Deception as a Means for Power among Collaborative Agents

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

Abstract

We have developed a method for measuring the social power that agents exercise upon one another due to the influence of communicated messages. Our social power is a measure of the damage that a communication inflicts upon another agent due to the agent's revision of beliefs and plans due to the received message. This measure of power was used in several experiments involving deceptive agents, as deception is used to gain power over others. Our results illustrate the effects of deceptive communication on agent decision making and ultimately agent performance.

1. Introduction

It is inevitable that competing agents will have plan collisions that will keep one or all agents involved from achieving their goals. If agents are equipped with limited or non-existent negotiation abilities, these power struggles are unavoidable. An alternative for agents to deal with such situations is communicated deception. A selfish deceptive agent, for instance, would be motivated to generate deceitful communication to influence others or gain power over others in an attempt to alleviate conflict. Deception, however, comes with a heavy cost, as distrust and its implications are unavoidable.

The amount of social power agents exercise using communication, deceptive or otherwise, can be measured as a difference between the agent's utility value before it has revised its beliefs based on the information contained in a received message, and the agent's utility value if the information in the message had been truthful. In other words, power as described here, is not perceived agent power in the traditional sense, such as that gained by physical force or leadership roles. This measure of power exercised by agents via communicated messages.

In a previous paper, we discussed preliminary experimental results using a simulator [4]. In our experiments in the Mars Terraforming Agents, two rational agents generated plans with the goal of gaining energy. Agents gained energy by pushing blocks so that adjacent squares had equal numbers of blocks. These leveled areas are then obtained or become owned by agents so that solar receptors can be installed to gain energy. The effects of deception on performance were examined between the two agents.

We have since performed several follow up experiments using the Mars Terraforming Agents simulator. Three different types of agents, where agent type is based on agent credibility (*trustworthiness*) or tendency to deceive. Experimental results show the relationship between agent deception, power, and performance in the simulator. Section 2 gives an overview of agent communication and deception, section 3 describes our model of power, while section 4 shows the experimental results.

2. Agent Communication and Deception

Agents used a simple communication language, which allows agents to only express information specifically about the terrain in their environment, or blocks in adjacent squares. This language was kept simple in order to reduce the complexities involved in agents' deception generation.

Definition 2.1: A *communication message*, denoted by *C*, is a four tuple.

$C = \langle T,L,A,A' \rangle$

Where *T* is *terrain data*, *A* is the sending agent, and *A'* is the receiving agent. Each round, an agent may decide to generate a *communication message*, which is then submitted to the simulator. The simulator then delivers the message to the requested destination agent or addressee. The addressee may or may not decide to read the message based on factors such as trust for the speaker, and the agent's need for information. If the addressee does decide to read the message, the addressee will use the *terrain data* that the *communication message* contains to revise its beliefs.

(a) *terrain data (T)*

3/0/0	6/0/0	4/1/20
1/0/0	**0/0/0 agent**	4/1/19
3/0/0	6/0/0	4/1/18

<4,8> 5 1

b. Location (*L*), sending agent (*A*), receiving agent (*A'*)

Figure 1. An example of a *communication message*, with *terrain data* shown as elevation/ownership/harvest. The agent has placed *terrain data* about its location and its 8 adjacent locations into the *communication message*. The simulator sets the sender field and sending location field.

Agents normally intend to collaborate with other agents by sharing *communication messages* containing *terrain data* that is believed to be truthful and of use to other agents. Agents expect the same in return. As all agents in the simulator are individualistic except for this behavior, deceptive communication is used by agents only for personal gain (as opposed to altruistic gain for instance). Deceptive agents generate false *terrain data* with the selfish intent for the false generated *terrain data* to become part of the addressee's beliefs and either lead them away from the deceiver's goals or hide the deceiver's goal related object. However, unintentional deception, which results from misunderstandings, is circulated among all agents regardless of agents' *trustworthiness*.

