From Desires, Obligations and Norms to Goals

Frank Dignum[1]
Utrecht University

David Kinny
The University of Melbourne

Liz Sonenberg
The University of Melbourne

Traditional models of agents based on Beliefs, Desires and Intentions usually only include either desires or goals. Therefore the process whereby goals arise from desires is given scant attention. In this paper we argue that the inclusion of both desires and goals in the same model can be important, particularly in a Multi-Agent System context, where other sources of individual motivation such as obligations and norms may be present. This leads us to propose an extended BDI architecture in which obligations, norms and desires are distinguished from goals and explicitly represented. In this paper we consider suitable logical representations for and properties of these elements, and describe the basic method of operation of the architecture, focusing on how goal generation and goal maintenance may occur.

Keywords: Agents, Desires, Obligations, Norms, Goals, Intentions.

1. Introduction

One of the most successful theoretical models of situated, rational agency is the BDI model, which takes the trinity of Beliefs, Desires and Intentions to be the key elements of an agent's mental state that serve as the basis for its decisions about when and how to act in the world. Conceptually, beliefs are statements of properties of its world (and of itself) that an agent takes to be true (distinct from knowledge by virtue of possibly being false); desires are actions that an agent wants to perform or situations that it prefers and wants to bring about; and intentions are those feasible

[1] Institute of Information and Computing Science, Utrecht University, P.O.Box 80089, 3508 TB Utrecht, The Netherlands, e-mail: dignum@cs.uu.nl, Home page: www.cs.uu.nl/~dignum

actions, plans or desired situations that an agent has selected and committed to performing or achieving.

Operationally, BDI agents repeatedly revise their beliefs to accomodate newly perceived information, and reason or deliberate, on the basis of their beliefs, desires and existing intentions, about what actions next to take and how to modify their intentions as time goes by and actions are performed. More abstractly, the elements B, D and I can be viewed as capturing an agent's informational, motivational, and deliberative states. The approach has its roots in work by Dennet (1987), Bratman (1987) and others, which explored the value of mentalistic descriptions and the critical role of intentions in tractable, practical reasoning, and effectively created a new field of study – BDI Agent theory. Perhaps the prototypical example of an agent architecture based on this theory is IRMA , proposed by Bratman, Israel and Pollack (1988). The BDI approach has been highly influential, in part because it is based on a simple, intuitive model of human practical reasoning, but perhaps especially because it has led to the development of sophisticated formal treatments of the model which can expose its intricacies and the consequences of particular choices about the properties of its elements in an exact and detailed manner.

Traditional formal treatments of the BDI model, such as those of Cohen and Levesque (1990), Rao and Georgeff (1991) and (Wooldridge, 2000), attempt to capture in a multi-modal logic framework the static and dynamic properties of beliefs, desires and intentions, and the relationships that are required to hold between these elements. Beliefs, for example, are traditionally axiomatized in a particular modal logic (weak S5), an approach which effectively distinguishes belief from knowledge and captures requirements for consistency, closure under consequence and introspection, etc., whereas intentions, whether defined as primary or derived modalities, are usually subject to logically weaker requirements such as consistency. Much of the logical apparatus in such frameworks is concerned with the relationships between modalities and their consequences for the dynamics of the system, e.g., by requiring that intentions should be believed logically possible and actually achievable in order to be adopted, and that they should be dropped when beliefs change in such a way as to bring them about or undermine these requirements.

Despite the sophistication of these formal models, desires have always been the Cinderella of the BDI trinity. Indeed, most formal treatments only contain either desires or goals, and often provide a logical representation of goals alone. Even in (Wooldridge, 2000) where Wooldridge mentions only desires as a modality and in the beginning of the book states that desires can be mutually exclusive (inconsistent), in his later formalization of the modality he assumes they have KD axiomatization (meaning that they cannot be inconsistent). In fact what he modeled is closer to goals than to desires even though he used the latter name. The dynamic process whereby goals arise from an agent's desires is often given scant attention or even swept under the carpet completely. While desires and goals are undoubtedly rather similar, the way in which an agent's desires influence, determine or perhaps are elevated to become goals is an important issue even within an individual agent, and one which becomes even more significant in the context of a Multi-Agent System (MAS), where other sources of motivation and influence on intentions arise. In particular, social relationships among groups of agents give rise to normative

conventions or rules of behaviour, called simply *norms*, and interactions between individual agents, such as the making of agreements and commitments, give rise to *obligations* that are expected to be be fulfilled. Failure to conform to norms or to fulfill obligations may lead to some form of penalty being imposed upon an agent.

In this paper we will argue that existing theoretical models of BDI agents are thus somewhat incomplete, lacking an adequate treatment of how goals arise from desires and the relationships that hold between them, and that they are also insufficiently rich to capture the various influences on goals and intentions that arise in complex MAS applications where obligations and norms are significant. As a consequence of these and other shortcomings, there continues to be a well recognized gap between BDI theory and the various practical agent architectures that adopt the BDI model. Judged in the light of their applicability, it might seem that theoretical BDI models are becoming less important, due to the looseness of their connection to practical systems, and because of their failure to guide research into new implementation directions in any obviously useful way.

One reaction to such criticisms might be to ignore the formalism and focus on refining implementations, however this would risk losing key intuitions and insights that could guide further development of the agent paradigm, and extensions of BDI agents with concepts like emotions, power relations, obligations, etc., would likely be *ad hoc*, lacking an adequate theoretical basis. Following this path, there would be no easy way of determining how such extensions might relate to the basic components of the BDI model, so the link between BDI logic and the implementation would remain unclear, and the implementation will not impose clear inherent restrictions or guidelines on these relations. The BDI logic might give some insights, but if there is no clear relation between the logic and the implementation, these insights are of dubious value.

Our long-term goal is instead to extend the BDI model, at a theoretical and practical level, with social concepts such as norms and obligations. Elsewhere we and others have argued that such concepts are important to "glue" together autonomous agents in a MAS (Castefranchi, 1998; Castelfranchi, Dignum, Jonker & Treur, 2000; Cavedon & Sonenberg, 1998; Dignum, Morley, Sonenberg & Cavedon, 2000; Singh, 1996). In this paper we will take two steps towards achieving this goal. This paper can be seen as an extension of (Dignum, et.al., 2000) where we concentrated on the technical aspects of the model. In this paper we concentrate on the argumentation for the model and its application in practical situations (while including some of the formalism). After introducing basic concepts of motivation and presenting an example (Section 2), we will revisit the relationships between goals, desires and intentions in the BDI model with a view to understanding the kinds of situations in which both goals and desires should be represented (Section 3). We will then look at ways in which norms and obligations might be added to this model and make some general comments about such formalisations (Sections 4 and 5). In particular, we will propose that for a process of goal generation, interactions between desires, obligations and norms can be resolved by preference orderings, and then go on to discuss how an agent might form intentions given certain desires, obligations, norms and goals.

