From the Theory of Mind to the Construction of Social Reality

Guido Boella (guido@di.unito.it)
Dipartimento di Informatica, Università di Torino - Italy
Cso Svizzera 185, I-10149 Torino, Italy

Leendert van der Torre (torre@cwi.nl)
CWI-Amsterdam and Delft University of Technology
Kruislaan 413, NL-1098 SJ Amsterdam, The Netherlands

Abstract

In this paper we argue that the hypothesis of the theory of mind advanced in cognitive science can be the basis not only of the social abilities which allow interaction among individuals, but also of the construction of social reality. The theory of mind is the attribution, via the agent metaphor, of mental attitudes, like beliefs and goals, to other agents. Analogously, we attribute mental attitudes to social entities, like groups, normative systems and organizations with roles. The agent metaphor explains the necessary abilities to deal with complex aspects of social behavior, like acting in a group, playing a role in an organization, and living in a reality organized in institutions which create regulative and constitutive norms to regulate behavior. To show the feasibility of this approach we provide a computational model of the construction of social reality based on multiagent systems.

Introduction

Interpreting other people's actions and intentions involves a mutual attribution of mental states so that the understanding of the people around us becomes coherent and intelligible. Our interpretive abilities should be viewed as a specific endowment of the human mind to understand others and ourselves in terms of mental states, like beliefs and goals. A new field of investigation, exploring the so-called *theory of mind*, has emerged as a major issue in cognitive science in the last two decades.

The ability to reason about mental states has been called a theory of mind because it shares some features with scientific theories: humans postulate unobservables, predict them from observables, and use them to explain other observables. Different views of the theory of mind have been proposed. Some scholars describe the underlying cognitive structure responsible for the theory of mind as an innate, dedicated, fast, automatic, at least partly encapsulated module, that is activated around three years of age.

A different view is proposed by [Wellman, 1990] who has argued for a theoretical model of the theory of mind: instead of seeing it as a mental mechanism, he conceives it as a naive theory, with axioms and rules of inferences.

A striking different hypothesis, suggested by [Gordon, 1986], is mental simulation: the idea that our capacity of psychological understanding depends on our ability to run cognitive simulations. According to this view, it is possible to infer other people's intentions and future actions by using our own mind as a model for theirs. This presupposes only a capacity of pretense and of putting oneself in the other's place, and is a more economical explanation. Even if this last model seems different from the preceding ones, some authors argue that the approaches are homogeneous if one regards simulation as one of several processes involved in attributing mental states (another being inference) and if one recognizes that both processes rely crucially on a conceptual framework of mental states and their relation to behavior.

The theory of mind enables social behavior by means of the attribution of mental attitudes to other people. Less attention, instead, has been devoted to study which abilities are necessary to deal with more complex aspects of social behavior, like acting in a group, playing a role in an organization, living in a reality organized in institutions which create regulative and constitutive norms to regulate behavior. In [Searle, 1995]'s terms, these are the abilities necessary to *construct the social reality* humans live in.

This paper addresses the following research question: how is it possible to pass from a theory of mind to the construction of social reality? Moreover, as a sub-question: how it is possible to explain social reality without introducing further primitive abilities with respect to the theory of mind?

As methodology we apply the *agent metaphor* underlying the theory of mind, where we interpret "agent" as an entity whose behavior is explained in terms of beliefs, desires and goals. We claim that like humans attribute mental attitudes to other humans, thus considering them as agents, humans conceive social reality by attributing mental entities to social entities like groups, roles, institutions, normative systems and organizations [Boella and van der Torre, 2004b]. Thus we say, metaphorically, that social entities are agents. The attribution of mental attitudes to social entities is used to conceptualize them, to reason about them and to predict their behavior as well as to understand how to behave cooperatively in a group, how to play a role in an organization or in a society regulated by norms.

