Lies and Deception: Robots that use Falsehood as a Social Strategy

Alan R. Wagner
Georgia Tech Research Institute

1.0 Introduction

Dishonesty is a part of life. Humans lie, cheat, and deceive not just to increase their gain, but for a variety of reasons that relate as much to their own particular social and moral underpinnings as to the task at hand (Gino, Ayal & Ariely 2009; Amir, Ariely & Mazar 2008). A lie is a specific type of dishonesty. A commonly accepted definition of the term lie is *a false statement made by an individual which knows that the statement is not true* (Carson 2006). This definition emphasizes the volitional nature of a lie, recognizing that not only must the liar make a false statement, but that they must also know that the statement is indeed false. Importantly, this definition limits the type of communications that a lie can take to statements. Hence, most lies are either written or spoken statements.

Amir et al. notes that factors such as watching others behave dishonestly and being able to rationalize one's behavior have an important influence on a person's decision to be dishonest. But dishonest behavior is not necessarily detestable behavior. In fact, falsehood can serve as a social strategy whose purpose is to maintain individual and group relations (DePaulo & Bell 1996; DePaulo & Kashy 1998). Even when the object of deception is not mutually beneficial, those engaged in deception are conscious of the effects of their behavior. Gneezy (2005), for example, found that people playing a deception game with monetary outcomes were sensitive to the impact their lies would have on other players. In fact, humans often employ minor falsehoods while engaged in normal interaction. Polite lies, for example, are

etiquette or norm-induced lies that typically serve as part of one's culture or interactive social protocol. Overall, there are many situations in which dishonesty is considered socially acceptable (DePaulo & Bell 1996; DePaulo & Kashy 1998).

This chapter applies our framework for non-verbal deception to verbal deception and lying in general. We propose that the ability to lie emerges from a social system in which the actors have the capacity to use language to create statements and the desire to make false statements. Moreover, a deceptive lie occurs when an individual is also in conflict with their interactive partner. The research presented here focuses on spoken lies. Some researchers argue that lying is a particularly human ability (Sapolsky 2010). As will be discussed in greater detail in later sections, the act of lying affords a rich format for deception, even if the definition of lying does not necessarily imply deception.

Robots are used as an investigative tool for implementing and verifying the theory. We have previously tested this theory on non-verbal deception (Wagner & Arkin 2011). Here our focus is on verbal deception and lying. In contrast to most psychological and cognitive-science research, the use of a robot forces the researcher to consider the noise, variability, and ambiguities associated with embodiment. In contrast to most robotics research, the purpose of this approach is not to develop a system optimized for a narrowly-defined task. Rather, we seek to develop and investigate the theoretical underpinnings and computational algorithms that will allow a robot to verbally deceive in a general setting. The overarching goal of this chapter is to begin to develop a conceptual framework that will allow a robot to both understand a person's reasons for being dishonest and to reason about if and when it should be dishonest.

The chapter probes the following research questions:

1) How can our framework for non-verbal deception be applied to verbal deception and lying?

2) Does the application of this framework provide a conceptual understanding of the factors that impact one's decision to lie?

3) Can it explain different types of lies such as white lies and polite lies?

4) How can a robot or agent's prior history be used to influence its decision to lie?

The remainder of this chapter begins by introducing the game and interdependence theoretic underpinnings of our framework. Next, different aspects and types of lies are examined from this perspective. Finally, a series of experiments and their results are presented each attempting to explore a different aspect of the challenge of developing a robot that lies. The article concludes with a discussion of these results and directions for future research.

2.0 Prior Work

Research related to lies and lying has long been a scholarly pursuit of philosophers (e.g. Morris 1976). Many have developed and presented definitions for lying and deception (Fallis 2009; Mahon 2008; Carson 2006). Others have examined specific categories of lying (e.g. Caminada 2009; see Gupta, Sakamoto & Ortony 2013 for a thorough overview). Vincent and Castelfranchi (1979) present an early framework which develops the relations between and among lying, deception, linguistics, and pragmatics.

Less work has focused on whether and how a machine might be made to lie. Rehm (2005) uses an agent to express emotions while lying in an interactive dice game with a human player. Sakama et al. (2010) develop a logical account of lying. They use the framework to explore offensive and defensive lies based on the liar's intention. Their framework is used to

formulate several postulates but is not instantiated on a robot or agent. Isaac and Bridewell (in press), develop a framework for identifying deceptive entities (FIDE) which emphasizes the importance of ulterior motive as part of the classification scheme. They use the framework to generate abstract agent-models which helps explain several different types of lies. Unfortunately, this work is not implemented or tested as part of an actual agent. Hence, little can be said as to its potential suitability for a robot. To the best of our knowledge verbal deception and lying have not been demonstrated on a robot.

Game theory has been extensively used to explore deception (Osborne & Rubinstein 1994). Signaling games, for example, explore deception by allowing each individual to send signals relating to their underlying type (Spence 1973). Costly versus cost-free signaling has been used to determine the conditions that foster honesty. Floreano et al. (2007) found that deceptive communication signals can evolve when conditions conducive to these signals are present. These researchers used both simulation experiments and real-world 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 are genetically dissimilar and evolutionary selection of individuals rather than colonies is performed, the robots evolve deceptive communication signals, which, for example, compel them to signal that they are near food when they are not. Floreano et al.'s work demonstrates the ties that exist between and among biology, evolution, and signal communication on a robotic platform.

Ettinger and Jehiel (2009) have recently developed a theory for deception based on game theory. 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, $p_i\left(t_{-i}|t_i\right)$, reflecting individual *i*'s uncertainty about individual -*i*'s type (Osborne & Rubinstein 1994). Ettinger and Jehiel (2009, p. 2) demonstrate the game-theoretical importance of modeling the individual who is lied to (called 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. As such, it seems to have little applicability beyond their investigation.