2.1 Unintentional Deception

Unintentional *deception* usually results from outdated beliefs regarding agents communicating truthful information about *terrain data* that is received and added to the beliefs of an addressee agent. If the addressee does not get a chance to verify its belief about this information using its sensors until after other agents have modified the information (which would make the addressee's belief outdated), the information is perceived as intentionally deceitful by the addressee. Unintentional *deception* is unavoidable at all times during the game, but increases as the game goes on and agents' non-null beliefs about areas become increasingly erroneous. Unintentional deceit also increases due to the spreading of false "rumors" about *terrain data* obtained from deceptive agents. This deceitful information may originate from an un*trustworthy* agent, or from a well-meaning honest agent with outdated or received false information.

2.2 Intentional Deception

As mentioned above, intentional deception is used by agents with the selfish intent to achieve personal goals at the expense of others. Agents deceive by generating false *terrain data* that if received via a *communication message* and accepted into the beliefs of the addressee, would coerce the addressee agent into avoiding a *terrain data* object desired by the deceiver (such as an area that is almost leveled). Deceivers determine the optimality of a deceptive *communication message* through simulation of the addressee's reaction to the deceit as in [2]. Deception is easy for agents in this simulator to generate if agents have knowledge of one another's strategies. This knowledge gives agents knowledge of which *terrain data* patterns increase or decrease other agents' utility values allowing them to predict through simulation the reactions of others with great accuracy.

The tendency for an agent to deceive intentionally in our model is determined by an agent's *trustworthiness* value. Agents with a *trustworthiness* value of 1.0 never intentionally deceive, while agents with a *trustworthiness* value of .1 deceive more frequently and more blatantly than agents with a *trustworthiness* value of .5.

3. Social Power

Social power in the Mars Terraforming Agent Simulator is a measure of the speaker agent's influence over an addressee's plan via a sent *communication message*. This influence leads the addressee agent to generate *plans* based on incorrect beliefs about *terrain data*. These *plan*s often cause damage to agents by leading agents to follow courses of action that are not as benefficial to them as a *plan* generated using correct beliefs would have been. This power of communication influence is intrinsic [3], as agents could choose to ignore messages and not be influenced by them. Although other notions of power exist in the simulator, such as the extrinsic power of an agent over another agent because of phsyical force imposed by performed actions, the power exercised by agents using communication is of primary relevance to this project so that the role of deception in agent decision making can be measured. Essentially, *power* as described here, is the power a *communication message* has over an agent rather than the power an agent has power over another agent. The agents themselves are not even aware of this power (although agents attempt to estimate the power a communication message could potentially exercise over another agent).

In [1], several formulas were described for computing the extrinsic power that an agent exercises over other agents because of dependencies resulting from having joint plans. We have made slight modifications to these formulas in order to compute the intrinsic power an agent excercises over another due to communication.

Definition 3.1: An agent *Aj* exercesises power over an agent *Ai* if for a *communication message* received by *Ai* from *Aj*,

$UTIL(Si^*) > UTIL(Si^{*'})$ and $UTIL(Si^{*'}) <= UTIL(Si^*)$

for all plans pi' where $Si^{*'} \hat{I} pi'$, and $Si^* \hat{I} pi^*$, and pi' was generated after *Ai* revised its beliefs based on *terrain data* received from *Aj*, while pi^* was generated without *Ai* revising its beliefs based on *terrain data* received from *Aj*.

Definitition 3.2 : The cost incurred by an agent *Ai* while exercising power over an agent *Aj* by means of a sent *communication message Cij*, denoted by function *Cost(Ai,Aj,Cij)*, is

Cost(Ai,Aj,Cij) = UTIL(Si) – UTIL(Si*')*

Where *Si*' Î pi'*, and *Si* Î pi**, and *pi*, *pi'* are *plan*s generated by *Ai*.