2. Types of Motivation

We will begin by describing and exploring informally the distinguishing properties of desires, obligations, norms and goals. Desires we take to be those forms of motivation, which arise internally within an agent, whereas obligations and norms are forms of motivation that arise from an agent's interactions and relationships with other agents. Goals also capture motivation, but we will regard them as arising from some form of decision process applied to various primary sources of motivation, which somehow filters and ranks them in order to determine which motivations will be acted upon, in what order. We will also assume a typical BDI scenario where an agent's goals, once adopted, serve as the basis for further deliberation and *means-end reasoning* processes which result in the agent making commitments to achieve particular goals by creating and acting upon intentions – abstract or concrete plans of action for how to achieve specific goals.

Both goals and desires are typically represented as (sets of) states or situations in which the desire or goal has been satisfied or achieved. Less typically but quite reasonably, one may have a desire or goal to perform a particular action or durative activity; such desires and goals are often represented in the former way as states in which a particular action has just been done by the agent. For example, a desire to drink coffee will often be represented as a desire to be in a state where one has just drunk coffee. Either type of representation may be taken as primary, but one advantage of a state-based representation of desires or goals is that they may be represented even if the actions which might satisfy them are not yet known.

The chief distinction between desires and goals concerns their feasibility and consistency: desires are usually taken to be a more primitive and in some sense less rational or predictable form of motivation, and it is accepted that it may be impossible in principle or in practice to achieve any or all of an agent's set of desires, although this fact may or may not be recognized by the agent. An agent's set of goals, by contrast, are usually required to be both individually and jointly achievable, so that the process by which new goals are generated is expected to take into account both the feasibility of achieving goals, given particular resources and present and expected situations, and also the extent to which their achievement is consistent with achieving existing goals, rejecting those candidate goals whose achievement is not believed possible or practicable. Of course, consistency of a set of goals may equally be maintained by abandoning existing goals when new goals are adopted, and this outcome may well occur in a situation where various degrees of urgency, utility or preference are associated with individual goals.

As mentioned above, obligations and norms differ from desires in that they arise from interactions and relationships with other agents. Obligations often arise from interactions between pairs of agents as a result of visible, explicit commissives such as promises, or agreements to requests or contracts, but may also arise due to the "rules of the game" which apply to an interaction, i.e., they may be explicit or implicit elements of an interaction protocol. Obligations may also arise in a multi-party setting such as team activity. For example, in a team situation where a joint goal has been established it is typically incumbent on a team member who recognizes that the joint goal is no longer feasible to promptly notify the other members of the team, to avoid futile activity on the part of others (Kinny, Ljungberg, Rao, Sonenberg, Tidhar & Werner, 1994).

Unlike desires, obligations are thus directly related to other individuals, and are also typically associated with penalties that apply when they are not fulfilled. There may be some mechanism, organization or other body which is responsible for enforcing the penalty, which may range from a transient and inconsequential loss of reputation, through a loss of privilege or a direct financial cost, to more substantial penalties such as expulsion from a team or organization. Obligations are thus explicit mechanisms for influencing the behaviour of agents and providing some stability and reliability in their interactions while allowing some flexibility[2]. They provide a level of "freedom of choice" with known explicit consequences.

Norms, as manifest in human societies and organizations, assist in standardising the behaviour of individuals, making it easier to cooperate and/or interact within that society. Commitment to membership of a group, organization or society means that an agent should tend to follow its norms, i.e., an agent will weigh up goals suggested by other sources against those suggested by societal norms. If agents are designed so they tend to follow norms, knowledge of these norms can allow for easier coordination, as certain behaviours of others can be anticipated with some degree of reliability. Norms may also be directly associated with a particular role within an organization rather than the organization as a whole. For example, a lecturer within a university is subject to certain normative rules of behaviour, which may not apply to other members of the university such as students.

The main difference between norms and obligations as we use them is that norms are more stable and abstract, and are inherent to a group of which an agent is a member, whereas obligations are usually a consequence of a direct action of the agent itself and entered into by choice, e.g., the obligation to pay when something is ordered. Just as an agent's interactions with different individuals may lead to distinct obligations, which perhaps may conflict, an agent's different roles and membership of different groups and organizations may impose various norms upon it, and these too may conflict. Indeed, norms may reasonably be regarded in many circumstances as representing abstract, standing obligations towards sets of individuals. The distinctions that we make between them in this paper are thus relatively crude ones; we are however including at least these two concepts as representatives of a whole spectrum of external motivational influence.

Consider the following scenario. I want my personal software agent to buy me a certain book (on agent mediated electronic commerce) as cheaply as possible. After communicating this desire to the agent, it has found the book listed at Amazon for $40, and on offer at an eBay auction, and it has also sent a request to the book broker El Cheapo, with a commitment to buy if El Cheapo can negotiate a price under $30. At eBay the price is still going up, and it has decided to bid $35 there, but it knows there is no guarantee of that being the winning bid. The search at El Cheapo has been going on for some time without any positive result. I then inform the agent that I really need the book by next Monday, and so it promptly places an order with Amazon to ensure that I have it in time. Just at that moment the results come in from eBay: it transpires that its bid of $35 at eBay has won the auction after all; moments later, before my agent can cancel the El Cheapo request, El Cheapo's

[2]We acknowledge that it is often useful to distinguish different types of obligations, but such distinctions are not important for the present paper.

agent reports back that it has found the book at a discounter not far from my home and negotiated a price of $28. My agent has now incurred three obligations. It should pay $40 to Amazon, it should pay $35 to eBay and it has committed to order the book for $28 through El Cheapo[3]. But I certainly don't need three copies of the book, and a quick check of my budget reveals that I can't afford to spend more than $50 in total. My agent must now somehow decide which of the three suppliers to pay, given this constraint.

Cancelling the Amazon order will legally require paying a $10 cancellation fee, but there's a chance that as I am a regular customer, they may waive the fee. Not paying at eBay will result in damage to my public reputation, but no fee. Not placing the order through El Cheapo may result in some hassle but no monetary cost, and any loss of reputation will be confined to El Cheapo and not seen by others. In terms of the penalties that will likely follow from fulfilling only one of the obligations, buying from Amazon seems the preferred solution for the agent. However, it also has some norms and desires involved in the decision. The norms that apply in this case are that all commitments should be fulfilled whenever possible, but where they conflict that public commitments should take priority, because loss of public reputation is most socially harmful. So, the agent's norms will dictate that fulfilling the obligation to eBay should have the highest priority[4]. Obviously, however, the agent's (and my) desire is to pay as little as possible, which can be achieved by ordering through El Cheapo even if a fine must also be paid to Amazon. So, according to its desires the agent prefers to order through El Cheapo, according to the severity of the penalties associated with not fulfilling its obligations it prefers to pay Amazon, but according to its norms it prefers to pay eBay. How should it decide what goals for making payment to adopt?

