

In search of simple and responsible agents

Henry Hexmoor and Gordon Beavers

Computer Science & Computer Engineering Department
 Engineering Hall, Room 313, Fayetteville, AR 72701
 {hexmoor, gordonb}@uark.edu

1 Introduction

An artificial agent is a computational entity capable of interacting with other agents and/or real-world entities. Being reactive is a standard property of agents. The agents considered here exhibit various degrees of sociability in the form of norms, roles, values, cooperation, motives, responsibilities, autonomies, and rights. Intentional agents have been modeled in multi-modal BDI logics, e.g. [5], with operators for belief, desire, and intention. This paper proposes the integration of social notions into BDI agent architectures to account for social decision-making. Although a large collection of notions is needed to explain the actions of complex social agents, this paper provides a somewhat simplified model of social agents built on a small set of agent properties (norms and values) and intentional notions (obligations). Since this model is a starting point for the investigation of social agents, it is expected that the model will be improved and expanded as the result of further research.

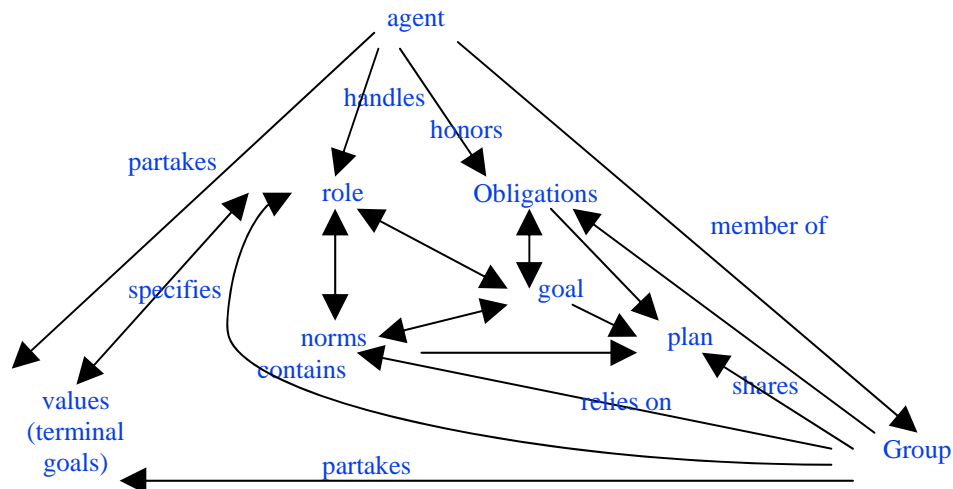


Figure 1 An agent as a member of a social group

Figure 1 shows some of the influences among social agents, groups and social notions that a model should take into account. The figure has been simplified to emphasize the most salient features of sociality. The reality is much more complex and many of the relationships shown are part of ongoing research projects, however, our aim is to set the stage to discuss issues at a more abstract level.

As an example of the potential complexity that the model presented here avoids, values (or guiding principles) can have varying scope in terms of the set of agents to which they apply. Principles may guide the actions of individuals, groups, societies, or even be global guides to behavior. In this paper, in order to reduce the computational complexity of the model, principles will be taken to be constraints that are determined by roles, so that principles can be modeled with a filter on possible worlds¹. These principles are terminal goals that any agent would be expected to observe when assuming the given role. Further developments might allow a group to set the principles to which its members will adhere with each agent helping to determine these principles, however, at this early stage in the development of social agents, principles will be determined off-line and will remain fixed. The values are things like “always cooperate with team members”. When the group adopts a joint intention, the members of the group will negotiate a division of responsibilities, which determines the roles assumed by each agent. An individual role will normally be fulfilled in a standard way, that is, each role will imply a set of norms that the agent is expected to comply with in addition to the principles that it will observe. Norms and principles are at opposite ends of an abstract to concrete continuum of entities that generate obligations. Although there are complex relationships among roles, norms, and goals so that an agent might be expected to weigh the alternatives against one another in order to settle on a consistent set of intentions at any moment, only the simplest relationships will be treated in this model. The simplicity of the model will enable agents to use the relationships to predict the behavior of other agents.

