Motivational Attitudes of Agents: On Desires, Obligations and Norms

Frank Dignum* David Kinny† Liz Sonenberg†

* Institute of Information and Computing Science, Utrecht University
3508 TB Utrecht, The Netherlands
dignum@cs.uu.nl

†Department of Information Systems, The University of Melbourne
Victoria 3010, Australia
{dnk,LizS}@staff.dis.unimelb.edu.au

1 Introduction

Many papers about agents mention BDI as a reference logical model for agents, but this model does not offer a thorough formal account of the connections between the different modalities of Beliefs, Desires and Intentions. Of course, work such as that of Rao and Georgeff [14] and of Cohen and Levesque [6] has pointed to some specific constraints, but does not offer a complete logical theory that explains all possible connections between, e.g., goals and intentions. Another point of concern often voiced is the long-standing gap between the BDI logical model and practical agent implementations. Judged by its applicability, it might seem that the BDI model is becoming less important, due to the looseness of its connection to practical systems, and because of its failure to guide research into 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 intuitions that could guide further development of the agent paradigm, and extensions of BDI agents with concepts like emotions, power relations, obligations, etc. would be *ad hoc*, lacking an adequate theoretical basis. There would be no way of determining how these extensions might relate with the basic components of BDI agents if the link between BDI logic and the implementation is not clear. The implementation will not impose any inherent restrictions or guidelines on these relations. The BDI logic might give some insights, but if there is no relation between the logic and the implementation, these insights are of dubious value.

Our long-term goal is to extend the BDI model with social concepts such as norms and obligations. Elsewhere we and others have argued that such concepts are an important tool to "glue" autonomous agents together in a multi-agent system [2, 3, 5, 7, 17]. In this paper we do two things towards this goal. After presenting a motivating example, we 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 2). We then look at ways in which norms and obligations might be added to this model and make some general comments about such formalisations (Sections 3 and 4). In particular, we propose that for goal generation, interactions between desires, obligations and norms can resolved by preference orderings, and discuss how an agent might form intentions given certain desires, goals, norms and obligations.

2 Sources of Motivation

Consider the following situation. I want my software agent to buy me a certain book as cheaply as possible. 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 dollars. At eBay the price is still going up, and it has decided to bid $35 there, but 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. Then I inform the agent that I need the book by next Monday, and it immediately places an order at Amazon to ensure that I have it in time. At that moment the results come in from Ebay: it transpires that its bid of $35 at eBay has won the auction after all; a moment later, before my agent can cancel the El Cheapo request, El Cheapo's 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 now has 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. But I don't need three copies, and a quick check of my budget reveals that I can't afford to spend more than $50 anyway, so the agent now must somehow decide which of the three suppliers to pay.

Cancelling the Amazon order will probably require paying a $10 cancellation fee, but there's a chance that they may waive the fee. Not paying at eBay will result in damage to my 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 these outcomes, fulfilling the obligation to Amazon seems the preferred solution for the agent. However, it also has some norms and desires involved in the decision. The norms in this case are that all commitments should be fulfilled whenever possible, but if 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 has the highest priority. Obviously, however, the agent's (and my) desire is to pay as little as possible for the book, 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 consequences of not fulfilling obligations it prefers to pay Amazon, but according to its norms it prefers to pay eBay. How should it decide what goals for payment to adopt?

We will discuss some rules for making such a decision in section 4, however this 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 also need to balance potentially conflicting obligations towards different parties. In more general cases it may even need, as a member of different communities, to balance conflicting norms.

In the BDI approach, the behaviour of an individual agent is shaped by the agent's state of knowledge of the world (its *beliefs*), the states of the world it seeks to bring about (sometimes termed *desires*, sometimes *goals*), and the execution of stored *plans* that have been designed to bring about certain world states in pre-specified circumstances. The plans selected by the agent for execution are referred to as its *intentions*, of which at any time there may be many. Our focus here will be on exploring how an agent's goals should arise from the various sources of motivation to which it may be subject.

From the literature [13, 14] it is clear that desires are a very abstract concept with few general logical properties, and that they are only loosely coupled with beliefs and intentions. Taking desires to be identical to goals, as has often been done in the past, makes them more tangible and practically usable. We suggest, however, that the main reason to consider desires as a separate interesting concept arises when they are not the sole source of motivation, i.e., when obligations or norms are also present, and goals must be determined by balancing desires against such external sources of motivation.

