Obligations in a BDI Agent Architecture

Gordon Beavers and Henry Hexmoor
University of Arkansas
Computer Science & Computer Engineering Department
Engineering Hall 313, Fayetteville AR 72701
{gordonb, hexmoor}@uark.edu
Phone: 501 575 2420 Fax: 501 575 5339

Abstract. *This paper presents a conceptual model that adds mental states, capable of dealing with social elements, to BDI architectures. It has long been recognized that social elements enhance multi-agent performance. Ideas for extending BDI architectures to incorporate norms, values and obligations are presented. Norms and values endow multi-agent systems with social attributes thereby facilitating the cooperative behavior of agents. The extension makes use of sets of modal operators for desires and obligations*

Keywords: Socially Aware Agents, BDI Architectures, Multiagent Systems

1. Introduction

The underlying assumption of BDI agent models is that agents in a dynamic environment autonomously determine their behavior by considering (i) their current situation (as reflected by their set of beliefs and background knowledge), (ii) various ways that the agent would want to change the current situation (the agent's set of desires), and (iii) what the agent is currently planning to do (the agent's set of intentions). An agent operating in a multi-agent system will find it advantageous to be able to reason about the mental states of other agents. Although considerable work has been undertaken in the past few years on norms, the consideration of values (principled behavior) for BDI agents has received little attention [1, 2, 3, 4]. In this paper we expand our discussion of our earlier consideration of agent architectures that incorporates norms, evaluations and principles as components of the decision making process of rational social agents [6, 7, 8]. Norms and principles generate obligations, which are modeled with modal operators. Norms and principles provide expected standards of behavior, thereby allowing other agents to reason about the probable behavior of an agent.

Effective multi-agent systems require the agents to model not only their own mental states, but also the mental states of other agents. Codified norms, values and standards for evaluation simplify the modeling of other agent's mental states. However, modeling the mental states of others also introduces the possibility for agents to exhibit duplicity since an agent may intentionally follow an inappropriate norm, or otherwise exhibit behavior that is not representative of the agent's current stance.

The structure of the paper is as follows: first suggestions are made for altering the semantics of desire, then principles and evaluations are introduced as ways of generating obligations and simplifying agent intention determination. A subsection is devoted to the consideration of norms since group intentions lead to the acceptance of norms. We then suggest how obligations can be incorporated in a BDI algorithm, and follow with a look at agent types that exemplify precedence orderings of the intentional notions.

2. Intentional Attitudes

We have added obligations to agents as a way to express the principles that guide the behavior of the agents, that is, as a way to give these artificial

agents different purposes or ultimate goals, and as a way to enforce norms. It is standard to use a KD45 axiomatization for belief and a KD axiomatization for both desire and intention. The KD axiomatization of desire requires that the set of desires be consistent which is not realistic. Similarly, one should not expect obligations to be consistent. Therefore, we deviate from standard BDI practice and stipulate that there be multiple modal operators both for desires and for obligations. We also add a modal operator for knowledge with the standard S5 axiomatization to represent unchanging information about norms, evaluations and values. In our system there is a difference between evaluations and values; evaluations will be treated as instrumental knowledge while principles will be treated as knowledge that generates obligations. Goals and principles, being motivational, are determinative of behavior regardless of the situation, while evaluations, being informational, enter into deliberative processes that determine intentions only contingently.

Values and norms will be modeled similarly, since both values and norms provide guides to behavior in the form of obligations, but serve different social purposes. Values (or principles) are general and abstract and serve to characterize the type of society (or agent) that obeys the principle. Norms tend to be particular and concrete and yield guidelines for behavior in specific situations. Obligations arise from values and norms. Values, norms and obligations can be used to make the computations that constitute agent reasoning more efficient by constraining the search space. We suggest that obligation have a KD4! axiomatization, where 4! is combination of two axioms "4" and "!" that taken together can be stated as the following:

$$Oi(b, t, \phi) \leftrightarrow Oi(b, t, Oi(b, t, \phi))$$

That is, agent b is obligated (denoted by operator "Oi" which is discussed in section 2) at situation t to see to it that ϕ obtains just in case agent b is obligated at situation t to see to it that agent b is obligated at situation t to see to it that ϕ obtains. It could be argued that an obligation to an obligation is not the same as an obligation and thus the two obligations should have different subscripts. This results in a needlessly complicated system.

