Calibrating Trust to Integrate Intelligent Agents into Human Teams

Katia P. Sycara
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
katia@cs.cmu.edu

Michael Lewis
Terri Lenox
Linda Roberts
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA. 15260

Abstract

Complex team tasks, such as joint operation planning, require that information and resources from distributed sources be exchanged and fused because no one individual or service has the collective expertise, information, or resources required. We are at the beginning of a major research program aimed at effectively incorporating intelligent agents into human teams. Our initial experiments use a low fidelity simulation of a target identification task, TANDEM, to investigate human-agent interaction for individual operators and human teams. In the TANDEM simulation subjects identify and choose an appropriate response to a series of air, surface, or submarine targets. Subjects perform a sequence of time critical information gathering and communications tasks in order to decide whether to shoot or clear each target. The data necessary to make a correct decision is distributed among human and in our experiments software agents. Subjects must communicate, coordinate, and negotiate with one another and their agents in order to amass the information needed to classify and dispose of each target. The most transparent presentation led to the best performance and greatest reliance on the simulated agent.

1. Introduction

Integrating agents into human teams presents many challenges including determining agent structure, agent roles, agent persistence, identifying the parameters controlling human-agent and mixed group interaction, and assessing team effectiveness and agent contributions to human-agent teams.

Our research in human-agent interaction is conducted in the context of a multi-agent coordination architecture developed by Sycara [1]. Our model of human-agent teams results from extending both our multi-agent architecture and the team training/performance model for human teams developed by Salas and colleagues [2]. Both approaches revolve around the development of shared models to enable members (both humans and machine agents) to understand the decision making situation, effectively communicate this understanding to other team members, and develop a unified approach to reaching a final team decision. To function effectively within such a team software agents must not only possess common interface languages and protocols for communicating among themselves but also models for discriminating and communicating situational distinctions salient to humans and the team's mission. We believe that co-training and mutual adaptation will be the key to successfully integrating software agents into teams of this sort.

Our first experiment substituted simulated software agents for two subordinate team members in order to investigate the effects of agent communication protocols on human decision making. Subsequent experiments will address interactions involving multi-human/multi-agent teams. In these experiments "Trust" will be manipulated as an intervening variable which enables human decision makers to assess the reliability and meaning of communications from software and human agents. These experiments test hypotheses relating the extent of processing of information, implicit justification through presentation, explicit justification through explanation, and locus of processing unreliability to the calibration of Trust in intelligent agents and other team members.

1.1 Trust, predictability, and coordination

The importance of predictability and modeling of other agents to coordination and reduction of communications has long been a touchstone for research in Distributed Artificial Intelligence (DAI)[3]. When we introduce software agents into human teams, those agents will need models of the team's tasks, the humans' roles, and mechanisms for inferring human and team goals from observable actions. Human members of the team are presented with the corresponding

This research was supported by ONR grant N-00014-96-1-1222

1060-3425/98 $10.00 © 1998 IEEE
Proc. 31st Annual Hawaii International
Conference on System Sciences

problem of identifying a software agent's level of competence, interpreting communications in the context of the agent's functionality and learning the language/protocols they must use to interact with the agent.

While many of the same problems arise when introducing human members into a team difficulties are accentuated for a software agent because:

1) they are heterogeneous w.r.t. human team members, requiring humans to explicitly model their behavior to predict and understand agent actions

2) as flexible automation software agents cannot effect their own integration within a team but must depend on the willingness of human team members to interact with them

3) software agents lack the rich understanding of context expected of humans making it more difficult to instruct them to perform new procedures or to alter their behavior in response to changes in team goals

The first two of these difficulties can be subsumed as aspects of *trust* combining both the sense of predictability (1) and of user acceptance (2). The third difficulty underlies our second research hypothesis that: 1) Difficulty of human-agent interaction will increase with the flexibility/generality of a software agent to a point at which this difficulty can exceed the potential benefits of agent use.

As flexibility is the key characteristic distinguishing "software agents" from other forms of software, our second hypothesis predicts a region of suitability within which software agents enhance team performance and beyond which they may degrade it. We hope to develop a practical framework for measuring and predicting where these boundaries may be and what methods of human-agent interaction and task allocation can be effective in extending them.