Our legal system holds the owners of software agents responsible for the actions of those agents, therefore, agents capable of considering their responsibilities could offer some protection to the owner of the agent. Such software agents might be agents involved in electronic commerce, automated teller machines, proxy email agents, or robot assistants. Likewise in a command and control situation, a commander is responsible for the actions of the agents under his/her/its control and therefore would have greater confidence in responsible agents capable of considering the repercussions of their actions. The model proposed here allows agents to consider their individual responsibilities and

¹ Possible world semantics is a logical formalism for modal logic. See [Chellas, 1984]. Intuitively speaking, possible worlds capture the various ways the world might develop. Since the formalism in [Wooldridge 2000] assumes at least a KD axiomatization for each of B, D, and I, each of the sets of possible worlds representing B, D and I must be consistent. Since it is unreasonable to assume that an agent's set of desires is consistent we adopt a slightly different semantics from that found in Wooldridge.

thereby the model makes it possible for the agents to account for their actions. Having responsible agents will provide a safeguard to the owner of the agent as well as help the agent arbitrate practical actions and to recognize legal violations by other agents.

Principles and norms guide the behavior of the agent through the generation of particular obligations. Responsibility is a general term covering principles, norms and the obligations that are generated by principles and norms. Responsible agents are true to their principles, obey the norms of behavior in specific situations, and take their obligations seriously. Varieties of responsibility include *responsibility to* concerning an agent's obligation to perform an action, *responsibility for* concerning an agent's obligation to see that a state of affairs obtains, *character responsibility* is the agent's obligation to behave in accordance with its principles, which are general and abstract, and its *norms* which are particular and concrete. The agent has an obligation to observe the norms that apply to a given situation. Whereas responsibilities tend to restrict the agent's choices, rights leave certain choices open and thus can be used to explore the limits of actions permitted to the agent. The relationships between rights and responsibilities regulate the agent's commitments.

2 VON-BDI architecture

In addition to the intentional notions of Belief, Desire, and Intention (BDI), Value, Obligation, and Norm (VON) are three notions that will prove useful in adding social properties to artificial agents. We take these notions as primitive and derive other notions from them. Taken together, they guide an agent's high-level behavior and help provide a level of predictability, accountability, and responsibility. Our first step is to attempt to clarify the meanings of the elements of VON. Values are understood as principles that govern the agent's behavior and which the agent will attempt to uphold as end-goals in the sense that principles yield obligations which the agent is expected to attempt to fulfill. Likewise, norms yield default behaviors that the agent is expected to observe whenever the agent finds itself in a situation to which the norm applies. We invoke a function that maps an agent a , a set of currently imposed values V , a set N of currently active norms, and a set of current Beliefs B to a set of obligations for the agent in that situation:

$$f: a \times V \times N \times B \rightarrow \{o_1, \dots, o_n\}.$$

In this model obligations reflect the influence of principles and norms on the behavior of the agent. We see Value and Norms on a strong to weak continuum of tenets to uphold. Principles will be viewed as more general and abstract, e.g., "do no gratuitous harm" while norms are considered more specific and concrete, e.g., "when in area A and 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 and

include them in the agent's set of beliefs. In natural agents some norms are characteristic behaviors that evolve over time in response to selective pressures while other norms are conventions selected through deliberation. Again, for simplicity in this early model, norms are to be determined off-line and thus are not alterable by the agents. The set of all obligations active at a given time may not be consistent; the intention determination function needs to select a consistent set of obligations to be honored. Since this is a computationally intensive task, we suggest a function running in the background continuously check for consistency of intentions and obligations.

Norms and values differ in specificity. "When fulfilling the role of an ATM, always offer a receipt for each transaction" is a norm that can be rephrased as an obligation "ATMs ought to offer receipts for transactions" or as a fact "ATMs offer receipts for transactions". The specificity and concreteness of this standard suggest that it is a norm. In contrast "always cooperate with team members" is more general and abstract, in part because what constitutes cooperation requires interpretation. So "always cooperate with team members" is a principle. Obligations are implemented as modal operators, with distinct obligations having distinct operators. If agent A and agent B are both on the same team as agent C, then "agent C ought to cooperate with agent A" and "agent C ought to cooperate with agent B" are distinct obligations. Distinct obligations yield distinct modal operators in order to accommodate conflict of obligation without having the logical system degenerate into triviality.

