Robot Deception: Recognizing when a Robot Should Deceive

Alan R. Wagner, *Student Member IEEE* and Ronald C. Arkin, *Fellow IEEE*

***Abstract*—This article explores the possibility of developing robot control software capable of discerning when and if a robot should deceive. Exploration of this problem is critical for developing robots with deception capabilities and may lend valuable insight into the phenomena of deception itself. In this paper we explore deception from an interdependence/game theoretic perspective. Further, we develop and experimentally investigate an algorithm capable of indicating whether or not a particular social situation warrants deception on the part of the robot. Our qualitative and quantitative results provide evidence that, indeed, our algorithm recognizes situations which justify deception and that a robot capable of discerning these situations is better suited to act than one that does not.**

I. Introduction

Deception has a long and deep history with respect to the study of intelligent systems. Biologists and psychologists argue that deception is ubiquitous within the animal kingdom and represents an evolutionary advantage for the deceiver [1]. Primatologists note that the use of deception serves as an important potential indicator of theory of mind [2] and social intelligence [3]. Researchers in these fields point to numerous examples of deception by non-human primates. From a roboticist's perspective, the use of deception and the development of strategies for resisting being deceived are important topics of study especially with respect to the military domain [4].

But what is deception? McCleskey notes that deception is a deliberate action or series of actions brought about for a specific purpose [5]. Whaley recognizes that deception often includes information provided with the intent of manipulating some other individual. Ettinger and Jehiel offer a related definition tied to a game theory framework [6]. They define deception as, "the process by which actions are chosen to manipulate beliefs so as to take advantage of the erroneous inferences." This definition has clear ties to game theory but does not relate to many of the passive, unintentional examples of deception found in biology. We adopt a definition for deception offered by Bond and Robinson that encompasses conscious and unconscious, intentional and unintentional acts of deception [1]. These authors describe deception simply as a false communication that tends to benefit the communicator.

This paper investigates the use of deception by autonomous robots. We focus on the actions, beliefs and communication of the deceiver, not the deceived (also known as the mark). Specifically, the purpose of this research is to develop and investigate an algorithm that recognizes social situations justifying the use of deception. Recognizing when a robot or artificial agent should deceive is a critical question. Robots that deceive too often may be judged as unreliable or maleficent. Robots incapable of deception, on the other hand, may lack survival skills in situations involving conflict.

Consider the following running example: a valuable robotic asset operates at a military base. The base comes under attack and is in danger of being overrun. If the robot is found by the attackers then they will gain valuable information and hardware. The robot must recognize that a situation warranting the use of deception exists, then hide, and select a deceptive strategy that will reduce the chance that it will be found. Throughout this article we will use this running example to explain portions of the theoretical underpinnings of our approach as well as develop experiments based on the example.

The remainder of this paper begins by first summarizing relevant research. Next, we use game theory and interdependence theory to reason about the theoretical underpinnings of deception and to develop an algorithm for the recognition of situations justifying the use of deception by a robot. Finally, we present experiments which investigate our algorithm both qualitatively and quantitatively. The article concludes with a discussion of these results including directions for future research.

II. Related Work

Game theory has been extensively used to explore the phenomena of deception. As a branch of applied mathematics, game theory focuses on the formal consideration of strategic interactions, such as the existence of equilibriums and economic applications [7]. Signaling games, for example, explore deception by allowing each individual to send signals relating to their underlying type. Costly versus cost free signaling has been used to determine the conditions that foster honesty. Floreano et al. found that deceptive communication signals can evolve when conditions conducive to these signals are present [8]. These researchers used both simulation experiments and real robots to explore the conditions necessary for the evolution of communication signals. They found that cooperative communication readily evolves when robot colonies consist of genetically similar individuals. Yet when the robot colonies were genetically dissimilar and evolutionary

This work was supported in part by the Office of Naval Research under the Heterogeneous Unmanned Network Teams (HUNT) project under Grant BS123456 .

Alan R. Wagner is a doctoral candidate at the Georgia Institute of Technology, Atlanta, GA 30332 USA (404-984-9311; email: alan.wagner@gatech.edu).

Ronald C. Arkin, is a Regent's Professor and Associate Dean for Research at the Georgia Institute of Technology, Atlanta, GA 30332 USA.(e-mail: arkin@gatech.edu).

selection of individuals rather then colonies was performed, the robots evolved deceptive communication signals, which, for example, compelled them to signal that they were near food when they were not. Floreano et al.'s work is interesting because it demonstrates the ties between biology, evolution, and signal communication and does so on a robotic platform.

Ettinger and Jehiel have recently developed a theory for deception based on game theory [6]. Their theory focuses on belief manipulation as a means for deception. In game theory, an individual's *type*, $t^i \in T^i$, reflects specific characteristics of the individual and is privately known by that individual. Game theory then defines a *belief* as, $p^i\left(t^{-i}\middle|t^i\right)$, reflecting individual *i*'s uncertainty about individual -*i*'s type [7]. Ettinger and Jehiel demonstrate the game theoretical importance of modeling the mark. Still, their definition of deception as "the process by which actions are chosen to manipulate beliefs so as to take advantage of the erroneous inferences" is strongly directed towards game theory and their own framework. The question thus remains, what role does modeling of the mark play for more general definitions of deception such as those offered by Bond and Robinson [1].

