# 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 (embodied or otherwise) that interacts 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. [8], 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.
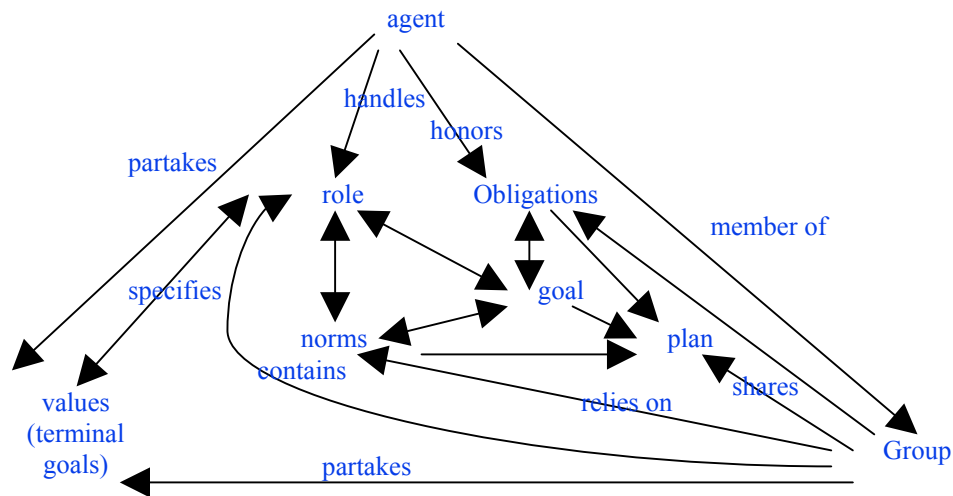
Figure 1 An agent as a member of a social group

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 worlds[1]. 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. 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 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 this section we present an architecture for an individual agent. This architecture relates the interaction feedback loops we saw earlier in greater detail. Figure 2 shows the following feedback loops that are internal to an agent: belief, intention, attention, conversation, obligation and role. We call these feedback loops revision functions. In all, six feedback loops are shown as rounded comparator circle attached to each[2]. Desire does not have a revision function. This is partly because inconsistent desires are allowed and the set of desires changes less often. Desires are reassessed when intentions and beliefs change. Revision functions monitor prevailing conditions and maintain an agent's commitments by implementing revisions only when necessary. For example, the attention revision function has an agent maintain its attention on particular agents until the condition that justifies the attention no longer obtains. Planning, like desire determination, is a function, but not a "revision function". This is because intention revision is responsible for guiding planning when it becomes necessary [4]. 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. Taken together, they guide an agent's high-level behavior and help to provide a level of predictability, accountability, and responsibility. Values are understood as principles that govern the agent's behavior and which the agent will attempt to uphold as

---

[2] Figure 2 contains an abbreviated rendering of Feedback loops with the comparator circle gathering inputs. The feedback arc in each loop is not shown for brevity.

end-goals. 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: $f: a \times V \times N \times B \rightarrow \{o1, \dots, on\}$.
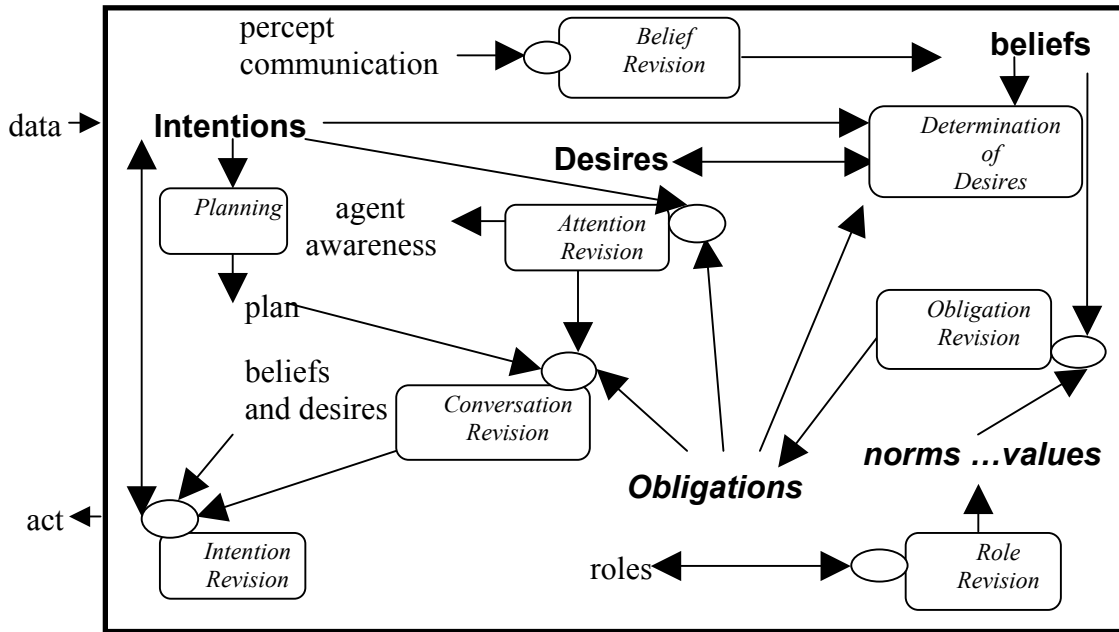
Figure 2 Intra-agent notions