Our initial experiments focus at the low flexibility end of the agent continuum controlling for "return on investment" impediments to agent use. Although our first domain, a low fidelity target identification simulation, involves a relatively simple agent task, task complexity and agent flexibility are not necessarily coupled. An autopilot, for example, may perform a complex and varying sequence of actions to achieve an approach and landing yet remain a stereotype, inflexibled instance of automation.

1.2 The Tandem Simulation

A low fidelity simulation (TANDEM) of a target identification task, jointly developed at the Naval Air Warfare Center - Training Systems Division [4] and the University of Central Florida was used to investigate human-agent interaction. In the simulation subjects must identify and choose an appropriate response to a series of air, surface, or submarine targets. Subjects perform a sequence of time critical information gathering and communications tasks in order to decide whether to shoot or clear each target. The data necessary to make a correct decision is distributed among three team members. Subjects must communicate with one another in order to amass the information needed by the coordinating node to make the final decision.

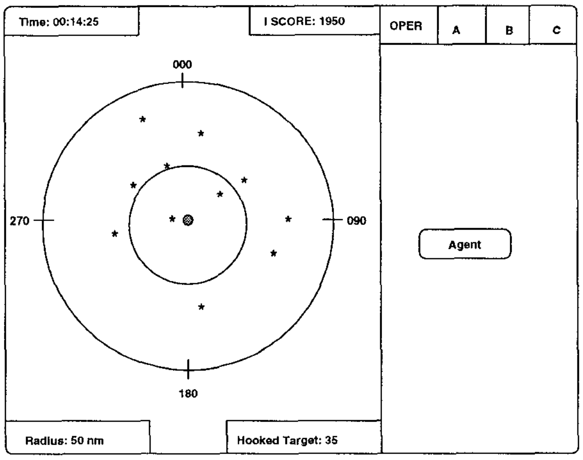


Figure 1. The TANDEM Display

The current experiment substituted simulated software agents for the two non-coordinating team members in order to investigate the effects of agent communication protocols on human decision making (see Figure 1). Subsequent experiments will address interactions involving multi-human/multi-agent teams.

1.3 Manipulating trust

"Trust" is treated as an intervening variable enabling the human decision maker to assess the reliability and meaning of communications from software agents. Following Muir [5], trust is considered to be multidimensional and varying in character from verifiable consistency to blind faith and teleology.

We hypothesize that effective human/agent performance requires a precise calibration of trust so that the decision maker can accurately interpret an agent's communications. This calibration depends on experience with an agent and "explanations" the agent may provide to support its messages.

An agent which displays the minimum value from a bank of sensors, for example, might report the full set of sensor readings as an "explanation" for its message. The decision maker can incorporate this "explanation" into her decision by considering the representativeness of the minimum reading and the likelihood that it indicates an aberrant condition or a sensor failure. A more sophisticated agent might monitor the performance of an internal combustion engine and alert the driver to an imbalance among parameters. Whether this is explained as "high oxygen reading" or "probable timing error" will depend on the sophistication of the user and the character of his "trust" (confidence or faith) in the agent. For an agent which identifies aircraft from radar signatures, presentation of raw or processed images may be ineffective as explanation, forcing the pilot to rely on faith in the agent's design and experience with its reliability.

Table 1 : Trust

Level	Trust Basis	Data
1	Predictable	Data
2	Technical Competence	Inference
3	Faith	Decision Making

Information presentations handled by agents can be classified into three types which roughly parallel the level of trust they must rely upon for interpretation: (1) non-integrative, (2) integrative, and (3) non-decomposable. Non integrative information involves numerical values and text which is self sufficient in meaning and interpretation. To check your savings and checking accounts to see that neither is overdrawn and to transfer money from savings if the checking balance is low is a task which can be supported by non-integrative presentation. If savings, checking, and money-fund accounts were involved with a minimum savings balance for maintaining an array of free services, high interest rates in the money-fund but only above a particular balance, and relatively low interest checking overdraft protection on a credit card, an integrative display showing the relation between the various quantities and balance thresholds (probably in analog form) might be needed to support the decision task. Non-decomposable information is complex verbatim information such as images, video, or computed result which needs to be presented in a particular form or modality.

The reported experiment addresses these hypotheses by pairing error-making and error-free software agents with differing levels of explanation to observe the effects on decision quality, reliance on agent provided information and reported confidence.