The core problem in this example is how to resolve conflicting obligations, given some preferences between them, in the context of various desires and norms, but the more general problem is how to resolve conflicts between arbitrary sets of desires, obligations and norms, where the agent may associate various preferences, utilities or urgencies with each of these elements, and certain overall resource constraints may apply. As in the example, norms may in fact be sufficiently abstract to guide how conflicts between other motivations should be resolved. Moreover, a solution to a concrete problem such as that facing my agent may in fact depend not just upon various motivational influences, but also upon knowledge of possible outcomes and possible future activities. For example, if the agent does not expect (me) to have future dealings at eBay or El Cheapo, the obligation towards Amazon may prevail over the desire to pay the lowest possible amount. However, if the agent regularly deals through eBay on my behalf and the cancellation fee to be paid to Amazon can be avoided, the desire to pay less and also fulfill a social norm by paying eBay might prevail over the the lowest cost option of ordering through El Cheapo.

In summary, the example reveals that an agent will often need to balance its own interests (desires) against the interests of other agents (obligations) or society (norms), and may also need to balance potentially conflicting obligations towards

---

[3] In the formalism described subsequently, these obligations are represented by the formulae: $O^{law}_{a,Am}(paid(40))$, $O^{com}_{a,eBay}(paid(35))$ and $O^{com}_{a,ElC}(paid(28))$.

[4] Represented by $N^{law}(G(\phi) \mid O^{law}_{a,Am}(\phi))$, $N^{com}(G(\phi) \mid O^{com}_{a,eBay}(\phi))$ and $N^{com}(G(\phi) \mid O^{com}_{a,ElC}(\phi))$

different parties. In more general cases it may even need, as a member of different communities, to balance conflicting norms. As we shall see, this is a complex and difficult problem in the general case, and we will propose in this paper approaches to solving only some simple special cases. Our main focus will be on providing a framework in which distinct types of motivation can be satisfactorily represented, and a general process of goal generation and maintenance based on these described.

3. Formalizing Desires and Goals

Although the BDI literature often mentions desires and sometimes considers their formal properties, little has been done to give a full formal account of this concept. Especially lacking is a formal treatment of the relationship between an agent's desires and its intentions. In this section we will give a brief formal analysis of desires and goals, contrasting them and discussing reasons why it is useful to represent them both explicitly. We assume here some familiarity with the standard axioms and possible worlds interpretation of modal logics.

The original report of Rao and Georgeff (1991) which introduced a BDI model for agents actually says nothing about desires! In the report desires are immediately translated to the more familiar and manageable concept of goals. Although in later reports and articles by these authors desires are sometimes mentioned, they are always actually used in the sense of goals. The main distinction made informally is that goals are required to be consistent, whereas desires need not be. This is also mentioned by Thomason (2000). He tries to model desires using default logics and makes an explicit link between desires and plans. However, he does not include goals in his theory anymore. Other work that explicitly deals with desires includes that of Linder (1996). In his thesis, problems (like unwanted consequences of desires) in formalizing desires (called wishes in this case) are mentioned, however, as they are largely irrelevant for his purpose, Linder just accepts them. Work by Kiss and Reichgelt (1992) takes a rather different approach which focuses on the degree of intensity associated with desires and their dynamics.

Goals are traditionally formalized in modal logics as obeying three key axioms: K, D, and necessitation. The K axiom requires closure under consequence, for goals this may be expressed as $(G(\phi) \wedge G(\phi \rightarrow \psi)) \rightarrow G(\psi)$. If $\phi \rightarrow \psi$ is a tautology (and thus $\vdash \phi \rightarrow \psi$) we can deduce with the necessitation rule that $\vdash G(\phi \rightarrow \psi)$. This implies a kind of omniscience on the part of an agent. I.e. all formulas that are implied by the current goal of the agent are also goals of the agent. Therefore the axiom is often weakened in BDI logics to $(G(\phi) \wedge B(\phi \rightarrow \psi)) \rightarrow G(\psi)$. Even this formulation is often considered unsatisfactory, since it requires an agent to have as goals the potentially undesirable side-effects of its real goals, e.g., a goal to have pain if it believes that pain is an inevitable consequence of its goal of having a tooth filled.

The D axiom, $G(\phi) \rightarrow \neg G(\neg \phi)$, captures the requirement of consistency, while the necessitation rule, $\phi \Rightarrow G(\phi)$ requires, perhaps unintuitively, that all tautologies are goals. Amongst various axiomatizations of goals the K and D axioms, at least, have been generally accepted. Goals are thus usually taken to obey a subset of the axioms of the KD45 (weak S5) system typically adopted for beliefs. Certain relationships are also usually taken to hold between goals and beliefs, e.g., that goals should be believed to be logically and practically possible.

There are, however, a number of problems in formalizing desires using a modal logic operator. As noted by Shoham and Cousins (1994), desires are motivationally weaker and less constrained than beliefs, goals and intentions. It is accepted that desires may often be inconsistent, hence the D axiom should not apply. It might seem that dropping the D axiom will let any logic collapse since from inconsistent desires it follows that everything is desired. However, these consequences might be prevented by introducing different D operators that represent desires in e.g. a specific context. Desires are allowed to be inconsistent between but not within contexts. These solutions are very similar to those proposed for formalizing inconsistent beliefs. (see (Meyer & van der Hoek, 1995) for a good overview). Another approach would be to abandon the traditional implication and instead use default logics (as was done in (Thomason, 2000)).

But if we adopt the K axiom the following deduction can be made:

$$D(\phi)\wedge D(\psi)\wedge D(\phi\rightarrow(\psi\rightarrow(\phi\wedge\psi))) \Rightarrow D(\psi)\wedge D(\psi\rightarrow(\phi\wedge\psi)) \Rightarrow D(\phi\wedge\psi)$$

And thus $(D(\phi)\wedge D(\psi)) \Rightarrow D(\phi\wedge\psi)$, which is arguably problematic as desires cannot, at face value, necessarily be combined. For example, although I may have the two individual desires:

$$D(\text{spend time with family}) \text{ and } D(\text{finish this paper})$$

I probably do not have the desire:

$$D(\text{spend time with family}\wedge\text{finish this paper})$$

in the sense of doing both concurrently. This suggests that the K-axiom should perhaps be given up if desires are modeled as a modality in a modal logic. Note that we do not assume the two desires to be inconsistent. That is, it is in principle possible to finish this paper while spending time with the family. However, this activity is not likely to be a desirable one for anyone involved.

The problem here seems to lie with the representation of the goals as activities and the interpretation of their conjunction as implying concurrent activity, and may be resolved if desires are instead represented by states rather than activities.

