90]). It is necessary that these agents have the goal to do so. Norms may act as a mechanism of goal-generation. Indeed, they represent a powerful mechanism for inducing new goals in people's minds in a cognitive way. How is this possible?

At first, Ns need being represented as beliefs in an A's mind. Let us start from beliefs about requests. A simple request, before succeeding and producing acceptance on the side of the required person, is nothing but a belief, namely a belief about someone's will. Such a belief can be expressed as follows:

(BEL x (GOAL y (GOAL x p))) (1)

where agent *x* believes someone else, agent *y*, has the goal that *x* has the goal that *p*. More specifically, *x* believes that what *y* requires of her is to do an action planned for *p*:

(BEL x (GOAL y ((DONE x a))). (2)

Now, two questions arise here:

1. what is the difference, if any, between this type of belief, that is, a belief about an *ordinary will*, and a belief about a normative or *prescriptive will*? Is a N always represented in people's minds as an expression of some particular external will?

2. how do we go from a belief about a normative will to the goal to comply with it, that is, to a normative goal?

Starting from the former question, one could say, that normative beliefs are beliefs about a general will affecting a class of agents. This equals to saying that there is a class of agents wanting some of them to accomplish a given action. This view is interesting and fits rather well the DAI field, where attention is now increasingly paid to shared mental states and collective action ([GRO90]; [LEV90]; [COH91], etc.; for

review and further analysis, see [RAO92]). However, it presents two drawbacks.

First, it relies upon a notion of collective or group's will as yet fundamentally distributive. In [RAO92], for instance, a social agent's wants and beliefs are defined as the *conjunction* of goals and beliefs of the group's members. The authors contrast this with the opposite view (held, for instance, by [SEA90]) of social entity as "irreducible" to their members. An alternative to both views exists, namely to say that individuals form a collective entity if *they objectively depend on one another to achieve one and the same goal or interest* (the latter being defined in [CON91a] as a worldstate neither wanted nor believed by the agents involved which nonetheless implies the future achievement of (one of) their goals). Furthermore, the group may achieve its common goal, or realize its common interest, by distributing tasks among its members. However, the task assignment may be accomplished by a specialized subcomponent of the group. In other words, the notion of *group will* can be "reduced" to the mental states (of some) of its members without necessarily supposing shared goals, or even joint intentions,

Secondly, and moreover, a normative belief seems to be grounded on something more than a general will. In particular, what seems to be implied is a notion of *obligation*, or duty. In this perspective, let us express the general form of a normative belief as follows:

$$(N\text{-}BEL\ x\ y_i\ a) =def \qquad (3)$$
$$(\Lambda_{i=1,n}(BEL\ x\ (OUGHT\ (DONE\ y_i\ a))))$$

where *(OUGHT (DONE y_i a))* stands for an *obligation for a set of agents y_i to do action a*. The question is: what relation does *(3)* bear with belief type *(2)*? This relation seems possible thanks to the notion of *normative will*. In other terms, a normative belief implies a belief about the existence of a normative will:

$$(N\text{-}BEL\ x\ y_i\ a) \supset \qquad (4)$$
$$\exists z(BEL\ x\ (GOAL\ z\ (OUGHT\ (DONE\ y_i\ a))))$$

where z belongs to a higher level set of agents y_j of which y_i is a subset: An agent x has a normative belief about action a if x believes that *someone wants* that it is *obligatory* for y_i to do a. This is a minimal condition, since z might even coincide with the whole superset y_j. In the latter case, the normative will coincides with the group's will. But this is not necessary. Suffice it to say that, in a normative belief, a subcomponent (individual or social[1]) of the group is mentioned to issue a request. In x's belief, what makes a request normative is the very fact that a given z is believed to want that y_i have an obligation to do a.

