M. Harbers, K. van den Bosch and J.-J. Meyer. (2011) A Theoretical Framework for Explaining Agent Behavior. Proceedings of SIMULTECH 2011

A THEORETICAL FRAMEWORK FOR EXPLAINING AGENT BEHAVIOR

Maaike Harbers[1,2], Karel van den Bosch[2], John-Jules Meyer[1]
[1]*Utrecht University, P.O.Box 80.089, 3508TB, Utrecht, The Netherlands*
[2]*TNO Human Factors, P.O.Box 23, 3769ZG, Soesterberg, The Netherlands*
maaike@cs.uu.nl, karel.vandenbosch@tno.nl, jj@cs.uu.nl

Keywords: Explanation, agent-based modeling, social simulation

Abstract: To understand emergent processes in multi-agent-based simulations it is important to study the global processes in a simulation as well as the processes on the agent level. The behavior of individual agents is easier to understand when they are able to explain their own behavior. In this paper, a theoretical framework for explaining agent behavior is proposed. By distinguishing different types and contexts of explanations, the framework aims to support the development of explainable agents. The framework is based on an examination of explanation literature, and experiences with developing explainable agents for virtual training. The use of the framework is illustrated by an example about the development of a negotiation agent.

1 INTRODUCTION

Social simulations provide the opportunity to investigate the relation between the behavior of individuals and emerging social phenomena like crowd behavior, cooperation and reputation. To fully understand the social phenomena that arise, not only the macro processes should be studied, but also the behavior of the single agents. For instance, a crowd can start to move because all agents are running towards something or because they are following one leader. Another example is cooperation, which may emerge because agents behave in an altruistic or self-interested way. More insight in the behavior of individual agents is facilitated when the agents are explainable, that is, able to explain their own behavior.

Some social patterns emerge out of agents modeled by a few simple if-then rules only, and their behavior can be explained by their rules and interaction with the environment. The behavior of more complex agents that, besides merely reactive behavior, also display proactive behavior is more variable, and harder to predict and understand. In particular in simulations with proactive agents, the similarities or contradictions in the explanations of different agents can help to understand the overall processes (Harbers et al., 2010b).

In order to explain agent behavior, it is important to choose an appropriate behavior representation model. For instance, it is easier generate understandable explanations of agent behavior when the underlying social and psychological processes are represented, rather than the chemical. To support the design of explainable agents simulating human behavior, we propose a theoretical framework for explaining agent behavior (Section 3). The resulting explanations of individual agents can be used to better understand emergent phenomena in social simulations. The framework is based on an examination of explanation literature (Section 2), and on our experiences with developing explainable agents for virtual training. We will illustrate the framework with an example (Section 4).

2 EXPLANATION RESEARCH

In psychological literature, Malle's framework about how people explain behavior is one of the most elaborate (Malle, 1999). The framework distinguishes four modes of explanation (see Figure 1). Unintentional behavior is explained by *causes*, e.g. she woke up because she heard the alarm. Intentional behavior is always preceded by an intention. Intentions themselves

only yield useless explanations like 'she did x because she intended to do x'. However, the *reasons* for an intention, i.e. the actor's beliefs and goals, do form useful explanations. The third explanation mode concerns *causal histories*, explaining the origin of beliefs and goals. The fourth mode, *enabling factors*, consider the capabilities of the actor, e.g. he finished the assignment because he worked hard. Malle states that most explanations for intentional behavior are reason explanations.

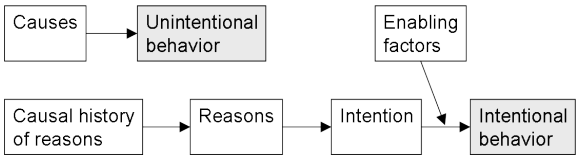

Figure 1: Malle's four modes of explaining behavior.

In argumentation for practical reasoning, argumentation about what is most sensible to do is studied. This is closely related to explanation of behavior, as arguments for a certain action can also be used to explain that action. Atkinson et al proposed the following argumentation scheme for practical reasoning (Atkinson et al., 2006). *In the circumstances R, we should perform action A, to achieve new circumstances S, which will realize some goal G, which will promote some value V.* The scheme provides a motivation of an action, but can also be used to explain an action. For instance, I go to the supermarket (A) because I am at home (R) and after going there I will be at the supermarket (S), where I can to buy food (G) so that I can feel healthy (V).

The explanation of agent behavior is studied in the domain of virtual training. In virtual training, agents play the role of team member, colleague or opponent of the user. Explaining the behavior of such agents gives the user insight in the reasons for others' actions and helps him to better understand played training scenarios. There are several proposals of explanation components that explain agent behavior in virtual training systems (Johnson, 1994; Van Lent et al., 2004; Gomboc et al., 2005). Developers noticed that training simulations differ in their 'explanation-friendliness', that is, the availability and suitability of information that is delivered to the explanation module (Core et al., 2006). At best, agent behavior is represented by goals and the preconditions and effects of actions, and behavior can automatically be imported. In the worst case, behavior is represented by procedural rules, and a manually built representation of the behaviors has to be made. In earlier work, we proposed an approach for explainable agents for virtual training in which explanation capabilities are integrated in the agent model (Harbers et al., 2010a). Though that poses certain requirements on agent design, the quality of explanations no longer depends on the explanation-friendliness of a training simulation.