2. Methods

TANDEM, A low-fidelity simulation of a target identification task was used in its single user mode for this experiment. In the TANDEM task subjects must identify and take action on a large number of targets (high workload) and are awarded points for correctly identifying the targets (type, intent, threat, etc.). and taking the correct action (clear or shoot). A maximum of 100 points is awarded per target for correct identification and correct action.

In the TANDEM task, users "hook" a target on their screen by left-clicking on the target or selecting "hook" from a menu and specifying a target's unique contact number. Only after a target is hooked can they access information relative to that target. In standard configuration TANDEM consists of three networked pc's each providing access through menus to five parameters relative to a "hooked" target. Their tasks involve identifying the type of contact (submarine, surface, or aircraft), its classification (military or civilian), and its intent (peaceful or hostile). Each of these decisions is made at a different control station and depends on five distinct parameter values, only two of which are available at that station. Subjects therefore must communicate among themselves to assure that they have all hooked the same target and subsequently exchange parameter values to classify the target. It the team finds a target to be military and hostile it is shot, otherwise it is cleared and the team moves on to another target.

In standalone mode all of the information is made available on a single pc with the station specific parameters accessed using three distinct menus. In the reported experiment information relative to four of the parameters needed to make the type (sub surface air) decision was made redundantly available through a "simulated agent". The typing decision was therefore largely determinable from the agent's presentation although in highly ambiguous classifications the user's fifth parameter might be needed to tip the balance.

Subjects could use either the standard TANDEM menus or the data presented to them by the software agent, or both. Due to the high workload conditions, subjects who relied solely the standard menus would be unable to complete all their assigned tasks.

Each agent provided one of three possible levels of information, corresponding to the three levels of trust identified above (Table 1) [5,6] and their associated levels of information presentation. These information levels help the subject identify the TYPE of target and include :

1) aggregated information -- a list of parameters and values; performs simple transformations (climb/dive rate , speed, signal strength ,and communication time);

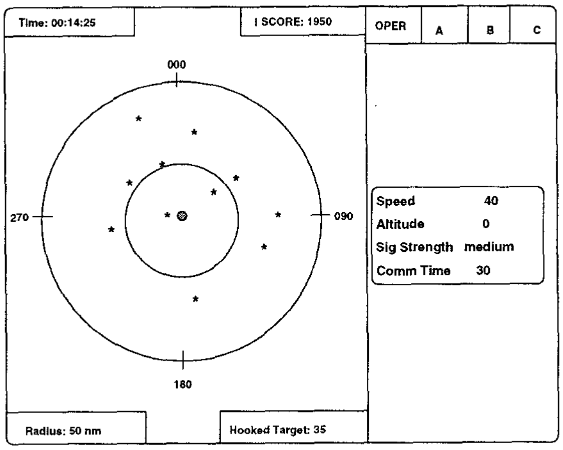
2) inferential information -- a table showing values in their correct cells with the underlying data shown; classifies values;

3) decision information -- a target classification message with a probability associated with it. E.g., target 35 has a .81 probability of being a submarine. Bases the decision using four out the five parameters associated with the target TYPE.

To manipulate the subjects' trust of the agents presentations, errors were introduced in all three levels (Table 2). In the control condition, both menus and agent were errorless (i.e., they will supply accurate values). In error conditions, errors were of three types: data errors, classification errors, or "bad" decision rules. Data errors occurred when the agent displayed different data than is shown on the corresponding menus. This type of error was explained to subjects as problems with the agents sensors. Classification errors occurred when the agent placed data in the wrong column of the display table. For example, a reading which should indicate a submarine is placed into the air column in the table. Decision Rule errors occurred when the agent used inaccurate rules to determine the type of target and a probability value. Classification and Rule errors were explained to the subjects as software problems. Level 1 can experience data errors only; while level 2 can experience data or classification errors; and level 3 can experience data errors, classification errors or "bad" decision rules. Only one source of error was presented during a TANDEM session. Buttons presses to access agent information, menu selections, target hooks, classifications, and final actions and times were collected for each subject.