For example, if my desires are represented as:

$$D(\text{to have spent time with family}) \text{ and } D(\text{to have finished this paper})$$

it seems reasonable that I also desire:

$$D(\text{to have spent time with family}\wedge\text{to have finished this paper})$$

i.e., that I desire to be in a state where both have been done, rather than desiring to do both together. The apparent problem with the K axiom can thus be resolved.

Such a solution, however, is not necessarily straightforward to encode, and the interpretation of desired combined states must be done rather carefully. For example, the standard approach of using a *done()* operator to transform activities into states,

e.g. D(*done*(spend time with family)), is normally interpreted to mean that the activity has just been done, and would thus seem to require in the case of a conjunction that both activities had just been done, implying some degree of concurrency. At the opposite extreme, an interpretation that regards D(*done*(activity)) as satisfied if the activity occurred at some time in the past seems to have the unsatisfactory consequence that when satisfied once, such a desire remains satisfied forever. What is needed is an interpretation of the desire modality that requires future activity in order to be satisfied but allows conjunctive desires to be satisfied when individual conjuncts are done sequentially rather than concurrently. To do this precisely may require explicit reference to time points or time intervals in the representation and interpretation of desires.

The necessitation rule normally holds for all modalities which have a Kripke semantics. Although not very intuitive, it does not hinder the theory much, and it is unavoidable if a possible world semantics is chosen for the modal logic. Giving up on non-normal logics by exploiting Rantala's "impossible worlds" approach has been attempted for the semantics of intentions, with limited success (Cavedon, Padgham, Rao & Sonenberg, 1995).

It is clear that a modality that can be used to model desires will be very weak. With only the K and necessitation axioms being valid for desires, one might wonder whether they should be modeled with a modal operator at all. One alternative is to represent an agent's desires by an explicit set of formulae, but this approach potentially introduces a problem of substitutability. For example, if $\phi \equiv \psi$ one would expect that $D(\phi) \rightarrow D(\psi)$, and consequentially that properties such as $D(\phi \wedge \psi) \leftrightarrow D(\psi \wedge \phi)$ would be valid. Such requirements may however be achieved by constraints upon the set of formulae adopted to represent desires.

Finally, we briefly consider the relationship between desires and beliefs, and desires and actions to be performed, which we assume to be captured by intentions. It seems that in this respect also there are only very weak links. One can certainly desire a situation which one believes impossible, or not desire a situation one believes inevitable, so there are no straightforward relationships between beliefs and desires other than ones based on introspection, such as $D(\phi) \rightarrow B(D(\phi))$, etc. Similarly one can intend to perform an action that is not desired at all, e.g., one might intend to finish writing a paper during the weekend while one desires not to work during the weekend at all. (However, one likely does not intend to do something if there is no motivation for it all.) On the other hand one might desire something while never intending to achieve it, such as to insult one's boss or the death of one's enemy. So, neither do desires always lead to intentions nor do intentions always stem from desires. The relation between intentions and desires seems to be more subtle than can be expressed in a general logical axiom such as $D(\phi) \rightarrow I(\phi)$ or $I(\phi) \rightarrow D(\phi)$. Relationships between goals and intentions of a similar form to these last two axioms have been discussed under the rubric of realism and weak realism by various authors (Cohen & Levesque, 1990; Rao & Georgeff, 1995; Wooldridge, 2000). In these works some consistent properties of goals with respect to intentions have been identified, but it appears that similar observations cannot be made for desires.

These various differences between the properties of desires and goals are one reason for explicitly distinguishing them within a logical framework. More importantly, capturing both within an extended BDI model provides a basis for

describing in detail a mechanism by which goals arise from desires, and also distinguishing between an agent's primary (internal) motivations, captured by desires, and those motivations which it has somehow decided "to go along with" or those which are logically forced upon it (by the K axiom or necessitation), captured by goals. In a situation where external motivations such as norms and obligations are represented, the ability to make such distinctions becomes more important. Motivations can be distinguished from the outcomes of a process which selects and maintains goals, and it can be made explicit how different sources of motivation are balanced and conflicts resolved. In the next section we will propose an extended agent model which allows such distinctions to be captured and such a process described.

4. B-DOING agents

As discussed above, an agent typically has intrinsic desires that it tries to satisfy by achieving some concrete goals, but these desires must be balanced with the norms of the society in which the agent operates and with the obligations it may have to other members of the society. Norms can be seen as the desires of the society, and obligations as a mechanism to balance the desires of individual agents. So, an agent in a MAS should not only take into account its own desires, but also the desires of other individual agents and the desires of the agent society as a whole. Desires, norms and obligations are motivational sources for the goals of the agent[5].

The influence of these sources might differ for different agents and also for different situations. Depending on (the beliefs of the agent about) the present situation an agent has to decide which goals to achieve next or whether it should give up or postpone a current goal for a new one. The way an agent weighs the different motivations can differ from one agent to another (e.g. selfish agents will hardly consider obligations and norms, while altruistic agents hardly look at their own desires). The weighing can also be influenced by the situation (e.g. an agent with very limited resources may mainly consider its own desires, because otherwise it may never get around to achieving them). Similar considerations apply to the formation of intentions to achieve selected goals, however mechanisms for doing this in a context sensitive way are rather well explored.

Although norms and obligations are important concepts to include in the theory of agents that operate in a multi-agent system, their inclusion also has the indirect effect of highlighting the operational differences between goals and desires. The process sketched above can be depicted as in Figure 1.

Here, we model the effect of motivational influences by a two stage process. First, a deliberation and goal maintenance process is responsible for balancing desires, obligations and norms, and determining which of the goals that might arise from these are sufficiently realistic, valuable and urgent, given its beliefs, to warrant adoption as goals. When new goals are adopted, goal consistency must be maintained, but this can of course also be done by dropping existing goals. In any case, existing goals may need to be dropped when an agent's belief's and motivations change. Operationally, goals may in fact be adopted without regard to

[5]For agents built around these notions: beliefs, desires, obligations, intentions, norms and goals, we affectionately adopt the term *B-DOING agents*.

their feasibility, only to immediately be abandoned if no means can be found to achieve them.