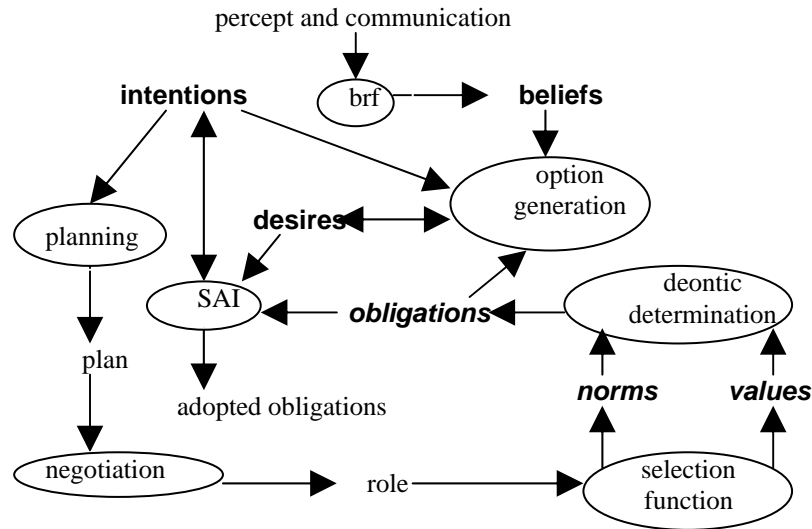


Figure 2 Intra-agent notions

Figure 2 shows the salient relationships among VON-BDI concepts. To be more precise, we describe these relationships in an algorithm later in this paper. The circles represent the processes or functions. For instance the function in the top-right circle has input old beliefs and new percepts and communications with the function producing an update on beliefs. Intentions are derived from the states of VON-BDI as determined by an intention determination function. These operations are explored below.

Determining desires is as difficult as determining obligations. Wooldridge [5, page 32] suggests using an “option generating” function with inputs a set of beliefs and a set of intentions and output a set of desires. However, Wooldridge does not inform us about how this function goes about determining the resulting set of desires. In the model developed here all the elements of VON-BDI are assumed to play a part in the determination of the revised set of desires, since it is reasonable to assume that current desires will influence future desires as will beliefs, intentions, values, obligations and norms. Wooldridge requires that beliefs, desires and intentions all be consistent sets. Unlike intentions, which agents normally attempt to keep consistent, agents do not require their desires to be consistent and thus our model differs from Wooldridge’s by having distinct modal operators for distinct desires. Beyond that, social forces provide an influence on an individual agent’s desires through obligations. This is shown as obligations feeding into the function that will determine the desires. We will continue to refer to this function as the option generator.

Imagine that an agent has a value to “protect oneself from danger” and also has the obligation to perform action α . Suppose that when the sensed data are right for executing α , the agent anticipates that it may come to harm from results of that action, so the value of protecting itself invokes another obligation, which is in conflict with the obligation to perform α . Agents that give more importance to their values are called *principled agents*. If the agent is principled it may well give greater importance to protecting itself.

2.1 Algorithm

In this section we give a revision of the algorithm on page 32 in [5]. We introduce two flags whose values are determined concurrently with, but outside the algorithm. First $\alpha := \text{fn1}(O)$ which checks the current set of obligations for consistency. Next, $\beta := \text{sound}(\pi, I, B)$ checks the current plan for consistency with current beliefs and intentions. The functions that set these flags are envisioned to be running continuously in the background. When a flag is set that value is communicated to the process in the foreground, namely the algorithm below. For simplicity we assume that an agent’s set of values and norms are completely determined by the role that the agent is currently fulfilling and thus V is immutable so long as the agent does not change roles. This algorithm extends that given in Wooldridge in two ways. First, we have introduced an obligation revision function (step

5) that updates the agent's obligations against its norms and values and in light of new percepts. Obligation revision considers the effects of the current beliefs on values and norms. Although we consider values and norms to be immutable in our agents (depending only on roles), their relevance to the current situation is in a constant state of flux.

```

1. B = B0;
2. I := I0;
3. while true do
4.   get next percept ρ;
5.   O := orf(O,ρ);           // obligation revision
6.   B := brf(B, ρ);         // belief revision
7.   D:=options(B,D,I,O);    // determination of desires
8.   I := filter(B, D, I,O);  // determination of intentions
9.   π := plan(B, I);        // plan generation
10.  if α and β is true execute(π) else re-compute π
11. end while

```

Figure 3. Deliberation algorithm

Our second extension to Wooldridge's algorithm is to make the *options* function (step 7) account for influences of obligations.

3 Interagent Sociality

Agents that work together must reciprocate in order to reach equilibrium levels of sociality [4]. This means agents must adjust their own social attitudes in order to experience a sense of fair exchange. Game theory calls agent actions that are equilibrium inducing *policies*. We borrow this notion from game theory to refer to an agent's mental attitude about social relationships. Here we will briefly outline a few of the attitudes about relationships that help establish equilibrium. For each social attitude such as Autonomy we will introduce notations that help us refer to a quantity of (or degree of) that attitude. Since our statements apply to all social notions, instead of repeating, we will label the social attitude as v , which is a member of the set of social notions $N = \{\text{Autonomy, Control, Power, Obligation, Delegation, Dependence}\}$. There are many works that discuss these social attitudes and have influenced us such as [2]. In this section we will introduce notations and state a number of useful definitions and conditions.