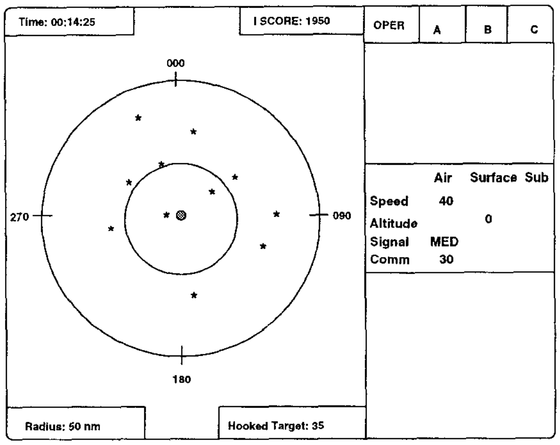
Table 2: Types of Experiments

Errors	Level 1 Agent	Level 2 Agent	Level 3 Agent
None	X	X	X
Data	X	X	X
Classification	X	X	
Decision Rule		X	



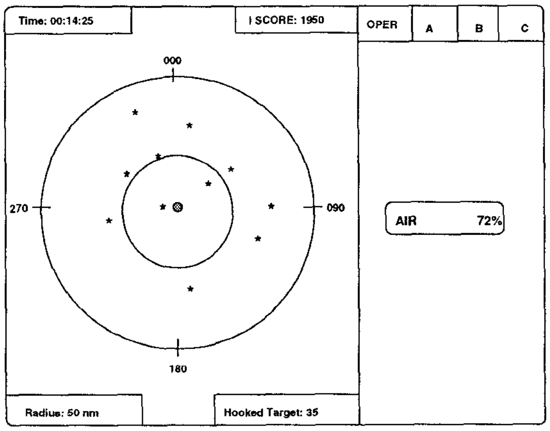
Speed	40
Altitude	0
Sig Strength	medium
Comm Time	30

Figure 2. Level 1 agent



	Air	Surface	Sub
Speed	40		
Altitude		0	
Signal	MED		
Comm	30		

Figure 3. Level 2 agent



AIR 72%

Figure 4. Level 3 agent

Independent Variables

- Level of agent information (aggregated, inference, decision)
- Presence/absence and type of errors

Dependent Variables

Dependent variables included task performance, reliance on information from the agent, and ratings of trust.

Task performance was measured as:

1) Correct identification of targets
2) Timeliness of target identification
3) Reliance on agents
4) Ratings of trust

Targets were distributed in several concentric rings on the screen. The circle closest to the center is referred to as the circle of fear and the amount of time a target spent in this circle before being identified was measured as penalty time. The number of targets identified while in this penalty circle, targets identified outside of the penalty circle, and targets hooked but not resolved were all measured.

Ratings of "trust" of simulated information agents using scales developed and validated by Muir [2] were also gathered from each subject. These ratings, on a scale of 1(low) to 5 (high) focused on issues of dependability, predictability, accuracy, reliability and an overall assessment of trust in the agent.

3. Results

3.1 Performance measures

Performance was analyzed using a repeated measures analysis of variance with session as the within subject factor and types of error and level of agent as between group factors. Effects of session were significant (p < .05) for each of the dependent measures reported. Where differences were found between groups, data was pooled across the two sessions and Post hoc analyses conducted using Tukey's HSD to identify reliable differences among the conditions.

Performance measures fell into three groups which can be roughly categorized as Group 1 - total targets engaged, Group 2 - targets engaged within penalty circle, and Group 3 - use of agent and correct target identification.

The number of targets processed (shot or cleared), the number of targets hooked, and the displayed score (a sum based on correct hooking, classification, and disposal) showed effects for level (p < .04), error (p < .04) and their interaction (p < .04). Subjects using agents without errors processed more targets than those affected by data errors (p=.003).

Table 3. Total targets engaged

Dependent Measures	Agent Level	Agent Error	Level x Error
N targets	.04	.001	.04
N hooked	NS	.001	.01
score	.02	.005	.008

A similarly sized, but non-significant, (p=.07) difference was found favoring target processing by subjects using agents committing rule errors over those committing data errors.

Scores were also higher for subjects using agents without errors (p=.029) and those making "rule" errors (p=.024) than using those prone to data errors. A parallel result for targets hooked finds subjects using agents without errors engaged more targets than those using data error-prone agents (p=.003). Level 2 (table) agents were found to lead to reliably (p=.027) higher scores than Level 3 (probability assignment) agents.

