What is a Friendly Agent?

Petra Funk and **Jürgen Lind**
German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
+49 681/302 2464
funk@acm.org lind@dfki.de

Abstract

In this paper, we are concerned with formalizing properties of social agents, aiming at a generic framework for agent description. Operationalizations of the formal definitions are the building blocks used to realize our main objectives. These are providing agents with a sense of self, introducing a social utility measure and defining social norms in agent societies.

Introduction

What is a friendly agent? This question comprises the main issues in our attempt to build or model social agents. In a multi-agent system, as opposed to a single-agent world, each agent is faced with the problem to cope with other agents. Therefore, it needs a set of attributes which can be used to describe other agents and their behavior. Accumulating these agent descriptions then leads to the social model of the agent society from the point of view of a particular agent.

Thus, the main emphasis of the title questions are:

- *What is a friendly* ***agent****?* Our first goal is to search for *general properties of agents.* By conceptualizing such general properties, we hope to find a *general model of agency* or at least provide a *generic framework for agent description.*

- *What is a* ***friendly*** *agent?* The term friendly is one of several properties, social agents may exhibit. It is a social concept and we will transform these concepts into a formal model, such that the agent can use this knowledge about agent properties during its reasoning.

  How can we *formalize or quantify* the properties we ascribe to agents, to make them usable for reasoning purposes?

Obviously, the new formal framework must be made available to the agents in order to improve their performance and so the final question is:

- *How can this knowledge be used?* In our approach, the formalizations of agents and their properties can be used in several ways. Firstly, we equip the agent with a model of itself by applying the generic framework for agent description, providing for a *sense of self.* Secondly, we introduce the concept of *social utility* which expresses the social value of a particular agent. Thirdly, we use the formal social attributes to describe *social norms* in agent societies.

This discussion supports our view of social agents—we believe that it is not enough to have several concurrently running and communicating agents on a machine or network to call it a "social system". Social agents require knowledge about themselves to model competence and reflection as well as knowledge about other agents, their goals, desires, abilities, etc.

What is a friendly *agent*?

As we said in the introduction, the emphasis on the term *agent* means, we collect properties that identify an arbitrary object in a multi-agent world as an agent. By describing these properties of agents, we aim at a *generic framework for agent description* (Funk 1997). We will not develop a new view of the agent world, but look at existing models and extract general, useful properties of agents. The most general definition by Russell and Norvig (Russel & Norvig 1996) characterizes an agent as

> "...anything, that can be viewed as *perceiving* its environment through *sensors* and *acting* upon that environment through *effectors*".

This definition is widely accepted by most researchers in the multi-agent system community. Unfortunately it is far to general for our purposes. It is well-suited for a single agent in a single-agent world. In a multi-agent system agents need to know about others, they must exhibit some kind of autonomy, and communicate with others. A definition, comprising some more of these features (Green *et al.* 1997) characterizes agents as

"...a computational entity which

- acts on behalf of other entities in an autonomous fashion
- performs its actions with some level of pro-activeness and reactiveness
- exhibits some level of the key attributes of learning, cooperation and mobility."

This definition is more appropriate for agents in a multi-agent system, it comprises cooperation and action on behalf of some other entity. In order to act correctly on someone's behalf, an agent requires knowledge about this entity. For cooperation with other agents knowledge about them is not only useful, but necessary. Such knowledge can be described by social concepts. Autonomy and pro-activeness or reactiveness are also required as key attributes of agency.

It is very hard to describe general properties of objects that make them agents, Weiß (Weiß 1995) addresses this by listing a minimum of properties an agent needs to have to be accepted as agent: reactivity, situatedness, pro-activeness and deliberation, rationality, mobility, introspection, veracity, and benevolence.

The last two definitions are more useful for characterizing agents in a multi-agent system than the first one. Still, they lack an important issue we think essential for agents in a multi-agent world: knowledge about others as well as self awareness. The underlying benevolence assumption allows for easier for communication and cooperation among agents, but is in fact unrealistic for real-world applications (Rosenschein & Genesereth 1985). This leads to posing our initial question with different emphasis:

What is a *friendly* agent?

Friendly, being one of several useful social properties of social agents, is in our human horizon, a positive personality trait. In fact, we seek help from friendly people, rather than from unfriendly ones. But, how can an agent use such social concepts for building a social model of its environment? Conceptualizing fuzzy properties of agents is not enough, we need to formalize these concepts in a way, that an agent can use them for its reasoning process. An approach to formalize properties of agents was taken in (Goodwin 1993). The author uses the Z specification language to formalize attributes such as "successful", "capable", "reactive", etc. Furthermore, his definitions are based purely on the externally accessible parts of the agent. However, the major problems with the definitions presented there are:

- they are given from a global perspective and not from the view of an individual agent within a multi-agent system and
- the focus of attention lies more on technical aspects of an agent and not on social issues.

We aim at formalizing *social terms* with emphasis on properties of agents in a multi-agent system from the perspective of an agent *within* the multi-agent system and not from a global point of view.