Figure 2 shows the salient relationships among VON-BDI concepts. These relationships are incorporated in an algorithm provided later. For instance the revision function for attention revision has obligations and intentions as inputs, and produces cues for the agent to be aware of certain other agents. Intentions are derived from the states of VON-BDI as determined by an intention revision function. In the model being developed all the elements of VON-BDI 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. In this model obligations reflect the influence of principles and norms on the behavior of the agent. We see Values 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 one 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.

Determining desires is as difficult as determining obligations. Wooldrige [8, 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 Woolridge'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.

It is standard to require 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 common BDI systems. Beyond that, social forces provide an influence on an individual agent's desires through obligations. Agents with plans might enter negotiation with other agents about obligations and roles. We discuss negotiation in a later section.

2.1 Algorithm

In this section we give a revision of the algorithm on page 32 in [8]. We introduce two flags whose values are determined concurrently with, but outside the algorithm. First $\alpha :=$ fn1(O) which checks the current set of obligations for consistency. Next, $\beta :=$ sound(π, 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, O is obligation
6.       B := brf(B, ρ);               // belief revision, B is beliefs
7.       A := af(I, O);                 // attention revision, A is agent awareness
8.       R := rrf(C, r);               // role revision
9.       D:=options(B, D, I, O);    // determination of desires, D is desires
10.      I := irf(B, D, I, O);          // intention revision, I is intentions
11.      π := plan(B, I);              // plan generation
12.      C := crf( π, O),               //conversation revision, C is conversational Context
13.      if α and β is true execute(π) else re-compute π
14. end while
```

Figure 3. Deliberation algorithm

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

3 Interagent Sociality

Our purpose is to make preliminary suggestions for measures of social characteristics affecting the ability of artificial agents to collectively accomplish a task. The typical situation will have these agents working together to collectively accomplish tasks requiring the coordination of their actions with one another. The coordination can be more or less rigidly determined. If the agent group is highly structured, with well-defined roles that are to be filled in prescribed ways, then the coordination is highly orchestrated and rigid. On the other hand, if the agents in the group are allowed to develop their own structure in response to the environment in which they find themselves, coordination may take an unexpected turn. Since our space is limited, we cannot consider the spectrum of coordination, and thus make some assumptions to constrain the discussion. We consider only what Tuomula [6] calls "plan-based cooperative joint action". This is the simplest of his cooperation types that is explicit and agreement-based between agents who share a plan.

In an earlier work [1], we have argued that in order for an agent to be a member of a team, the agent must recognize itself as a team member and have the intention to cooperate with other team members in the achievement of the common goal. Tuomela [6] gives the following necessary and sufficient conditions for agents a and b "intentionally acting together" in performing X:
  1) X is a collective action type, viz., an "achievement-whole" divided into a's and b 's parts;
  2) a and b each intend to perform X together and act in accordance with this intention;
  3) a and b each believe that the other will do its part in the performance of X; and
  4) a 's and b 's intentions in 2) are in part because of the beliefs in 3).

We leave open the question whether or not the participating agents have come to an agreement about how the goal is to be achieved. It is possible that a powerful manager has imposed intentions and the corresponding roles. This leaves the agents no room for dissent. Agreements also need not be negotiated in the case that a plan has been adopted where the plan specifies actions for each of the team members.

Agents that work together must reciprocate in order to reach equilibrium levels of sociality [7]. This means agents must adjust their own social attitudes in order to experience a sense of fair exchange. In game theory, agent actions that are equilibrium inducing are called *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 = {Autonomy, Control, Power, Obligation, Dependence}. There are many works that discuss these social attitudes and have influenced us such as [2]. In this section we 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 $v_{experiences}$.