By the original definition in [1], where conflicts among multiple agents' plans were the mechanism for power, an agent might choose a plan that has a cost to itself, but a greater cost to the other agent, in order to gain power over the other agent. In our system however, this cost is always 0 since sending a *communication message* will have no immediate effect on the plan generation of the sender (although eventually it could indirectly have costs to the agent if the *communication message* causes him to be distrusted). The receiving agent, however, may revise his beliefs to contain the *terrain data* communicated, which may damage that agent by leading him to generate a less beneficial plan.

Definition 3.3: The damage, denoted by a function *Damage(Ai,Aj,Cij)*, inflicted by an agent *Aj* on an agent *Ai* by means of a *communication message Cij*, is

Damage(Aj,Ai,Cij) = UTIL(Si) – UTIL(Si*')*

Where *Si*' Î pi'*, where *pi'* is a plan generated by *Ai* where *Ai* has not revised its active beliefs based on *Cij* and *S** Î *pj**, and *pi* is a plan generated by *Ai* where *Ai* has revised its active beliefs based on *Cij*.

Definition 3.4 : Amount of power, denoted by function *Power*, is the total power an agent *Aj* exercises an agent *Ai* by means of a *communication message Cij* received by *Ai* from *Aj*.

power(Ai,Aj,Cij)=Damage(Ai,Aj,Cij) - Cost(Ai,Aj,Cij)

It should be noted that an agent can (and often does) have negative power over another agent if the *communication message* is helpful (causes negative *damage*) to the addressee. Thus, while utilites range from 0 to 1.0, the *amount of power* exercised by an agent can range from –1.0 to 1.0. Positive power indicates that an agent is influencing another agent into acting against its strategy. In contrast, an agent exercising negative power over another agent has influenced the other agent into acting "super-normally". In other words, enhancing the agent's normal behavior by providing more useful (desirable) information allowing the agent to improve upon its original plan. Following this, if an agent *A* has *x* power over agent *B*, then *B* does not necessarily have *–(x)* power over *A*. In fact, *B* could exercise *x* power over *A* as well.

For an example scenario where one agent has power over another agent, consider a case where an agent *A* has generated a *plan left->north->dig* where the *UTIL(SA*)=.5*, where *SA** is the possible world or state of the environment created if the agent performs the dig action. On the next turn, *A* receives a *communication message* from *A'* and decides to revise its beliefs using the received *terrain data*. Since the revised beliefs conflicted with the beliefs upon with the original plan was based, the agent generates a new *plan* based on its new beliefs: *left->south->dig*, where *UTIL(SA*')=.7*. Unfortunately for *A*, the data received from *A'* was outdated, and *UTIL(SA*'')*, where *SA*''* is a state based on the actual *terrain data* rather than either agent's beliefs about *terrain data*, is only *.3*. Thus, *A'* inadvertently has *.5-.3 = .2* power over *A* at that moment.

4. Results

Two sets of experiments were performed. In the first set, *trustworthiness* was varied in order to measure the effects of the resulting deceptive behavior. In the second set of experiments, the performance of agents with only partial information about the strategies of other agents was measured against varying *trustworthiness* values among agents in order to determine the role this knowledge plays in generating *deception* that is effective in influencing others.

4.1 Experiment Set One

In the first set, five experiments using five agents each were recorded. In each experiment, *trustworthiness* parameters are were varied in order to determine the effects of deception on performance. *Trustworthiness* parameters for agents in experiments were varied from 1.0, to .5, to .1, having the effect of increasing deceptive behavior with each experiment. Three experiments were performed in which all participating agents had the same *trustworthiness* values. Two additional experiments were performed in which agents 3-5 were 1.0 *trustworthy* or the "honest" group, while agents 1-2 were the deceptive group (both being .1 or .5). The five experiments are denoted *all agents 1.0 trustworthy*, in which all agents only communicate data believed to be truthful (0 deceivers), *agents 1-2 .5 trustworthy*, in which agents 1 and 2 have *trustworthiness* values of .5 and agents 3-5 are totally honest (*trustworthiness* of 1.0), *all agents .5 trustworthy*, in which all agents have *trustworthiness* of .5, *agents 1-2 .1 trustworthy*, in which agents 1 and 2 have *trustworthiness* values of .1 and agents 3-5 are totally honest, and finally, *all agents .1 trustworthy*, in which all agents had *trustworthiness* values of .1. Twenty trials were perfomed for each experiment, with each trial lasting 1000 time-steps. The *terrain data* matrices (game board) are size 12x12.