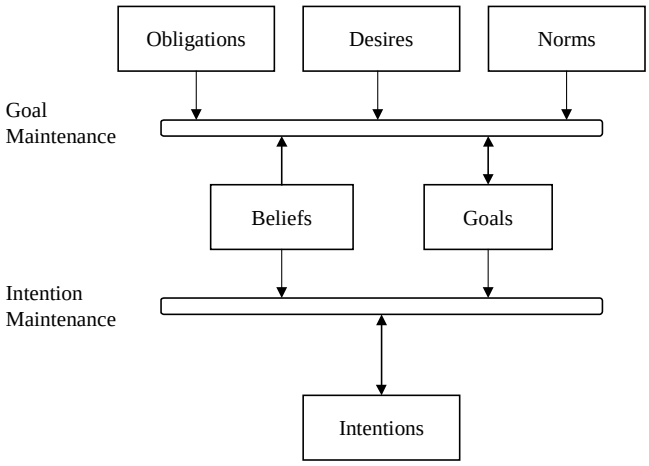

Figure 1: Decision processes of a B-DOING agent

In the second stage, a means-end reasoning and intention maintenance process is responsible for determining which of an agent's feasible goals it should commit to by forming new intentions, and how its existing intentions might need to be modified, due to changed beliefs and goals. Although commitments in this context are often seen as a commitment by an agent to a goal (Cohen & Levesque, 1990), if they are seen as a measure of an agent's resistance to changing its attitudes, they may also apply to beliefs, and to the intentions of an agent, i.e. the chosen means to a goal (Bratman, 1987; Kinny & Georgeff 1991). Committing directly to goals means that current goals may influence the choice of new goals in ways beyond the requirements of maintaining consistency. Hence the bidirectional arrow between the Goals and Goal Maintenance process. If such commitment is strong then the chance that an agent will change its goals based on new desires, obligations and norms is small. A similar argument can be made to show that commitment to current intentions influences which new intentions will be adopted.

An aspect of the goal maintenance process not captured in the figure are the many possible direct relationships between desires, obligations and norms, which are depicted as independent influences. For instance, as in the example, it might be a norm in a society to fulfill one's obligations, or one might have desires relating to individual obligations or norms. Certain general relationships between these motivations may also apply or be reasonably assumed in specific domains. We leave exploration of these issues as a subject for future work.

4.1 B-DOING logic

Having argued concepts of beliefs, desires, goals, intentions, norms and obligations are important to model agents in a multi-agent system, in this section we will offer some ideas as to how these concepts should be defined in a logical framework. A sketch of a possible formal logical description has been given in previous work (Dignum, et.al. 2000), but the resulting logic is very complicated. Here we will

just give pointers to the formalization of beliefs, goals and intentions, which have been discussed extensively elsewhere already, and discuss a possible formalization of norms, obligations and desires. Note that this is only one possible way to do so, and for the argument of this paper it does not really matter which logical theory is used as long as the requirements that we define for the concepts are met.

Beliefs, goals and intentions

The formalisation of these concepts is quite conventional. A belief ϕ of an agent i is represented as $B_i(\phi)$. For the beliefs of agents we will use a KD45 axiomatization relative to each agent. This induces a standard relation in the Kripke semantics which we will not further expand upon.

A goal ϕ of an agent i is represented as $G_i(\phi)$. Although goals have been defined in the literature in terms of other primitives (e.g. (Cohen & Levesque, 1990; van Linder 1996)) we define them here as a primitive notion and indicate some relations with desires, norms and obligations below. We assume a (minimal) KD axiomatization to ensure goal consistency. The semantics of $G_i(\phi)$ is defined in the standard way in the Kripke model.

Finally, the intentions of an agent are represented in a manner similar to its goals. $I_i(\phi)$ stands for the intention of an agent i to achieve ϕ. Just like goals, intentions have a KD axiomatization and a straightforward semantic representation.

Desires

A desire ϕ of an agent i is represented in the logic as $D_i(\phi)$. Although desires are rather opaque and very few useful axioms hold, their semantics may be captured by a primitive modality with a Kripke semantics represented by a relation between possible worlds, provided one employs a suitable interpretation for compound formulae ϕ to avoid problems with the K axiom, as discussed in Section 2. We will not be explicit on the details of how that is done, but will take the modality to obey the axioms K and necessitation.

Related to desires there are also preference orderings $>_{D(i,w)}$ on possible worlds indicating their relative desirability. This preference ordering is relative to the present state w of each agent i, i.e. it might change depending on the state of the agent. We require the following constraint at least to hold between each agent's desires and its preference ordering:

$\forall i,w,w',w''$: **if**
$\forall \phi$ [M,$w \models D_i(\phi)$ **and** M,$w' \models \phi$ **then** M,$w'' \models \phi$] **and**
$\exists \phi$ [M,$w \models D_i(\phi)$ **and** M,$w'' \models \phi$ **and** M,$w' \not\models \phi$] **then**
$w'' >_{D(i,w)} w'$

which states that worlds that satisfy more simultaneous desires must be more preferable. More specific constraints on the preference ordering are all domain dependent. One could extend this to some kind of metric that determines how "close" a state comes to satisfying a desire.

If all desires were consistent it would be possible to determine a situation in which all of them would be fulfilled. This situation could then be adopted as a goal, or if not realizable from the current situation, one as close to it as possible could be chosen. As desires do not have to be consistent this "ideal" situation may not exist, and so the preference ordering may not necessarily have a top element.

Norms and obligations

The formal semantics of obligations (and norms) is based on Prohairetic Deontic Logic (PDL) (Torre & Tan, 1999). We have extended this logic to directed obligations (including the agents as subscripts in the operator) and also included norms in a similar way as the obligations. Of course in the total logic that describes our agents this is only one part. PDL is a logic of dyadic obligation defined axiomatically in terms of a monadic modal preference logic. Only dyadic obligation is defined, i.e., all obligations are conditional, $O(p|q)$, however unconditional obligation can be represented using a tautology for the condition, $O(p)=O(p|q\vee\neg q)$. PDL allows the representation of *contrary-to-duty* obligations (obligations that hold in in sub-ideal circumstances) without contradiction, yet true deontic conflicts (conflicting obligations) imply inconsistency.

We extend PDL to allow for *multiple modalities* to denote norms and obligations from different sources. Norms for different groups or societies are distinguished, as are obligations directed to different individuals or within different organisational contexts. These modalities are represented:

$N^x(p|q)$ denotes "It is a norm of the society or organisation *x* that *p* should hold when *q* holds".

$O^x_{a,b}(p|q)$ denotes "If *q* holds, agent *a* is obliged to *b* that *p* should hold". *x* is the society or organisation that is responsible for enforcing the penalty.

We take the view that obligations from the same source must be consistent, but it is allowable for obligations from two *different* sources to conflict. For example, one can't simultaneously have two obligations to Bill: one to achieve ϕ and the other to achieve $\neg\phi$. However, one can have an obligation to Bill to achieve ϕ and an obligation to Chris to achieve $\neg\phi$.