The amount of v the agent wishes to exert is denoted by v_{exerts}.
The actual amount of v the agent achieves is denoted by $v_{accomplishes}$.

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.

$$v_{min} <= v_{experiences} <= v_{max}$$

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 ($v_{exerts-ba}$) compliments the amount agent b experiences from agent a ($v_{experiences-ab}$), agent a is said in general to have normal exchange condition.

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

Definition 4: Exchange Frustration
If the exchange normality condition (as in definition 3) does not hold, i.e., $v_{experiences-ab} + v_{exerts-ba} = x$, the agent experiences exchange frustration by the amount of $|1 - x|$

The amount of exchange frustration for the experiencing agent varies in proportion to discrepancy between exertion and experience.

From above definitions, four combinations are possible combining internal normalcy and frustration and exchange normalcy and frustration. Internal-normalcy-exchange-frustration is clearly an interaction issue. Whereas, internal-frustration-exchange-normalcy is discomfort that the agent must deal with, perhaps by negotiation with the agent who is exerting the influence. Internal-frustration-exchange-frustration is doubly complex since not only is the interaction faulty, but also the agent is experiencing an internal discomfort. Perhaps a parallel attempt can be made to remedy the discrepancies.

Definition 5: External Normality
When the amount of a social attitude exerted by an agent **a** is between a maximum and a minimum level, the agent's social attitude in general has external normal condition.

$$v_{min} <= v_{exerts} <= v_{max}$$

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

Definition 6: External Frustration
If the internal normality condition does not hold, the agent experiences external frustration.

A frustrated agent may consider changing its role or rank in the group.

Definition 7: Efficacy Normality
When the amount of a social attitude exerted by an agent **a** equals the amount of that attitude accomplished by **a**, the agent **a** is said to have efficacy normality.

$$v_{accomplishes-ab} = v_{exerts-ab}$$

The exerting agent will be satisfied if it manages to produce the effect of its attitude in intended agent.

Definition 8: Efficacy Frustration condition
If the efficacy condition does not hold, the agent experiences frustration.

The amount of exchange frustration for the exerting agent varies in proportion to the discrepancy between exertion and experience.

From the last four definitions, four combinations are possible combining external normalcy and frustration and efficiency normalcy and frustration. External-normalcy-efficiency-frustration is clearly an interaction issue. Whereas, external-frustration-

efficiency-normalcy is discomfort that the agent must deal with, perhaps by negotiation with the group so it returns to its comfort range. External-frustration-efficiency-frustration is doubly complex since not only the interaction is faulty, the agent is experiencing an external discomfort. Perhaps a parallel attempt can be made to remedy both discrepancies.

So far we have been discussing limits and quantities of an agent's attitude when dealing with a single social attitude. Next, we turn to interactions among agents involving multiple social attitudes. We will sketch how the relationship between two agents involving one type of social attitude leads to indirect levels in another social attitude.

Control and autonomy are related. When agents consent to a balance of control between them, their balance of autonomy is complimentary to their control. I.e., control affects autonomy in the sense that as control is imposed on an agent, that agent's autonomy decreases. Autonomy does not affect control. There may not be a control relationship between an agent that has greater autonomy and another with lower autonomy. Over a common set of choices, the autonomy of the agent who has agreed to be controlled is lower than the agent who is controlling choices. Exerting control is representative of the degree to which power is asserted. There is a direct relationship between experiencing control and the amount of power that an agent feels imposed. "Actual control" is the amount to which an agent concedes to power asserted. If social power is disputed, control is not established but a sense of defiance exists. Naturally, the degree of autonomy is affected by control to the extent the agent allows or resists control. The relationship between autonomy and power is complimentary.

"Trust" is an ambiguous term, and furthermore, trust and control can interact in various and complex ways. See Castelfranchi and Falcone for a discussion [3]. For simplicity we consider only "strict trust" which takes trust and control as opposite and complementary notions in the sense that the exercise of control represents a lack of trust, and likewise, if agent a has a high degree of trust in agent b, then agent a will refrain from attempting to control agent b. That is, if agent a trusts that agent b will successfully accomplish a task without interference or assistance, then agent a will not attempt to control agent b, however, to the extent that agent a lacks trust in agent b's ability or intention, agent a will attempt to guide or control agent b.

