

Coping with Deception

Derrick Ward and Henry Hexmoor
 Computer Science and Computer Engineering Department
 Fayetteville, Arkansas 72701
 daward@uark.edu, 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.*

1. Introduction

In the movie “Liar, Liar”, Jim Carrey plays lawyer Fletcher Reed, a man who never has time for his son Max. At Max’s birthday party, he wishes that his father could never tell a lie. When the wish becomes true, Fletcher’s life quickly spirals out of control. The movie satirizes the social norms of common deception by showing that truthfulness is not necessarily beneficial and lies are not necessarily harmful. Fletcher and those who knew him had been used to dealing with his deceptions. 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 altruism and trustworthiness. As we will see from our experiments, the optimal amount of deception varies with the environment and task at hand.

Previous work in this area includes Castelfranchi’s experiments on help pilfering in the GOLEM environment [1]. Experiments on deceit between agents cooperating, or pilfering help from one another in order to achieve a goal, were then simulated. Deceit about obtaining help was divided into 3 categories: deception about capabilities, deception about personality, and deception about goals. All possibilities for each help deception case were examined between two agents.

The results showed that (1) in order to perform deception an agent must have beliefs about the beliefs of the agent they are deceiving, and (2) deception is not necessarily non-cooperative toward the agent being deceived.

In contrast, Carley and Prietula performed experiments to determine the effects of trust and rumor on rational cooperative agents [2]. Each agent’s task was to find certain blocks stacked in isles in a warehouse. To achieve this goal, the agents communicated advice amongst themselves [3].

If this advice is determined to be untruthful by the advisee, it may modify its perceived trust for the advisor agent and begin to communicate rumors about the advisor. The rumors generated by the advisee will in turn cause other agents to modify their perceived trust for that original advisor agent. Trust models for the agents could vary by adjusting the trust adjustment rates. For instance, agents that update all of their trust each round were known as reactive, agents that updated only a portion of their trust per round were known as forgiving, and agents that never changed their trust were either always distrusting or always trusting. Finally, some agents could be deceptive, intentionally communicating false rumors and advice. The simulation was then run using five agents with varying trust models, numbers of deceptive agents, rumors or no rumors, and environment stability. Different benchmarks were also used: amount of work performed, information withheld from agents not trusted, conflict with other agents, and duration of information coalitions. The results showed that forgiving agents perform better and withhold more information than reactive agents in unstable environments with deceptive agents. In other words, as stability increased and deception decreased, the performances by reactive and forgiving agents became similar.

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.

A is an agent, t is the current time-step, $gain$ is a domain specific function that approximates how beneficial the verified communication is, and $investment$ is a domain specific approximation of how much work the communication caused the agent to perform. The $rate$ is a value in the range 0.0 to 1.0, which specifies what percentage of an agent's trust should be modified in each time step. Low $rates$ produce an agent who is reluctant to changes in trust, medium $rates$ cause the agent to "forgive" others, while large $rates$ cause an agent to have reactive trust beliefs towards others. The verification of other agent's utterances must be performed by comparing information in the agent's communication table, with beliefs perceived directly from the environment. For simplicity, no rumors about other agents' trust are communicated.

Agents have three discrete levels of deception. Correspondingly, there are three categories of deception devices, the means of deceiving another agent, which can be deployed:

- (1) Deceptions about goals,
- (2) Deceptions about beliefs, and
- (3) Passive deception or withholding of information.

For example, a category 1 deception device might consist of a communication by the deceiver about false elevations for squares on the grid. A category 2 deception device might consist of a fake plan to help another agent level an area.

The lower an agent's own trustworthiness, the more types of deception devices that agent is willing to use. The frequency an agent is willing to rely on deception also increases with that agent's lack of trustworthiness. Altruism is the motivation for the deception, determining whether to use devices that could help the team, or devices that would just help the agent. Whether an agent will resort to deceit is determined by a utility function for each deception generated. This utility must be computed by A for each agent A', since a deception object is generated for these agents based on the beliefs of A about each other agent's beliefs.

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 $\text{Bel}(B, \text{Bel}(A))$ indicates B's beliefs about A's beliefs. Each function returns a value in the range 0.0 to 1.0. Thus, if the $efficacy$ is high, but $plausibility$ is low, the product will cause those two functions to balance one another. If $efficacy$, $plausibility$, and $trustworthiness$ are all values approaching 1, then the

$trustworthiness$ will equally weight the other two values. As a result of this, a trustworthy agent is not as likely to deceive, and vice versa. The deception device d with the maximum deception value is the one chosen to communicate. No deception device will be communicated if the maximum value is < 0.5 . Our development of efficacy and plausibility is inspired by [6].