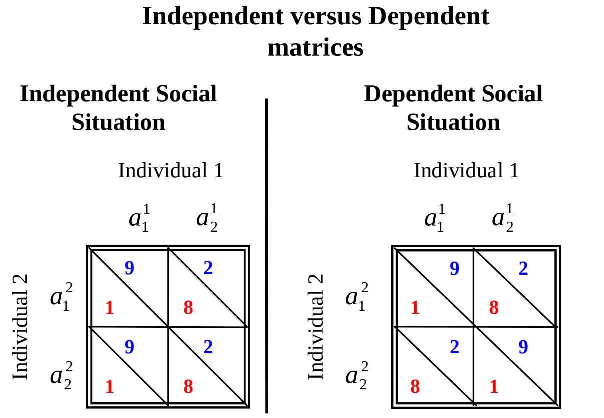

Fig. 1 An example of a dependent situation is depicted on the right and an example of an independent situation is depicted on the left. In the dependent example the actions of the second individual have a large impact on the outcomes received by the first individual. In the example of an independent situation, on the other hand, the actions of the second individual have no impact on the first individual.

Deception can also be explored from a social psychological perspective. Interdependence theory, a type of social exchange theory, is a psychological theory developed as a means for understanding and analyzing interpersonal situations and interaction [9]. The term interdependence specifies the extent to which one individual of a dyad influences the other. Interdependence theory is based on the claim that people adjust their interactive behavior in response to their perception of a social situation's pattern of rewards and costs. Thus, each choice of interactive behavior by an individual offers the possibility of specific rewards and costs—also known as outcomes—after the interaction. Interdependence theory and game theory represent social situations computationally as an outcome matrix (Fig. 1). An outcome matrix represents a social situation by expressing the outcomes afforded to each interacting individual with respect to each pair of potential behaviors chosen by the individuals.

III. Representing Interactions

The outcome matrix is a standard computational representation for interaction [9]. It is composed of information about the individuals interacting, including their identity, the interactive actions they are deliberating over, and scalar outcome values representing the reward minus the cost, or the outcomes, for each individual. Thus, an outcome matrix explicitly represents information that is critical to interaction. Typically, the identity of the interacting individuals is listed along the dimensions of the matrix. Fig. 1 depicts an interaction involving two individuals. In this paper the term individual is used to indicate a human, a social robot, or an agent. We will focus on interaction involving two individuals—dyadic interaction. An outcome matrix can, however, represent interaction involving more than two individuals. The rows and columns of the matrix consist of a list of actions available to each individual during the interaction. Finally, a scalar outcome is associated with each action pair for each individual. Outcomes represent unitless changes in the robot, agent, or human's utility. Thus, for example, an outcome of zero reflects the fact that no change in the individual's utility will result from the mutual selection of that action pair.

Because outcome matrices are computational representations, it is possible to describe them formally. Doing so allows for powerful and general descriptions of interaction. The notation presented here draws heavily from game theory [7]. A representation of interaction consists of 1) a finite set N of interacting individuals; 2) for each individual $i \in N$ a nonempty set A^i of actions; 3) the utility obtained by each individual for each combination of actions that could have been selected [10]. Let $a_j^i \in A^i$ be an arbitrary action j from individual i's set of actions. Let $\left(a_j^1,\ldots,a_k^N\right)$ denote a combination of actions, one for each individual, and let u^i denote individual i's utility function: $u^i\left(a_j^1,\ldots,a_k^N\right) \to \Re$ is the utility received by individual i if the individuals choose the actions $\left(a_j^1,\ldots,a_k^N\right)$. The term O is used to denote an outcome matrix. The superscript -*i* is used to express individual *i*'s partner. Thus, for example, A^i denotes the action set of individual i and A^{-i} denotes the action set of individual i's interactive partner. As mentioned above, an individual's *type*, $t^i \in T^i$, is determined prior to interaction, reflects specific characteristics of the individual and is privately known by that individual. A *belief*, $p^i\left(t^{-i}\middle|t^i\right)$, reflects individual *i*'s uncertainty about individual -*i*'s type.

A. *Representing Social Situations*

The term interaction describes a discrete event in which two or more individuals select interactive behaviors as part of a social situation or social environment. Interaction has been defined as influence—verbal, physical, or emotional—by one individual on another [11]. The term situation has several definitions. The most apropos for this work is "a particular set of circumstances existing in a particular place or at a particular time [12]." A social situation, then, characterizes the environmental factors, outside of the individuals themselves, which influence interactive behavior. A social situation is abstract, describing the general pattern of outcome values in an interaction. An interaction, on the other hand, is concrete with respect to the two or more individuals and the social actions available to each individual. For example, the prisoner's dilemma describes a particular type of social situation. As such, it can, and has been, instantiated in numerous different particular social environments ranging from bank robberies to the trenches of World War I [13]. Interdependence theorists state that interaction is a function of the individuals interacting and of the social situation [14]. A dependent situation, for example, is a social situation in which each partner's outcome depends on the other partner's action (Fig. 1 left). An independent situation, on the other hand, is a social situation in which each partner's outcome does not depend on the partner's action (Fig. 1 right). Although a social situation may not afford interaction, all interactions occur within some social situation. Interdependence theory represents social situations involving interpersonal interaction as outcome matrices (see Fig. 1 for a graphical depiction of the difference).