2.1 Principles and Evaluations

We shall use the term "principle" instead of "value" since "principle" has fewer positive connotations and we want to leave open the possibility that an agent with socially unacceptable behavior characteristics might prove useful in some context. The term "principle" is being used in the sense of a guiding rule or standard. Principles help determine intentions through agent obligations.

The term "value" is used in natural language ambiguously. We shall focus on the following uses of *value* as a noun: first, *worth in usefulness or importance to the possessor; utility or merit*, and second, *a principle, standard, or quality considered worthwhile or desirable.* We call the former "(instrumental) evaluations" which include judgments of the form "x is a good y" or "x is good for ying", and the latter "principles" which include judgments of the form judgments of the kind "x is good". An example of the first kind is A="this knife is good for chip carving", while B = "honesty is good" is an example of the second kind. We assume that principles are determined off-line and thus are unchanging and intrinsic to the type of agent, while evaluations are determined either on-line or off-line, but are independent of and extrinsic to the agent. Principles relate to other ideas such as "ultimate goal", "root motivation", "essence", or "telos". Agents of a particular type are obliged to observe certain principles. Principles need not be fixed and unchanging and may change as the agent changes from being of one type to another. Principles are general guides to behavior. The principles that guide an agent's behavior are part of the agent's background, and as such are not reasoned about (although they are used in reasoning), principles are unquestioned, persistent, normative, and terminal or ultimate in the sense that there is nothing more basic that the agent can turn to in order to resolve conflicts of purpose. Evaluations, on the other hand, are distinct from the introspective, deontological cognitive activity of determining the "right" thing to do. For example "this knife is good for chip carving", "that agent is honest", and "honesty is a

virtue" are evaluations. Such beliefs are only of intermittent and contingent use, that is, if one agent is considering relying on another agent then an evaluation of that agent's honesty is relevant information for this purpose. On the other hand "an agent ought to be reliable" has the content of the command "be reliable" stated as a fact and as such is a constant guide to behavior, that is, it is like a goal and remains in force regardless of what other goals might be adopted.

Principles for BDI agents be modeled using the knowledge operator. The socially aware intention determination function (see below) will consider this knowledge in generating intentions and obligations. Obligations frequently have penalties associated with the failure to meet the obligation.

How are principles to be incorporated into a BDI architecture, i.e. into a goal seeking knowledge representation system? Both statements A="this knife is good for chip carving", and B = "honesty is good" are representative of beliefs, but statement B seems to represent a motivational attitude (perhaps with an affective, emotional component) or even to represent a goal as well, while statement A is more like a factual belief under rational control. Although it is desirable to keep the number of modal operators small the introduction of norms and principles seems best handled by the introduction of new operators. Evaluations will be taken to be either knowledge or belief depending on the strength and generality of the evaluation. For example, the evaluation "chlorine is an effective anti-bacterial agent" being well-confirmed will be taken as knowledge, while the evaluation "agent a is unreliable because it did not respond to the last request made of it" will be taken as a belief because of its inferior warrant. Norms and principles that are built into the system as background, will be considered as knowledge. We introduce a distinct modal operator "Oi" for each distinct obligation. So the logic that models agents will be a BDI+KO logic.

We see principles and norms on a strong to weak continuum in terms of the obligations generated. Principles are more general and abstract, e.g., "do no gratuitous harm" while norms are more specific and concrete, e.g., "when moving at a speed of 1 meter per second or faster make sure that there are no obstacles within a range of five meters". Norms are determined by roles and designate a range of behaviors that are consistent with the agent's having adopted a given role. When an agent accepts a role, the agent is expected to acquire the set of norms that are appropriate to the role.