To to make these notions more precise and to provide a first step towards a computational model for simulation or analysis, we summarize the logical multiagent framework developed in [Boella and van der Torre, 2004b] illustrating how the formal model of an agent can be used to describe both the behavior of an agent and its ability to attribute mental attitudes to other agents, either real or socially constructed.

The paper is organized as follows. First, we motivate the agent metaphor. Then we apply it to different types of social entities: groups, normative systems and organizations with roles. Afterwards, we present the formal model. Conclusion ends the paper.

The agent metaphor

Social reality, to which groups, normative systems and organizations with roles belong, is a complex phenomenon and it is not directly accessible to our bodily experience. So it is plausible that to conceptualize and reason about it humans resort to analogical reasoning starting from some better known source domain which has a structure rich enough to be informative when mapped onto the new target domain of social reality.

First of all, to proceed in our analysis, we must identify a suitable domain to start from. As source domain in this paper we use the notion of agent, which is at the basis also of the theory of mind. This idea is also proposed by [Dennett, 1987]: attitudes like belief and desire are folk psychology concepts that can be fruitfully used in explanations of rational human behavior. For an explanation of behavior it does not matter whether one actually possesses these mental attitudes: we describe the behavior of an affectionate cat or an unwilling screw in terms of mental attitudes. Dennett calls treating a person or artifact as a rational agent the *intentional stance*:

> "Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do."

Predicting the actions of other agents is a necessity when agents interact in a common environment where they compete for resources. This requirement has been put forward by [Goffman, 1970], who argues that human actions are always taken in a situation of *strategic interaction*:

> "When an agent considers which course of action to follow, before he takes a decision, he depicts in his mind the consequences of his action for the other involved agents, their likely reaction, and the influence of this reaction on his own welfare" [Goffman, 1970, p. 12].

To predict the reaction of other agents it is necessary to have a model of their decision making process. The most economical way is to use the same decision model an agent uses to take decision himself. In the field of agent theory this idea has been formalized by [Gmytrasiewicz and Durfee, 1995] as *recursive modelling*:

> "Recursive modelling method views a multiagent situation from the perspective of an agent that is individually trying to decide what physical and/or communicative actions it should take right now. [...] In order to solve its own decision-making situation, the agent needs an idea of what the other agents are likely to do. The fact that other agents could also be modelling others, including the original agent, leads to a recursive nesting of models."

Recursive modelling considers the practical limitations of agents, since they can build only a finite nesting of models about other agents' decisions. [Gmytrasiewicz and Durfee, 1995] uses a quantitative model of decisions, while in [Boella and van der Torre, 2004b] we use the attribution of mental attitudes to recursively model the behavior of other agents in a qualitative model.

To make predictions about behavior, mental attitudes are attributed to already existing entities, i.e., other agents in the environment. Nothing prevents, however, that mental attitudes are attributed to entities which do not exist yet. Consider the case of expectations about the behavior of another agent. Since complex behavior can be better described by beliefs and goals, as Bratman argues, expectations can be explained by in terms of beliefs and goals too. But these mental attitudes are not attributed to the agent whose behavior is expected, since the expectations can be different from what he is predicted to do. Thus expectations describe in terms of beliefs and goals a fictional entity which is the desired representation of some real agent according to someone else. However, since this fictional entity is attributed beliefs and goals, we metaphorically consider it as an agent too.

Expectations are different from predictions of behavior: both can be given in terms of beliefs and goals, but the entity whose behavior is predicted is not requested to know what it is predicted to do. Expectations, instead, have a public character: they are known by the other agents who are associated with a desired representation of their behavior in terms of beliefs and goals. Moreover, they are also expected to act in the desired way due to their knowledge of what they are expected to do [Castelfranchi, 1998]. If an agent must be aware of what he is expected to do, he is requested to understand the description given in terms of beliefs and goals, i.e., to be an agent, too.

Using the attribution of mental attitudes to describe expectations is a first step towards a construction of social reality based on the theory of mind. While predictions describe what an agent is believed to do, expectations describe something which different from what is believed to happens: they describe a desired behavior. In the next section, the second step is to attribute mental attitudes to fictional entities which do not have a counterpart in the reality, in order to describe the behavior of these entities.