To be noted, what is here proposed is a notion of *abstract* obligation, one which goes beyond any *personal* (or supposedly such) *will*. In coercion, for example, the coercive agent does not (nor is believed to) have the goal that the coerced agent believes in an abstract obligation to do something. All he needs is to persuade other that she is forced (namely threatened) to do *what the coercive agent wants*. On the contrary, a normative will is believed to create a mental state of abstract obligation, independent of any personal want or need. In other words, z is not (believed to be) happy with y_i's doing a in virtue of z's personal request (be it coercive or not). By default a normative will is one which wants you to have an obligation to do something, and not simply the corresponding goal.

Furthermore, a normative will is usually believed in turn to be *grounded on norms*, to be norm-based. More specifically, a strong sense of normative will occurs when that will is characterized as *held to* issue Ns:

$$(N\text{-}BEL\ x\ y_i\ a) \supset \qquad (5)$$
$$\exists z(BEL\ x(OUGHT(GOAL\ z\ (OUGHT(DONE\ y_i\ a)))))$$

A weaker or milder meaning is that of entitled, or *legitimate* will, a legitimate goal being defined as follows:

$$(L\text{-}GOAL\ x\ p) =def \qquad (6)$$
$$\forall y \exists q(GOAL\text{-}CONFL\ x\ y\ p\ q) \supset$$
$$(OUGHT\ \neg(GOAL\ y\ q))$$

in words: p is a legitimate goal of x's iff forall agents y that happen to have a given goal q conflicting with p, y ends up with having an obligation to give up

[1] As in [RAO92], the definition of a social entity should be recursive. Therefore, what has been said with regard to the group at large applies to its subcomponents as well. In case z in turn is a multi agent subcomponent, its will might be shared or not among its members. Its task (say, to legiferate) might be accomplished in such a way that not all members share the same goals.

q. In other words, a legitimate goal is watched over by a norm. In this sense, rights and legitimacy are said to give assistance to, and go to the rescue of, those who cannot defend themselves against aggression and cheat. Consequently, an entitled will is that which a norm protects from any conflicting interest.

To sum up, for an agent to have a normative belief is sufficient to believe that there is an obligation for a given set of agents to do a given action. At a more careful examination, however, the obligation is believed to imply that:

- a given action is *prescribed*, that is, requested *by*

- *a norm-based will*, be it held to issue that request, or simply entitled to do so.

Of course, a N-belief does not imply that a deliberate issuing of a N has *in fact* occurred. Social norms are often set up by virtue of functional unwanted effects. However, once a given effect is believed to be a social norm, an entitled will is also *believed* to be implied, if only an anonymous one ("You are wanted/expected to (not) do this...", "It is generally expected that...", "This is done so...", etc.).

This equals to saying that the present model of N-beliefs is recursive. A request is believed to be normative if it ultimately traces back to some norm. This is not to say that Ns are "irreducible" objects. Of course, the origins of Ns call for an explanation which unavoidably brings into play the community of agents, their interests and their interactional practice (cf. [ULL77]). However, in the agents' representations there is no need for keeping a record of such history. In the agents' beliefs a N is always represented as a legitimate, even a norm-driven, prescription. The present model tries to give an account of this evidence.

3. The route of Ns in the mind.

Turning to question 2. raised above, a normative belief is only one of the ingredients of normative reasoning. *Norms*, indeed, *are hybrid configurations of beliefs and goals*. Actually, as defined so far, a normative belief is only descriptive: it does not "constrain" or regulate the believer and his decisions. Indeed, an observer 's description of a society's rules does not influence in any relevant way her decisions. What is needed for an agent to regard herself as subject to, addressed by, a given N?

3.1. The pertinence belief.

First another belief is needed, namely a pertinence belief: For x to believe that she is addressed by a given N, x needs to believe that she is a member of the class of As of that N:

$$(P\text{-}N\text{-}BEL\ x\ a) =def \quad (7)$$
$$(\Lambda_i{=}1,n\ (N\text{-}BEL\ x\ y_i\ a)) \wedge (V_{j=1,n}(BEL\ x\ (x = y_j)))$$

where *P-N-BEL* stands for normative belief of pertinence.

Now, x's beliefs tell her not only that there is an obligation to do action a, but also that the obligation concerns precisely herself.

3.2. The normative goal.

Still, *(7)* is not much less "descriptive" than *(3)*. We do not see any normative goal, yet.