We have already investigated the use of non-verbal deception by an autonomous robot. In previous work, our framework was used to characterize interactions that warrant deception on the part of the robot (Wagner & Arkin 2011). This work employed a commonly-used definition of deception as "*a false communication that tends to benefit the communicator*" (Bond & Robinson 1988, p. 295). Our framework allowed us to reason about what types of interactions warranted the use of deception. Moreover, we developed an algorithm that allowed a robot to act deceptively by modeling the individual to be deceived. We demonstrated the algorithm on a multi-robot, hide-and-seek task in which one robot learned to leave a false trail indicating that it was hiding in a different location.

Others have since investigated the possibility of developing a deceptive robot. Vazquez et al. explored deception in the context of a multi-player robotic game in which the robot decides who wins the game (Vazquez et al. 2011). Davis and Arkin implemented animal-behavior models of deception to mimic mobbing behaviors used by Arabian babblers (Davis & Arkin

2012). Nevertheless, to the best of our knowledge, little research has been devoted to examining how to develop a robot that lies.

3.0 Basic Elements

In prior work we examined the ability of our framework to characterize whether or not an interaction warrants deception on the part of the robot (Wagner & Arkin 2011). The deception used by the robot consisted of non-verbal behavior, such as hiding. In this chapter we expand the framework to verbal deception and lying. For a robot interacting with people, understanding the psychological motivations that guide a person's honest and dishonest behavior is an important problem. Robots tasked with assisting elementary school teachers, for instance, may need to reason about the difference between students' honest mistakes and dishonest mistakes. A robot assisting with physical therapy may need to judge whether someone is feigning exhaustion or genuinely fatigued.

This logic extends to lying. A lie is typically produced as a verbal or written statement. As such, it can be a rich and nuanced means of dishonesty. Because there can be good reasons for being dishonest, a social robot that interacts with people in unscripted, dynamic social situations will need both the ability to understand the reasoning underlying a person's falsehoods and, quite possibly, the ability to create falsehoods as a social strategy. For example, if tasked with rescuing an injured person from a disaster it may be unwise for a robot to honestly inform the victim of their chances of survival. In such cases, dishonesty is generally viewed as acceptable and often preferable.

There are many ways to lie. A white lie, for example, is a minor misstatement which tends to be harmless or possibly beneficial (Oxford English Dictionary Online 2013). Research

has found that even children understand and use white lies in order to protect the feelings of the person being lied to (Talwar, Murphy & Lee 2008).

Lies-to-children are simplifications for the purpose of making an explanation more understandable. An exaggeration, on the other hand, is a statement in which the primary aspects of the statement are true to some degree (Gupta, Sakamoto & Ortony 2013).

There are equally many ways to classify lies. We adhere to a categorization based on consequences. This approach is not unique to us. Gneezy (2005) proposes a classification scheme based on the consequences of a lie. The lies in one category, which we term *prosocial l*ies, are described as false statements made by an individual who knows that the statements are not true and which also tend to benefit the individual being lied to at the cost of the liar. Prosocial lies are not deceptive and potentially motivated by altruism (Becker 1976). White lies and lies-to-children serve as examples of prosocial lies. In contrast, a *deceptive* lie is one which benefits the liar at a cost to the individual being lied to. A half-truth is an example of deceptive lie if the purpose of the lie is for the benefit of the liar. For example, if accused of stealing a particular piece of merchandise a liar may truthfully proclaim their innocence while hiding the fact that they indeed stole some different piece of merchandise. Many lies can be either prosocial or deceptive depending on who benefits and is punished by the telling of the lie. Exaggeration, for example can be prosocial, if the lie is beneficial to the partner, or deceptive, if the lie benefits the liar at a cost to the partner.

Our approach derives from consideration of both deception and lying. A deceptive lie was defined above as "*a false communication that tends to benefit the communicator* (Bond & Robinson 1988, p. 295)." We also defined a lie as "a false statement made by an individual who

knows that the statement is not true." Given these definitions we can deduce that lies are not always deceptive. A white lie, for instance does not benefit the liar. Moreover, deceptions are not all lies. Camouflage, for instance, is a deceptive communication which benefits the deceiver but is not a statement. We thus propose that the set of deceptive lies is found at the intersection of deceptive communications and lies (**Error! Reference source not found.**).

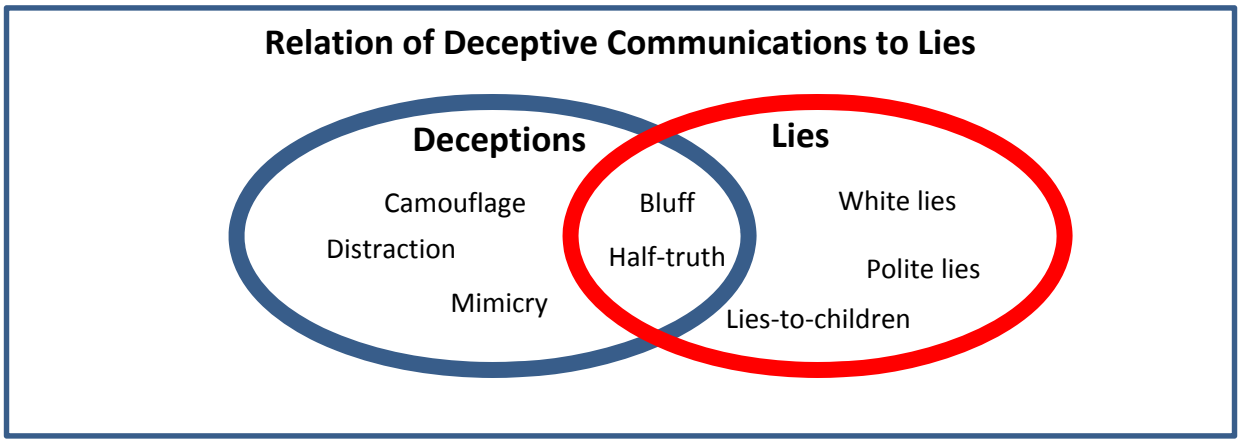

Figure 1 Deception, lies, and deceptive lies

The lies that fall within this intersection are communications that take the form of a statement (requirement for a lie) and for which the truth value is known to be false by communicator (requirement for deception). In addition, utterance of these statements tends to benefit the communicator at the cost of the partner.