The use of analogical reasoning to exploit the theory of mind to construct social reality is cognitively plausible. For example, [Lakoff and Johnson, 1980] argue that metaphorical reasoning explains complex cognitive abilities in terms of other more basic abilities. Metaphors, as Lakoff and Johnson argue, are not only a form of figurative use of language, but they are at the basis of the cognitive ability of humans. Our minds use metaphors to understand and reason about concepts which we have no direct bodily experience of. For example, the domain of time is conceptualized and talked about by means of spatial notions and expressions. In the "time-as-space" metaphor, space is the *source* domain which is mapped to the *target* domain of time: the first is better known to us so that we can attribute its properties to the less known domain of time.

Our agent metaphor maps beliefs and goals of agents onto the features of the social entities we want to understand.

The construction of social reality

We apply the agent metaphor to explain how humans conceptualize, reason and talk about social entities, like groups, institutions, normative systems and organizations with roles. Social entities, which we cannot have a direct bodily experience of, are conceptualized via the agent metaphor: they are described as they were agents by attributing them mental attitudes. Social entities exist only as far as the members of a community collectively attribute them a functional status [Searle, 1995]. In our model this status is defined in terms of the mental attitudes attributed to social entities. In contrast, the theory of mind is the attribution of mental attitudes to already existing entities.

A similar view is supported also by [Tuomela, 1995] with his analysis of collectives like groups, institutions and organizations:

> The possibility of ascribing goals, beliefs, and actions to collectives relies on the idea that collectives can be taken to resemble persons. [. . .] Following commonsense examples, I will accept [...] that both factual and normative beliefs can be ascribed (somewhat metaphorically) to groups, both formal and informal, structured and unstructured.

The analogy underlying the agent metaphor, however, is not complete, since it must respect the constraints on the target domain: in particular, social entities are not capable of performing actions, but they only act indirectly via their members and representatives.

In the following sections we apply the agent metaphor to groups, normative systems and organizations structured into roles, detailing the mapping between beliefs and goals and the features of the different social entities.

Groups as agents

In the model presented in [Boella and van der Torre, 2004a], we explain cooperative behavior by considering the group as an agent: a group exists because it is collectively attributed by all its members mental attitudes like beliefs, desires and goals. Its beliefs represent the knowledge about how to achieve their shared goals. Its goals and desires represent the shared goals of its members as well as their preferences about the means to fulfill their goals and about costs they incur into. Note that to the group are attributed as motivations not only the shared goals, but also some private desires of the agents, so to minimize the costs for each agent; otherwise, the partners would not agree to stay in the group.

Following [Bratman, 1992] we consider as key features of shared cooperative activity the following behaviors of the members: *commitment to the joint activity*, *commitment to the mutual support* and *mutual responsiveness*. In our model Bratman's conditions are realized since agents of a group coordinate with each other, in the following way:

a. When they take a decision, they consider first the goals of the group and they try to maximize their fulfillment. Hence, they are committed to the joint activity.

b. When they take a decision, they include in it some actions which contribute to the efforts of their partners to reach the goals of the group. Hence, they are committed to mutual support.

c. When they take a decision, they recursively model the decisions of their partners and their effects under the assumption that the partners are cooperative, too. Hence, they are mutually responsive to each other.

In more detail, when an agent evaluates a decision, he first considers which goals and desires of the group are fulfilled by his decision and which are not (a); only after maximizing the fulfillment of these motivations he includes in his decision some actions fulfilling also his private goals. When agents base their decisions on the goals and desires of the group we will say that their agent type is cooperative. This classification of agents according to the way they give priority to desires, goals or obligations is inspired by the BOID agent architecture presented in [Dastani and van der Torre, 2002]. Taking into accounts the motivations of other agents, and, thus also the goals and desires of the group, is a cognitive ability called *adoption*: "having a state of affairs as a goal *because* another agent has the same state as a goal", [Castelfranchi, 1998]. According to him, adoption is a key capability for an agent to be social: social agents must be able to consider the goals of other agents and to have attitudes towards those goals. Hence, sociality also in this case presupposes a theory of mind.