Notation:

The maximum amount of v the agent allows itself to tolerate is denoted by v_{\max} .

The minimum amount of v the agent allows itself to experience is denoted by v_{\min} .

The amount of v the agent actually experiencing is denoted by $v_{\text{experiences}}$.

The amount of v the agent wishes to exert is denoted by v_{exerts} .

The actual amount of v the agent affects is denoted by $v_{\text{accomplishes}}$.

Definition 1: Internal Normality

When the amount of social attitude an agent a experiences is between a maximum and a minimum level, the agent's social attitude in general has normal internal condition.

$$v_{\min} \leq v_{\text{experiences}} \leq v_{\max}$$

The normality condition specifies that an agent may tolerate sociality within its acceptable range and this is perhaps part of the agent's inherent personality. However, when this is violated the agent experiences frustration.

Definition 2: Internal Frustration

If the internal normality condition does not hold, the agent experiences internal frustration.

A frustrated agent may consider announcing its frustration or act it out by changes in a related social notion to rectify the frustration. We will give examples of this later in this section. But first, let's introduce more relationships.

Definition 3: Exchange Normality

When the amount of social attitude an agent b exerts on agent a complements the amount agent b experiences, agent b is said in general to have normal exchange condition.

$$v_{\text{experiences-ab}} + v_{\text{exerts-ba}} = 1$$

Definition 4: Exchange Frustration

If the exchange normality condition does not hold, the agent experiences exchange frustration.

Definition 5: Efficacy Normality

When the amount of a social attitude exerted by an agent equals the amount accomplished, the agent is said to have an efficacy normality.

$$v_{\text{accomplishes}} = v_{\text{exerts}}$$

Definition 6: Efficacy Frustration condition

If the efficacy condition does not hold, the agent experiences frustration.

Condition 1: Autonomy and Control ranges

Upper and lower ranges of Autonomy and Control for any agent a are complimentary, I.e., $\text{Autonomy}_{\max-a} = 1 - \text{Control}_{\max-a}$ and $\text{Autonomy}_{\min-a} = 1 - \text{Control}_{\min-a}$

This condition specifies the complimentary nature of autonomy and control ranges. Furthermore, the actual values of autonomy and control are complimentary as stated in the following condition.²

Condition 2: Autonomy and Control are complimentary, i.e., $\text{Autonomy}_{\text{experiences}} = 1 - \text{Control}_{\text{experiences}}$ and $\text{Autonomy}_{\text{exerts}} = 1 - \text{Control}_{\text{exerts}}$, and $\text{Autonomy}_{\text{accomplishes}} = 1 - \text{Control}_{\text{accomplishes}}$

Condition 3: If agent a assumes an obligation (i.e., responsibility) for some action to agent b³, then agent b has dependence on agent a for that action, i.e, $\text{Obligation}_{\text{ab}} \rightarrow \text{Dependence}_{\text{ba}}$ regarding an action.

Condition 4: If agent a delegates an action to agent b⁴, then agent b has obligation (i.e., responsibility) to agent a for that action, i.e, $\text{Delegation}_{\text{ab}} \rightarrow \text{Obligation}_{\text{ba}}$ regarding an action.

Condition 5: If an agent b depends on an agent a or b delegates an action to a, then agent b's autonomy is diminished. i.e., $\text{Dependence}_{\text{ba}} \vee \text{Delegation}_{\text{ba}} \rightarrow \Delta - \text{Autonomy}_{\text{b}}$

Condition 6: If agent a depends on agent b for some action or a delegates an action to b, then agent b has power over agent a by that amount for that action, i.e, $\text{Dependence}_{\text{ab}} \vee \text{Delegation}_{\text{ab}} \rightarrow \Delta + \text{Power}_{\text{ba}}$ regarding an action.

Condition 7: Changes in either Power or Control influence the other proportionately, i.e., $\Delta \pm \text{Power}_{\text{ab}} \leftrightarrow \Delta \pm \text{Control}_{\text{ab}}$.

Increase or decrease in Power between two agents influences similar changes in their Control relationship and vice versa.

Condition 8: Changes in either Delegation or Control influence the other proportionately and complimentary. I.e., $\Delta \pm \text{Delegation}_{\text{ab}} \leftrightarrow \Delta \mp \text{Control}_{\text{ba}}$.

Condition 8 is directly supported by the work of Tuomla [3].