Table 4. Targets Engaged Within Penalty Circle

	Error
non-penalty targets engaged	.001
penalty targets engaged	.001
penalty time	.04

A series of effects were found that were only associated with the presence/absence of errors for several measures associated with a targets presence inside the circle of fear (i.e., penalty circle - - radius closest to the center) . These measures include the number of non-penalty targets engaged (p=.001), number of penalty targets engaged (p= .001) and the penalty time (p=.040). (Penalty time is the time targets are in the penalty circle.)

Subjects had larger mean penalty times when dealing with decision rule errors than when dealing with data errors; and larger mean penalty times when dealing with decision rule errors compared to no errors.

In particular, post hoc comparisons of the total penalty time showed significant differences : 1) between no errors and decision rule errors (p=.009); and 2) between data errors and decision rule errors (p < .01).

Table 5. Use of Agent and Correct Identification

	Level	Error
agent activation	.032	.01
correct identification	.065	.001

Subjects willingness to activate an agent depends on the level of the agent (1, 2, 3) and the presence/absence of errors. We found significant results for both level and errors for agent activation and correct target identification measures for between subjects, but none for the interaction of level and error. For agent activation, the significance levels were p = .032 for agent level and p =.010 for error. For correct identification of targets, effects were found for both level (p=.065) and error (p <

Subjects activated the agents more often in the no error condition than in the data error condition, but activated the third level agent (probability statement) more often than either level 1 (list of data values) or level 2 (table) agents. In particular, post hoc comparisons of the number of times an agent was activated by a subject, showed significant differences : 1) between the no error and data error conditions (p=.021) ; 2) between agent level 1 and agent level 3 (p=.012); and 3) between agent level 2 and agent level 3 (p=.034).

3.2 Ratings of Trust

Ratings of trust in automation using scales developed by Muir [5] were collected at the conclusion of the experiment. Subjects ratings on 10 of the 11 scales were lower (p < .05) for agents committing errors. Ratings of trust were not affected by the level of agent.

4. Discussion

Overall, the level 2 (table) agent seems to provide the best support for the target identification task Although subjects consulted the level 3 (probability assignment) agent more than either of the other two agents, their scores were lower than subjects using level 2 agents and errors in the probability assignment led to longer penalty times. Regardless of their source, errors affected subjects performance, reliance on agents and ratings of trust in a similar manner.

Contrary to our expectations, the level of the agent did not appear to affect the subjects' ratings of trust or penalty times. We suspect that the subjects' unfamiliarity with the task and the structure of the agents may have obscured the expected effects of proper explanation and calibration of trust

In planned experiments, we will investigate these issues in a group context using more complex and realistic tasks and extended training and evaluation. Greater task complexity will allow us to better separate the effects of data transparency from trust in an agent's competence for level 2 agents. Extending our study of trust as intended to the performance of teams will introduce additional issues including including the relative efficacy and acceptance of passive critiquing and active advocacy for level three agents.

5. References

[1] K. Sycara, K. Decker, A. Pannu M. Williamson, and D. Zeng, "Distributed Intelligent Agents", *IEEE Expert: Intelligent Systems and their Applications*, Vol. 11, No. 6, December 1996.

[2] R. Guzzo, E. Salas, and Associates, *Team Effectiveness and Decision Making in Organizations*, Jossey-Bass Publishers, San Francisco, CA, 1995.

[3] Durfee, E. *A unified Approach to Dynamic Coordination: Planning Actions and Interactions in a Distributed Problem Solving Network*, PhD thesis, COINS, University of Massachusetts, Amherst, MA 1987.

[4] J. L. Weaver, B. B. Morgan, Jr., and J. Hall, *Team Decision Making in the Command Information Center : Development of a Low- Fidelity Team Decision Making Task for Assessing the Effects of Teamwork Stressors*, Naval Training Systems Center, Technical Report 92-xxx, February 1993.

[5] B. M. Muir, Trust in Automation, Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems, *Ergonomics*, Vol. 37. No. 11, 1994, pp. 1905- 1922.

[6] J. Lee and N. Moray, Trust, Control Strategies and Allocation of Function in Human- Machine Systems, *Ergonomics*, Vol. 35. No. 10, 1992, pp. 1243-1270.