Norms, as manifest in human societies and organizations, assist in standardising the behaviour of individuals. Commitment to membership of a society means that an agent will tend to follow its norms, i.e., an agent will weigh up goals suggested by other sources against those suggested by societal norms. Similar considerations apply to obligations, which are associated with specific enforcement strategies that involve punishment of violators. Obligations are explicit mechanisms for influencing the behaviour of agents and providing some stability and reliability in their interactions while allowing some flexibility. They provide a level of "freedom of choice" with known explicit consequences.

The main difference between norms and obligations as we use them is that norms are more abstract and are inherent to the society of which the agent is a member. Obligations are (usually) a consequence of a direct action of the agent itself and entered by choice (e.g. the obligation to pay when something is ordered). The distinctions that we make in this paper are thus relatively crude; we are however including at least these two concepts as representatives of a whole spectrum of external motivational influence.

3 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 which functions in a multi-agent system 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 three motivational sources for the goals of the agent[1]. The influence of these sources might differ for different agents and also for different situations.

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 differences between goals and desires. The situation sketched above can be depicted as in Figure 1. Here, we model the effect of motivational influences by a two stage process. First, a goal maintenance process is responsible for determining which of the goals that might arise from an agent's obligations, desires and norms are sufficiently realistic, valuable and urgent, given its beliefs, to warrant adoption as goals.

In the second stage, an 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.

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

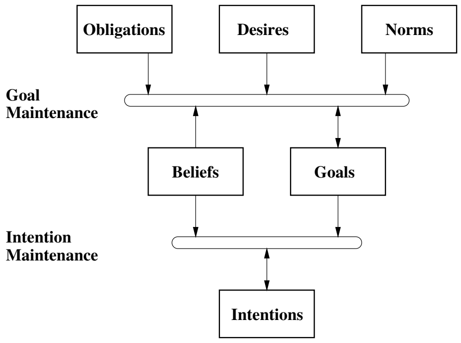

Fig. 1. Decision processes of a B-DOING agent

Commitment to goals means that current goals also influence the choice of new goals. If commitment to goals is high then the chance that an agent will change its goals based on new desires, obligations and norms is small.

An aspect of the goal maintenance process not captured in the figure are the direct relationships between desires, obligations and norms, which are depicted as independent influences. However, there are many possible relationships between these concepts. For instance, it might be a norm in a society to fulfill one's obligations.

B-DOING logic

Having argued that the concepts of belief, desire, goal, intention, norm and obligation 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 [7], but the resulting logic is very complicated. Here we will just give pointers to the formalization of goals, beliefs and intentions, which have been discussed 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.

Goals, beliefs 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 (see e.g. [6, 13]) we will define them as a primitive notion here and indicate some relations with desires, norms and obligations later on. We assume a (minimal) KD axiomatization to ensure the consistency of the goals. 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 $D_i(\phi)$. As desires are almost completely opaque and the K-axiom does not hold, their semantics cannot easily be represented by a Kripke semantics where desires are represented by a relation between possible worlds. We will avoid the problems with opaqueness by defining desires to be an explicit (infinite) set of formulas, denoted by $DES_i(s)$, that are desired by each agent i in each state s. This set of formulas should comply with the following constraint:

$$\forall \phi, \psi, s, i : \phi \in DES_i(s) \wedge \Box(\phi \rightarrow \psi) \text{ then } \psi \in DES_i(s)$$

which states that the set of desires is closed under implication by necessity. Although the K-axion does not hold for desires in general, making the set of desires closed under implication by necessity at least avoids the problem of distinguishing desires that only differ syntactically but are logically equivalent. E.g. if $\phi \wedge \psi \in DES_i(s)$ then also $\psi \wedge \phi \in DES_i(s)$. The semantics of $D_i(\phi)$ is now given as follows:

$$M, s \models D_i(\phi) \textbf{ iff } \phi \in DES_i(s)$$

Related to the set of desires there is a preference ordering on the set of possible worlds indicating their relative desirability. This preference ordering is relative to the present state of the agent, i.e. it might change depending on the state of the agent.

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) [18]. 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 societies are distinguished, as are obligations to different individuals or within different organisational contexts. 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 p and the other to achieve $\neg p$. However, one can have an obligation to Bill to achieve p and an obligation to Chris to achieve $\neg p$:

- $N^z(p|q)$ – it is a norm of the society or organisation z that p should be true when q is true.
- $O^z_{ab}(p|q)$ – when q is true, individual a is obliged to b that p should be true. z is the society or organisation that is responsible for enforcing the penalty.

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 fulfil its obligation towards b and defines the preference ordering for the operator O^x_{ab}.