Ideally, agents are expected to not be frustrated. However, when frustrated they may either try to remedy it directly or to seek other social attitudes for indirect relief. Consider an agent who is frustrated by lack of control. The agent may choose to act out by expression of dissatisfaction with control or other attempts to rectify the situation directly.

² Our statement is too rigid. We envision agents that may tolerate deviations.

³ We assume obligation with perfect consent.

⁴ We assume delegation with perfect agreement. Furthermore, both agents are benevolent.

Here we pointed out the complimentary relationships with control. Such a frustrated agent might act out by causing imbalances in autonomy. An agent who is frustrated by obligation, may not choose to act out by expression of dissatisfied dependence or attempts to rectify the situation directly but instead act out by changing its obligations.

4 Toward Guarantees

Let's summarize the structure we have sketched, Figure 4. The agent internally reasons about its values and norms and that leads to its adoption of obligations. That type of consideration influences the agent's social relationships in various ways. We choose to focus on one particular influence illustrated in Figure 2, obligations (i.e., responsibility). Obligations affect agent's dependence as well as autonomy. We have argued that autonomy depends upon ability and social permissions (i.e, obligations) [1]. The bottom half of Figure 4 shows the relationships we discussed in the previous section.

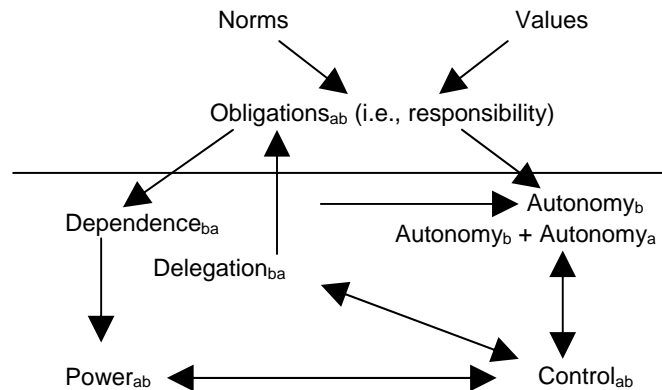


Figure 4 Exploring relationships for guarantees

We envision four possible approaches that can be used for building predictable behavior. Each approach focuses on parameterizing a different social attitude in Figure 4. The first method is to parameterize control. Consider a sphere of social control between two agents in which one agent sets goals and monitors the other agent. Control can be designed to be in various levels, e.g., master-slave, supervisory, recommender levels. The tighter we set the control the more we can rely on the subordinate agent's behavior. The controlling agent is responsible for the behavior of the other agent. A second approach is to set parameters at the power level. If two agents have a differential power relationship, they can affect one another's behavior. Command and control authority relationships are one example of establishing power relationships. By setting agent a in charge of agent b, a directly controls the actions of b.

A third method of adjustment is with setting specific obligations (i.e., responsibilities) for the agent. An agent might be given specific obligations about certain tasks to perform on behalf of a chosen agent (or the human user in case the agent interacts with a human) and that affects its autonomy and control with respect to the agent (or the user). Value and norm adjustment is a fourth method we are proposing. Although this is the least direct method of controlling behavior, it can be used to design an agent who will uphold certain general principles, e.g., Azimov's three laws of robotics.

We believe we have set up the foundation for an delineating the relationships among social attitudes. Much more work remains. The links among social attitudes are naturally defeasible since the agent might find it necessary to violate them.

5 Conclusion

Agents must maintain complex relationships of social attitudes. The resulting web of relationships provides cohesive forces in the group. We outlined basic relations among social attitudes and pointed out a research direction that can be used to develop mechanisms to adjust individual attitudes. We have shown how that can be used in developing methods that will guarantee individual behavior and system performance.

Acknowledgements

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

References

1. G. Beavers and H. Hexmoor, 2001. Teams of Agents, In Proceedings of the **IEEE Systems, Man, and Cybernetics Conference**.
2. C. Castelfranchi, M. Miceli, A. Cesta, 1992. Dependence relations among autonomous agents. In Proceedings of **MAAMAW'92**, Elsevier Science Publishers B. V., Amsterdam, pages 215-227, 1992.
3. R. Tuomela, 2000. **Cooperation: A Philosophical Study, Philosophical Studies Series**, Kluwer Academic Publishers.
4. W. Walsh and M. Wellman. Efficiency and **Equilibrium** in Task Allocation Economies with Hierarchical Dependencies, In The International Joint Conferences on Artificial Intelligence Workshop on **Agent-Mediated Electronic Commerce**, August 1999.
5. M. Wooldridge, 2000. **Reasoning about Rational Agents**, The MIT Press.