Amount of power as shown in figures 2,4, is computed as the average power an agent exercised each time-step of a game. Performance is measured using *AGOPT*, the average gain in

ownership (of *terrain data* elements) per time-step. This particular measurement of performance is used since agents have the goal of gaining the ownership of as many *terrain data* elements as possible in order to harvest them and obtain energy.

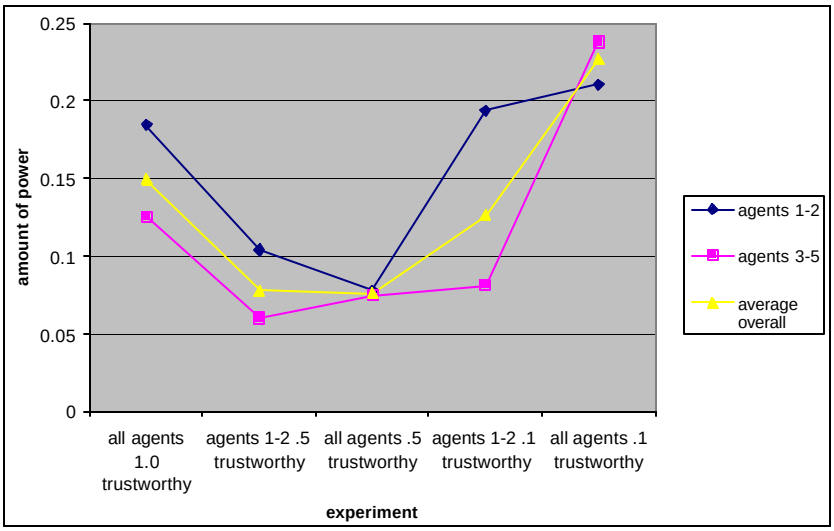

Figure 2. Average amount of power each agent exercised over all other agents.

In experiments with deceiving agents, as deception increased, agents were able to gain more power over others (figure 2). This is especially apparent in the experiments with mixed agents; in particular experiment *agents 1-2 .1 trustworthy*, in which the deceptive agents exercised *138.512%* more power than the honest agents. This occurs since the lower an agent's *trustworthiness* value, the more likely that agent is to use deception to exploit other agents, thereby exercising greater power over others. Similarly, the decreased *trustworthiness* results in decreased reciprocal behavior, specifically sharing helpful information with another agent (which would result in negative power). In experiment *all agents 1.0 trustworthy*, however, this does not apply, as agents were distrusted *88.143%* less than the deceptive agents in other experiments due to high *trustworthiness*, allowing agents to exercise small amounts of power between one another via unintentional deception for longer periods of time than usual. In fact, agents in this experiment could potentially have exercised the largest amount of power over all experiments, but because of the large amount of negative power exercised due to the reciprocal behavior of agents, the average power is still lower than that of *all agents .1 trustworthy*.

