Managing Helpful Behavior in Collaborative Activities of Heterogeneous Agent Groups

Ece Kamar
SEAS, Harvard University
Cambridge, MA 02138
kamar@eecs.harvard.edu

1 Introduction

This thesis aims to provide a foundation for designing computer agents able to work better with people and with other agents in heterogeneous groups. When agents work together on a collaborative activity, in addition to performing their share of the activity, they may be able to help one another and thus improve the collective utility. However, helpful actions typically result in some costs that may include resources consumed in communicating, lost opportunities to do other activities, the need for group members to adapt their individual plans to the helpful act or its effects, or interruption costs. The thesis specifically focuses on investigating the question of how, when and what kinds of helpful behavior should emerge when agents collaborate, taking into account the costs of a helpful action. It considers collaborative activities that take place in settings in which there is uncertainty about agents' capabilities and about the state of the world. To ensure that helpful behavior improves the overall benefit of the collaboration, I have designed decision theoretic mechanisms that manage helpful behavior by considering the costs and utilities to both the agents and people participating in the collective action. For example, these mechanisms facilitate agents' communication within a collaborative group as a type of helpful behavior. They increase efficiency of collaboration by better estimating the utility of helpful behavior.

A particular focus of this thesis is designing general agent strategies for helpful behavior in settings involving computer agents and people. I investigate the factors that influence people's perception of helpful behavior (interruptions in particular), when they collaborate with computer agents or people. Subsequently, I aim to revise the decision theoretic mechanisms for helpful behavior with respect to these findings. Thus, this work enables the design of agents that consider both the utility of helpful acts and people's willingness to initiate or respond to these acts while managing interactions.

My thesis makes three contributions. First, it incorporates decision-theoretic mechanisms for managing helpful behavior into existing formalizations of collaborative activity. Second, it provides an investigation of the way people perceive the usefulness of helpful actions when proposed by a computer agent. Third, it proposes incentives for facilitating collaboration among self-interested agents. In addition to these theoretical and empirical contributions, my findings are applied to several real-life application domains with different characteristics. The remainder of this abstract describes each contribution in detail.

2 Agent Strategies for Helpful Behavior

This work considers the design of agent strategies for deciding whether to help other members of a group that is engaged in collaborative activity. The decision-making strategies take into account that agents may have only partial information about their partners' plans for sub-tasks of the collaborative activity; the effectiveness of helping may not be known a priori; and, helping actions have some associated cost. Existing formalizations of collaborative activity fail to adequately model helpful behavior in a way that ensures improvements in the efficiency of collaboration. To address this gap, I have expanded the SharedPlan formalization of collaborative activity [Grosz and Kraus, 1996] with a set of rules that formalize the way that helpful behavior arises from the commitments and intentions of the participants in a collaborative activity [Kamar *et al.*, 2009]. I have designed decision-theoretic mechanisms that deploy these rules to capture the utility of performing a helpful act or communicating with group members, so that agents in a collaborative activity help each other when appropriate. The novelty of this work is presenting a formalized and general model of helpful behavior that makes decisions based on the utility of a helpful act, in contrast to axiomatic methods. I investigate techniques for the tractable integration of a BDI model (i.e., SharedPlans) with the decision-theoretic mechanisms and propose a novel probabilistic representation of agents' beliefs for the recipes selected for the group activity in the form of an AND-OR tree. This representation is exponentially more compact than an exhaustive representation, and makes reasoning about helpful behavior tractable despite the partial information. I have tested the mechanism using a multi-agent test-bed with configurations that varied agents' uncertainty about the world, their uncertainty about each others' capabilities or resources, and the cost of helpful behavior. In all cases, the decision-theoretic mechanism outperformed axiomatic methods.

3 Human Perception of Interaction Opportunities

In this part of my thesis, I propose a new methodology for managing interactions between computer agents and people

when they collaborate, with a special focus on managing interruptions. This novel methodology takes into account the costs and benefits to both people and computer agents in human-computer collaborations on tasks being done in fast-paced environments [Sarne and Grosz, 2007]. Thus, this methodology combines concepts both from the adjustable-autonomy [Scerri *et al.*, 2003] and the interruption management literature [Horvitz *et al.*, 2003]. It also takes into account the possible mismatch between an agent's calculation of utility of an interruption and the person's perception of it. To empirically investigate human perception of interaction opportunities, I have designed and developed an abstract game, which provides an analogue of human-computer interactions in collaborative task settings [Kamar *et al.*, 2007]. Using multi-agent decision-making techniques, I created a computer agent that accurately captures the utility of an interruption in this game setting, and initiates interaction when the expected utility is positive [Kamar and Grosz, 2007]. Analysis of human subjects' responses to interruptions in the abstract game show that both the magnitude of interruption outcome and the type of partner that a subject collaborates with influence the likelihood that people accept interruptions. The results of this analysis will be used to revise decision making mechanisms for helpful behavior in future work.

