Towards Collaboration between Human and Social Agents that mind Human Social Personality

Henry Hexmoor and Srinivas Battula
Department of Computer Science & Computer Engineering
University of Arkansas
Fayetteville, AR 72701
{hexmoor, sbattul}@uark.edu

Keywords: Adjustable Autonomy, Trust, Intelligent Agents, Social Agents, and Simulated Battlespace.

Abstract

We focus on collaboration issue of mixed initiative interaction between a human user and a group of agents. Cognitive requirements for agent design as well as interface are discussed. We present a methodology for eliciting user preferences. We then illustrate adjustment of autonomy and trust in an implemented system with a number of Unmanned Combat Aerial Vehicles (UCAVs) either under autonomous agent control or a human remote pilot control is presented.

1 INTRODUCTION

Recent research in mixed initiative and human-computer interaction has focused on user-interactions where either interface agents with specialized agents observe human actions and guess human preferences [14, 16] or methods where humans monitor and control agent actions [19]. The kinds of agents we have in mind are not mere human assistants [8] as would be the case with personal assistants that do not have an independent task. The agents we have in mind have independent tasks. An example is running an automated factory. We believe successful collaboration between a human user and a system of such agents centers around cognitively oriented agent designs and interactions. First, human users need to be able to express aspects of their personalities that affect their interaction in general as well as express preferences for specific action or decision [17]. Second, agents need to be designed to alter their interactions to suit the human personalities as well as express themselves to human users in cognitively appropriate manner. Third, human users and agents need to communicate with intentional notions such as intentions and obligations.

An important motivation for exploring personality in our human-agent interface is the *similarity-attraction hypothesis* where individuals prefer to interact with others with similar personality. Psychologists have identified two major interpersonal personality dimensions, the dominance/submissiveness dimension, and the warmth/hostility dimension [9,10,11]. The interest in developing agents with personalities is most seriously pursued by the community in believable agents and aLife [13,17]. In this paper we will not suggest a personality model. Instead we focus on a few salient social notions determined by personality traits for interaction. Along the dominant-submissive personality continuum we will look at control autonomy, and power. Along the personality continuum warmth-hostility continuum we will look at benevolence and trust.

Other than personality, we are interested in how an agent and a human user can share intentional notions such as intention, attention, and obligation. Figure 1 shows the components we will consider. The Figure shows modes of adjustments we envision being available for a human user. We will refer to these adjustable levels in interaction as *cognitive dials*. The basic idea in the diagram is that when one dial is tweaked, it changes setting in other categories. The dials are a means of giving humans a quick and easy method to read and/or set those parameters. Cognitive dials as nominated here are preliminary, and since they are not independent as described herein, the relationships between them are overly complex at this stage of development and thus not suitable for direct implementation [15]. Cognitive dials affect interactions by changing the decision-making of agents. A user might wish agents to implement new values and norms and abandon others. The user might wish to adjust an agent's level of sociality or benevolence. A benevolent agent takes the welfare of others into account and this might directly affect its autonomy. The user may wish to change how agents consider the relationship between social notions. For example, the user might equate power and control levels in agents. Agents must be designed to account for human dialed changes in interaction, and thus the design of agent architectures must account for such properties.

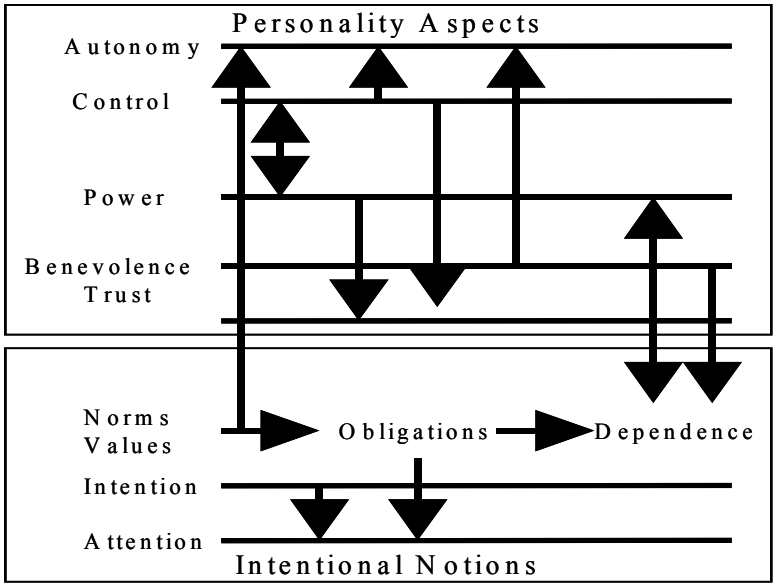

Figure 1 A partial list of "Cognitive dials" of human-agent interactions

In the remainder of this paper we will outline cognitive adjustments of agents in more detail in section 2. In section 3 we will discuss how agents should learn human preferences. In section 4 we will preset a simulation testbed that shows collaboration and exemplifies autonomy and trust as two concepts of improving collaboration. In closing, we will offer some concluding remarks.