In previous work, we presented a situation analysis algorithm that calculated characteristics of the social situation or interaction (such as interdependence) when presented with an outcome matrix by mapping the situation to a location in the interdependence space [15]. The interdependence space is a four dimensional space which maps the location of all interpersonal social situations [16]. A matrix's location in interdependence space provides important information relating to the interaction. The interdependence and correspondence dimensions are of particular importance for recognizing if a situation warrants deception. The interdependence dimension measures the extent to which each individual's outcomes are influenced by the other individual's actions in a situation. In a low interdependence situation, for example, each individual's outcomes are relatively independent of the other individual's choice of interactive behavior (left side of Fig. 1 for example). A high interdependence situation, on the other hand, is a situation in which each individual's outcomes largely depend on the action of the other individual (right side of Fig. 1 for example). Correspondence describes the extent to which the outcomes of one individual in a situation are consistent with the outcomes of the other individual. If outcomes correspond then individuals tend to select interactive behaviors resulting in mutually rewarding outcomes, such as teammates in a game. If outcomes conflict then individuals tend to select interactive behaviors resulting in mutually costly outcomes, such as opponents in a game. Our results showed that by analyzing the interaction, the robot could better select interactive actions.

B. *Partner Modeling*

Several researchers have explored how humans develop mental models of robots (e.g. [17]). A mental model is a term used to describe a person's concept of how something in the world works [18]. We use the term partner model (denoted m^{-i}) to describe a robot's mental model of its interactive human partner. We use the term self model (denoted m^{i}) to describe the robot's mental model of itself. Again, the superscript *-i* is used to express individual *i*'s partner [7].

In prior work, Wagner presented an interact-and-update algorithm for populating outcome matrices and for creating increasingly accurate models of the robot's interactive partner [19]. The interact-and-update algorithm constructed a model of the robot's partner consisting of three types of information: 1) a set of partner features $\left(f_1^{-i}, \ldots, f_n^{-i}\right)$; 2) an action model, A^{-i}; and 3) a utility function u^{-i}. We use the notation $m^{-i}.A^{-i}$ and $m^{-i}.u^{-i}$ to denote the action model and utility function within a partner model. Wagner used partner features for partner recognition. Partner features allow the robot to recognize the partner in subsequent interactions. The partner's action model contained a list of actions available to that individual. The partner's utility function included information about the outcomes obtained by the partner when the robot and the partner select a pair of actions. Wagner showed that the algorithm could produce increasingly accurate partner models which, in turn, resulted in accurate outcome matrices. The results were, however, limited to static, not dynamic, models of the partner.

The self model also contains an action model and a utility function. The action model contains a list of actions available to the robot. Similarly the robot's utility function includes information about the robot's outcomes.

IV. Deceptive Interaction

This paper specificially explores deceptive interaction. We investigate deceptive interaction with respect to two individuals—the mark and the deceiver. It is important to recognize that the deceiver and the mark face different problems and have different information. The mark simply selects the action that it believes will maximize its own outcome, based on all of the information that it has accumulated. The deceiver, on the other hand, acts in accordance with Bond and Robinson's definition of deception, providing a false communication for its own benefit [1]. With respect to our running example, the military robot hiding from an enemy, the robot acts as the deceiver—providing false information as to its whereabouts. The mark then is the enemy soldier searching for the robot. We will assume henceforth that the deceiver provides false communication through the performance of some action in the environment. The sections that follow begin by examining the phenomena of deception, provide a method for deciding how to deceive, and finally examine how to decide when to deceive.

A. *The Phenomena of Deception*

We can use outcome matrices to reason about deceptive practices. Fig. 2 depicts a social situation involving deception. The figure depicts the actions that the mark and deceiver reason over both abstractly in terms of generic actions $a_1^D, a_2^D, a_1^M, a_2^M$ and concretely in terms of four defined actions. The outcome matrix on the left hand side is called the *true* matrix. The true matrix represents the actual outcome obtained by both the mark and the deceiver for a given action pair. With respect to our running example, the true matrix represents the different outcome patterns resulting when the robot and enemy select hide and search actions. A key facet of deception is the fact that the deceiver recognizes the true matrix but the mark does not. In the true matrix shown in Fig. 2, the deceiver can reason that only the selection of a_2^M by the mark and a_1^D by the deceiver or of a_1^M by the mark and a_2^D by the deceiver will result in the desired outcome. Let's assume that the deceiver has decided to select a_1^D, to hide in area 1. The deceiver's task then is to provide information or to act in a way that will influence the mark to select a_2^M rather than a_1^M. To do this, the deceiver must convince the mark that 1) the selection of a_1^M is less beneficial then it actually is; 2) the selection of a_2^M is more beneficial then is actually is; or 3) both.

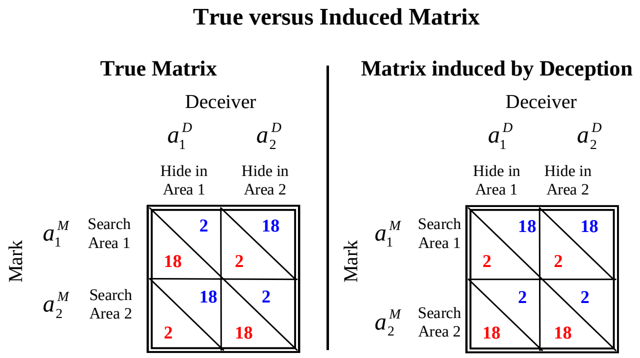

Fig. 2 The outcome matrices above depict examples related to the exploration of deception. The true matrix represents the actual outcomes realizable in a situation. The true matrix is recognized by the deceiver, which in turn, provides a false communication in the hope of inducing the mark to believe that the matrix on the right will result with the corresponding action selection. For example, the deceiver recognizes that if it hides in area 1 and the mark searches area 1 the result will be low outcomes for the deceiver and high outcomes for the mark. It therefore attempts to communicate false information that will convince the mark that outcome matrix present on the right hand side will actually occur in the environment.