In addition to this fundamental work on agent design and decision making, I have developed algorithms that analyze the benefit of interruptions in two real-life application domains. These algorithms drawn on my findings from the theoretical and empirical work to applications of managing interruptions in office [Kamar and Horvitz, 2007] and mobile [Kamar *et al.*, 2008] domains.

4 Collaboration of Self-Interested Agents

In heterogeneous groups of agents, self-interested agents with conflicting or aligned goals can collaborate on a collective activity, and they may help each other if doing so improves their individual utilities. This work investigates the design of computational methods for developing and sustaining collaboration among self-interested agents, managing helpful behavior among them, and providing incentives to them to promote helpful behavior. It compares different payment mechanisms in terms of fairness, computational tractability, efficiency, budget-balance and incentive compatibility, and focuses on the challenges that may arise due to the uncertain and sequential nature of many multi-agent domains.

I have performed an initial investigation of these ideas on a real-life application domain of ridesharing. This work presents fair and efficient computational mechanisms for the collaboration among self-interested people aimed at minimizing the cost of transportation and the impact of travel on the environment [Kamar and Horvitz, 2009]. The mechanism has been empirically evaluated on hundreds of real-life GPS traces collected from a community of commuters, and indicated significant reductions on number of commutes and on total cost for varying preferences of users. The evaluation has also pointed out important challenges for applying mechanism design ideas to real-life domains for future work.

5 Future Work

In future work, I plan to enrich the general agent strategies for helpful behavior in SharedPlans with rules about who, when and how to help. This extension will also include more comprehensive empirical analysis of helpful behavior in general collaborative settings. The work on understanding human perception of interaction opportunities will be continued to investigate the effect of trust on people's willingness to accept interruptions, and to explore design of tools and interfaces that better represent the usefulness of interruptions. By better understanding the factors affecting people's perception of interruptions, I aim to provide foundation for building agents that interact with people better. I have recently started the work on building a theoretical model of collaboration for self-interested agents. This extension will broaden the applicability of my work on helpful behavior to heterogeneous groups of agents in which agents do not necessarily share a common utility function.

References

[Grosz and Kraus, 1996] B.J. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[Horvitz *et al.*, 2003] E. Horvitz, C. Kadie, T. Paek, and D. Hovel. Models of attention in computing and communication: from principles to applications. *Commun. ACM*, 46(3):52–59, 2003.

[Kamar and Grosz, 2007] E. Kamar and B. J. Grosz. Applying MDP approaches for estimating outcome of interaction in collaborative human-computer settings. *MSDM*, pages 25–32, 2007.

[Kamar and Horvitz, 2007] E. Kamar and E. Horvitz. Jogger: Personal adaptive filter for reminders. 2007.

[Kamar and Horvitz, 2009] E. Kamar and E. Horvitz. Preferences, mechanisms, and collaboration in transportation: The ABC of ridesharing. *To Appear In IJCAI*, 2009.

[Kamar *et al.*, 2007] E. Kamar, B.J. Grosz, and D. Sarne. Modeling User Perception of Interaction Opportunities in Collaborative Human-Computer Settings. In *AAAI*, volume 22, page 1872, 2007.

[Kamar *et al.*, 2008] E. Kamar, E. Horvitz, and C. Meek. Mobile opportunistic commerce: mechanisms, architecture, and application. In *AAMAS-Volume 2*, pages 1087–1094, 2008.

[Kamar *et al.*, 2009] E. Kamar, Ya'akov Gal, and B. J. Grosz. Incorporating Helpful Behavior into Collaborative Planning. *AAMAS*, 2009.

[Sarne and Grosz, 2007] David Sarne and Barbara J. Grosz. Estimating information value in collaborative multi-agent planning systems. In *AAMAS*, page 48, 2007.

[Scerri *et al.*, 2003] P. Scerri, D. Pynadath, W. Johnson, P. Rosenbloom, M. Si, N. Schurr, and M. Tambe. A prototype infrastructure for distributed robot-agent-person teams. In *AAMAS*, pages 433–440, 2003.