We now follow [18] for describing a preference semantics of the conditional norms and obligations. Refer to [18] for an extensive explanation of the choice of operators.

Start with three sets of monadic modal operators $\overleftrightarrow{\Box}$, $\Box^N_x$, and $\Box^O_{a,b,x}$. The formula $\overleftrightarrow{\Box}p$ 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 $\Box^N_x p$ can be read as "p is true in all worlds that are preferred according to the norms of society x". The formula $\Box^O_{a,b,x}p$ 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 $\Diamond p \equiv \neg\Box\neg p$.

$M, w \models \overleftrightarrow{\Box}p$ iff $\forall w' \in W$ if $POS(w, w')$ then $M, w' \models p$
$M, w \models \Box^N_x p$ iff $\forall w' \in W$ if $SW(w, x) \leq SW(w', x)$ then $M, w' \models p$
$M, w \models \Box^O_{a,b,x}p$ iff $\forall w' \in W$ if $PW(w', a, b, x) \leq PW(w, a, b, x)$ then $M, w' \models p$

The $\Box^N_x$ and $\Box^O_{a,b,x}$ are S4 modalities, while the $\overleftrightarrow{\Box}$ is an S5 modality. Assume that if $SW(w, x) \leq SW(w', x)$ or $PW(w', a, b, x) \leq PW(w, a, b, x)$ then also $POS(w, w')$.

From the monadic operators $\Box^N_x$ and $\Box^O_{a,b,x}$, 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. Introduce $\succ^O_{a,b,x}$ similarly.

$$p \succ^N_x q \;\equiv\; \overleftrightarrow{\Box}((p \wedge \neg q) \rightarrow \Box^N_x \neg(q \wedge \neg p))$$
$$p \succ^O_{a,b,x} q \;\equiv\; \overleftrightarrow{\Box}((p \wedge \neg q) \rightarrow \Box^O_{a,b,x} \neg(q \wedge \neg p))$$

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:

$$Id^N_x(p|q) \;\equiv\; \overleftrightarrow{\Box}(q \rightarrow \Diamond^N_x(q \wedge \Box^N_x(q \rightarrow p)))$$
$$Id^O_{a,b,x}(p|q) \;\equiv\; \overleftrightarrow{\Box}(q \rightarrow \Diamond^O_{a,b,x}(q \wedge \Box^O_{a,b,x}(q \rightarrow p)))$$

Finally, define a *norm*, $N^x(p|q)$, or *obligation*, $O^x_{ab}(p|q)$, 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.

$$N^x(p|q) \;\equiv\; ((p \wedge q) \succ^N_x (\neg p \wedge q)) \wedge Id^N_x(p|q)$$
$$O^x_{ab}(p|q) \;\equiv\; ((p \wedge q) \succ^O_{a,b,x} (\neg p \wedge q)) \wedge Id^O_{a,b,x}(p|q)$$

4 Decision making and priorities

In the previous section we have shown that the motivational inputs for goal generation all induce a preference ordering on the possible worlds and thus on the possible goals. In general we can distinguish two situations in which the agent has to make a decision about which new goals to choose. The first situation is characterised by the fact that the agent has either achieved or dropped its former goals, and can, in principle, choose new goals freely among all possible alternatives. The second situation arises when an agent is trying to achieve some goals, but either a new opportunity or obligation arises (through some event in the world) or its goals become less "interesting". For example, the goal to work in order to earn money can become uninteresting if one wins a million dollars in a lottery. In this second situation the agent first decides upon the best alternative goal, which is 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 this will be important in an implementation of a B-DOING agent.

We have not discussed the role of commitment in the previous sections, because we did not want to introduce it as a separate modality in the logic. 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 can determine how often the agent will reconsider its goals and intentions and how big a change between the actual situation (considering both mental as well as physical parameters) and the planned situation will be needed to cause a change, as in [12]. Due to space limits, we will not further expand upon this concept here, but assume it as a design characteristic of an agent.

Combining preferences

In this section we will expand on the combination of the different preference orderings induced by the obligations, norms and desires of the 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. Unfortunately, social choice theory [1, 8] points out that it is not possible to find such an aggregation mechanism if it has to possess a number of intuitive properties. These properties are:

- *collective rationality*. The aggregate preference ordering is a function of the separate preference orderings.
- *Pareto principle*. I.e. the ordering of the goals agrees with uncontested strict preferences.
- *independence of irrelevant alternatives*. Other formulas do not influence the final ordering.
- *nondictatorship*. There is no preference ordering that solely determines the ordering on the goals.
- *conflict resolution*. If two formulas are comparable in one of the preference orderings they also are comparable in the resulting ordering of goals.