First, let us express a N-goal as follows:

$$(N\text{-}GOAL\ x\ a) =def \quad (8)$$
$$(P\text{-}N\text{-}BEL\ x\ a) \wedge (GOAL\ x\ (DONE\ x\ a))$$

A normative goal of a given agent x about action a is therefore a goal that x happens to have as long as she has a pertinence normative belief about a. Ultimately, x has a normative goal in so far as and because she believes to be subject to a N. Therefore, a N-goal differs, on one hand, from a simple constrain which reduces the set of actions available to the system (cf. [SHO92]), and, on the other, from other ordinary goals.

With regard to behavioral constrains, a N-goal is less compelling: An agent endowed with N-goals is allowed to compare them with other goals of hers and to some extent freely choose which one will be executed. Only if endowed with N-goals an agent may legitimately be said to comply with, or violate, a N. Only in such a case, indeed, she may be said to be truly normative.

With regard to ordinary goals, a N-goal is obviously more compelling: when an agent decides to give it up, she knows she both thwarts one of her goals

and violates a N [2].

Now, the question is: How and why does a N-belief come to interfere with *x*'s decisions? What is it that makes her "responsive" to the Ns concerning her? What is it that makes a normative belief turn into a normative goal?

3.2.1. Goal- and norm-adoption.

There seem to be several ways of accounting for the process leading to normative goals (N-goals) as well as several alternative ways of constructing a N-agent. There also seems to be a correspondence between the process from a belief about an ordinary request to the decision of accepting such a request, which we called (cf. [CON91b]) *goal-adoption*, and the process from a N-belief to a N-goal, which by analogy will be called here *norm-adoption*.

	Goal-Adoption	**Norm-Adoption**
1. Conditional Action	**Slavish**	**Automatic**
	$(BEL\ x\ (GOAL\ y\ (DONE\ x\ a))) \supset \Diamond (DONE\ x\ a)$	$(P\text{-}N\text{-}BEL\ x\ a) \supset \Diamond (DONE\ x\ a)$
2.Instrumental Adoption thanks to (A1)	**Self-interested**	**Utilitarian**
	$\forall p \exists q.(BEL\ x\ ((OBTAIN\ y\ p) \supset \Diamond (OBTAIN\ x\ q))) \supset \Diamond (GOAL\ x\ (OBTAIN\ y\ p))$	$\forall a \exists p.((P\text{-}N\text{-}BEL\ x\ a) \wedge (BEL\ x\ ((DONE\ x\ a) \supset \Diamond (OBTAIN\ x\ p)))) \supset \Diamond (N\text{-}GOAL\ x\ a)$
3.Cooperative Adoption thanks to (A1))	**Co-interested**	**Value-driven**
	$\forall p \exists q.(BEL x((OBTAIN y\ p) \supset \Diamond q)) \supset \Diamond (GOAL\ x\ (OBTAIN\ y\ p))$	$\forall a \exists q.((P\text{-}N\text{-}BEL x\ a) \wedge (BEL\ x\ ((DONE\ a) \supset \Diamond q))) \supset \Diamond (N\text{-}GOAL\ x\ a)$
	with *q*'s being (in *x*'s beliefs) commonly wanted by *x* and *y*	with *q*'s being (in *x*'s beliefs) a worldstate positively value by both *x* and the normative source: $(BEL\ x\ (q = v_{(x\ z)}))$ with *v* standing for any value.
4. Terminal Adoption	**Benevolent**	**Kantian**
	$(\Lambda_y=1,n\ (GOAL\ x\ (OBTAIN\ y\ p_y)))$ with p_y being the set of y's goals.	$(\Lambda_x=1,n(N\text{-}GOAL\ x\ n_{(x)}))$ with a_x being the set of N-actions required of x

Table 1: **The route of Ns in the mind.**

In situation 1. (**conditional action**), we find some sort of production rule: in goal-adoption (G-A), anytime a request is received by a system endowed with such a rule, a goal that *a* be done is fired. Analoguously, in N-adoption (N-A), anytime a N-belief is formed a N-goal is fired. Now, this is a rather cheap solution: *no* reasoning and *autonomy* are allowed. It is simple machinery that could be of help in cutting short some practical reasoning, but is insufficient as far as the modelling of normative reasoning is concerned. However, such a rule seems to account for a number of real-life situations. Think, as far as *slavish* G-A is concerned, of the habit of giving instructions when asked by passengers, and in the case of *automatic* N-A, of the routine of stopping at the red light (of course, in situations 1, it is hard to differentiate G-A from N-A).