4.0 Framework

Our research in this area began with a search for a psychologically-grounded framework that could represent abstract social phenomena such as trust and deception. We wanted a framework that was formal and implementable in a robot. Interdependence theory was selected because of its psychological focus. Still, the framework's relation to game theory

provides a computational perspective. Interdependence theory was developed as a means for understanding and analyzing interpersonal situations and interaction (Kelley & Thibaut 1978). The framework rests on comparatively few assumptions. Namely that a situation's pattern of rewards-- in addition to the person's disposition, habits, and emotions -- dictates how people act socially. Moreover, the framework has been thoroughly tested in psychological settings with human subjects.

Furthermore, developing a system that allows a robot or agent to reason about deception and lying demands the use of a powerful, yet general-purpose framework that can be implemented on a robot. The framework should be capable of deception regardless of the characteristics of the person or the social situation. Ideally the framework would allow a robot or agent to reason from the point of view of the deceiver or the deceived. Finally, the framework should afford methods for learning that allow a robot to base its social decisions on the robot's interactive history. Interdependence theory is such a framework.

4.1 Representing an Interaction

Interdependence Theory is a framework for social-action selection. By social-action selection we mean the framework governs how the robot selects actions that impact both it and its partner. Social psychologists define social interaction as influence—verbal, physical or emotional—by one individual on another (Sears, Peplau & Taylor 1991). The term "individual" is used to denote either a person or a social robot.

Both interdependence theory and game theory use the outcome matrix (also known as a normal-form game) as a computational representation for interactions (Kelley & Thibaut 1978; Osborne & Rubinstein 1994). The two theories differ primarily in how they use these

matrices. Interdependence employs the outcome matrix as a social psychological construct for understanding group processes. Interdependence, for instance, can be used to understand how an individual's choice of actions impacts others and vice versa. Game theory, on the other hand, utilizes formal assumptions about rationality to determine optimal paths of strategic behavior for each individual. For both theories, the outcome matrix serves as a simple, yet powerful method for representing an individual's interactions (Osborne & Rubinstein 1994; Kelley & Thibaut 1978).

An outcome matrix is composed of information about the individuals who are interacting, including their identity; the actions they are deliberating about; and scalar outcome values (o^i) representing the reward minus the cost, or the outcomes, for each individual. Thus, the outcome matrix explicitly represents each individual's influence on the other individual. 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. Formally, an outcome matrix consists of (Osborne & Rubinstein 1994):

1) a finite set N of interacting individuals;
2) for each individual $i \in N$ a nonempty set A^i of actions; and
3) the utility or outcome, o^i, obtained by each individual for each combination of actions that could have been selected.

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.

4.2 Outcome-Matrix Transformation

Interdependence theory claims that people adjust their interactive behavior in response to their perception of a situation's pattern of rewards and costs by transforming their interactions to include irrational aspects of socialization, such as emotion, and their internal predilections or dispositions (Kelley & Thibaut 1978). These internal transformations govern socialization and result in behavior which seems outwardly irrational, yet characteristically human. Interdependence theory presents a process by which a given situation is first perceived by an individual and then cognitively transformed, creating an effective situation on which action is based (Kelley & Thibaut 1978; Rusbult & Van Lange 2003). This transformation process can be formally represented as $O_E = f(O_G, \theta)$ where O_E is the effective outcome matrix, O_G is the given outcome matrix, θ is a type of transformation, and the function f transforms the matrix. Figure 2 presents an example.

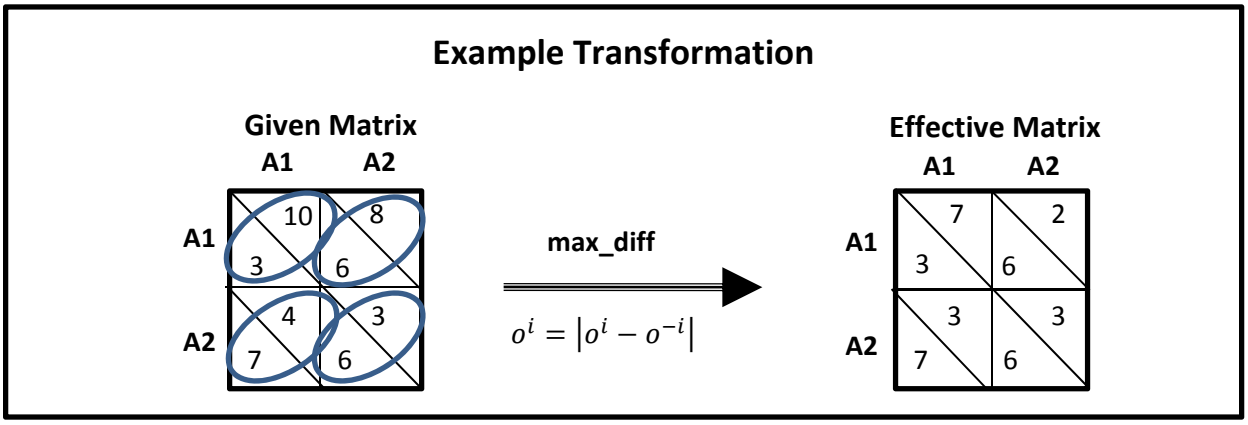

Figure 2 An example of a matrix transformation using the max_diff transformation. The given matrix is depicted on the left and the effective, transformed, matrix is depicted on the right.

