Towards Deception in Agents

Derrick Ward
Computer Science & Computer Engineering Department
Engineering Hall, Room 313
Fayetteville, AR 72701
+01-479-575-6197
daward@uark.edu

Henry Hexmoor
Computer Science & Computer Engineering Department
Engineering Hall, Room 313
Fayetteville, AR 72701
+01-479-575-2420
hexmoor@uark.edu

ABSTRACT

We have implemented a test-bed with agents who collaborate and communicate, potentially with deceptive information. Agents range from benevolent to selfish. They also have differing trustworthiness levels. Deception handling can vary from agents with mood swings about being deceived, to agents who react minimally to being deceived. We show the effects of different trustworthiness values in different scenarios on group performance. Our agents are designed in BDI paradigm and implement a possible-world semantics.

Categories and Subject Descriptors

H.1. [MODELS AND PRINCIPLES]: General.

General Terms

Experimentation, Theory

Keywords

Trust, Deception, Autonomy

1. Introduction

Without making a moral judgment, deception is commonplace with its detection and handling being crucial in successful interactions. In this paper, we have developed a methodology to examine a range of deception handling for artificial agents, each designed with parametric values for trustworthiness. As we will see from our experiments, the optimal amount of deception varies with the environment and task at hand.

We will begin by describing the simulated test-bed we have implemented. In section three, we will describe our model of trust and deception. In section four, we will describe a series of experiments performed with our simulator that show the empirical results from dealing with deception. In section five, we draw conclusions about deception handling.

--

AAMAS'03, *JULY 14–18, 2003, MELBOURNE, AUSTRALIA.*
ACM 1-58113-683-8/03/0007.

2. Simulated Terraforming Mars

The simulation environment consists of a two dimensional grid. Each element of the grid contains a stack of blocks. There can be 0 through 9 blocks per square (elevation). The agents' goal is to push blocks so that adjacent squares will have the same number of blocks. When four or more adjacent squares known by an agent have the same value, the team gains ownership of those squares.

The agents may only plant solar receptors in owned leveled areas. The energy contained within a receptor can only be harvested once that receptor has charged. Blocks may only be pushed into adjacent squares if the elevation of the square the agent is located in is greater than or equal to the square that is the potential destination of the pushed block.

Agents are modeled in the belief-desire-intention paradigm [2], in which each agent holds sets of beliefs and desires. Agents' beliefs are based on that agent's perceptions of their world. The beliefs, in this case, are about the block levels contained in the grid elements, locations of other agents, trustworthiness of other agents, and utterances from other agents. Desires are the agents' goal states. The goal states for all agents are based on the need to obtain energy. For now, all agents have the same desires regardless of differences in parametric values. Finally, the intentions of the agents consist of a plan selected from a state space using depth first search.

The utility function, UTIL(S) for a state S, consists of the weighted sum of smaller more specific utility functions. The function is of the form:

$$UTIL(S) = [W1 * UTILmove(S) + W2*UTILdig(S) + W3 * UTILplant(S) + W4 * UTILharvest(S)] / 4$$

Wi is a weight between 0 and 1.0 determined by the agent's desired strategy.

3. A Model of Deception

Our agents work as a team, negotiating their plans for simultaneous gains toward sub-goals. By working on individual plans that are compatible, agents can reduce the total work overhead involved. In such group-work, it may be in the interest

of one of the agents to lie about its goals, so as to be given a smaller share of the joint plan to execute.

So far, we have modeled three different types of deception devices for our simulator. These are: (1)Deceit about goals, (2)Deceptions about beliefs, and (3)Passive deception or withholding of information. The deception utility function value for each device d generated for an agent B to deceive an agent A is of the form:

$$\text{Deceive}(d,B,A) = (1 - \text{trustworthiness}(B)) * \text{efficacy}(d, \text{Bel}(B, \text{Bel}(A))) * \text{plausibility}(d, \text{Bel}(B, \text{Bel}(A))$$

where Bel(B, Bel(A)) indicates B's beliefs about A's beliefs. Each function returns a value in the range 0.0 to 1.0. The functions are equally weighted, so that no matter how beneficial the lie could potentially be, if it is not plausible, or if the agent desires to be trustworthy, the lie will be less likely to be communicated (and vice-versa). No deception device will be communicated if the maximum value Deceive(d,B,A) for any d is < 0.5. Our development of efficacy and plausibility is inspired by [1].