4. Experiments and Discussions

This section reports on the experiments performed using the terraforming agent simulator. All experiments were performed using two agents and 400 time steps. The results were then averaged over ten trials. Trustworthiness values were changed from experiment to experiment, and two different scenarios were used. In scenario 1, agents are not allowed to share owned areas: if an agent detects land, it must be owned by that individual agent in order for it to be used by that agent for planting or harvesting. In Scenario 2, agents are allowed to share owned areas: if an agent detects leveled land, it may plant and harvest on that land regardless of which agent the land was originally claimed by since land is owned by the team, not an individual agent.

As of this writing, we have tested deception using only type 1 devices, which are based on deception about agent's beliefs about the current state of the environment. Two different deception devices within this category were used: (a) land ready to be harvested owned by the deceiver and (b) land that needs to be leveled. Deceptive device (a) has a completely different meaning in the two scenarios. In scenario 1, this device will cause other agents to stay away from the land the deceiver claims is owned and ready to be harvested. This deception appears to be selfishly motivated by the deceiver. In scenario 2, this same device appears altruistic since as shown below, it tricks the other agents into pursuing the harvest, and in the meantime possibly leveling more land.

We plotted each agent's trust values, and the sum of all squares owned by all agents at a given time. The trust rate used for all agents was a relatively reactive value of 0.3333. The deceptive agents were given trustworthiness values of 0.1 making them near compulsive liars. All non-deceptive agents were totally trustworthy with values of 1.0. The agents begin the game with total trust in all other agents (a value of 1.0 for each other agent). All agents used the same altruistic strategy in these experiments, although as stated, the different scenarios may cause deceptive agents to act selfishly.

Figures 5a-c show the results of agents operating in scenario 1, non-shared areas. As expected, the best performance occurred when there were no selfish deceptive agents as shown in Figures 5a. In the cases

that one or both of the agents were deceptive, performance was somewhat less than with honest agents, but with little performance difference between themselves with two deceptive agents having the better performance of the two. The performances shown in figures 5b and 5c are practically identical.

In scenario 2, shared areas (Figures 6a-c), normally the agents are totally incompetent, usually unable to gain ownership of more than a few parcels of land (Figure 6c). This is because the strategy used by the agents is unsuitable for dealing with the shared land scenario: agents will simply crowd around the first area owned and harvest and wait to harvest again. The agents either eventually ran out of energy and died, or spent all their time harvesting just to survive. In this case, deception can help in two ways. First, deception causes agents' trust values to decrease, which is good in this case because it decreases the communication between agents. Without communication, the agents will not be as likely to know about other leveled areas and will seek to dig their own areas. Secondly, the agents can trick others into exploring the landscape causing them to find more areas to level. On the other hand, the results shown in Figure 6b show that too much deception is detrimental to the team. With little trust for each other, the agents will not communicate at all, reducing the potential benefits of being able to trick others. The middle case for deception in this scenario, shown in Figure 6a, ends up being the best, but not quite as good as any of the performances with non-shared areas. This was almost the exact inverse of the performance by the agents in scenario one where the performance was the worst with one deceptive agent. In conclusion, although the agent strategies are inadequate for this scenario, by tricking each other they

are able to perform almost as well as they could with a good strategy, and much better than without any deception at all.

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 performance. Instead of individual gains from deception, we have focused on group performance as our objective measure of effects of deception. 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. We would like to run this same experiment with agents deceiving about their intents (deception type 3 in section 3). 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. Finally, we would like to measure not only the achievement of goals (ownership), but also cohesion, the ratio of the number agents that are following the current team plan divided by the number of agents not following the current team plan [7]. This would give us a clearer picture of how deception affects the cooperation among agents.