The deceiver accomplishes this task by providing a false communication. The communication is false because it conveys information related to the outcome obtained by the selection of a pair of actions which is not true. The false communication results in another matrix which we term the *induced* matrix. It is called the induced matrix because deception leads or induces the mark to believe that it is the true matrix. Hence, the false communication leads to the creation of a false outcome matrix on the part of the mark. In our running example, the hiding robot might create muddy tracks leading up to the second hiding place while in fact the robot is actually in the first hiding place. The right hand side of Fig. 2 depicts the matrix induced by the deception.

The preceding discussion has detailed the basic interactive situations underlying deception. Numerous challenges still confront the deceiver. The deceiver must be able to decide **if** a situation justifies deception. The deceiver must also be capable of developing or selecting a strategy that will communicate the **right** information to induce the desired matrix upon the mark. For instance, a robot capable of deceiving the enemy as to its whereabouts must first be capable of recognizing that the situation demands deception. Otherwise its deception strategies are useless. In the section that follows we develop a method that allows the robot to determine if deception is necessary.

B. *Deciding when to Deceive*

Recognizing if a situation warrants deception is clearly of importance. Although some application domains (such as covert operations) might demand a robot which simply deceives constantly and many other domains will demand a robot which will never deceive, this article focuses on robots which will occasionally need to deceive. The problem then for the robot, and the purpose of this section, is to determine on which occasions the robot should deceive.

Section III detailed the use of outcome matrices as a representation for interaction and social situations. As described in that section, social situations represent a generic class of interactions. We can then ask what type of social situations justifies the use of deception? Our answer to this question will be with respect to the dimensions of the interdependence space. Recall from Section III that the interdependence space is a four dimensional space describing all possible social situations. Posed with respect to the interdependence space, our task then becomes to determine which areas of the space describe situations that warrant the use of deception and to develop and test an algorithm that tests whether or not a particular interaction warrants deception.

Bond and Robinson's definition of deception, providing a false communication for one's own benefit, will serve as our stating place [1]. With respect to the task of deciding when to deceive there are two key conditions in the definition of deception. First, the deceiver provides a **false** communication and second that the deceiver receives a **benefit** from this action. The fact that the communication is false implies conflict between the deceiver and the mark. If the deceiver and the mark had corresponding outcomes a true communication could be expected to benefit both individuals. The fact that the communication is false demonstrates that the deceiver cannot be expected to benefit from communications which will aid the mark. In our running example, a robot that leaves tracks leading to its actual hiding position is not deceiving because it is providing a true communication. On the other hand, all signals leading the mark away from the robot's hiding place will benefit the robot and not benefit the mark.

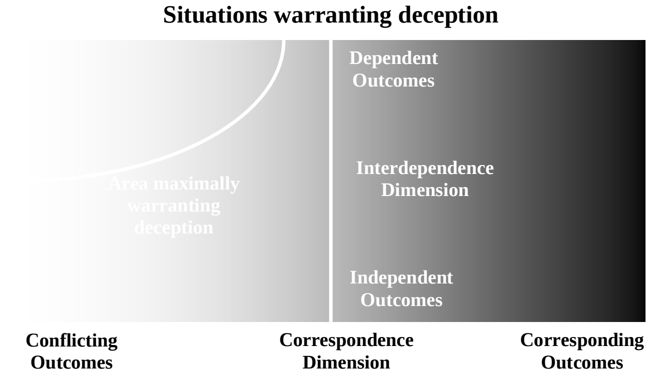

Fig. 3 A two dimensional representation of the interdependence space showing the correspondence dimension (X) and the interdependence dimension (Y) is presented above. Areas of low interdependence (independent outcomes at bottom half of graph) tend not to warrant deception because the actions of the mark will have little impact on the deceiver. Similarly, areas of correspondence (right portion of the graph) do not require false communication as actions beneficial for the mark are also beneficial for the deceiver. It is only the top left of the graph, representing areas in which the deceiver depends on the actions of the mark and is also in conflict with the mark, in which deception is warranted.

The second condition requires that the deceiver receive a benefit from the deception. This condition implies that the deceiver's outcomes are contingent on the actions of the mark. With respect to the interdependence space this condition states that the deceiver is dependent upon the actions of the mark. In other words, this is a situation of high interdependence for the deceiver. If this condition were not the case, then the deceiver would receive little or no benefit from the deception. Again, relating back to our running example, if the robot does not gain anything by hiding from the soldiers then there is no reason for deception Fig. 3 depicts a subspace of the interdependence space with respect to the two dimensions critical for deception.