3 THE FRAMEWORK

Though it is impossible to represent all explaining factors of agent behavior in a model, it is possible to choose an agent model and represent behavior tactically, such that most concepts in the representation can be reused for explanation. To guide this agent development process it is helpful to be aware of different possible explanations for an action. In this section we therefore present a theoretical framework for explaining agent behavior. The framework distinguishes five different ways to explain an action of an agent, and on top of that, different contexts of explanation. Compared to Malle's framework and Atkinson's argumentation scheme discussed in the previous section, our framework distinguishes more types of explanations.

3.1 Five questions

To discuss the various ways to explain an agent action, we introduce the following five subquestions of the question: *Why did you perform this action?*

- *What goal did you try to achieve?*
- *Why do you want to achieve this goal?*
- *Why does this action achieve that goal?*
- *Why did you perform the action at this particular time point?*
- *Why did you perform this action and not another?*

The first question considers the goal behind an action, or in other words, it refers to the desired effects of the action. An explanation is for instance, I called a friend to wish him a happy birthday. Both Malle's and Atkinson's frameworks distinguished goals as an explanation or argumentation type.

The second question, why do you want to achieve this goal, concerns the reasons behind a goal. For instance, I called my friend because I know that he appreciates phone calls for his birthday. In Malle's framework such explanations are called causal history explanations, and they are similar to values in Atkinson's scheme. In a goal hierarchy, a goal above a goal provides a reason for a goal, however, as we will see later, these are not always useful.

The third question, why does this action achieve that goal, can be answered by domain knowledge, e.g. terminology or the function of a tool. The domain

knowledge required in our example is rather common, but an explanation of this type would be: I called my friend because calling someone allows one to talk to that person. This category is not distinguished in the frameworks of Malle and Atkinson.

The fourth question concerns the timing of an action. Possible answers to this question are the event that triggered the action, or the events that made it possible to perform the action. In our example, I called my friend because today is his birthday. The timing of an action may be explained by an enabling factor such as distinguished in Malle's framework, but Malle's enabling factors explanations do not involve triggers like someone's birthday.

The fifth and last question asks why this particular action was performed and not another. The answer may concern multiple possibilities, e.g. I called my friend because I did not have the time to visit him, or preferences, e.g. I called my friend because I believe that calling is more personal than sending an email. Explanations referring to multiple possibilities are similar to the enabling factors in Malle's framework, but preferences are not. Explanations with preferences are similar to values in Atkinson's scheme.

3.2 Contexts of explanation

We have distinguished five different questions, but often there are multiple possible answers to these questions. For instance, I leave a note at your desk because I want you to find it, but also because I want to remind you of something. Both explanations in the example contain a goal. To account for different possible explanations of the same type we introduce the notion of an explanation context. An explanation uses concepts in a certain domain or from a certain description level. Explanation contexts are for instance the physical context, psychological context, social context, organizational context, etc. A physical context of explanation refers to explanations in terms of positions of agents and objects, and physical events in the environment. A psychological context refers to characteristics of the agent such as personality traits, emotions, preferences and values. A social context refers to aspects like mental states attributed to other agents, and trust in others. An organizational context refers to an agent's role, its tasks, its power relation to others, procedures, rules and norms. The two explanations for putting a note at your desk, 'so you will find it' and 'to remind you of something', concern a physical and social context, respectively.

3.3 Use of the framework

The purpose of the framework is to support explainable agent development. Developers can use the framework to determine which questions within which explanation context(s) an agent should be able to answer. Being aware of an agent's explanation requirements will facilitate the choice for an agent model, and subsequently, design choices that must be made within the model. The development is an interaction process between subject matter experts and programmers, where subject matter experts have knowledge about desired explanation types and processes that bring about certain behavior, and programmers know which agent models and architectures are available to represent agent behavior.

It may happen that some information needed for explanation is not necessary for the generation of behavior, or simply does not fit in the agent's behavior representation. In that case, extra information needs to be added to the behavior representation, like justifications for reasoning steps are added to expert systems. Then still, choosing an appropriate representation will facilitate the addition of explaining elements to the model.

4 ILLUSTRATION

In this section we illustrate the use of the proposed framework with an example about the development of an agent for virtual negotiation training[1]. The training scenario involves a negotiation about terms of employment and involves two players, a human player who has the role of employer and a virtual agent playing the future employee. The scenario focuses on the joint exploration phase of the negotiation, in which negotiation partners explore each others' wishes. An often made mistake is that people only explore each others' preferences on issues, e.g. the height of a salary, and forget to ask about the other's interests, e.g. the need of enough money to pay the mortgage. By exploring someone's interests, alternative solutions can be found that are profitable for both partners, e.g. a lower monthly salary but with a yearly bonus.