To illustrate the inherent difficulties of this task, consider the following situation: every human being has a more or less clear idea of any descriptive attribute. The problem we are faced with is, that this idea is made up of a large number of personal values, desires and experiences. Besides the difficulty to identify all factors which make up the term, the set of factors may be inconsistent or even contradictory. Therefore, we need abstract definitions of the terms used in everyday life. Consider the following attempts to define the attribute "cooperative" from the point of view of a particular agent:

Definition 1 (cooperative) *An agent is cooperative if it's actions do not conflict with my goals.*

Obviously, this definition is very handy at first sight, because it enables the agent to classify the other agents easily. On the other hand, does this definition capture what a human calls "cooperative"? What about the case when an agent to be classified does not even know that there are other agents present? Clearly we do not want to call an agent "cooperative" which is not aware of other agents. Consequently, we need a more elaborate definition:

Definition 2 (cooperative) *An agent is cooperative if it knows about my goals and its actions do not conflict with them.*

This extension of definition 1 fixes the bug, but it also introduces an additional problem: how can the classifying agent tell whether the other agent knows about its goals? Furthermore, what about an agent which knows about my goals but its own goal do not force it to cooperate simply because there is no potential conflict? Again such an agent is not what we call "cooperative". Therefore, a further refinement is needed:

Definition 3 (cooperative) *An agent is cooperative if it knows about my goals, its own goals are potentially conflicting with mine and its actions lead to the least possible conflicts.*

Although this definition does not capture the entire term (e. g. it does not include joint action selection of the two agents), it may suffice as an approximation. If we trace the development from definition 1 to definition 3, we see that the prerequisites have changed dramatically from *no knowledge about the other* to *mutual goal recognition and conflict detection*.

This examples illustrates the problems of the task to define social attributes from the point of view of a single agent. Definitions which are able to capture the major part of a social concept need powerful support mechanisms (such as goal recognition techniques) to be used efficiently. It is therefore very important to identify a set of useful attributes and to provide formal definitions for them on various levels of abstraction.

What for?

Now that we have outlined the need of social attributes and the difficulties of obtaining operational definitions of them, we will present our ideas for the usage of these attributes. The main ideas are:

- Sense of self. As has been stated by Marsh (Marsh 1995), a very important feature of a socially intelligent agent is a sense of self—the ability to see oneself as others see one. This allows an agent to behave in way it believes acceptable to others. A generic framework for agent description enables an agent to build a model of itself and to reason about this model. Thus, the agent is able to reflect and introspect about its capabilities and actions. This closely resembles to the well known concepts of reasoning about knowledge in multi-agent systems (Fagin *et al.* 1996). We will aim at providing for a modality connected to social intelligence and thus equip agents with social reasoning capabilities.

- Assessment of other agents. Until now, the attributes have been defined free of any value judgment. To enable the agent to perform social reasoning, the other agents are assigned their *social utility* for the assessing agent. This utility function is not fixed over time, it depends on the situative context of the agent.

- As building blocks for social norms. Social norms and social modeling have been discussed by a number of authors (e. g. (Shoham & Tennenholtz 1992), (Marsh 1995), (Chaib-draa 1997)). In our work we use the definition given in (Lind 1997):

  Definition 4 (Norms) *Norms are common rules for agent behavior which are oriented at widely agreed social values. They seek to determine agent behavior in situations, where it is not determined in any other way. Therewith, they support expectations.*

  The social attributes are needed in this respect to describe and communicate social norms.

Conclusion

In the previous sections, we have outlined the need for an explicit representation of social knowledge. By conceptualizing properties of social agents we aim at a generic framework for agent description. In order to assess agents and their social attitude in a multi-agent system the social concepts must be made operational. Finally this new knowledge and reasoning can be used for giving the individual agent the self awareness needed in a social system, enabling it to asses other agents and describe and communicate social norms as well as meeting these norms. This will lead to what we understand by socially intelligent agents.

References

Chaib-draa, B. 1997. Connection between micro and macro aspects of agent modeling. In *MAAMAW'97*.

Fagin, R.; Halpern, J. Y.; Moses, Y.; and Vardi, M. Y. 1996. *Reasoning About Knowledge*. MIT Press.

Funk, P. 1997. Representation of Self and Others. Phd thesis proposal, Universität des Saarlandes. In preparation.

Goodwin, R. 1993. Formalizing Properties of Agents. Technical Report CMU-CS-93-159, Carnegie Mellon University.

Green, S.; Hurst, L.; Nagle, B.; Cunningham, P.; Sommers, F.; and Evans, R. 1997. Software agents: A review. IAG report, Trinity College Dublin, Broadcom Èreann Research, Intelligent Agents Group.

Lind, J. 1997. Learning social behavior. Phd thesis proposal, Universität des Saarlandes. In preparation.

Marsh, S. 1995. Exploring the socially adept agent. In *DIMAS'95*, 301 – 308.

Rosenschein, J. S., and Genesereth, M. R. 1985. Deals among rational agents. In *IJCAI'85*, 91–99.

Russel, S., and Norvig, P. 1996. *Artificial Intelligence, A Modern Approach*. Prentice-Hall.

Shoham, Y., and Tennenholtz, M. 1992. On the synthesis of useful social laws for artificial agent societies. In *Proc. of the National Conference on AI*, 276–281.

Weiß, G. 1995. Adaptation and learning in multiagent systems: Some remarks and a bibliography. In *Adaption and Learning in Multi-Agent Systems*, number 1042 in Lecture Notes in Artificial Intelligence, 1–21. Springer-Verlag.