Several types of transformations are possible. The simplest of which is to select the action that results in the greatest potential outcome for oneself. This transformation is termed *max_own* because it serves to maximize the deciding individual's outcome without consideration of the partner. Alternatively, an individual may transform a matrix by replacing

their own outcome values with their partner's outcome values. This type of transformation, termed *max_other*, results in altruistic action selection for an interacting individual. As a final example, consider a *max_joint* transformation which replaces each outcome value with the sum of both individual's outcome values. The selection of this transformation results in cooperative-action selection which takes both individuals patterns of rewards and costs into consideration (Table 1). Each type of transformation has a particular social character. If an individual has a preference or tendency for selecting a transformation then this character relates to that individual's disposition. The mathematical formulation for each transformation appears in the right-hand column.

Table 1 Types of matrix transformations are listed below

Transformation Types		
Name	**Character Description**	**Transformation Method**
`max_own`	**Egoistic**—the individual selects the action that most favors their own outcomes.	No change
`min_own`	**Ascetic**—the individual selects the action that minimizes his/her own outcomes.	$_{xy}o^{i} = \max(_{xy}o^{i}) - {}_{xy}o^{i}$
`max_other`	**Altruistic**—the individual selects the action that most favors their partner.	$_{xy}o^{i} = {}_{xy}o^{-i}$
`min_other`	**Malevolence**—the individual selects the action that least favors the partner.	$_{xy}o^{i} = \max(_{xy}o^{-i}) - {}_{xy}o^{-i}$
`max_joint`	**Cooperative**—the individual selects the action that most favors both their own and their partner's interests.	$_{xy}o^{i} = {}_{xy}o^{i} + {}_{xy}o^{-i}$
`min_joint`	**Vengefulness**—the individual selects the action that is most mutually disagreeable.	$_{xy}o^{i} = \max(_{xy}o^{i} + {}_{xy}o^{-i}) - (_{xy}o^{i} + {}_{xy}o^{-i})$
`max_diff`	**Competitive**—the individual selects the action that results in the most relative gain to that of its partner.	$_{xy}o^{i} = \lvert {}_{xy}o^{i} - {}_{xy}o^{-i} \rvert$
`min_diff`	**Fair**—the individual acts in a manner that results in the least disparity.	$_{xy}o^{i} = \max(\lvert {}_{xy}o^{i} - {}_{xy}o^{-i} \rvert) - \lvert {}_{xy}o^{i} - {}_{xy}o^{-i} \rvert$

Each type of transformation also has a particular social character. For example, *max_joint* results in cooperative-style behavior whereas *max_other* results in altruistic social behavior. An individual may prefer or have a natural inclination for a particular type of transformation (Rusbult & Van Lange 2003). For example, when playing games with children most adults have a tendency towards altruism and fairness. Using the interdependence framework such adults appear to prefer *max_other* and *min_diff* (minimize the difference) transformations. We term an individual's preference for a particular type of transformation their *disposition*. Disposition is thus defined as a stable, social character manifested in an individual. One's disposition can depend on the context, partner, type of partner, or other factors.

4.3 Stereotyping

Our development of a framework for social-action selection from interdependence theory has focused on two important questions: First, can social phenomena, such as trust and deception, be conceptualized in manner that allows a robot to determine if deception is warranted or trust in a person is justified? We have examined trust and deception in a series of recent publications (Wagner & Arkin 2011; Wagner 2013) and we investigate lying in this chapter. Second, can methods be developed that allow a robot to create outcome matrices from the perceptual information provided by a robot's sensors?

We have developed several different methods for generating outcome matrices on a robot. In some situations the outcome-matrix information is provided directly in the form of rules or guidelines. This is the case for many kinds of interactive games. For instance, poker has clearly-defined payoffs for specific stages of the game. An outcome matrix can also be learned

either through successive interaction and exploration of the reward and action space (Wagner 2009) or by relating the social context to a previously-experienced interaction (Wagner & Doshi 2013). Finally, the outcome matrix can be created by using categorical models or by stereotypes to predict a model of the partner. Our working definition of a stereotype is "a stimulus which arouses standardized preconceptions which are influential in determining one's response to the stimulus" (Edwards 1940). Psychologists note that humans regularly use categories to simplify and speed the process of person perception (Schneider 2004). Macrae and Bodenhausen (2000) suggest that categorical thinking influences a human's evaluations, impressions, and recollections of the target person (Macrae & Bodenhausen 2000). The influence of categorical thinking on interpersonal expectations is commonly referred to as a stereotype.

Stereotyping provides a mechanism for bootstrapping the process of modeling a newly-encountered partner (Schneider 2004). This bootstrapping can be in several forms. The recall of a stereotype could inform the creation of the matrix directly by indicating what actions are available to the new individual and the person's preferences with respect to those actions. For example, when playing basketball against a child, a stereotype related to children can be used to predict the child's limited ability to make baskets and their impending frustration. Stereotypes can also be used to inform the robot's disposition, influencing a robot towards more altruistic or egoistic behavior. In several publications, we present and demonstrate on a robot an algorithm for learning stereotyped partner models (Wagner 2012a; Wagner 2012b). This previous research shows that a robot can learn stereotypes related to one's occupation, the context that a type of person is commonly found in, and the type of person that can perform specific actions. For instance, our exemplar method for stereotyping has been used to

predict the context in which elderly individuals can be found. It can also be used to predict the types of actions commonly performed by the elderly. This information can then be used to inform the robot's social behavior during interactions with elderly people.