2 COGNITIVE DIALS

In some interactions between collaborative humans and agents, it is advantageous to give the human the ability to adjust the level of agent's autonomy. For example, allowing the human collaborator to change the balance of a shared decision making task between herself and the agent by choosing who will perform a task that can be performed by either can enhance the performance of the human/agent group. Later in this paper we will show an implementation of agent autonomy adjustment by a human user. Beyond autonomy, human users might want to adjust the level and type of control of an agent. A user, who desires a higher level of control over the agent, might want more frequent feedback from the agent and also want to delegate more tasks to the agent.

The notions of control and autonomy are inversely related in the sense that an increase in control directly affects autonomy by inducing higher autonomy for the controller and lower autonomy for the controlled. Changing autonomy, however, does not affect control relationships. This is because changes in autonomy, broaden or narrow range of an agent's choices. In contrast, changes in control are interpreted as changes in constraints over choices. Changes of autonomy and control might be offered to a human user as part of its interface to an agent. These changes might be global or only with respect to a specific task. A human user might instruct agents to obey certain levels of autonomy and control globally. The ability to adjust autonomy and control allows human users to tune agents to work in ways that are more comfortable and transparent to the human user.

Besides autonomy and control, a human user might want to alter power relationships among agents. Power and control are directly related [2,18]. By setting agent a in charge of agent b, a has control over decisions with respect to b. A simplified model of power is the assignment of ranks that denote a certain level of power. A user might set ranks for agents to establish control among agents. If the set of shared decisions (or tasks) is manageably small, a user might set ranks among agents separately for each decision. For instance, deciding whether or not to mount an attack might have a completely different rank order than a decision about the movement of certain supplies.

A human user might want to alter trust levels among agents. For simplicity we consider here 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. If we can set up a power relationship, and thereby a control relationship, then why is it necessary to alter trust levels? In general, exercising control incurs a cost that is prohibitive at times. If an agent decides it can sufficiently trust another agent, it can save

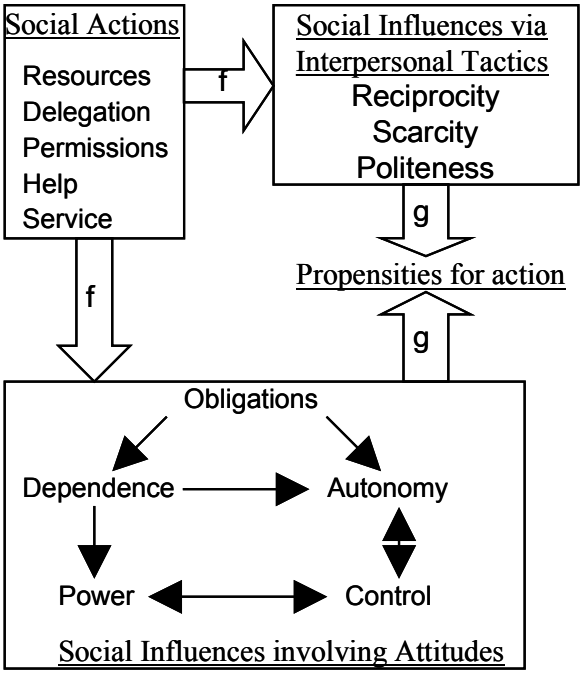

Figure 2 Social actions and influences; values and norms are not shown in the figure

the cost of controlling the agent. The controlled agent can similarly reason about its decisions and if there is adequate trust, it can act independently when the control instructions are not available.

Human users can choose to introduce values and norms into agent behavior as a means to increase agent safety, predictability, and effectiveness. In our previous work we have argued that by setting values and norms, we can control an agent's commitments and obligations. We plan to further design the influence of values and norms on an agent's obligations. Obligations further affect an agent's autonomy and dependence and subsequently an agent's power and control.

3 INTENTIONAL NOTIONS: RESPONDING TO HUMAN REFERENCES

We envision agents to adjust their understanding of a human's preferences for interaction so that their interaction is more cognitively appealing to the human user. Since in general it is difficult to extract the human-desired characteristics of interaction and often these desired methods don't fit a mathematical model, our approach is for agents to use simple human feedback to learn normative interactions that are preferred. Therefore, we have explored methodologies to capture values and norms favored by human users. Whereas values and norms are specific patterns of interactions, social attitudes such as autonomy are more enduring. Agents observing human actions can attempt to determine human attention and intention. If that fails, agents resort to assessing norms and values.