The semantics of each modality is based on a preference ordering over worlds, unique to the modality, and an equivalence relation, *Pos*, common to all modalities, that is used to interpret "possibility". The preference ordering of norms is based on the social benefit of situations, while the preference ordering of obligations is based on the punishments for their violation. For each society *x*, each state *w* has a social worth $SW(w,x)$ which defines the preference ordering for the operator N^x. In the same way, for each state *w* there is a value of that world for an individual *a*, with respect to its relation to individual *b* and society *x*: $PW(w,a,b,x)$. This value can be seen as the "cost" of the punishment in case *a* does not fulfill its obligation towards *b*, and defines the preference ordering for the operator $O^x_{a,b}$.

The actions of an individual can have an impact (benefit or cost) in at least three ways: on individuals' personal utility functions, on the utility for all members of a society and on the degree of social cohesion. It is obviously possible to include all of

these in individuals' utility functions, but it is useful to distinguish these various components. For example, participating in a school working bee has a cost to the individual, a direct benefit to the members of the society (the benefit roughly being the product of the work done per individual and the number of individuals doing the work), and a more intangible benefit of bringing the school community closer together (which would not be as significant if the members instead contributed money to the cause). An action may have no direct benefit to the individual or the other individuals in the society, but result solely in an increase in the social cohesion. A norm is thus a statement that the benefits of an action to the members of the society and to social cohesion outweigh the cost to the individual of the action.

We now follow (Torre & Tan, 1999) in describing a preference semantics of conditional norms and obligations. Refer to (Torre & Tan, 1999) for an extensive explanation of the choice of operators. We start with three sets of monadic modal operators *Nec*, Nec^N_x , and $Nec^O_{a,b,x}$. The formula *Nec*(*p*) can be read as "p is true in all possible worlds" defined in terms of the access condition, *Pos*, which is required to satisfy the minimal constraints below. The formula $Nec^N_x(p)$ can be read as "p is true in all worlds that are preferred according to the norms of society x". The formula $Nec^O_{a,b,x}$ (p) can be read as "p is true in all worlds that are preferred according to the obligations of a towards b with respect of society x". As usual $\Diamond p \equiv \neg Nec \neg p$. The operators' semantics is defined:

M,*w* $\models$ *Nec*(*p*) **iff** $\forall w' \in$ W if *Pos*(*w*,*w'*) then M,*w'* $\models p$

M,*w* $\models$ $Nec^N_x(p)$ **iff** $\forall w' \in$ W if $SW(w,x) \leq SW(w',x)$ then M,*w'* $\models p$

M,*w* $\models$ $Nec^O_{a,b,x}$ (p) **iff** $\forall w' \in$ W if $PW(w',a,b,x) \leq PW(w,a,b,x)$ then M,*w'* $\models p$

The Nec^N_x and $Nec^O_{a,b,x}$ are S4 modalities, while the *Nec* is an S5 modality. We assume that if $SW(w,x) \leq SW(w',x)$ or $PW(w',a,b,x) \leq PW(w,a,b,x)$ then *Pos*(*w*,*w'*).

From the monadic operators Nec^N_x and $Nec^O_{a,b,x}$, we define binary "betterness" relations for the norms and obligations: $p \succ^N_x q$ states that "*p* is preferred according to the norms of society *x* to *q*". More precisely, it holds in a world *w* if for all possible worlds w_1 where $p \wedge \neg q$, and w_2 where $\neg p \wedge q$, w_2 is not preferred to w_1. The relation $\succ^O_{a,b,x}$ is defined similarly.

$$p \succ^N_x q \equiv Nec((p \wedge \neg q) \rightarrow Nec^N_x \neg(q \wedge \neg p))$$

$$p \succ^O_{a,b,x} q \equiv Nec((p \wedge \neg q) \rightarrow Nec^O_{a,b,x} \neg(q \wedge \neg p))$$

Also from the monadic operators, define $Id^N_x(p|q)$ and $Id^O_{a,b,x}(p|q)$. [We use the non standard notation *Id* rather than *I* to avoid later confusion with intentions.] These state that of all the worlds that are possible from the current world, (i) in all

the maximally preferred (ideal) worlds where q holds, p also holds, and (ii) in all infinite chains of increasingly preferred worlds, p eventually holds:

$$Id^{N}_{x}(p|q) \equiv Nec(q \rightarrow \Diamond^{N}_{x}(q \wedge Nec^{N}_{x}(q \rightarrow p)))$$
$$Id^{O}_{a,b,x}(p|q) \equiv Nec(q \rightarrow \Diamond^{O}_{a,b,x}(q \wedge Nec^{O}_{a,b,x}(q \rightarrow p)))$$

Finally, define a *norm*, $N^{x}(p|q)$, or *obligation*, $O^{x}_{a,b}(p|q)$, to be that not only is $p \wedge q$ preferred to $\neg p \wedge q$ but also the preferred (or ideal) q-worlds all satisfy p.

$$N^{x}(p|q) \equiv ((p \wedge q) \succ^{N}_{x} (\neg p \wedge q)) \wedge Id^{N}_{x}(p|q)$$
$$O^{x}_{a,b}(p|q) \equiv ((p \wedge q) \succ^{O}_{a,b,x} (\neg p \wedge q)) \wedge Id^{O}_{a,b,x}(p|q)$$

5. Goal Generation and Maintenance

In the previous section we have shown that the motivational inputs for the goal generation and maintenance process all induce a preference ordering on the possible worlds and thus on the possible goals. In general we can distinguish two cases in which an agent has to make decisions about which new goals to choose. The first is characterised by the fact that an agent has either achieved or abandoned any former goals, and can, in principle, choose new goals freely among all possible alternatives, as ranked by appropriately balancing its motivations. We will suggest some simple rules for how this case may be handled below.

The second case is when an agent has existing goals, and new desires arise internally, new opportunities or obligations arise through some event in the world, or one or more current goals become less motivated. The latter can be caused by the fact that it becomes less feasible or more costly to achieve a goal, e.g., because an intention has failed, or because the goal itself has become less valuable, i.e., the agent's preference orderings have changed, due to a changed in its state. For example, a goal to earn money can become unmotivated if one wins a lottery. In this second case an agent must not only balance preferences to determine potential new goals, but possibly also either filter these candidate goals or abandon some existing goals to ensure overall goal consistency.

Naïvely, both cases may be dealt with uniformly by a goal generation process which determines new goals from prioritized motivations and state without regard to existing goals, and then merely replaces existing goals, if any, with new ones. Such an approach may incur an undesirably high computational cost, but in any case is only applicable when an agent has no commitment to its existing goals. More usually, an agent would be somewhat committed to existing goals, particularly when it has active intentions based upon them. Under these circumstances it may first decide upon preferred alternative goals, which are then compared with its current goals. If the agent has a strong commitment to its current goals the alternative has to be really much better to cause a current goal to be dropped.