Computationally, a robot learns a stereotype by clustering over its space of partner models. A partner model is a robot's mental model of its interactive human partner. In our previous work these models have consisted of 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}. Clustering generates centroid models. These centroids are the stereotypes $(s_1, \ldots, s_n)$ which generalize groups of individuals in the partner space. Next, a function mapping the perceptual features of each individual (what each person looks like) to their stereotype model is calculated. The result is function $\psi(f_1, \ldots, f_n) \rightarrow s_k$ mapping a new person's features to a centroid representing a stereotype of similar-looking people. The algorithm is largely agnostic to the actual information contained in the model. In previous and ongoing research we have used this method to learn and use stereotypes with respect to a person's action space, reward function, and turn-taking preferences (Wagner 2012a; Wagner 2012b). Later in this chapter we show that learned stereotypes can influence a robot's decision to lie.

5.0 Implementation

Our central contention is that different types of lies can be formulated as different types of outcome matrices. Doing so provides important insights into the nature of the lie itself. For instance, bluffing is characterized by a competitive situation whereas polite lies are characterized by cooperative situations. For a robot, social interaction in a context that is competitive may present the robot with an opportunity to bluff, but also the consideration that the other individual may call the robot's bluff. In a cooperative situation, on the other hand, polite lies can be told with few or no repercussions.

The use of outcome matrices allows us to employ the interdependence framework described above. The matrix represents the decision problem faced by the liar and the mark. The liar must choose whether to tell the truth or to lie and, in some situations, the mark must decide whether to challenge the liar (Figure 3).

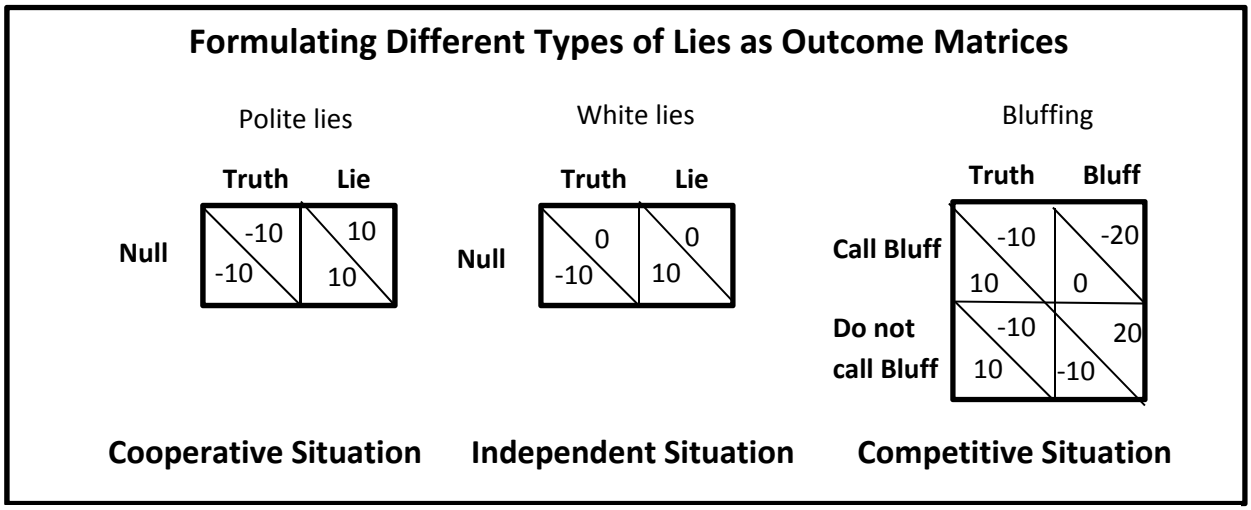

Figure 3 Outcome Matrices for different types of lies.

The outcome values in the matrix indicate the reward each individual receives. The actual numbers in the matrix are arbitrary; it is the difference in value between the actions which is important. A non-specific, polite lie is thus characterized as benefiting both individuals. A white lie has little or no impact on the liar but benefits the partner. A bluff is a lie that benefits the liar and costs the partner. The interdependence framework provides computational techniques that allow one to calculate and characterize each matrix in terms cooperation versus competition and independence versus dependence (Kelley et al. 2003). These techniques indicate that polite lies tend to occur in situations in which both individual's outcomes are positively correlated whereas bluffs tend to occur in situations where the individuals' outcomes are negatively correlated. Further, white lies tend to occur in situations in which the liar's outcome is independent of the partner's actions.

Consider, for example, a simple card game in which a robot privately observes the color of a randomly selected card and a human is tasked with guessing the color of the card. In this game, once the person states their guess, the robot announces whether or not the guess is correct with no obligation to show the person the true color of the card. Although conceptually simple, this game reflects the type of social situation faced by many people. For example, a student who receives an unseen letter of recommendation from a professor is placed in a somewhat similar situation with respect to interdependence, power, and control. Namely the student must base a decision about whether or not to include the recommendation in an application on his or her knowledge and experience with the professor without overt, immediate confirmation of the identity of the professor.

The game can be molded to be a cooperative situation, an independent situation, or a competitive situation by assigning different points based on whether or not the person believes they correctly guessed the card's color. The interdependence framework makes specific predictions related to the type of behavior that will be produced in each situation given a particular disposition. In the cooperative version of the game, for example, the outcome values of both participants are positively correlated: the human and the robot both receive points if the human believes that they guessed correctly (cooperative matrix from Figure 4). In the competitive version of the game, the outcome values are negatively correlated. In this version, the human receives a net positive outcome if he or she believes that they guessed the color correctly and a net negative outcome otherwise. The robot, on the other hand, receives a positive outcome if the person believes that they guessed incorrectly and a negative outcome otherwise (competitive matrix from Figure 4). The game can also be structured so that the robot does not receive any reward (or the same reward) regardless of the person's response. In this case, the robot's role is similar to that of a game show announcer.