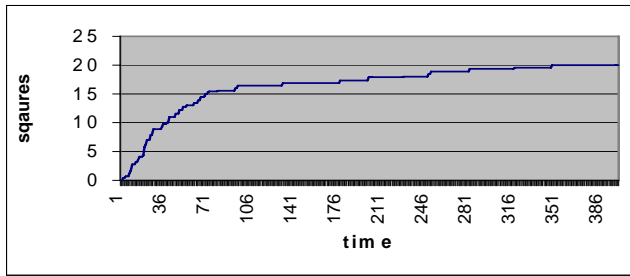


Figure 5a - Ownership, No Deception, No Shared Areas

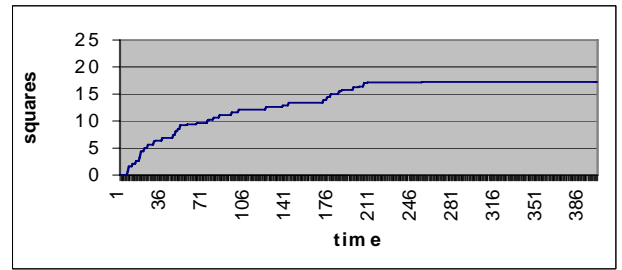


Figure 5b - Ownership, One Deceptive Agent, No Shared Areas

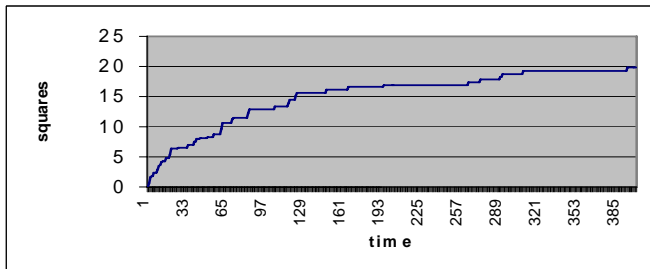


Figure 5c - Ownership, Two Deceptive Agents, No Shared Areas

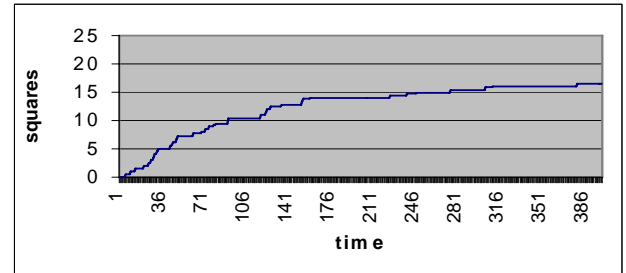


Figure 6a - Ownership, One Deceptive Agent, With Shared Areas

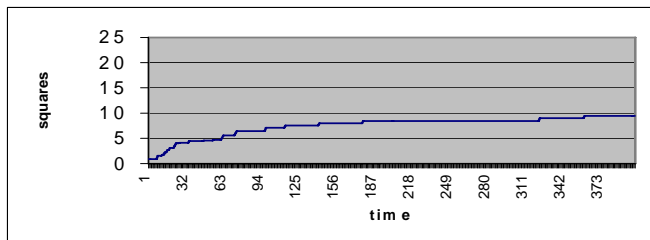


Figure 6b - Ownership, Two Deceptive Agents, Shared Areas

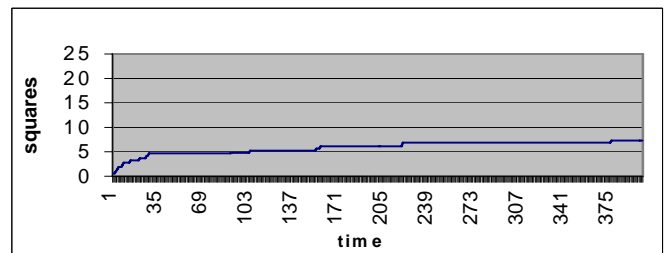


Figure 6c - Ownership, No Deception, Shared Areas

Acknowledgements

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

References

[1] C. Castelfranchi, R. Falcone, and F. de Rosi, "Deceiving in GOLEM: how to strategically pilfer help", In: *Deception, Fraud and Trust in Multiagent System*", Y. H. Tan and C Castelfranchi (Eds), Kluwer Publishing, 1998.

[2] M. Prietula, K. Kathleen Carley, "Boundedly Rational and Emotional Agents Cooperation", *Trust and Rumor*, In *Trust and Deception in Virtual Societies*. Edited by Cristiano Castelfranchi and Yao-Hua Tan, 2000.

[3] P. Gmytrasiewicz, E. Durfee, "Toward a Theory of Honesty and Trust Among Communicating Autonomous Agents", *Group Decision and Negotiation*, 2:237-258, 1993.

[4] M. Wooldridge, "Reasoning about Rational Agents", *The MIT Press*, 2000.

[5] A. Birk, "Boosting Cooperation by Evolving Trust", *Applied Artificial Intelligence*, 14, pages 669-784, C. Castelfranchi, R. Falcone, B. Sadighi F., and Y.Tan (Eds), 2000.

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

[7] G. Kaminka G., M. Tambe, "Robust Agent Teams via Socially-Attentive Monitoring", *Journal of AI Research*, 105-147, 2000.