An agent, to understand the impact of his decisions on his partners and, thus, on the goals of the group, has to recursively model what his partners will decide and how their decisions will affect the group's motivations (c). First, by using recursive modelling, the agent understands whether the group's performance can be improved by including in his decisions some actions which contribute to his partners' efforts (b). Second, the agent understands whether his decision conflicts with the predicted decisions of the other agents. Third, he understands when he needs to inform the partners when their goal has been achieved, or to proactively inform them about his decisions.

Our approach departs from the idea due to [Bratman, 1992] that shared cooperative activity is defined by individual mental states and their interrelationships, without collective forms of attitudes that go beyond the mind of individuals and without further mental states characterizing cooperative behavior. Bratman's "broadly individualistic" approach contrasts also with [Tuomela, 1995], who introduces *we-intentions* - "we shall do G" - which represent the internalization of the notion of group in its members, and [Searle, 1990] for whom "collective intentional behavior is a primitive phenomenon".

Normative systems as agents

In [Boella and van der Torre, 2004b], we use the agent metaphor of attributing mental attitudes to normative systems in order to explain normative reasoning in autonomous agents. The normative system is considered as an agent playing a game with the bearer of the obligation.

We start with a well known definition.

> *Normative systems* are sets of agents (human or artificial) whose interactions can fruitfully be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave [...]. Importantly,

the norms allow for the possibility that actual behaviour may at times deviate from the ideal, i.e. that violations of obligations, or of agents rights, may occur [Jones and Carmo, 2001].

This definition of Carmo and Jones does not seem to require that the normative system is autonomous, or that its behavior is driven by beliefs, desires and goals.

Our motivation for using the agent metaphor in [Boella and van der Torre, 2004b] is inspired by the interpretation of normative multiagent systems as dynamic social orders. According to [Castelfranchi, 2000], a social order is a pattern of interactions among interfering agents "such that it allows the satisfaction of the interests of some agent A". These interests can be a shared goal, a value that is good for everybody or for most of the members; for example, the interest may be to avoid accidents. We say that agents attribute the mental attitude 'goal' to the normative system, because all or some of the agents have socially delegated goals to the normative system; these goals are the content of the obligations regulating it, we will call them *normative goals*.

Moreover, social order requires *social control*, "an incessant local (micro) activity of its units" [Castelfranchi, 2000], aimed at restoring the regularities prescribed by norms. Thus, the agents attribute to the normative system, besides goals, also the ability to autonomously enforce the conformity of the agents to the norms, because a dynamic social order requires a continuous activity for ensuring that the normative system's goals are achieved. To achieve the normative goal the normative system forms the subgoal to consider as a violation the behavior not conform to it and to sanction violations.

Thus, in [Boella and van der Torre, 2004b] we define obligations in this way: an agent $\mathbf{a}$ is obliged by a normative agent $\mathbf{b}$ to do x in context c, $O_{\mathbf{ab}}(x, s \mid c)$ or else he is sanctioned with s, iff:

1. If agent $\mathbf{b}$ believes that condition c holds, it wants x holds too and that agent $\mathbf{a}$ adopts x as his decision.
2. If agent $\mathbf{b}$ believes $\neg x \wedge c$ then it has the goal that $\neg x$ is considered as a violation and to sanction $\mathbf{a}$ with s.
3. Agent $\mathbf{a}$ has the desire not to be sanctioned.

Hence, to reason about what is obligatory for him, an agent has to recursively model the behavior of the normative agent to understand whether he will be considered a violator and sanctioned. Note that obligations are modelled only by means of motivations, which formalizes the possibility that a normative system does not recognize that a violation counts as such, or that it does not sanction it. Both the recognition of the violation and the application of the sanction are the result of autonomous decisions of the normative system who is considered as an agent (but acts via its representatives).