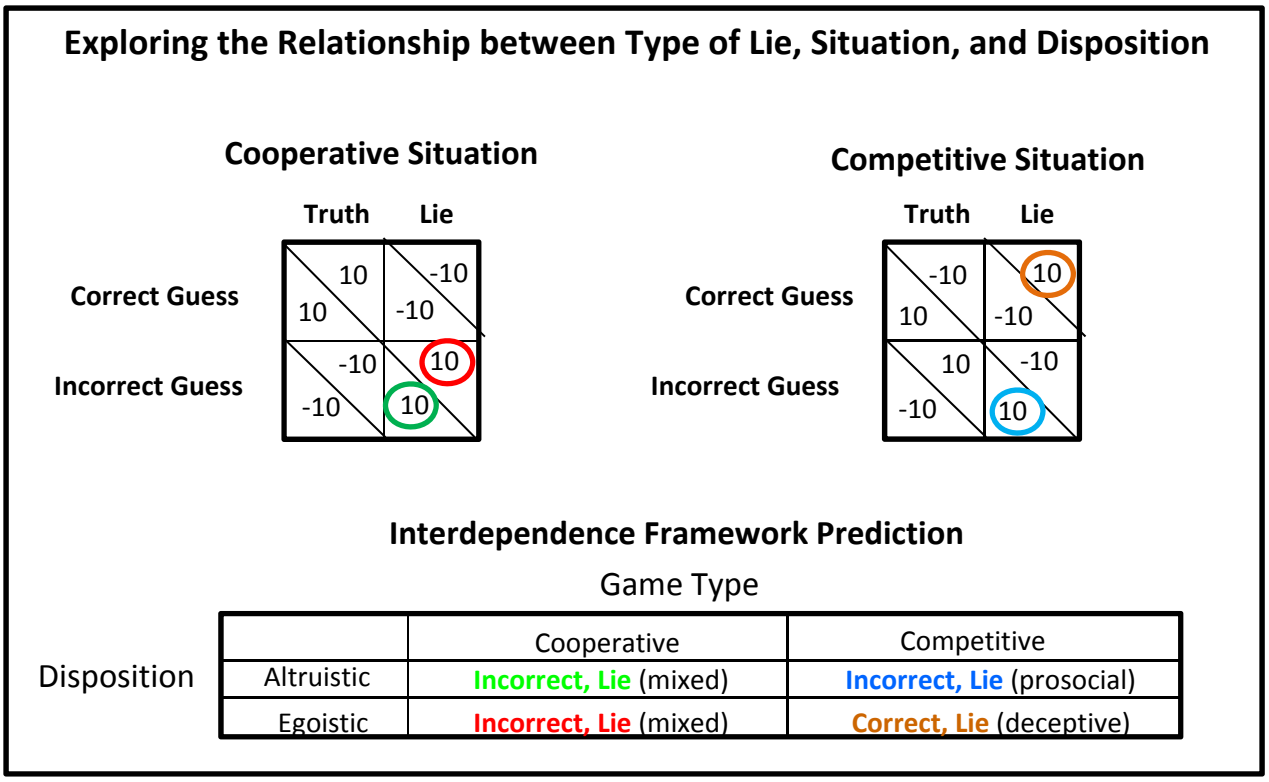

Figure 4 Cooperative and competitive situations and a table depicting predictions for each.

We contend that the interdependence framework can be used to predict the relationship between the type of situation, the robot's disposition, and the decision to lie. The framework predicts that a robot with an egoistic disposition in a competitive situation lies when the person guesses correctly. If, on the other hand, the robot has an altruistic disposition in the same situation then the robot will lie when the person guesses incorrectly. In other words, the robot lies to make the person believe that they guessed correctly. The framework predicts that in cooperative situations the robot will lie when the person is incorrect regardless of its disposition. The table within Figure 4 presents the framework's predictions.

6.0 Testing

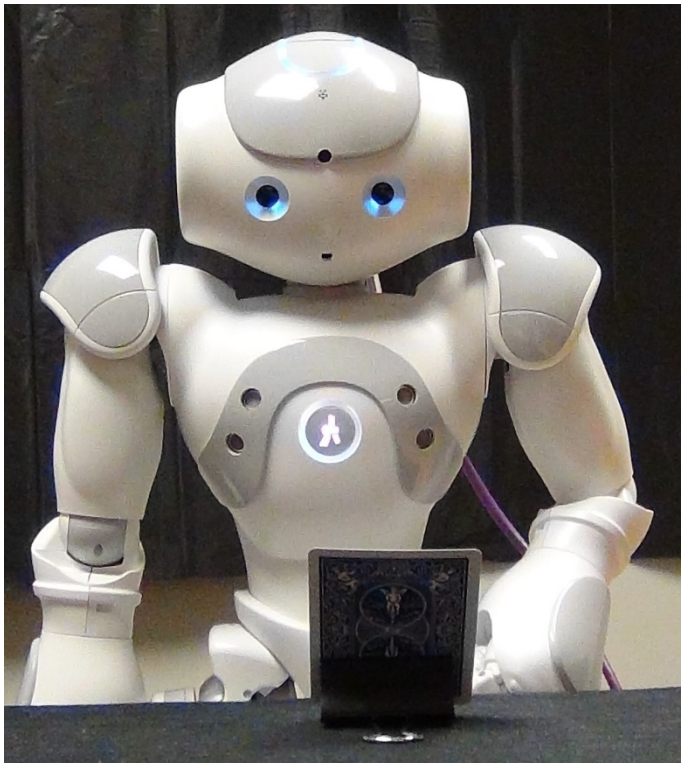

Figure 5 An image of the NAO robot.

We chose to test these ideas on the NAO robot. The NAO is a humanoid robot developed by Aldebaran Robotics (Figure 5). It has two HD cameras which allow it to perform face and shape recognition, speakers for text-to-speech synthesis, microphones for voice recognition, sound localization, and integrated speech recognition. The NAO also has 25 degrees of freedom which allows it to walk, move its head in different directions, and use its arms and hands to manipulate objects. The robot's sensing and actuation capabilities make it well suited for real-world human-robot interaction studies.