Figure 2 shows our first version of the future employee agent, modeled as a goal hierarchy. Note that only the agent's goals and actions (in gray) are displayed, and not its beliefs. The actions in the hierarchy can be explained by their underlying goals. For instance, the action to propose 40 hours per week is

[1]The training was developed at the TU Delft as part of the pocket negotiator project

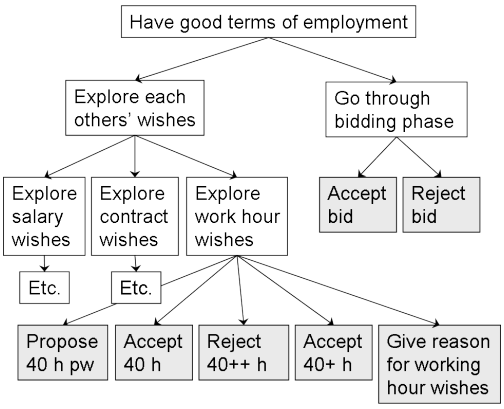

Figure 2: First model of the negotiation agent.

explained by the goal that you want to explore each other's wishes on working hours because you want to explore all wishes. The acceptance of a good bid is explained by the goal that you want to go through the bidding phase. These explanations seem rather useless. After examining the goal hierarchy according to the framework, we realized that we had used a procedural context, whereas we wanted explanations from a psychological perspective.

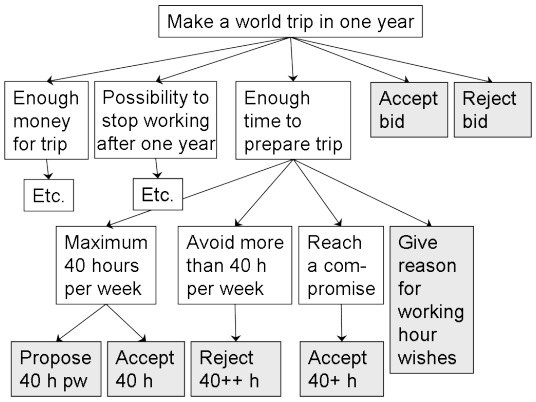

Figure 3: Second model of the negotiation agent.

Figure 3 shows a new version of the goal hierarchy in which the agent's personal preferences and goals are taken into account. The actions (gray) in this model are the same as in the original one. But though both models generate exactly the same behavior, the explanations of the actions are different. In the new model, for instance, the action to propose 40 hours per week is explained by the goal that you want to work with a maximum of 40 hours per week because you want to have enough time to prepare your trip. And the action of accepting a good bid is explained by the goal that you want to make a world trip in a year. On face validity, these explanations are much more useful than the previous ones.

5 CONCLUSION

To summarize, we have presented a theoretical framework for explaining agent behavior with different explanation types and contexts. When modeling an explainable agent, the framework can be used as a guide for the choice for an appropriate agent model, and choices about the representation of the agent's behavior. This should result in agents that can provide useful explanations in the domain or applications for which they are intended, and lead to better understanding of individual agent behavior in social simulations. A next step will be to aggregate explanations of individual agents into one, more global explanation about emergent phenomena in social simulations.

ACKNOWLEDGEMENTS

The authors thank Joost Broekens for his contribution to the development of both negotiation agents in Section 4. This research has been supported by the GATE project, funded by the Netherlands Organization for Scientific Research (NWO) and the Netherlands ICT Research and Innovation Authority (ICT Regie).

REFERENCES

Atkinson, K., Bench-Capon, T., and McBurney, P. (2006). Computational representation of practical argument. *Synthese*, 152(2):157–206.

Core, M., Lane, H., Van Lent, M., Gomboc, D., Solomon, S., and Rosenberg, M. (2006). Building explainable artificial intelligence systems. In *AAAI*.

Gomboc, D., Solomon, S., Core, M. G., Lane, H. C., and van Lent, M. (2005). Design recommendations to support automated explanation and tutoring. In *Proc. of BRIMS 2005*, Universal City, CA.

Harbers, M., Bosch, K. v. d., and Meyer, J.-J. (2010a). Design and evaluation of explainable BDI agents. In *Proceedings of IAT 2010*, volume 2, pages 125–132.

Harbers, M., Bosch, K. v. d., and Meyer, J.-J. (2010b). Explaining simulations through self explaining agents. *Journal of Artificial Societies and Social Simulation*, 12(1)(4).

Johnson, L. (1994). Agents that learn to explain themselves. In *Proceedings of the Conference on AI*, pages 1257–1263.

Malle, B. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1):23–48.

Van Lent, M., Fisher, W., and Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proc. of IAAA 2004*, Menlo Park, CA. AAAI Press.