Eliciting user preferences requires modeling the user. Trust and understanding between user and agent can be increased by allowing the agent to get further input from the user about his preferences and desires [3,4]. Preference modeling is being widely used in building automated travel schedules. For instance, Greg Linden's Automated Travel assistant starts with minimal information about the user's preferences, and preferences are inferred incrementally by analyzing the feedback given by the user [12]. Learning from the users feedback as well as giving more appropriate information in subsequent iterations improves performance in the system.

As in our approach, Maes suggested agents that collect a case history of scenarios [14]. She has devised a case retrieval mechanism based on the distance between the current state and each of the past situations the system has stored in its memory using a weighted sum of several relevant features.

Interaction between humans and agents can be seen an exchange of social influences. Influences might also be indirect and of indefinite duration. One type of indirect influence is via changes of attitudes. This is shown as the box in the lower part of Figure 2. These are perceived changes in social relationships that affect an agent's ties. The Figure shows Autonomy, Dependence, Obligations, Control, and Power and salient relationships we see among them. Let's refer to the set of influences as I. Let's refer to the set of social actions as A. We define function f that maps the agent's current beliefs B, a set of currently

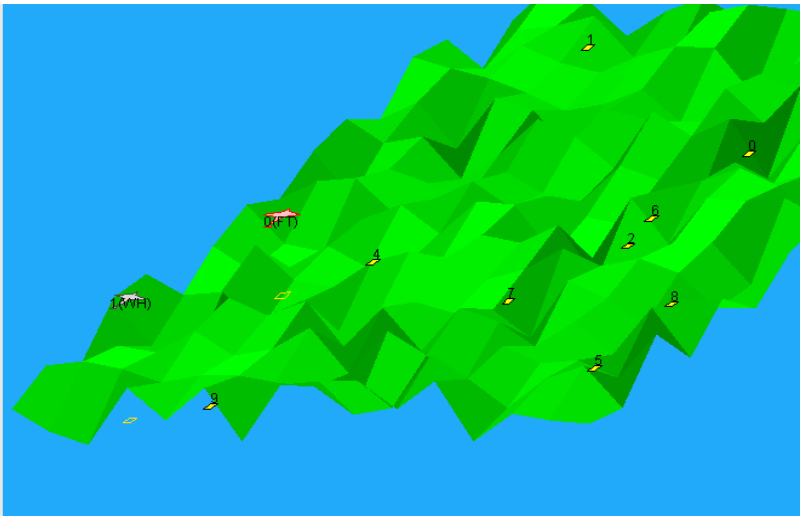

Figure 3 Testbed Screen

active values V, a set N of currently active norms, and a set of social actions A to a set of influences I:

$$f: B \times V \times N \times A \rightarrow I.$$

Agents use influences that result from social actions they experience in their action selection. In addition to social influences, action selection accounts for means end analysis and rationality principles that are governed by the agent's endogenous sources. How action selection is affected by social influences is a complex issue that is beyond our current scope and is denoted as function g in Figure 2. Agents can project such a propensity for action in deciding to perform a social action. The reasoning might also include a chain effect where one agent produces an influence in another, which in turn produces an effect in another and so on. An agent can intend such a proliferation of influences and intentionally start such a chain reaction. This in fact is commonplace in a team setting.

Let's consider a variation in function f where an agent performs a social action (say help), the human user senses a social influence and provides feedback between –10.0 (to indicate disapproval) to +10 .0 (to indicate approval). So now we have:

$$B_{shared} \times V_{human} \times N_{human} \times A_{agent} \rightarrow I_{human} \times \text{Feedback}.$$

The agent can consider the human feedback as the reward. The agent who performed the social action (say help) can now conclude the following simple rule:

$$B_{shared} \rightarrow A_{human\text{-}desired} \times \text{Desirability}.$$

Initially, desirability is unknown and can be randomly assigned for pairs of behavior and action. But as feedback is provided the agent can compute desirability by considering past values of desirability and the amount of reward:

$$\text{Desirability} = \text{Feedback} + (\lambda * \text{Desirability})$$

This formulation of desirability is similar to update function in Q learning. λ is usually set to low levels such as 0.25. We consider the learned rule to be the agent's perception of a norm of its interaction with the human. In this methodology we simply suggest recording some rule-like interactions and have not applied machine-learning techniques (For application of machine learning in user modeling see [1,20]).