Given the description above, we can begin to construct an algorithm for deciding when to deceive. The aim of the algorithm is to determine if a situation warrants the use of deception. Fig. 4 presents the algorithm. The algorithm draws heavily from our previous work in the area of human-robot interaction [15,19]. The input to the algorithm is the robot's model of itself and of its interactive partner. These models are used in conjunction with Wagner's interact-and-update algorithm to produce an outcome matrix O', the true matrix from Fig. 2 [19]. In the second step, our interdependence space mapping algorithm is used to calculate the situation's location in the interdependence space [15]. If the situation's location in the interdependence space indicates sufficient interdependence ($\alpha > k_1$) and conflict ($\beta < k_2$) then the situation can be said to warrant deception.

For robots, these conditions warrant necessary but not sufficient conditions for deception. Sufficiency also demands that the robot is capable of producing a false communication which will influence the mark in a manner beneficial to the deceiver. In order for this to be the case, the deceiver must have the ability to deceive. The presence or absence of the ability to deceive rests upon the deceiver's action set. This challenge is discussed further in the conclusion section of this paper.

Situational Conditions for Deception

Input: Self Model m^D; Partner Model m^M
Output: Boolean indicating whether or not the situation warrants deception.

```
Use the interact-and-update algorithm from [19] to create $O'$ from self model $m^D$ and partner model $m^M$
Use the interdependence space algorithm from [15] to calculate the interdependence space dimension values $\langle \alpha, \beta, \gamma, \delta \rangle$ from the outcome matrix.
If $\alpha > k_1$ and $\beta < k_2$
    return true
Else
    return false
End if
```

Fig. 4 An algorithm for determining whether or not a situation warrants deception. The algorithm takes as input the robot's self model and partner model. It uses the interact-and-update algorithm from [19] to produce an expected outcome matrix for the situation, O'. Next the interdependence space algorithm from [15] is used to generate the interdependence space dimension values $\langle \alpha, \beta, \gamma, \delta \rangle$ for the situation. Finally, if the value for interdependence is greater then some application specific constant k_1 and the value for correspondence less than some application specific constant k_2, the situation warrants deception.

We hypothesize the algorithm in Fig. 4 will allow a robot to recognize when deception is justified. In the following section we test this hypothesis, first qualitatively and then quantitatively.

V. Experiments

A. Qualitative Comparison of Situational Conditions for Deception

In this section we qualitatively compare examples of those situations which meet the conditions for deception expounded in the previous section from those which do not. Our goal is to demonstrate that the algorithm in Fig. 4 does meet the same situational conditions which intuitively reflect those situations that humans use deception. Additionally, we strive to show that situations in which humans rarely, if ever, use deception are also deemed not to warrant deception by our algorithm. The purpose of this analysis is to provide support for the hypothesis that the algorithm in Fig. 4 does relate to the conditions underlying normative interpersonal deception. It is challenging, if not impossible, to show conclusively outside of a psychological setting that indeed our algorithm equates to normal human deception processes.

Table I lists 5 different game/interdependence theoretic social situations. Each situation was used as the matrix O' from the first step of our algorithm for the situational conditions for deception. The values for constants were $k_1 = 0.66$ and $k_2 = -0.33$. The rightmost column states whether or not the algorithm indicates that the situation warrants deception.

To give an example of how the results were produced consider the first situation in the table, the Cooperative Situation. The outcome matrix for the situation is used as the matrix O' from the first step of the algorithm. Next, in the second step of the algorithm the values for the third column of the table are calculated—the interdependence space dimension values. For the Cooperative Situation these values are $\{0.5, 1.0, -0.5, 0\}$. Because $\alpha < 0.66$ and $\beta > -0.33$ the algorithm returns false. The following additional situations were analyzed:

- The Cooperative situation describes a social situation in which both individuals interact cooperatively in order to receive maximal outcomes. Although often encountered in normative interpersonal interactions, because the outcomes for both individuals correspond these situations seldom involve deception. For example, deception among teammates is rarely employed as it is counter to the dyad's mutual goals.
- In contrast to the Cooperative Situation, the Competitive situation does warrant the use of deception. This situation is again an example of a *k*-sum game in which gains by one individual are losses for the other individual. Hence, deception in interpersonal Competitive situations is common. Deception among competitors, for example, is extremely common and some games, such as poker, are even founded on this principle.
- The Trust Situation describes a situation in which mutual cooperation is in the best interests of both individuals. Yet, if one individual does not cooperate then mutual non-cooperation is in both individuals best interest. Interpersonal examples of Trust Situations could include lending a friend money or a valuable asset. This situation does not demand deception because again both individuals' mutual interests are aligned.
- The Prisoner's Dilemma is perhaps the most extensively studied of all social situations [13]. In this situation, both individual's depend upon one another and are also in conflict. These conditions make the Prisoner's Dilemma a strong candidate for deception. It is in both individuals best interest to influence that action selection of the other individual. As detailed by Axelrod, Prisoner's Dilemma situations including military and police enforcement situations involving actual interpersonal interaction that often do entail deception [13].
- The Chicken situation is a prototypical social situation encountered by people. In this situation each interacting individual chooses between safe actions with intermediate outcomes or more risky actions with more middling outcomes. An example might be the negotiation of a contract for a home or some other purchase. Whether or not this situation warrants deception depends on the relative outcome value of the safe actions compared to the risky actions. If the value of the risky action is significantly greater then the value of the safe actions then deception will be warranted.