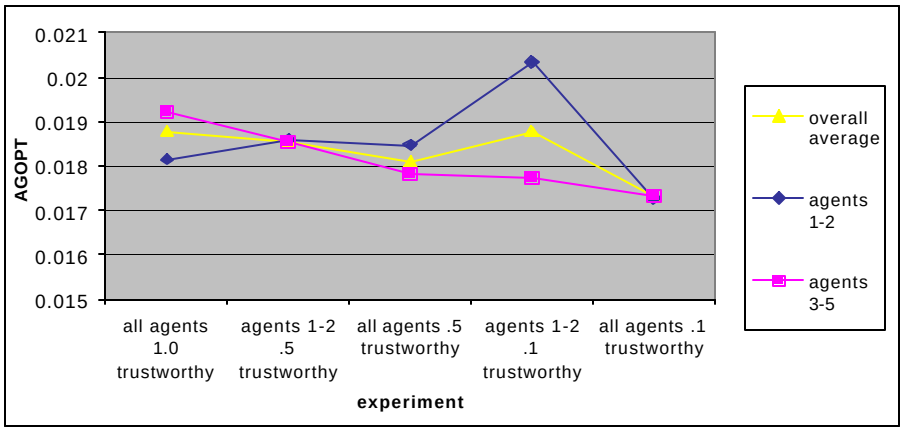

Figure 3. Average agent performance per game.

Comparing the progression of power among experiments as deception increases to the progression of performance

among experiments as deception increases, it is clear that increased power does not necessarily translate into increased performance for agents. When agents intentionally exercised power over others, their performance increased unless the agents exercising power in turn had significant power (an amount of power similar to what they are exercising) exercised on them. Hence, in experiment *all agents .1 trustworthy*, despite the fact that agents maintained the largest amount of power among all experiments, agents performed the poorest overall experiments. As before, the results of experiment *all agents 1.0 trustworthy* do not apply, as agents exercised small relatively harmless amounts of power over one another for long periods of time.

4.2 Experiment Set Two

In the second set, two experiments from the first set have been duplicated in order to determine how a deceiving agent's knowledge of other agent's strategies affects their ability to exercise power (see section 2.2 for more details on deception generation). Agents 3-5 have been given a slightly different strategy than used before for the generation of plans. Neither group has knowledge of each other's strategies.

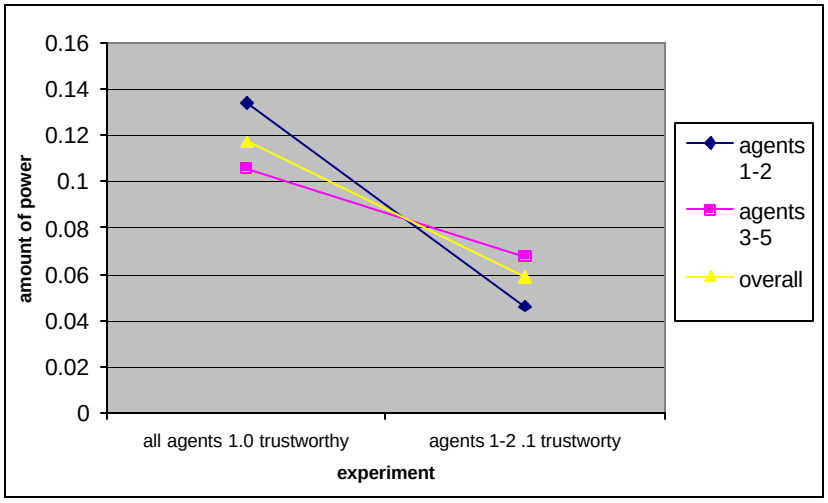

Figure 4. Average amount of power each agent exercised over all other agents.

Figure 4 shows how power exercised by both groups of agents decreased *49.61%* overall in experiment *agents 1-2 .1 trustworthy*, mostly due to a decrease of *65.63%* in the amount of *power* exercised by agents 1-2. There are two reasons for this: (1) agents 1-2 are *distrusted 169%* more often than agents in experiment *all agents 1.0 trustworthy* and (2) deceivers are not able to simulate optimality of deceptive *communication messages* effectively. The latter of these factors is the most important, since in the equivalent experiment from the first set of experiments (*agents 1-2 .1 trustworthy*), agents 1-2 were distrusted a similar amount of time, but were still able to gain large amounts of *power* over others due to the usage of *deception* during the period they were still trusted. The disadvantage to deceivers is especially apparent when compared to the first experiment, in which deceivers exercised *320.38%* more *power* than shown here.