Instrumental evaluations are important in determining how an intention is to be achieved during planning and decision-making. Principles are important in determining which intentions will be attempted. Evaluations explicitly represent established relations that can be used in means-ends reasoning. Agents have bounded rationality and limited computational capacity so having an evaluation ready makes intention determination easier.

2.2 Norms

There are complex relationships among values, norms, obligations and the BDI components of agents. An agent may have several intentions; each might invoke a role, and each role might invoke a set of (defeasible) norms with the result that an agent has a conflicting set of norms to be followed.

Norms are standard or canonical ways of reacting to what are recognized as recurring situations. Norms have been developed to enhance the reliability of communication, to make an agent's actions predictable and verifiable, to facilitate coordination of agent actions, to enhance social stability, and for other purposes. Norms develop, in part, because similar situations are repeatedly encountered.

Norms and principles are introduced to give a means to express socially motivated behavior. Agents have knowledge about the relevant norms in a multi-agent environment and decide whether or not to follow any particular norm based on a cost/benefit analysis. The agent must be able to respond to changing conditions and this requires that the agent be able to drop current norms, obligations and intentions in favor of norms, obligations and intentions that better fit the changed environment.

Social laws are norms that are frequently related to evaluations and principles. [5] contend that minimal and simple social laws will tend to be more effective in guiding agent behavior. Social laws simplify the coordination of behavior of multiple agents. These social laws can be of two types: first, laws determined off-line which the agents should obey but which they had no part in determining, and second, laws negotiated by the agents themselves on-line. Informally, a social law is minimal if it imposes fewer constraints than any other comparable social law that accomplishes the same purpose. The object is to find the simplest specification that guarantees that the agents will complete a specified task. Minimal laws (which place fewer constraints on an agent's actions) leave the agent better able to adapt to a changing environment. Note that the simplest specification does not always lead to the most efficient behavior. However, it is likely that simple laws will be effective because they are more easily understood and followed by the agent. "Safety Goals" are conditions that should always obtain, and therefore will fail to be part of an agent's goals only in the most extreme cases. A social law is a norm that helps determine the goals of all the agents to which the law applies. A social law is "useful" if first, it does not make any agent goals unobtainable ("liveness condition"), and second, it does not remove any safety goal ("safety condition"). Liveness requires a bit more explanation, namely, liveness requires that after the application of the useful social law, each agent have an available strategy to achieve each of its goals regardless of the strategy adopted by the other agents, provided that there was such a strategy before the application of the useful social law.

[5] refines the idea of a minimal social law with "minimally preserving social law" (MPSL). Intuitively, these will be the social laws for which the prohibited strategies for individual agents cannot be reduced without permitting some combination of strategies for agents that is prohibited. The distinction is subtle, so we clarify with an example below. We note that there need not be a unique minimally preserving social law and that identification of such laws is computationally expensive. Note that although every minimal social law (MSL) is a MPSL, the converse is not the case. A MSL has the fewest constraints necessary to accomplish its purpose and thus is minimal and preserving. A MSPL, on the other hand, need not be minimal in the sense of eliminating the least number of strategies. A MSPL specifies criteria for individual agents while a MSL has a more global view and specifies criteria over combinations of agents. An example provided by [5] will clarify the distinction. Consider a two agent system in which, at the current state, each of the agents x and y has both of the strategies a and b available, but that only the joint strategies (a, a), (a, b), and (b, b) allow both x and y to achieve their goals while the strategy (b, a) does not allow either x or y to achieve their respective goals. A social law that prevents the strategy (b, a) is a MSL (and a MPSL), while a social law that prevents agent x from adopting strategy b is a MPSL (but not a MSL).

3. A BDI Algorithm

The model of BDI agents that we are considering contains two types of obligation. The first type of obligation is to act in accordance with the agent's purpose, teleology, design, or `personality'. Its purpose will be reflected in the set of values that the agent is committed to. The second type of obligation is to act in accordance with norms. Normally obligations of the first type will be stronger than obligations of the second type. Introduction of the notion of the strength of an obligation introduces the need for a representation of strength so that the intention determination function may consider the relative strengths of obligations. Reasoning about these obligations will use multiple modal operators Oi, each of which has a KD4! axiomatization. Obligations can be general ("do no gratuitous harm") or specific ("keep your promise to meet agent B at 3:00 p.m."). Obligations can be to perform an action or to see that a situation is the case (a proposition is satisfied). Values are part of the agent to the extent that we do not expect values to change as the agent progresses through the temporal structure, unless the agent becomes a different type of agent. Conflict of values in a particular situation may require a reconsideration of the agent's purpose.