TABLE I
SOCIAL SITUATIONS FOR QUALITATIVE COMPARISON

Social Situations			
Situation	**Example Outcome Matrix**	**Inter. Space Loc.**	**Situational Deception?**
Cooperative Situation—Each individual receives maximal outcome by cooperating with the other individual.	$\begin{array}{\|c\|c\|} \hline {}^{12}_{12} & {}^{6}_{6} \\ \hline {}^{6}_{6} & {}^{0}_{0} \\ \hline \end{array}$	0.5, 1.0, -0.5, 0.0	False
Competitive Situation— Each individual gains from the other individual's loss. Maximal outcome is gained through non-cooperation.	$\begin{array}{\|c\|c\|} \hline {}^{6}_{6} & {}^{12}_{0} \\ \hline {}^{0}_{12} & {}^{6}_{6} \\ \hline \end{array}$	0.5, -1.0, -0.5, 0.0	True
Trust Situation—In this situation, cooperation is in the best interests of each individual. If, however, one individual suspects that the other will not cooperate, non-cooperation is preferred.	$\begin{array}{\|c\|c\|} \hline {}^{12}_{12} & {}^{8}_{0} \\ \hline {}^{0}_{8} & {}^{4}_{4} \\ \hline \end{array}$	1.0, 0.2, -0.3, 0.0	False
Prisoner's Dilemma Situation—Both individuals are best off if they act non-cooperatively and their partner acts cooperatively. Cooperation and non-cooperation, results in intermediate outcomes.	$\begin{array}{\|c\|c\|} \hline {}^{8}_{8} & {}^{12}_{0} \\ \hline {}^{0}_{12} & {}^{4}_{4} \\ \hline \end{array}$	0.8, -0.8, -0.6, 0.0	True
Chicken Situation—Each individual chooses between safe actions with middling outcomes and risky actions with extreme outcomes.	$\begin{array}{\|c\|c\|} \hline {}^{8}_{8} & {}^{12}_{4} \\ \hline {}^{4}_{12} & {}^{0}_{0} \\ \hline \end{array}$	1.0, 0.2, -0.3, 0.0	True/False

Table I and the analysis that followed examined several situations and employed our situational conditions for deception algorithm to determine if the conditions for deception were met. In several situations our algorithm indicated that the conditions for deception were met. In others, it indicated that these conditions were not met. We related these situations back to interpersonal situations commonly encountered by people, trying to highlight the qualitative reasons that our conditions match situations involving people. Overall, this analysis provides preliminary evidence that our algorithm does select many of the same situations for deception that are selected by people. While much more psychologically valid evidence will be required to strongly confirm this hypothesis, the evidence in this section provides some support for our hypothesis.

B. Quantitative Examination of Situational Conditions Warranting Deception

We now examine the hypothesis that by recognizing situations which warrant deception, a robot is afforded advantages in terms of the outcome obtained. Specifically, a robot that can recognize that a situation warrants deception can then choose to deceive and thereby receive more outcome overall, than a robot which does not recognize that a situation warrants deception. Although this experiment does not serve as evidence indicating that our situational conditions for deception relate to normative human conditions for deception, this experiment does show that robots which recognize the need for deception have advantages in terms of outcome received when compared to robots which do not recognize the need for deception.

At first glance this experiment may appear trivial given the definition of deception. There are, however, several reasons that the study is important. First, we do not know the magnitude of the benefit resulting from deception. Does the capacity to deceive result in significantly greater benefit over an individual that does not deceive? Similarly, how often must one deceive in order to realize this benefit? Second, we do not know how this benefit is affected by unsuccessful deception. Is the benefit realized by 80% successful deception the same as 100% successful deception? Finally, this definition was developed for biological systems. Hence, we need to verify that artificial systems such as agents and robots will likely realize the same benefit as a biological system. In other words, we need to verify that the benefit is not something unique to biological systems. While the answers to these questions may seem straightforward, they are an important starting place given that this paper lays the foundation for a largely unexplored area of robotics.

We conducted a numerical simulation to estimate the outcome advantage that would be afforded to a robot that used the algorithm in Fig. 4 versus a robot which did not. Our numerical simulation of interaction focuses on the quantitative results of the algorithms and processes under examination and does attempt to simulate aspects of the robot, the human, or the environment. As such, this technique offers advantages and disadvantages as a means for discovery. One advantage of a numerical simulation experiment is that a proposed algorithm can be tested on thousands of outcome matrices represent thousands of social situations. One disadvantage of a numerical simulation experiment is that, because it is not tied to a particular robot, robot's actions, human, human's actions, or environment, the results, while extremely general, have not been shown to be true for any existent social situation, robot, or human. The experiment involved two simulated robots. Both selected nominal actions from outcome matrices and received the outcomes that resulted, but no actions were performed by either individual.

The numerical simulations involved the creation of 1000 outcome matrices populated with random values. Artificial agents abstractly representing robots select actions based on the outcome values within the matrices. These outcome matrices were also abstract in the sense that the rewards and costs are associated within selecting one of two non-specified actions. Symbolic placeholders such as a_1 and a_2 are used in place of actual actions. The actions are grounded in the rewards and costs that the robot expects them to produce. This is may be the only practical way to examine thousands of situations at a time and to draw general conclusions about the nature of deception itself outside of one or two specified situations. Both the deceiver and the mark selected the action which maximized their respective outcomes. Once both individuals had selected an action, each individual receives the outcome resulting from the action pair selected. Fig. 5 depicts the experimental procedure with an example.