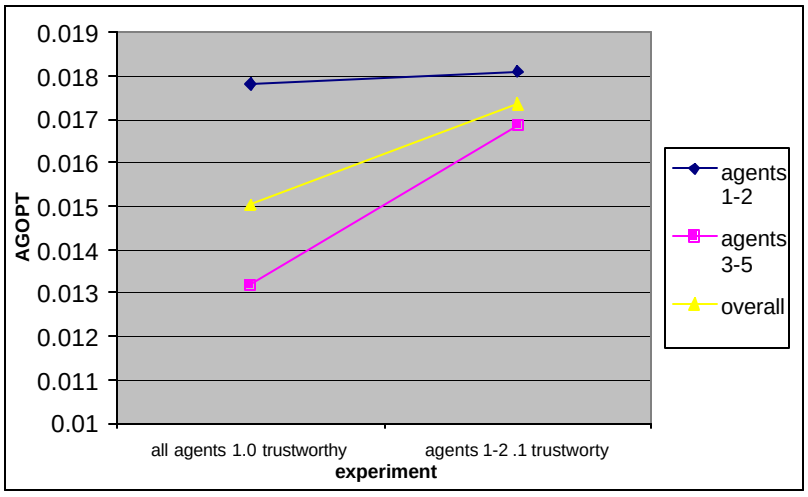

Figure 5. Average agent performance per game.

These effects are noticeable in ownership levels of agents as well (figure 5), with agents 3-5 gaining *27.95%* more ownership in experiment *agents 1-2 .1 trustworthy* than in *all agents 1.0 trustworthy* rather than losing land to the deceivers as in the first set. The gain occurs since honest agents were able to benefit from the truthful *terrain data* received from deceivers, while being little effected by deceiver's damaging deception tactics. The ownership gain of agents 1-2, however, was not significantly effected by the amount of *distrust* and lack of *power* in relation to *all agents 1.0 trustworthy* (in this set). However, deceivers were not able to achieve the huge ownership gain at the expense of honest agents that was achieved previously. Overall, agents gained *7.56%* fewer squares than in the equivalent experiment in the first set, most of which would have been the gain achieved by the deceptive agents.

5. Conclusion

Although deception is an effective means for agents to exercise power over others via verbal communication, this power is not with out its negative consequences on the deceiver and the society as a whole. At one extreme, too much deception allowed all agents to have power over one another, leading deceivers to interfere with one another's plans. In a mixed society however, deceivers were able to gain significant power over honest agents, benefiting the deceivers' performance at a cost to the honest agents. Further results showed that these figures are overly optimistic for deceivers, as without complete knowledge about others' strategies, their abilities to deceive effectively was drastically diminished. These results indicate that our model of power is an effective way to measure how agent communication, and specifically agent deception, has on other agent's mental states and ultimately agent performance.

In future work, we will examine how the effects of deceptive agents' knowledge of other agents' strategies on their ability to exploit and gain power over others due to deception. Other issues not dealt with here include the role agent models of deception and counter-deception play on power exercised between agents.

Acknowledgements

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

References

[1] Brainov, S., Sandholm, T., *Power, Dependence, and Stability in Multi-Agent Plans*, American Association for Artificial Intelligence (AAAI'99), Orlando, AAAI Press, 1998, pp11-16.

[2] Carofiglio, V., de Rosis, F., "Ascribing and Weighting Beliefs in Deceptive Information Exchanges", User Modeling 2001, LNAI 2109, Springer, 2001, pp222-224.

[3] Hexmoor, H., "A Model of Absolute Autonomy and Power: Toward Group Effects", Journal of Connection Science, Volume 14, No. 4. Taylor & Francis Ltd., 2003.

[4] Ward, D., Hexmoor, H., "Deception in Agents", Accepted for publication, KIMAS '03, 2003.