While regulative norms like obligations are defined in terms of goals, beliefs in our metaphorical mapping define constitutive rules. Constitutive rules have of the form "such and such an X counts as Y in context C" where X is any object satisfying certain conditions and Y is a label that qualifies X as being something of an entirely new sort: an institutional fact. Examples of constitutive rules are "X counts as a presiding official in a wedding ceremony", "this bit of paper counts as a five euro bill" and "this piece of land counts as somebody's private property". According to [Searle, 1995], they are at the basis of the construction of social reality. In our model, a bit of paper counts as money if the community collectively attributes to the normative system the belief that the bit of paper is money [Boella and van der Torre, 2004b].

Constitutive norms provide an abstract classification of reality which regulative rules can refer to, in the same way as the goals of an agent refer to the believed state of the world when an agent has to decide what to do.

We take here full advantage of the metaphor, as also [Tuomela, 1995] argues for collectives like groups:

> "The notions of goal, belief, and action are linked in the case of a group to approximately the same degree as in the individual case. In the latter case their interconnection is well established; given that the person-analogy applies to groups [...], these notions apply to groups as well."

Roles as agents

In this section, we apply the agent metaphor to explain the structure of organizations in terms of roles, like director, employee, *etc.*, i.e., the representatives through which a social entity acts.

Roles are defined in sociology as *descriptions of expected behavior*. Again, as descriptions of behavior, roles can be defined in terms of belief and goals. As expectations these mental attitudes are attributed to a fictional agent which represents how the real agent playing the role should behave. But how do roles differ from mere expectations discussed in the previous section?

The difference between mere expectations towards someone and roles rests in who is attributing mental attitudes to these fictional agents defining the expected behavior. To have an expectation it is sufficient that a single agent attributes mental attitudes to another agent. In contrast, roles are always parts of a social entity, like an organization, which defines them to describe its own structure: they are social roles. Hence roles are defined via the attribution of mental attitudes, but these mental attitudes are attributed by a social entity (the organization) to another social entity: the role. This means that the entity defining a role needs to be considered as an agent: attributing mental attitudes to other entities and being able to behave directed by beliefs and goals are fundamental features of agents.

This view of roles nicely fits our general approach based on the agent metaphor. Social entities like normative systems, organizations and groups are considered as agents and attributed mental attitudes: social entities, *qua agents*, are able to define roles by attributing them mental attitudes.

In the metaphorical mapping the role's expertise is represented by beliefs of the agent and his responsibilities as the goals of the agent. To play a role an agent has to adopt the goals representing his responsibilities and to carry out them according to the beliefs representing his expertise: the player has to act *as if* he had the beliefs and goals of the role. Hence, to play a role it is necessary to understand descriptions of behaviors in terms of beliefs and goals and to figure out by recursive modelling which is the expected behavior. Again, to play a role the theory of mind is necessary.

The formal model

In this section we sketch the formal model of multiagent systems which makes precise our theory and can be used to validate it. Full details can be found in other papers like [Boella and van der Torre, 2004a, Boella and van der Torre, 2004b]. To support the extension of the agent metaphor to social entities we need a way to describe agents' behavior in terms of mental attitudes, and to model how agents can attribute mental attitudes to other agents in order to foresee their decisions by means of recursive modelling.

First of all, we need a simple language to describe states of affairs. For this reason, we introduce a set of propositional variables X and we extend it to consider also negative states of affairs: $Lit(X) = X \cup \{\neg x \mid x \in X\}$ are the literals built out of X.

To represent mental attitudes like beliefs B, desires D and goals G we use a rule based formalism: in this way we capture their conditional nature. The rules represent the relations among propositional variables existing in conditional beliefs, desires and goals of the agent: $Rul(X) = 2^{Lit(X)} \times Lit(X)$ is the set of pairs of a set of literals built from X and a literal built from X, written as $l_1 \wedge \ldots \wedge l_n \rightarrow l$, and, when $n = 0$, $\top \rightarrow l$.