4 UCAV TESTBED: PRELIMINARY RESULTS OF ADJUSTING AUTONOMY AND TRUST

We have tested our methodology for discovering intentional notions between human and agents in a testbed of a team of Unmanned Combat Aerial Vehicles (UCAVs). This testbed was also developed to explore teamwork and explicit reasoning about roles [5,6,7]. However, in this section we will not focus on the methodology or teamwork. Instead we will report on our preliminary work on adjusting autonomy and trust.

In the testbed three or more fighter aircraft agents have the mission to deliver a bomb over a remote designated site. There is a one to one relationship between agents and planes. Artificial agents control all the planes except one, which is controlled by a human operator. The human operator controls his/her plane in the field along with the other planes, and will have similar visual and auditory sensing as well as similar flight maneuvering capabilities. The system is implemented in Java. We simulate Surface to Air Missile sites (SAMs), which are randomly generated each time when the program starts running. Figure 3 shows the main simulator screen with one agent-controlled plane flying close together followed by a human-controlled plane.

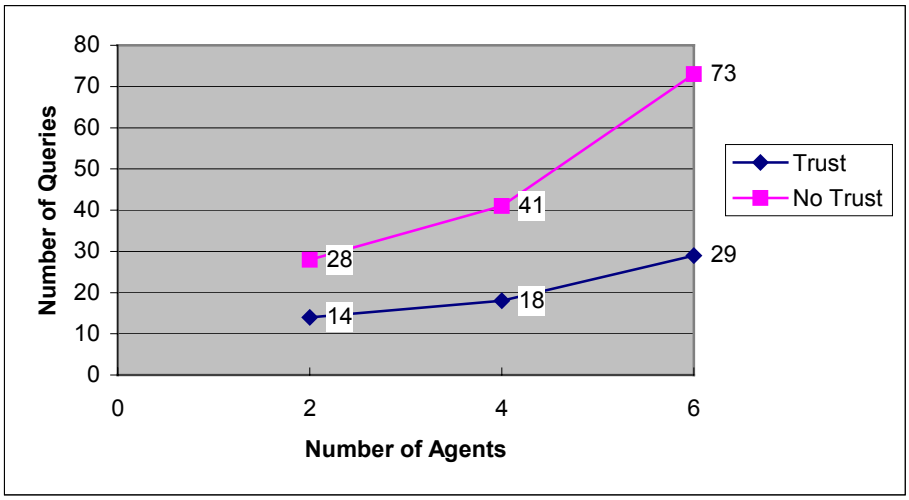

Figure 4 Trust among agents is useful in reducing human supervision

In this paper we highlight two user-agent features. The first is user control of agent autonomy. Second is user control of trust levels among agents. Agent actions are the first five actions listed below plus a default action:

- Attack SAM – an agent performs this action when it sees a SAM
- Avoid SAM – an agent performs this action when its hit-probability exceeds a certain level.
- Help – an agent performs this action to help another agent
- Offer Help – an agent offers help when another agent needs help
- Get Help – an agent performs this action when it needs help from another agent
- Fly to Target – this is the default action performed by all the agents during the simulation run

As agents are created either full autonomy or no autonomy is given for each of the five actions. An agent performs an action when a particular situation arises. Full autonomy means the agent does not have to ask permission from the user. i.e., independently without the human intervention. An agent that doesn't have autonomy must acquire permission from the human to perform an action.

Even with only a few restrictions of actions, the human user can be swamped with requests for permission. This makes the system unmanageable since the human cannot keep up with the questions. To alleviate the human from questions we rely on trust among agents. If two agents trust each other and are in a similar situation then the decision of the human user to perform an action can be shared by both agents. For example, Agent1 might not have the permission to perform the "Avoid SAM" action and seeks permission from the human. Agent2 trusts Agent1 and is in a similar situation, and also doesn't have the permission to perform "Avoid SAM" action. Agent2 can obtain the human permission from Agent1. The human user can still exercise his control over the agents by increasing or decreasing the trust between the agents. The similarity in situation is with respect to certain parameters of an environment in which the agent operates and is different in different environments. Agents interact with other agents they trust before asking the human user for permission, which reduces the number of queries to the human controller. This increases the number of interactions between the agents and reduces the human-agent interaction. Experimental results demonstrating the number of interactions in the trust and no trust conditions are shown in Figure 4. Naturally, the number of queries is proportional to number of agents. However, in absence of trust, each agent interacts with the user independently and the number of interactions is high.