Consider a simplistic view of the relationship between obligation and dependence. To motivate the relationship consider an example of automobile rental where a clerk agent a and a customer (renter) agent b interact. a must first determine the reliability of b in being a good customer and in a typical case b will present its valid driver's license to a. a then establishes trust in b based on this evidence. Trust is the catalyst for a contract between a and b. As the contract is signed, four simultaneous things will happen. b will have an *obligation* to a about the terms of the contract and a will come to *depend* on b to follow

through with the terms of the contract. a will have an *obligation* to rent a vehicle to b and b will *depend* on having this car from a.

Let's turn to a simplistic view of a relationship between obligation and delegation with the above automobile-renting example. If an agent c is delegated by a to find a vehicle to give to b and c agrees, c will have a kind of obligation to a. This obligation may not be a legal, but rather a kind of moral one that should be further formulated. A caveat is that a's trust in b must be high enough before a decides to delegate the task to c.

When two agents enter a consenting delegation, the delegee may experience diminished autonomy. However, the delegating agent may also experience a lower level of autonomy. If two agents share a choice set and a delegation action gives permission over a part of the choice set to delegee, both delegee and delegator agents agree to lower their freedom over the shared choice set, thereby, lowering their autonomy.

Tuomela gives an immediate connection between control and "dependence", with control is as an inverse of dependence" [6, page 212]. If an agent a has control over another agent b, b depends on a. In a game theoretic sense, for a shared choice set, if a has freedom of strategy and by its choices forces b into a strategy, a has control over b and b depends on a. Tuomela goes further and says that social control is connected to social power. If an agent a has control over another agent b, a has power over b, and vice versa [6]. Changes in either Power or Control influence the other proportionately.

When there is a control relationship between two agents, the controlling agents may experience social permission to delegate a task to the controlling agent. The controlled agent may feel a deontic pressure to be receptive to delegated tasks.

Ideally, agents should 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 its inability to control another agent, a lack of effective control. The agent may choose to act out by an expression of dissatisfaction with control or other attempts to rectify the situation directly. Here we point 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., a kind of responsibility). Obligations affect an agent's dependence as well as autonomy. We have argued that autonomy depends upon ability and social permissions [1]. The bottom half of Figure 4 shows the relationships we discussed in the previous section.
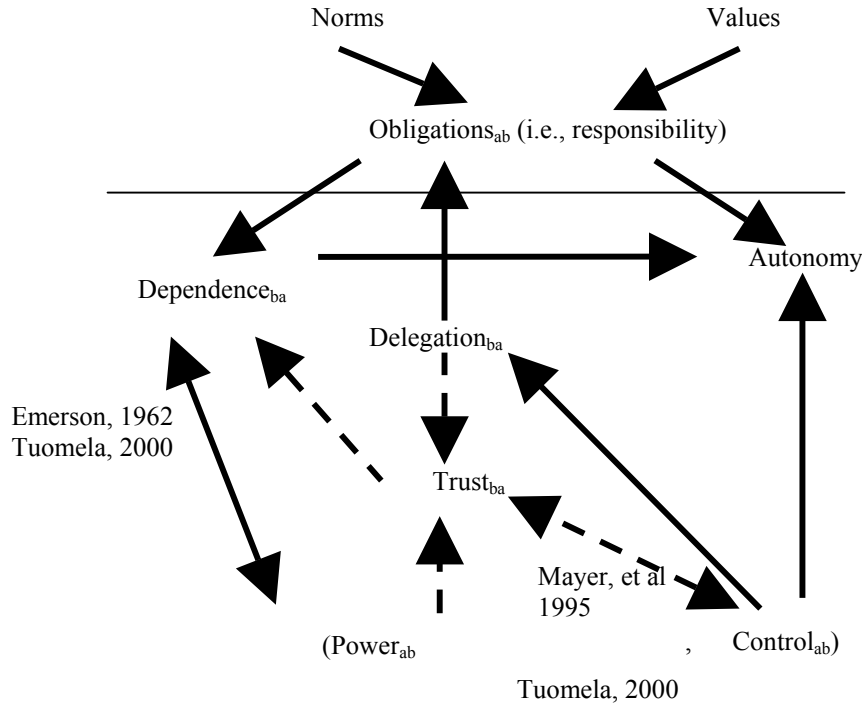
Figure 4 Exploring relationships for guarantees

We envision four possible approaches that can be used for building predictable behavior. Each approach focuses on adjusting a different social attitude in Figure 4. The first method is to adjust control. Consider a sphere of social control between two agents in which one agent sets goals and monitors the other agent. We stipulate for agents that control can be designed to be at various levels, e.g., master-slave, supervisory, recommender levels. This is much more complex in humans. 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 adjust 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.

We believe we have set up the foundation for delineating 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

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. Castelfranchi and R. Falcone 2002. Tuning the Agent Autonomy: the relationships between Trust and Control, In Autonomy, Control, and Delegation workshop in AAAI 2002, Edmonton.
4. N. Lacey, H. Hexmoor, and G. Beavers, 2002. Planning at the Intention Level, In Proceedings of the 15th International FLAIRS Conference (FLAIRS-2002), Pensacola, Florida.
5. R.C. Mayer J.H. Davis, F.D. Schoorman, 1995. An integrative model of organizational trust, Academy of Management Review, Vol.20, N°3, pp. 709-734.
6. R. Tuomela, 2000**.** *Cooperation: A Philosophical Study, Philosophical Studies Series*, Kluwer Academic Publishers.
7. 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.
8. M. Wooldridge, 2000. *Reasoning about Rational Agents,* The MIT Press.