---

[2] Intuitively, she gives up both the expected consequences of the action prescribed (any worldstate supposedly convenient to the agent or otherwise positively valued) *and in addition* sustains the costs of N-transgression. Although required, a formal treatment of both aspects is beyond the scope of this work.

In situations 2 (**instrumental adoption**), *greater autonomy* is allowed: adoption is subject to restrictions. In G-A, on the base of this rule, *x* will *self-interestedly* adopt only those of *y*'s goals which *x* believes to be a sufficient condition for *x* to achieve some of hers. Typically, but not exclusively, this rule depicts situations of *exchange*. An *utilitarian* N-A rule says that forall Ns, *x* will have the corresponding N-goals if she believes she can get something out of complying with them. (Think of the observance of Ns for fear of punishment, need of approval, desire to be praised, etc..)

Cooperative, or *cointerested*, goal **adoption** occurs whenever an agent adopts another's goal to achieve a common goal. N-adoption is cooperative when it is *value-driven*, that is, when the agent autonomously shares both the end of the norm and the belief that the latter achieves that end. This type of N-A can be seen as some sort of moral cooperation since the effect of the norm is shared (in the N-addressee's beliefs) by the aaddressee and the normative source.

The last situation is **terminal adoption.** This is not a rule, but a *meta-goal* which is defined, in the case of G-A, as *benevolent* (*x* is benevolent with regard to *y* when she wants the whole set of *y*'s goals to be achieved), and, in the case of N-A, *"Kantian"* "(*x* wants to observe the whole set of Ns addressing herself as ends in themselves).

In situation 1, the rule is a typical production rule. Its output is an *action*. In situations 2 and 3, the rules output some specific *goals*. In the case of N-A, the agent ends up with a new type of goal, namely a *normative goal*.

As seen at the beginning, this implies *x*'s belief that she is requested to do *a* by a normative will. But it implies two further beliefs as well, namely that the normative source is *not acting in its own personal interests;* and that *other agents are subject to the same entitled request* (in a normative belief, a set of norm addressees is always mentioned). Now, these further aspects play a relevant role, especially within the process leading from N-goals to N-actions.

A N-goal, in fact, is not sufficient for an agent *to comply with* a N. Several factors occurring within the process leading from N-goals to N-actions might cause the agent to abandon the goal and transgress against the norm. Among the others (more urgent conflicting goals; low expected chances of being caught red-handed, etc.), what is likely to occur is a confrontation with other As of a given N. As known, a high rate of transgressions observed discourages one's compliance. Viceversa, and for the same reason, it is possible to show that if one has complied with a given N, one will be likely to influence other agents to do the same (normative equity). Indeed, it can be argued (cf. [CON92]) that *normative influencing* plays a rather relevant role in the spreading of normative behavior over a population of autonomous agents.

4. Conclusive remarks and future research

In this paper, the necessity of a cognitive modelling of norms has been argued. It is proposed to keep distinct the normative choice from any norm-like behavior, that is that behavior which appears to correspond to norms. Such a difference is shown to be allowed only thanks to a theory of norms as a two-fold object (internal, that is, mental and external, or societal).

Some instruments, still rather tentative, have been proposed for a formal treatment of the "internal side" of Ns. In particular, a view of Ns as a complex mental object has been attempted. This object has been shown to consist of other more specific ingredients, namely goals and beliefs. Two notions of normative belief and goal have been provided and discussed, and aspects of the process of norm-adoption examined and confronted with the process of adopting another agent's goals.