The notation (Oi, b, t, w, α) represents that at the situation indexed by t in the possible world w, agent b has the obligation indexed by i to perform action α, while the notation Oip(b,t,ϕ) represents that at time t agent b has the obligation indexed by i to see to it that proposition ϕ is satisfied at some future time point.

```
B := B0;
I := I0;
while true
        get next percept ρ;          // update situation
        B := brf(B,ρ);               // revise beliefs
        O: = orf(B,I);               // revise obligations
        D := options(B,I,O);         // revise desires
        I := SAIDF(B,D,I,O);
   // Socially  Aware Intention Determination function
        π := plan(B,I,O);            // devise plan
        while not (empty(π) or succeed(I,B) or
                    impossible (I,B)  or unsound(π,I,B))
                a := head(π);
                execute a;
                π := tail(π);
                get next percept ρ;
                B := brf(B,ρ);
                if reconsider(I,B) then
                        D := options(B,I,O);
                        I := SADI(B,D,I,O);
                        O := orf(B,I);
```

The procedure above assumes that there is a preference ordering of situations according to desirability relative to the current situation. Situations that satisfy more desires are preferable, perhaps considering "closeness" of satisfaction. This procedure could be improved by making explicit how roles, and thus norms, are determined. As it is, this activity is hidden in the obligation revision function, orf. Socially Aware Intention Determination Function (SAIDF) is a function that suggests an intention most compatible with currently adopted B, D, I, O. As we will suggest in the next section, this function can be tuned to maximize the number of obligations satisfied by the adopted intentions.

We are now ready to suggest some correspondences among Values, Norms, and Obligations, and Beliefs, Desires, and Intentions in the following section.

4. **Social Classification of Agents**

BOID [1] is an agent architecture that extends BDI with obligations which is similar to our architecture. Additionally, we include notions of values and norms as components that guide reasoning about obligations. In the BOID architecture, agents are categorized by the method of conflict resolution that arises among the four modalities. In all agents, agent beliefs are given the highest status, which amounts to agents who must believe in whatever is intended, obliged, and desired. Such agents are called *realistic*. They denote this by prefixing B (for belief) to other intentional notions, as in BXXX (X denotes D, I, or O). The remaining six orderings among desires, intentions, and obligations define other agent types. Social agents give obligations second highest priority, denoted by BOXX. Desires have the second most priority in selfish agents, denoted by BDXX. Stable agents consider intentions to their second highest priority, denoted by BIXX. By and large, we agree with the BOID architecture classification of agents but since we consider values and norms we are able to extend it. In our conceptualization, values and norms are notions that supercede beliefs, obligations, desires, and intentions, denoted by VNBXXX and NVBXXX.

Relating intentional notions that link the formal properties of a logical system to the informal properties of an agent is called *correspondence* [10]. For example, a correspondence involving belief and desire is the property of an agent's being realistic about desires in the sense of not desiring any state that is not believed attainable, called *belief-desire weak realism*. This is formally achieved by requiring that "if an agent believes a proposition ϕ then that agent does not desire not ϕ," so desires that are believed to be impossible are eliminated. See chapter 5 in [10] for a discussion of correspondences involving beliefs, desires, and intentions. We will not discuss those correspondences but rather suggest some additional classification of agents, some of which are captured by correspondences that are made possible by the introduction of obligations.