In the framework of this paper we will not distinguish the two cases, but undoubtedly the distinction will be important in an implementation of a B-DOING agent. We have not discussed the role of commitment in presenting the logic because we did not want to introduce it as a separate modality. Rather, we take commitment (in this context) to be an aspect of an agent's decision processes, measured by the resistance to change of its current attitudes; one which also must determine how often the agent will reconsider its goals and intentions and how big a change between the actual situation and the planned situation will be needed to cause a change in them, as in (Kinny & Georgeff, 1991). We will not further expand upon commitment strategies here, but assume them to be a design characteristic of an agent.

5.1 Combining preferences

In this section we will expand on the combination of the different preference orderings induced by the desires, obligations and norms of an agent. As said before, these preference orderings have to be combined into one ordering on possible goals. In order to achieve this it would be nice to have a general, intuitive mechanism, but unfortunately, social choice theory (Arrow, 1963) reveals that it is not possible to find such an aggregation mechanism if it has to possess certain intuitive properties:

- *collective rationality*: the aggregate preference ordering is a function of the separate preference orderings.
- *Pareto principle*: the aggregate ordering agrees with uncontested strict preferences in the separate orderings.
- *independence of irrelevant alternatives*: only the separate preference orderings influence the aggregate ordering.
- *nondictatorship*: there is no one preference ordering that solely determines the aggregate ordering.
- *conflict resolution*: if two formulae are comparable in one of the preference orderings they also are comparable in the aggregate ordering.

The above properties seem to be all very intuitive. However, by a well known theorem due to Arrow (1963) they are not simultaneously satisfiable in general. Although the above result is given in the context of combining preferences of independent agents it is shown in (Doyle & Wellman, 1991) that this impedement exists in general for combining any two (or more) preference orderings, based on similar arguments as given by Arrow.

One way of solving this problem is by dropping one of the properties given above. The property that is often chosen is that of nondictatorship. In that case the combination rule becomes simple: we order the modalities in some way and always let the preference of the highest valued modality prevail over the other two. E.g. the agent might put obligations over norms and norms over desires. This results in the so-called dutiful agent. Or one might let the norms prevail over the other two and get a social agent. Or if the desires always prevail one gets a hedonistic agent.

Another way of "solving" the problem is by mapping all three preference orderings into a common scale. In this case one might use the more classical decision theoretic models based on utility functions (see e.g. (Boutilier, Dean & Hanks, 1999; Thomason & Horty, 1996)). An advantage in combining the different preferences by mapping them on one scale is that one can distinguish between

degrees of preference. If some situation is very desirable it will get a very high score and thus outweigh the costs of e.g. violating obligations. A disadvantage, however, is that the combination is static. One will assign some weight to each preference and this weight will be used in all situations where the preferences have to be combined. However, one would like this weight to depend on the current situation as well. E.g. the desire to gain $10 is getting less importance the more income I have. The importance of not violating an obligation towards someone becomes more important the more I know someone and like/trust him. So, we would like to have a more flexible way of combining the preferences than just mapping them all on one scale.

One way of achieving this is to use some domain independent heuristics to combine the three preference orderings related to the different motivational attitudes. Of course the rules that we give are over simplistic, but are meant as an indication of what type of rules one would like instead of the fixed hierarchy discussed above. We assume that a choice for a single goal has to be made between mutually exclusive alternatives. Each source will have zero, one or more preferred alternatives. Obligations and norms might have no preferred alternative if there are no obligations or norms applicable to any of the alternatives.

The first rule is an obvious special case:

> **Rule 1:** An agent should choose one of its most preferred desires as a candidate goal whenever there are no norms or obligations applicable to any of the alternatives, or when they are applicable, they prefer the same alternative as the desires.

The more interesting case occurs when the desires, norms and obligations prefer different alternatives. In the general case there is very little to say about how the decision should be made, but if we limit ourselves to choosing between two alternatives then the following very simple default rule can be used:

> **Rule 2:** An agent should choose as a candidate goal an alternative that is preferred by at least two of the three sources.

Although this is a very simple majority-rule heuristic one can also argue in favor of it for the specific cases. Of course when an alternative is preferred from all three viewpoints it is clear that it should be chosen. If an alternative is most desirable and also most preferred with respect to the obligations of an agent, but not with respect to the norms, then one might argue that the conflict between the norms and the obligations be resolved in favour of the obligations. For it seems that the agent has committed to some obligations that cannot be fulfilled in a normative way. Because the agent has chosen to incur the obligations, it apparently regards the obligations as more important than the norms. Because they are still in line with its own desires it should choose for that alternative again.

If an alternative is most desirable and also most preferred with respect to the norms, but not with respect to the agent's obligations, then one might argue that the desires of the agent are in line with that of the whole society. It will not fulfil the obligations in the preferred way, however, this will affect primarily the beneficiary of the obligation. Apparently the agent values the norms of the society higher at this

moment, because they are in line with its own desires. This argument would lead the agent to pay eBay in the example of section 2.

If an alternative is most preferred with respect to both obligations and norms, but not the most desirable, then the agent should comply to the preferences of the society. Unless the agent does not depend on agents from the society to fulfil its goals, just following its own desires in this case would incur sanctions from the other agents and might thus hamper achieving its own goals in the future.

If an agent has a choice from three or more alternatives, as was the case in our example, we can still use the above rule if one of the alternatives is the preferred one according to at least two of the three sources. The rule does not work, however, if all sources prefer a different alternative (as was the case in our example). In that case one could either design the agent in a simple way such that it would order the alternatives in a fixed order and, for instance, always choose an alternative preferred by its obligations first, one preferred by its norms second and one preferred by its own desires third. Using these rules means that an agent does not have to do any sophisticated reasoning, but also that it will not distinguish between the consequences of different situations.

5.2 Reasoning about preferences and consequences

A more complex approach is to decide based on reasoning about the consequences of choosing each alternative. Returning to the example, if we represent the agent's obligations by $O^{law}_{a,Am}(paid(40))$, $O^{com}_{a,eBay}(paid(35))$ and $O^{com}_{a,ElC}(paid(28))$, and the norms which apply (instances of more general rules) by $N^{law}(G(\phi)|O^{law}_{a,Am}(\phi))$, $N^{com}(G(\phi)|O^{com}_{a,eBay}(\phi)))$ and $N^{com}(G(\phi)|O^{com}_{a,ElC}(\phi)))$, such a decision may occur as:

$$O^{law}_{a,Am}(paid(40)) \wedge O^{com}_{a,eBay}(paid(35)) \wedge O^{com}_{a,ElC}(paid(28)) \wedge$$
$$D_a(paid(minimum)) \wedge N^{com}(G(\phi)|O^{com}_{a,b}(\phi))) \wedge$$
$$\neg paid(Am,40) \wedge \neg paid(eBay,35) \wedge paid(ElC,28) \rightarrow$$
$$do(Am,PayFine) \wedge BadReputation(eBay) \wedge paid(minimum)$$

Note that this is a simplified presentation, because we did not include all the consequences of not fulfilling the norms of fulfilling obligations (this would require some more notation about achieving goals which we do not want to add at here). However, in the same vein as the rule above the agent could reason about the consequence of each choice. Each choice gives rise to a certain state in which all the consequences of the choice are true. By comparing the "scores" of the states for each choice the agent can determine the most preferable situation.