Acknowledgement

We would like to thank Dr. Gianni Amati for his helpful reading the paper.

References

[BIC90] Bicchieri, C Norms of cooperation. *Ethics, 100*, 1990, 838-861.

[CAS91] Castelfranchi, C. & Conte, R. Problemi di rappresentazione mentale delle norme. Le strutture della mente normativa, in R. Conte (ed.) *La norma. Mente e regolazione sociale.* Roma, Editori Riuniti, 1991, 157-193.

[COH90] Cohen, P. R. & Levesque, H. J. Persistence, Intention, and Commitment, in P.R Cohen, J. Morgan & M.A. Pollack (eds.) *Intentions in Communication*. Cambridge, MA, MIT, 1990.

[COH91] Cohen, P. & Levesque, H.J. *Teamwork*.. TR-SRI International, 1991.

[CON91a] Conte, R. & Castelfranchi, C. Mind is not enough. Precognitive bases of social action. In N. Gilbert (ed.), Proceedings of the Simulating Societies Sympsium '92, London, UCL Press, in press; TR-IP-PSCS, 1991.

[CON91b] Conte, R., Miceli, M. & Castelfranchi, C. Limits and levels of cooperation. Disentangling various types of prosocial interaction. In Y. Demazeau & J.P. Mueller (eds.) *Decentralized AI-2*. North-Holland, Elsevier, 1991, 147-157.

[CON92] Conte, R. & Castelfranchi, C. Minds and Norms: Types of normative reasoning. In C. Bicchieri & A. Pagnini (eds.), *Proceedings of the 2nd Meeting on "Knowledge, Belief, and Strategic Interaction"*, Cambridge University Press, in press; TR-IP-PSCS, 1992.

[DAV83] Davies,R. & Smith P.G. Negotiation as metaphor for distributed problem-solving. *Artificial Intelligence, 20*, 1983, 63-109.

[GAL90] Galliers, J.R. The positive role of conflict in cooperative multi-agent systems, in Y. Demazeau, & J.P. Mueller (eds) *Decentralized AI*. North-Holland, Elsevier, 1990.

[GAS91] Gasser, L. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligenc, 47*, 1991, 107-138.

[GRO90] Grosz, B.J. & Sidner, C.L. Plans for discourse, in P.R. Cohen, J. Morgan, & M.E. Pollack (eds.) *Intentions in communication*. Cambridge, MA, MIT Press, 1990.

[GOU60] Gouldner, A. The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*, 1960, 161-179.

[LEV90] Levesque, H.J., Cohen, P.R., & Nunes, J.H.T. On acting together. *Proc. of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, 1990, 94-99.

[LEW69] Lewis, D. *Convention*. Cambridge, MA. Harvard University Press, 1969.

[MOS92] Moses, Y. & Tennenholtz, M. On Computational Aspects of Artificial Social Systems. *Proc. of the 11th DAI Workshop*, Glen Arbor, February 1992.

[RAO92] Rao, A. S., Georgeff, M.P., & Sonenmerg, E.A. Social plans: A preliminary report, in E. Werner & Y, Demazeau (eds.) *Decentralized AI - 3*. North Holland, Elsevier, 1992.

[ROS88] Rosenschein, J.S. & Genesereth, M.R. Deals among rational agents, in B.A. Huberman (ed.) *The ecology of computation*. North-Holland, Elsevier, 1988.

[SEA90] Searle, J.R. Collective intentions and actions, in P.R. Cohen, J. Morgan, & M.E. Pollack (eds.) *Intentions in communication*. Cambridge, MA, MIT Press, 1990.

[SHO92] Shoham, Y. & Tennenholtz, M. On the synthesis of useful social laws for artificial agent societies. *Proc. of the AAAI Conference*, 1992, 276-281.

[ULL77] Ulman-Margalit, E. *The emergence of norms*. Oxford, Clarendon, 1977.

[WER89] Werner, E. Cooperating agents: A unified theory of communication and social structure, in M. Huhns and L. Gasser (eds.), *Distributed artificial intelligence, Vol. 2*, Kaufmann and Pitman, London, 1989.