Efficacy is the potential benefit for the deceiver if the agent to be deceived accepts the deception into its beliefs.

Plausibility is the potential acceptance by B of a deception by communicated by A. Currently, this is simply computed as a function of the difference between what A wants B to believe and what B actually believes, as well as the physical distances between the agents on the grid. If the agents are close together and A attempts to deceive B, it is less plausible that B will believe A because information about the deception object is easier for B to directly access. Computing the difference in beliefs allows A to reject deceptions that will make him look too deceptive.

4. Experiments and Discussions

Figure 1 shows the performance of agents based on 10 trials with two agents, where each trial consists of 400 time-steps. Three different groups of agents were used: (1) No deceptive agents, (2) one deceptive agent, (3) two deceptive agents. Average Gain of Ownership per Timestep (AGOPT) is used as the performance metric.

Finally, two different scenarios were used for agent ownership. In scenario 1, agents could not plant/harvest in other agent's owned squares. In scenario 2, owned areas can be used by all. This has the effect of causing apathy towards digging and leveling land among the agents, hence the lower performance.

The results show that when there are deceptive agents, the deception is only effective when one agent is honest and trusting towards the deceiver. In other cases, there is either no deception, or the agents are both deceptive. These cases had similar results since deception is either non-existent or ineffective. For instance, when both agents are deceptive, the agents quickly become distrustful of one another and cease communication.

In scenario 1, when deception was effective on one agent, it simply had the effect of leading the agent away from its goals, distracting the agent from completing its partially completed tasks. On the other hand, in scenario 2, the inverse occurred. The deception, when effective, tricked the otherwise relatively inactive agent into traveling to unleveled areas. When in these new surroundings, the agent was more likely to have maximized dig utilities, which leads to more work being accomplished.

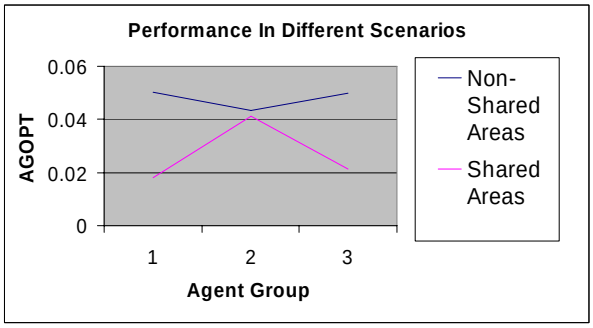

Figure 1: Performance

5. Conclusion and Future work

We have developed an agent simulator to test effects of deception using our suggested model of deception. Our preliminary results show that agents communicating deceit between one another can have positive as well as negative effects on group performance. Empirical results can guide the way to appropriate levels of deception, which can be introduced to the system.

In the future, we would like to test performance using fuzzy trustworthiness values. At the present, trustworthiness values are crisp: all agents are either compulsive liars or totally trustworthy at all times. In the results shown here, agents only deceived one another about their current beliefs. Since all of the results presented in this paper were achieved using only two agents, we would like to see how the effects vary over a larger group of agents.

6. Acknowledgements

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

7. References

[1] Carofiglio V., and de Rosis, F. Ascribing and Weighting Beliefs in Deceptive Information Exchanges, M. Bauer, P.J. Gmytrasiewicz, J.Vassileva (Eds.), User Modeling 2001, LNAI 2109, (2001), 222-224, Springer.

[2] Wooldridge, M. Reasoning about Rational Agents, The MIT Press, 2000.