In fact my agent was able to reason in a more sophisticated way about possible outcomes: it first successfully negotiated that the Amazon fine be waived, then offered to pay eBay, only to discover that the goods in question had been misdescribed and hence the bid could be cancelled, so it finally ordered via El Cheapo and so fulfilled all its norms, obligations and desires.

6. Conclusions

Our long-term goal is to extend the BDI model, at a theoretical and practical level, with social concepts such as norms and obligations. Towards this aim, in this paper we have attempted to do two things: (i) by critically reviewing the logical properties of goals and desires, and their relationships with the other basic modalities in the BDI model, we have given an account of situations in which both goals and desires should be represented; (ii) further, we have argued that to represent complex external influences on an individual agent's behaviour, norms and obligations have a role to play, and to capture the interaction between these and the internal motivations for goals, explicitly representing desires as well as goals becomes important. In support of this analysis we have proposed a way in which norms and obligations might be added to the BDI model, while making the point that the particular logic referred to is just one way of accomplishing this. The main point in this logic is the fact that desires, obligations and norms can all be based on preference orderings, which can be used to direct the choice of goals, and hence future actions.

Clearly a great deal of further work is required before the framework presented here can be regarded as comprehensive. One major point will be to combine the different parts of the logical framework described in this paper. Especially of interest is the combination and nesting of the different modal operators. We see the contribution being in the conceptual analysis, supported by the example formalism, rather than resting on the details of the formalism itself.

Acknowledgements

We thank the anonymous reviewers for their useful comments. This work was partially funded through a grant from the Australian Research Council.

7. References

Arrow, K.J.(1963). *Social Choice and Individual Values*. Yale University Press.

Boutilier, C., Dean, T. and Hanks, S. (1999). Decision-Theoretic Planning: Structural Assumptions and Computational Leverage. *Journal of Artificial Intelligence Research*, 11, 1-94.

Bratman, M.E. (1987). *Intentions, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. E., Israel, D. J. and Pollack, M. E. (1988). Plans and Resource-Bounded Practical Reasoning. *Computational Intelligence*, 4(4), 349–355.

Castelfranchi, C.(1998). Modeling Social Action for AI Agents. *Artificial Intelligence*, 103, 157–182.

Castelfranchi, C., Dignum, F. , Jonker, C. and Treur, J. (2000). Deliberate Normative Agents: Principles and Architectures. In N. Jennings and Y. Lesperance (Eds.) *Intelligent Agents VI (LNAI-1757*) (pp. 364-378) , Springer-Verlag.

Cavedon, L., Padgham, L., Rao, A. and Sonenberg, E. (1995). Revisiting rationality for agents with intentions. In Proceedings of the Eighth Australian Joint Conference on AI, *Bridging the Gap*, (pp. 131–138), World Scientific Publishers, Australia.

Cavedon, L. and Sonenberg, L. (1998). On social commitments, roles and preferred goals. In *Proceedings of ICMAS'98*, pp. 80–87, Paris.

Cohen, P. and Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42, 213–261.

Dennet, D. C. (1987). *The Intentional Stance*. MIT Press.

Dignum, F., Morley, D., Sonenberg, L. and Cavedon, L. (2000). Towards socially sophisticated BDI agents. In *Proceedings of the Fourth International Conference on MultiAgent Systems*, (pp. 111–118), Boston.

Doyle, J. and Wellman, M. (1991). Impediments to Universal Preference-Based Default Theories. *Artifical Intelligence*, 49(1-3), 97–128.

Hogg, L. and Jennings, N. (2000). Variable Socialability in Agent-based decision making. In N. Jennings and Y. Lesperance (Eds.) *Intelligent Agents VI (LNAI-1757*) (pp. 305-318) , Springer-Verlag.

Jennings, N. (1993). Commitments and Conventions: the foundation of coordination in Multi-Agent systems. *Knowledge Engineering Review*, 8(3), 223–250.

Jennings, N. and Campos, J. (1997). Towards a Social Level Characterisation of Socially Responsible Agents. *IEEE Proc. on Software Engineering*, 144(1), pp. 11–25.

Kinny, D. and Georgeff, M. (1991). Commitment and Effectiveness of Situated Agents. In *Proceedings of IJCAI'91*, (pp. 82–88), Sydney.

Kinny, D., Ljüngberg, M., Rao, A., Sonenberg, E., Tidhar, G. and Werner, E. (1994). Planned Team Activity. In *Artificial Social Systems* (*LNCS 830)*, (pp. 227–256), Springer-Verlag.

Kiss, G. and Reichgelt, H. (1992). Towards a Semantics of Desires. In *Decentralized AI 3 – Proceedings of MAAMAW'92*.

Linder van, B. (1996). *Modal Logics for Rational Agents*, PhD thesis, Utrecht University.

J.-J.Ch. Meyer and W. van der Hoek (1995), *Epistemic logic for AI and computer science*, Cambridge: Cambridge University Press.

Rao, A.S. and Georgeff, M.P. (1991). Modeling rational agents within a BDI architecture. In *Proceedings of the KR'91*, (pp. 473-484), Morgan Kaufman.

Rao, A.S. and Georgeff, M.P. (1995). BDI Agents: From Theory to Practice. In *Proceedings of ICMAS 95*, San Francisco.

Shoham, Y. and Cousins, S. B. (1994). Logics of Mental Attitudes in AI. In *Foundations of Knowledge Representation and Reasoning* (*LNAI 810)*, Springer Verlag.

Shoham, Y. and Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73, 231–252.

Singh, M. (1996). Multiagent Systems as Spheres of Commitment. In *ICMAS Workshop on Norms, Obligations, and Conventions*, Kyoto, Japan.

Thomason, R. (2000). Desires and Defaults: A Framework for Planning with Inferred Goals, In A. G. Cohn, F. Giunchiglia and B. Selman (eds), *KR2000*, (pp. 702-713), Morgan Kaufmann.

Thomason, R. and Horty, J. (1996). Nondeterministic Action and Dominance: Foundations for Planning and Qualitative Decision. In Y. Shoham, (ed) *Proc. of the 6th Conf. on Th. Aspects of Reasoning about Knowledge*, Morgan Kaufmann.

Torre, van der L. and Tan, Y.-H. (1999). Contrary-To-Duty Reasoning with Preference-based Dyadic Obligations. *Annals of Mathematics and AI*, 27, 49-78.
Wooldridge, M. (2000). *Reasoning about Rational Agents*. MIT Press.