5 CONCLUSION

The work described in this paper highlights critical methods for closing the cognitive gap between human and agents in collaborative work. With proliferation of agents that take up the bulk of dirty, dull, and dangerous tasks, humans in the loop are tasked with supervision. Agents in mixed initiative interaction must adjust their autonomy to allow the human the upper hand. Furthermore, agents must understand human preferences and humans must be able to impart cognitive desires as well as monitor agents. We have outlined a few cognitive aspects of interaction that provide the human with information and ability to guide agents. We have described agents designed with methods that elicit human preferences on intentional notions such as norms and values. We have also illustrated adjustment of autonomy and trust in a simulated Unmanned Combat Aerial Vehicles (UCAVs).

ACKNOWLEDGEMENT

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

REFERENCES

[1] Carberry, S. and L. Schroeder, 2001. "Recognizing and conveying attitude and its underlying motivation", In *2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction, User Modeling 2001*, Sonthofen, Germany.

[2] Conte, R.; R. Falcone; G. Sartor, 1999. "Agents and Norms: How to fill the gap". *AI and Law***,** *special issue on Agents and Norms.*

[3] Fleming, M. and R. Cohen, 1999. "User modeling in the Design of Interactive Interface Agents", In *Proceedings of The Seventh International Conference on User Modeling*, pages 67-76.

[4] Hexmoor, H.; H. Holmback; L. Duncan, 2001. "Detecting, Expressing, and Harmonizing Autonomy in Communication Between Social Agents", *AAAI spring symposium on Robust Autonomy*, Stanford, AAAI press.

[5] Hexmoor, H. 2001a. "Stages of Autonomy Determination", *IEEE Transactions on Man, Machine, and cybernetics*- Part C (SMC-C), Vol. 31, No. 4, November 2001.

[6] Hexmoor, H. 2001b. "A Cognitive Model of Situated Autonomy", In *Advances in Artificial Intelligence*, Springer LNAI2112 -pages 325-334, Kowalczk, Wai Loke, Reed, and William (eds).

[7] Hexmoor, H. and X. Zhang, 2002. "Socially Intelligent Combat Air Simulator", In *proceedings of The Seventh Pacific Rim International Conference on Artificial Intelligence* (PRICAI-2002), Tokyo, Japan.

[8] Huhns, M. and M. Singh, 1998. "Personal Assistants," In IEEE Internet Computing, Vol. 2, No. 5: Sept-Oct 1998, pp. 90-92, IEEE press.

[9] Kassin, S. 2001. *Psychology*, Third Edition, Prentice-Hall.

[10] Kiesler, D.J. 1983. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. Psychological Review, 90, 185-214.

[11] Larson, C.U. 1998. *Persuasion: Reception and Responsibility,* 9th edition. Boston: Wadsworth.

[12] Linden, G.; S. Hanks; N. Lesh, 1997. "Interactive Assessment of User Preference Models:The Automated Travel Assistant", In Anthony Jameson, Cécile Paris, and Carlo Tasso (Eds.), *User Modeling: Proceedings of the Sixth International Conference, UM97.* Vienna, New York: Springer Wien New York.

[13] Loyall, A.B. and J. Bates "Personality-Rich Believable Agents That Use Language", In *Proceedings of the First International Conference on Autonomous Agents*, February 1997, Marina del Rey, California.

[14] Maes, P. 1994. "Agents that Reduce Work and Information Overload". *Commun. ACM* 37,7, 31-40

[15] Ohguro, T.; K. Kuwabara; T. Owada; Y. Shirai, 2001. "FaintPop: In touch with the social relationships", In *International Workshop on Social Intelligence Design, The 15th Annual Conference of JSAI*, Japan.

[16] Moldt, D. and C. Von Scheve, 2001. "Emotions and Multimodal Interface-Agents: A Sociological View", In H. Oberquelle, R. Oppermann, J. Krause, (Eds): *Mensch & Computer 2001*. Tagungsband der 1. fachübergreifenden Konferenz. Stuttgart: Teubner Verlag, 287-295

[17] Rousseau, D. 1996. "Personality in Computer Characters", In *proceedings of the 1996 AAAI Workshop on Entertainment and AI / A-Life*, AAAI Press, Portland, Oregon, August 1996, pp. 38-43.

[18] Tuomela R., 2000**.** *Cooperation: A Philosophical Study***,** Philosophical Studies Series**,** Kluwer Academic Publishers.

[19] Schneiderman, B. 1992. *Designing the User Interface: Strategies for Effective Human-Computer Interaction,* ACM Press.

[20] Webb, G.; M.J. Pazzani; D. Billsus, 2001. *Machine Learning for User Modeling, User Modeling and User-Adapted Interaction* 11: 19-29, Kluwer Academic Publishers, the Netherlands.