Starting from a set of literals representing a state - for example a set of observations - and a set of belief rules, it is possible to incorporate the consequences of belief rules, using a simple logic of rules called out: $out(E, S)$ is the transitive closure of a set of literals $S \subseteq Lit(X)$ under the rules E. For details see the reusable input/output logic in [Makinson and van der Torre, 2000]. $out(B_a, S)$, e.g., represents the beliefs of agent a which derive from the observations S and the application of its belief rules on S.

Mental attitudes are represented by rules, even if they do not coincide with them: $MD : B \cup D \cup G \rightarrow Rul(X)$. To resolve conflicts among motivations $M = D \cup G$ we introduce a priority relation by means of a function $\geq$: $A \rightarrow 2^M \times 2^M$ from the set of agents to a transitive and reflexive relation on the powerset of the motivations containing at least the subset relation. We write $\geq_a$ for $\geq(a)$.

Different mental attitudes are attributed to the agents A by the agent description relation $AD : A \rightarrow 2^{B \cup D \cup G \cup A}$. We write $B_a = AD(a) \cap B$, $A_a = AD(a) \cap A$, for $a \in A$, *etc.*

As discussed in the previous sections, in our model there are different sorts of agents in the set of agents A. Besides real agents $RA \subseteq A$ (either human or artificial) we consider as agents in the model also socially constructed agents like groups, normative systems, organizations and roles SA ($RA \cap SA = \emptyset$ and $RA \cup SA = A$). Roles are described as agents but they are also associated with agents playing the role, $PL : SA \rightarrow RA$.

This does not mean that agents SA exist in the usual sense of the term. Rather, social entities exist only as they are accepted as such by other agents (either real or not): considering a social entity as an agent allows to describe its behavior in terms of mental attitudes. Agents are in the target of the AD relation for the this reason: groups, normative systems and organizations exist only as profiles attributed by other agents. The AD relation induces an exists-in-profile relation specifying that an agent $b \in SA$ exists only as some other agents attribute to it mental attitudes: $\{a \in RA \mid b \in A_a\} \neq \emptyset$.

To model actions of agents we adopt a simple solution: the set of variables whose truth value is determined by an agent (decision variables representing actions) is distinguished from the set of variables which are not controllable (the parameters P). The parameters P are a subset of the propositional variables X. The complement of X and P represents the decision variables controlled by the different agents. Hence we associate to each agent a subset of $X \setminus P$ by extending again the agent description relation $AD : A \rightarrow 2^{B \cup D \cup G \cup A \cup (X \setminus P)}$.

We can now define a multiagent system as $MAS = \langle RA, SA, X, P, B, D, G, AD, MD, \geq, PL \rangle$.

Games among agents

The advantage of the attribution of mental attitudes to social entities is that standard techniques developed in qualitative decision and game theory can be applied to interaction among agents: either real agents or socially constructed ones, whose behavior can be recursively modelled and predicted using the mental attitudes attributed to them. Here we consider a simple form of games between two agents **a** and **b** in A. For example, **a** and **b** can be two partners in a group, or **b** can be a normative system and **a** is predicting whether his behavior will be sanctioned.

The set of decisions Δ is the set of subsets $\delta = \delta_{\mathbf{a}} \cup \delta_{\mathbf{b}} \subseteq Lit(X)$. For an agent $a \in A$ and a decision $\delta \in \Delta$ we write δ_a for $\delta \cap Lit(X_a)$: the decision of a is the set of actions it performs in a certain situation. When agent **a** takes its decision $\delta_{\mathbf{a}}$ it has to minimize the unfulfilled motivational attitudes it considers relevant: its own desires $D_{\mathbf{a}}$ and goals $G_{\mathbf{a}}$, but also the desires $D_{\mathbf{b}}$ and goals $G_{\mathbf{b}}$ of the group it belongs to or of the normative system which **a** is subject to or of the role it is playing. But when it considers these attitudes, it must not only consider its decision $\delta_{\mathbf{a}}$ and the consequences of this decision; it must consider also the decision $\delta_{\mathbf{b}}$ of its interactant **b** and its consequences $out(B_{\mathbf{a}}, \delta)$. So agent **a** recursively considers which decision agent **b** will take depending on its different decisions $\delta_{\mathbf{a}}$: $out(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}}))$.