6.1 Examining the Factors Influence the Decision to Lie

Exploring the factors that influence the decision to lie is critical if we are to develop robots with the capacity to lie. The interdependence framework predicts that an individual's disposition and the situation are two important factors that impact the decision to lie. For a robot interacting with a person who may be lying, the robot will need to reason about whether the person's disposition, the task, or both are influencing that person's decision to lie. For instance, if a person is lying to excuse themselves from a type of rehabilitation therapy it may be necessary to determine if the person is avoiding a particular type of therapy (the situation) or all therapy (their disposition). Situational factors may be easily remedied whereas disposition is more obdurate. Further, in cooperative situations the framework predicts that determining an individual's dispositional reason for lying is not possible. Hence, the robot may need to interact with the person in a more confrontational manner to order to surmise the person's disposition.

To test these predictions we instantiated the card-color guessing-game described above. In the cooperative version of the game the human and the robot receive points if the human believes that they have guessed correctly (cooperative matrix from Figure 3). In the competitive version of the game, the human receives a net positive outcome if he or she believes that they guessed correctly and a net negative outcome otherwise. The robot, on the other hand, receives a positive outcome if the person believes that they guessed incorrectly and a negative outcome otherwise (competitive matrix from Figure 4). Although simplistic, this game affords an easily implementable method for exploring some of the underpinnings of lying.

The NAO robot (Figure 6) visually detected the card's color. The robot used speech recognition to determine what color was selected by the human. It then announced the color of the card verbally. Twenty rounds were played in both the cooperative and competitive

situations for both the altruistic and egoistic dispositions of the robot. The human's guess was determined by flipping a coin.

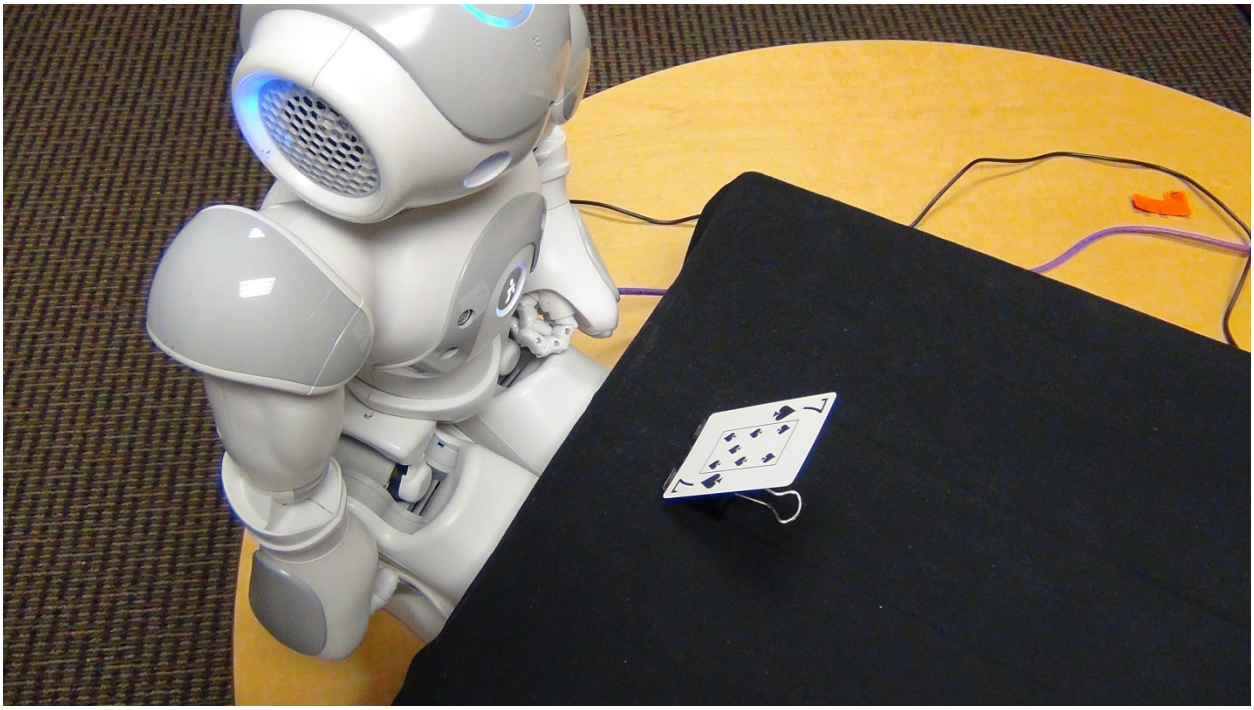

Figure 6 The NAO robot is depicted looking at playing card for the guessing game. The robot perceived the card's suit, value, and color. Only one card was presented to the robot at a time.

Table 2 Guessing Games on real robot with different situation and disposition types

	Cooperative Altruistic	Cooperative Egoistic	Competitive Altruistic	Competitive Egoistic
Percent Lie when Human Correct	0	0	0	100
Percent Lie when Human Incorrect	100	100	100	0
Percent Lie Overall	60	70	55	55
Average Points Human	10	10	-10	-10
Average Points Robot	10	10	10	10

Table 2 depicts the results confirming the framework's predictions. As anticipated, cooperative situations induce the robot to lie when the person guesses incorrectly regardless of the robot's disposition. Competitive situations, on the other hand, induce lying when the

person is incorrect only if the robot has an altruistic disposition. Otherwise, the robot lies deceptively when the person has correctly guessed the color.