The above properties seem to be all very intuitive. However, by a well known theorem due to Arrow [1] the above constraints are not simultaneously satisfiable.

Although the integration of preferences is not perfectly possible in the general case, we will sketch a number of general rules that might be used as heuristics in the case of combining the three preference orderings related to the different mental attitudes in a domain independent way. In principle we have three sources: desires, obligations and norms. We assume that a choice for a goal has to be made between mutual exclusive situations. 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.

From the above discussion we can easily formulate the first rule:

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

The more interesting case occurs when the desires and norms and/or 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 two possible alternatives then the following very simple (default) rule can be used:

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

Although this is a very simple numerical 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 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 (only) affects mainly 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. 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.

A more complex approach is to let the agent 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 $O^{law}_{a\ ElC}(paid(28))$, and the norms which apply 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 process may occur as:

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

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 this place). However, in the same vein as the rule above the agent could reason about the consequence of each choice and 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.

Although the above arguments can be undercut by counterexamples in each case, it seems that the rules could function well as a heuristic. However, we should emphasise that every solution inherently depends on domain specific properties, such as the fact that we combine three preference orderings and the nature of these preferences.

5 Conclusions

Our long-term goal is to extend the BDI model 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 relationships between the basic modalities in the BDI model, we have elicited 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 future actions.

Clearly a great deal of further work is required before the framework presented here can be regarded as comprehensive. 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 very useful comments. This work was supported by a grant from the Australian Research Council.

References

1. K.J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1963.
2. C. Castelfranchi. Modeling Social Action for AI Agents *AI* vol.103, pp. 157–182, 1998.
3. C. Castelfranchi, F. Dignum, C. Jonker and J. Treur. Deliberate Normative Agents: Principles and Architectures, In *Proceedings of ATAL-99*, Orlando, 1999, pp. 206–220.
4. L. Cavedon, L. Padgham, A. Rao and E. Sonenberg. Revisiting rationality for agents with intentions, In *Bridging the Gap*, pp. 131-138, World Scientific, 1995.
5. L. Cavedon and L. Sonenberg. On social commitments, roles and preferred goals, In *Proceedings of ICMAS'98*, Paris, July 1998, (ed) Y Demazeau, pp 80–87.
6. P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial Intelligence*, vol.42, pages 213–261, 1990.
7. F. Dignum, D. Morley, L. Sonenberg and L. Cavedon. Towards socially sophisticated BDI agents In Proceedings of the Fourth International Conference on MultiAgent Systems July, Boston, USA, pp 111-118
8. J. Doyle and M. Wellman. Impediments to Universal Preference-Based Default Theories, *Artifical Intelligence*, Vol.49, Nr.1-3, pages 97-128, 1991.
9. L. Hogg and N. Jennings. Variable Socialability in Agent-based decision making, In N. Jennings and Y. Lesperance (eds.) *Proceedings of ATAL-99*, Orlando, 1999, pages 276- 290.
10. N. Jennings. Commitments and Conventions: The foundation of coordination in Multi-Agent systems. *Knowledge Engineering Review*, vol. 8(3), pages 223-250, 1993.
11. N. Jennings and J. Campos. Towards a Social Level Characterisation of Socially Responsible Agents. *IEEE Proc. on Software Engineering*, vol.144, 1, pp.11-25, 1997.
12. D. Kinny and M. Georgeff. Commitment and Effectiveness of Situated Agents. In *Proceedings of IJCAI'91*, Sydney, 1991, pages 82–88.
13. B. van Linder. Modal Logics for Rational Agents, PhD thesis, 1996, University of Utrecht.
14. A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI architecture. In: R. Fikes et.al. (eds.), *Proceedings of the 2nd CKRR*, Morgan Kaufman, pp. 473-484, 1991.
15. A.S. Rao and M.P. Georgeff. BDI Agents: From Theory to Practice. *Proceedings of ICMAS 95*, San Francisco, 1995.
16. Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* 73(1995)231-252.
17. M. Singh. Multiagent Systems as Spheres of Commitment. In *ICMAS Workshop on Norms, Obligations, and Conventions*, Kyoto, Japan, December 1996.
18. L. van der Torre and Y.-H. Tan. Contrary-To-Duty Reasoning with Preference-based Dyadic Obligations. Submitted to *Annals of Mathematics and AI.*
19. M. Wooldridge. *Reasoning about Rational Agents*, MIT Press, 2000