On the decisions Δ we require that their closures under the beliefs $out(B_{\mathbf{a}}, \delta)$ and $out(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}})))$ do not contain a variable and its negation: a decision of an agent cannot lead to a situation which is believed inconsistent.

Note that there is no restriction to the possibility that decisions include decision variables which do not contribute to the goals of the agent. In particular, the decisions can contain decision variables contributing to the goals to be achieved by the partners of the agent in a group, or decision variables aiming at respecting the obligations of the normative system.

Given a decision $\delta_{\mathbf{a}}$, a decision $\delta_{\mathbf{b}}$ is optimal for agent **b** if it minimizes the unfulfilled motivational attitudes in $D_{\mathbf{b}}$ and $G_{\mathbf{b}}$ according to the $\geq_{\mathbf{b}}$ relation. The decision of agent **a** is more complex: for each decision $\delta_{\mathbf{a}}$ it must consider which is the optimal decision $\delta_{\mathbf{b}}$ for agent **b**. More formally:

- the unfulfilled motivations of decision δ according to agent $\mathbf{a} \in A$ be the set of motivations whose body is part of the closure of the decision under belief rules but whose head is not.
  $U(\delta, \mathbf{a}) = \{m \in M \mid MD(m) = l_1 \wedge \ldots \wedge l_n \rightarrow l, \{l_1, \ldots, l_n\} \subseteq out(B_{\mathbf{a}}, \delta) \text{ and } l \notin out(B_{\mathbf{a}}, \delta)\}$

- the unfulfilled motivations of decision $\delta = \delta_{\mathbf{a}} \cup \delta_{\mathbf{b}}$ according to agent **b** be the set of motivations whose body is in the observable part of the closure of the decision under belief rules, but whose head is not:
  $U(\delta, \mathbf{b}) = \{m \in M | MD(m) = l_1 \wedge \ldots \wedge l_n \rightarrow l, \{l_1, \ldots, l_n\} \subseteq out(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}}))) \text{ and } l \notin out(B_{\mathbf{b}}, \delta_{\mathbf{b}} \cup (out(B_{\mathbf{a}}, \delta_{\mathbf{a}})))\}$

- a decision δ is *optimal* for agent **b** if and only if there is no decision $\delta'_{\mathbf{b}}$ such that $U(\delta, \mathbf{b}) >_{\mathbf{b}} U(\delta_{\mathbf{a}} \cup \delta'_{\mathbf{b}}, \mathbf{b})$. A decision δ is optimal for agent **a** and agent **b** if and only if it is optimal for agent **b** and there is no decision $\delta'_{\mathbf{a}}$ such that for all decisions $\delta' = \delta'_{\mathbf{a}} \cup \delta'_{\mathbf{b}}$ and $\delta_{\mathbf{a}} \cup \delta''_{\mathbf{b}}$ optimal for agent **b** we have that $U(\delta', \mathbf{a}) >_{\mathbf{a}} U(\delta_{\mathbf{a}} \cup \delta''_{\mathbf{b}}, \mathbf{a})$.

Decision making

The agents value decisions according to the desires and goals which have been fulfilled and which have not. The agents can be classified according to the way they give priority to the different possible motivations: private desires and goals and desires and goals of the group or of the normative system or of the role they play that can be adopted. We define agent types as they have been introduced in the BOID architecture [Dastani and van der Torre, 2002].