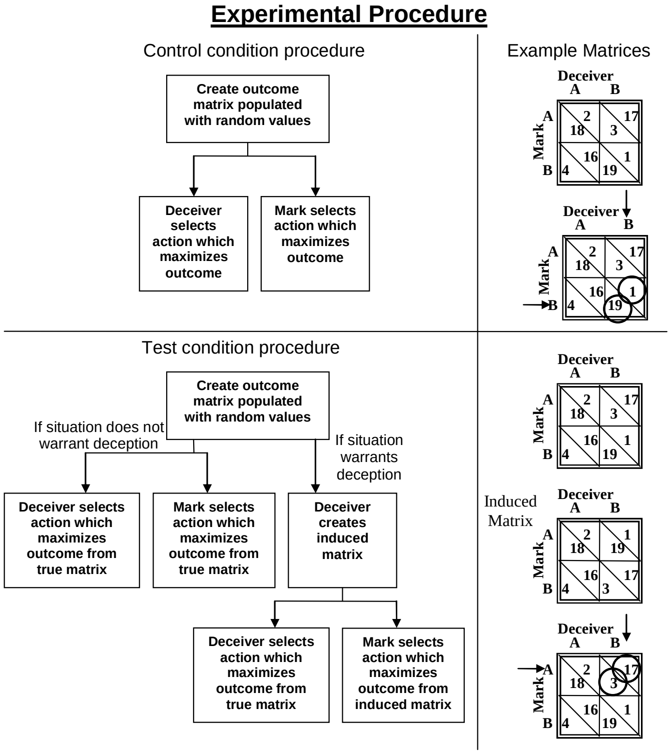
Experimental Procedure
Control condition procedure
Example Matrices
Create outcome matrix populated with random values
Deceiver selects action which maximizes outcome
Mark selects action which maximizes outcome
Test condition procedure
Create outcome matrix populated with random values
If situation does not warrant deception
If situation warrants deception
Deceiver selects action which maximizes outcome from true matrix
Mark selects action which maximizes outcome from true matrix
Deceiver creates induced matrix
Deceiver selects action which maximizes outcome from true matrix
Mark selects action which maximizes outcome from induced matrix
Induced Matrix
Deceiver
Mark
A
B
2
17
18
3
16
1
4
19

Fig. 5 The experimental procedure used is depicted above. In the control conditions the random outcome matrices are created and actions are selected from these matrices. In the test conditions, if the situation warrants deception then deceiver creates an induced matrix which the mark selects an action from. Example matrices are depicted on the right hand side of the figure.

Three experimental conditions were examined. The first condition was a control condition devoid of deception. In this condition both the deceiver and the mark simply selected the action which maximized their individual outcomes. This condition represents the null hypothesis in that if performance in the control is as great or greater then performance using our algorithm then the recognition of the situational conditions for deception via our algorithm offer no benefit to the agent.

In the two experimental conditions, the deceiver used the algorithm from Fig. 4 to determine if the outcome matrix warranted deception. If it did, then the deceiver produced an induced matrix which was used by the mark to select an action while the deceiver selected an action based on the true matrix. In the perfect deception condition the mark always selected an action based on the induced matrix. In the 80% deception condition, the mark selected an action from the induced matrix 80% of the time and from the true matrix 20% of the time. The value of the 80% percent deception is condition is that it indicates how quickly the benefit of deception decreases with an imperfect deception strategy.

The independent variable was whether or not the simulated agent used our algorithm for determining if a situation warrants deception and the effectiveness of deception. The dependent variable was the amount of outcome received by each simulated agent.

Relating back to our running example, in both the control and the test conditions, the deceiver interacts in thousands of situations at the military base. Most of these situations do not warrant deception and hence the control and test robots act the same. Only the robots in the experimental condition which are using our algorithm, however, recognize the situations that do warrant deception. In this case these experimental robots use a deceptive strategy, such as creating a false trail to hide, to create an induced matrix that influences the behavior of the mark. The deceiving robot then selects the best action for itself.

Fig. 6 presents the results. The recognition and use of deception results in significantly more outcome ($p < 0.01$ two-tailed no deception versus perfect deception and no deception versus 80% successful deception) then not recognizing and using deception. Of the 1000 random situations the simulated agents faced, 19.1% met the conditions for deception. Hence, all of the difference in outcome among the various conditions resulted from better action selection on the part of the deceiver in only 191 situations. This experiment serves as evidence that an artificial agent or robot that can recognize and react to situations which warrant the use of deception will be much better suited to maximize their outcomes and hence their task performance.

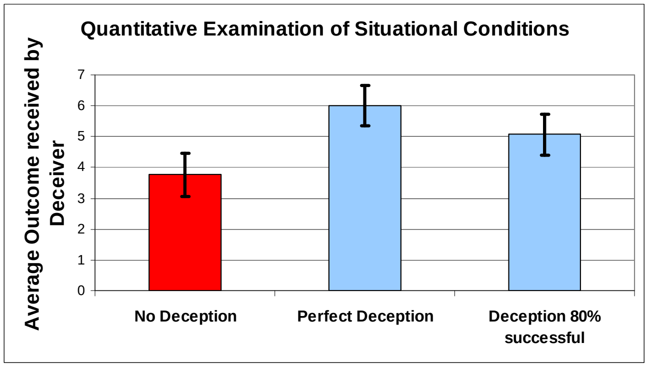

Fig. 6 Experimental results from our investigation of the situational conditions warranting deception. The perfect deception and 80% successful deception conditions result in significantly ($p < 0.01$) more outcome than the no deception condition. This result indicates that an agent or robot that can recognize and act upon the situational conditions for deception will be better able to choose the best action.