We offer the following characterizations of types of agents as a starting point for the consideration of correspondences. Agents that give precedence

to obligations over desires will be called *principled agents*. Principled agents may give precedence to obligations derived from norms. Such agents are *socially conscious*, i.e., NVXXX. On the other hand, if obligations derived from principles have precedence then the agent is *conservative principled agent,* i.e. VNXXX. If, in addition, desires are required to be a subset of the obligations derived from principles, the agent is called a *fundamentalist principled agent*. This is putting D after O which is not quite the same as BOXX. Also we said that principles prescribe obligations, which is not distinguished in BOID. If desires are required to be a subset of the obligations, the agent is *self-sacrificing*. Recall that BOXX is classified as social agents in BOID. If the set of obligations arising from norms is not a subset of the obligations arising from principles, the agent has an "outlaw" aspect. *Masochistic agents* attempt to make the intersection of desires and intentions empty. *Proselyte agents* attempt to get other agents to conform to their values. *Duty bound agents* adopt a set of intentions that is a maximal subset of the union of obligations. An *unmotivated agent* adopts intentions that are not desired, i.e., XXBIDX. *Introspective agents* are such that if the agent has a desire or obligation then it believes that it does. *Responsible agents* require the set of desires to be a superset of the set of obligations, i.e., D before O. *Legalistic agents* have a preference for obligations arising from codified principles. *Socially responsible* agents give precedence to obligations arising from social norms. The preceding classification is preliminary and we will build on this future work.

5. Conclusion

This paper has put forward suggestions about how to incorporate social notion into BDI systems through the introduction of norms and principles, which induce obligations. The use of norms and principles not only simplifies the intention determination process for individual agents, but also facilitates the development of agent social characteristics by allowing agents to better model their own intentional states and the intentional states of other agents. We have proposed augmenting the BDI agent control loop with an obligation revision function. Lastly, we briefly outlined salient correspondences between BDI and obligations arising from orientation of agents toward their values and norms. There is a strong correlation between inter-agent understanding of agent behavior and the mechanisms of sociality in the social organization of agent groups. Agents organized in sophisticated social groups have greater opportunity to develop richer behaviors that in turn lead to more efficient and effective methods of achieving agent goals, individually and collectively.

Acknowledgement

This work is supported by AFOSR grant F49620-00-1-0302.

References

[1] Boman M., 1999. Norms in Artificial Decision Making, *Artificial Intelligence and Law*, 7(1):17–35.

[2] Broersen, J., Dastani, M. Huang, Z., Hulstijn, J. and Van der Torre, L. 2001. An Alternative Classification of Agent Types based on BOID Conflict Resolution In: Proceedings of the 13th *Belgian-Dutch Conference on Artificial Intelligence* (BNAIC'01).

[3] Conte R., Castelfranchi C., and Dignum F. 1999. Autonomous Norm-Acceptance. In *Intelligent Agents V*: Proceedings of ATAL 98.

[4] Fasli M., 2000. Towards Circumspect BDI Agents: Preliminary Report. In Proceedings of the *International Conference on Artificial Intelligence*, IC-AI'00, Las Vegas, Nevada, pp. 573-579, CSREA Press,

[5] Fitoussi D. and Tennenholtz M. 2000. Choosing social laws for multiagent systems: minimality and simplicity, In *Artificial Intelligence*, 119(2000) 61-101.

[6] Hexmoor H. and Beavers, H., (In Press, 2002), In Search of Simple and Responsible Agents, In Proceedings of *GSFC/JPL Workshop on Radical Agents*, MD.

[7] Hexmoor H., (In press, 2002), Adaptivity in Agent-based Systems via Interplay between Action Selection and Norm Selection, Based on *International Workshop on Self-Adaptive Software*, (IWSAS-01), Robert Laddaga, Howard Schrobe, and Paul Robertson (Eds).

[8] Hexmoor H. and Zhang X., 2001, Norms, Roles, and Simulated RoboCup, In *2nd workshop on Norms and Institutions* in multiagent systems, (Agents 2001), Montreal, CA, ACM press.

[9] Meyer, J-J , Wiereinga, R. J., Dignum, F. P. M., 1996. The Role of Deontic Logic in the specification of Information Systems, Technical report UU-CS-1996-55, Utrecht University.

[10] Wooldridge, M. 2000. **Reasoning about Rational Agents**, The MIT Press.