For example, cooperative agents give priority to the desires and goals of the group; they pursue their private goals only if they do not prevent the achievement of the group's objectives:

Selfish agent A selfish agent always tries to minimize its own unfulfilled desires and goals:

Given decisions $\delta, \delta' \in \Delta$, if $U(\delta, a) \geq_a U(\delta', a)$ then $U(\delta, a) \cap (D_a \cup G_a) \geq_a U(\delta', a) \cap (D_a \cup G_a)$

Cooperative agent A cooperative agent always tries to minimize the unfulfilled desires and goals of the group **b**, before minimizing its private goals and desires:

Given decisions $\delta, \delta' \in \Delta$, if $U(\delta, a) \geq_a U(\delta', a)$ then $U(\delta, a) \cap (D_{\mathbf{b}} \cup G_{\mathbf{b}}) \geq_{\mathbf{b}} U(\delta', a) \cap (D_{\mathbf{b}} \cup G_{\mathbf{b}})$

Similar definitions can be provided for agents who give precedence to goals with respect to desires, agents who adopt as their goals the obligations they are subject to, *etc*.

Conclusion

In this paper we discuss the role of the theory of mind in the construction of social reality. We argue that the attribution of mental attitudes proper of the theory of mind can be fruitfully used to conceptualize social entities like groups, normative systems, organizations and roles. This agent metaphor is a conceptually economical and cognitively plausible way to explain a complex aspect of reality and it is supported also by philosophers like [Tuomela, 1995]. Furthermore, we provide a computational model of the agent metaphor based on multiagent systems. This model, which is only summarized in this paper, allows to explain various aspects of social reality, from groups [Boella and van der Torre, 2004a] to legal reasoning [Boella and van der Torre, 2006]. See these papers for further details and references.

References

[Boella and van der Torre, 2004a] Boella, G. and van der Torre, L. (2004a). Groups as agents with mental attitudes. In *Procs. of AAMAS'04*, pages 964–971. ACM Press.

[Boella and van der Torre, 2004b] Boella, G. and van der Torre, L. (2004b). Regulative and constitutive norms in normative multiagent systems. In *Procs. of KR'04*, pages 255–265. AAAI Press.

[Boella and van der Torre, 2006] Boella, G. and van der Torre, L. (2006). A game theoretic approach to contracts in multiagent systems. *IEEE Transactions on SMC*.

[Bratman, 1992] Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, 101:327–341.

[Castelfranchi, 1998] Castelfranchi, C. (1998). Modeling social action for AI agents. *Artificial Intelligence*, 103:157–182.

[Castelfranchi, 2000] Castelfranchi, C. (2000). Engineering social order. In *Procs. of ESAW'00*, pages 1–18, Berlin. Springer Verlag.

[Dastani and van der Torre, 2002] Dastani, M. and van der Torre, L. (2002). A classification of cognitive agents. In *Procs. of Cogsci'02*, pages 256–261.

[Dennett, 1987] Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press, Cambridge, MA.

[Gmytrasiewicz and Durfee, 1995] Gmytrasiewicz, P. J. and Durfee, E. H. (1995). Formalization of recursive modeling. In *Procs. of ICMAS'95*, pages 125–132.

[Goffman, 1970] Goffman, E. (1970). *Strategic Interaction*. Basil Blackwell, Oxford.

[Gordon, 1986] Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1:158–171.

[Jones and Carmo, 2001] Jones, A. and Carmo, J. (2001). Deontic logic and contrary-to-duties. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, pages 203–279. Kluwer.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. UC Press, Chicago.

[Makinson and van der Torre, 2000] Makinson, D. and van der Torre, L. (2000). Input-output logics. *Journal of Philosophical Logic*, 29:383–408.

[Searle, 1990] Searle, J. (1990). Collective intentionality. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in communication*. MIT Press, Cambridge (MA).

[Searle, 1995] Searle, J. (1995). *The Construction of Social Reality*. The Free Press, New York.

[Tuomela, 1995] Tuomela, R. (1995). *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press.

[Wellman, 1990] Wellman, H. (1990). *The child's theory of mind*. MIT Press, Cambridge (MA).