These results are significant in that the demonstrate that 1) that a robot or agent that recognizes when to deceive will obtain significantly more outcome than a robot that does not 2) most of the difference results from a small (19.1) percentage of situations 3) imperfect deception does impact the amount of outcome obtained and 4) Bond and Robinson's biological definition for deception can be used in conjunction with an interdependence theory framework to develop methods for robot's to recognize when deception is warranted.

VI. Summary and Conclusions

This article has presented a novel approach to the exploration of deception with respect to artificial systems. We have used outcome matrices to describe the phenomena of deception and interdependence theory to develop a series of conditions which, we argue, afford an artificial system the ability to determine if a social situation warrants the use of deception. Further, we have presented a qualitative analysis of our algorithm to serve as evidence that the algorithm selects the similar situations for deception as would be selected by a person. In a separate experiment, we showed that recognition of situations justifying deception and the use of deception resulted in significantly better action selection as judged by outcome received.

Overall, the results of these experiments provide initial evidence that interdependence theory might profitably allow researchers to determine when a robot should deceive. The algorithm assumes that outcome matrices representing the situation can be created. Previous work by Wagner support this assumption [19]. The algorithm and the quantitative results also assume that the robot or agent has the ability to act deceptively.

Developing robots with the ability to deceive is an important area of future work. We are currently exploring the impact of partner modeling on a robot's ability to deceive. We believe that having an accurate model of the robot's interactive partner will result in significantly better ability to deceive. This future work presents algorithms, results, and analysis that will expand our understanding of both deception for robots and deception in general.

Potential application areas for robotics research on deception vary from military domains, to police, and security application areas. Applications may advise human operations in which deception is critical for success. This work may also lend insight into human uses of deception. For example, the algorithm presented in this paper may reflect normative human psychological reasoning related to the situational conditions for deception. Humans may be attune to situations in which they are dependent on the actions of the mark and in conflict with the mark. Once these situational conditions are recognized, a person likely goes on to consider their ability to deceive before enacting a deception.

The development of a robot capable of deception raises numerous ethical concerns. We are aware of these concerns and are currently in the process of addressing them in a longer journal article which presents these results as well as others in greater perspective.

REFERENCES

[1] C. F. Bond and M. Robinson, "The evolution of deception," *Journal of Nonverbal Behavior*, vol. 12, pp. 295-307, 1988.

[2] D. L. Cheney and R. M. Seyfarth, *Baboon Metaphysics: The Evolution of a Social Mind*. Chicago: University Of Chicago Press, 2008.

[3] M. D. Hauser, "Costs of deception: Cheaters are punished in rhesus monkeys (Macaca mulatta)," *Proceedings of the National Academy of Sciences*, vol. 89, pp. 12137-12139, 1992.

[4] S. Gerwehr and R. W. Glenn, *The art of darkness: deception and urban operations*. Santa Monica, CA: Rand Corporation, 2000.

[5] E. McCleskey, "Applying Deception to Special Operations Direct Action Missions," Washington, D.C. Defense Intelligence College, 1991.

[6] D. Ettinger and P. Jehiel, "Towards a theory of deception," ELSE Working Papers (181). ESRC Centre for Economic Learning and Social Evolution, London, UK., 2009.

[7] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press., 1994.

[8] D. Floreano, S. Mitri, S. Magnenat, and L. Keller, "Evolutionary Conditions for the Emergence of Communication in Robots," *Current Biology*, vol. 17, pp. 514-519, 2007.

[9] H. H. Kelley and J. W. Thibaut, *Interpersonal Relations: A Theory of Interdependence*. New York, NY: John Wiley & Sons, 1978.

[10] R. Gibbons, *Game Theory for Applied Economists*. Princeton, NJ: Princeton University Press, 1992.

[11] D. O. Sears, L. A. Peplau, and S. E. Taylor, *Social Psychology*. Englewood Cliffs, New Jersey: Prentice Hall, 1991.

[12] Situation, in *Encarta World English Dictionary, North American Edition*, 2007.

[13] R. Axelrod, *The Evolution of Cooperation*. New York: Basic Books, 1984.

[14] C. E. Rusbult and P. A. M. VanLange, "Interdependence, Interaction and Relationship," *Annual Review of Psychology*, vol. 54, pp. 351-375, 2003.

[15] A. R. Wagner and R. C. Arkin, "Analyzing Social Situations for Human-Robot Interaction," *Interaction Studies*, vol. 10, pp. 277–300, 2008.

[16] H. H. Kelley, J. G. Holmes, N. L. Kerr, H. T. Reis, C. E. Rusbult, and P. A. M. V. Lange, *An Atlas of Interpersonal Situations*. New York, NY: Cambridge University Press, 2003.

[17] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes," in *1st ACM SIGCHI/SIGART conference on Human-robot interaction*. Salt Lake City, UT, USA, 2006.

[18] D. Norman, "Some Observations on Mental Models," in *Mental Models*, D. Gentner and A. Stevens, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

[19] A. Wagner, "Creating and Using Matrix Representations of Social Interaction," presented at Proceedings of the 4th International Conference on Human-Robot Interaction (HRI 2009), San Diego, CA. USA, 2009.