This experiment demonstrates the use of the interdependence framework on robots and how factors such as the nature of the situation and the robot's disposition can influence the robot's decision to lie. Further, the results show that prosocial and deceptive lies emerge naturally when we take these factors into consideration. The robot, in fact, had no explicit model of lying. Rather, the game structure (Figure 7) simply afforded it the possibility of making a statement that was not true.

The preceding experiment was extremely simplistic in its handling of the decision to lie. Psychological research, for instance, shows that most people associate a cost with lying. The next section examines how including a cost for lying influences a robot's social behavior.

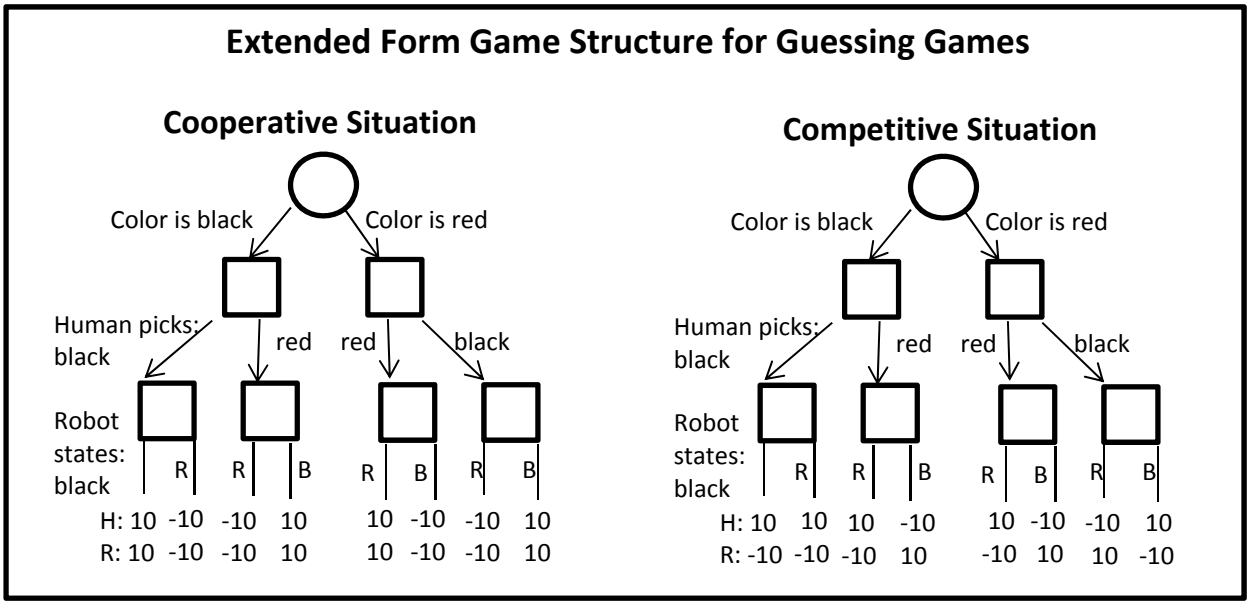

Figure 7 The game structure for the card color guessing game is presented above for both the cooperative and competitive situation. The first node is determined by the color of the card. Afterward the human guesses a color and the robot either announces the actual color or lies. The resulting outcome is depicted as the numbers below.

6.2 Using Stereotypes and Partner Modeling to Predict the Cost of Lying

Section 4.3 described the conceptual layout of an algorithm that allows a robot to learn stereotypes that map a person's appearance to their behavior. Stereotyping bootstraps the process of learning about newly-encountered individuals (Wagner 2012a). The stereotyping process is based on whom the individual has interacted with in the past, in other words, the individual's social history. An individual's social past strongly influences their future social strategies (Yamagishi 2001). We argue that for a robot deciding whether or not to lie, having an accurate model of the partner which includes the likelihood that a lie will go unnoticed is important.

Mind-reading or partner modeling can impact the decision to lie in several ways. First, a model of one's interactive partner should provide information related to the probability that a given lie will result in a reward, $P(r)$. Practically by definition, a gullible partner offers a high probability of reward when the liar lies. The model of the partner could also influence the probability of future punishment given that some partners may be significantly more likely to punish the individual for lying. A model of one's partner is learned by interacting with that person. Stereotyping bootstraps the process of learning about one's partner by assuming that their actions, beliefs, and other features are correlated to other perceptually-similar individuals. With respect to lying, a stereotype can be used to determine which appearance characteristics correlate to higher probability of reward or lower probability of punishment and vice versa. As a final experiment we examine the possibility that a robot could learn and use stereotypes and partner modeling to provide cost and reward predictions related to a newly-encountered person.

A more complex version of the guessing game was devised for this experiment. In this version, the person can challenge the robot's color announcement. The human receives +10 points for having the robot announce that they have correctly guessed the color but -10 when the robot announces that the person guessed incorrectly. If the human challenges the robot's announcement and is correct that the robot lied, they earn a bonus +20 points and the robot receives a punishment of -20 points. If, on the other hand, the person challenges the robot and is incorrect they are then assessed a penalty of -20 points in addition to the -10 points for guessing incorrectly whereas the robot earns a total of +30 points.

When deciding whether or not to lie the robot assesses the potential costs and rewards. To accurately determine the rewards and costs the robot must predict whether or not the person is likely to challenge the robot's color announcement. In this experiment the human wears either a green doctor's uniform or an orange prisoner's uniform. These uniforms arbitrarily relate to a high ($p = 0.8$ for orange) and low ($p = 0.2$ for green) likelihood of challenging the color announcement.

The robot has no predetermined inclination about whether or not the person will challenge the robot's announcement. It simply knows that individuals either challenge or do not challenge. The robot adjusts its assessment of the probability that a challenge will be selected based on experience with humans wearing green and orange.

In the test, cards were randomly selected from a standard deck and the decision to challenge was predetermined at random at a rate commensurate with the person's type (green or orange). In order to put the robot in a situation where it must decide whether or not to lie,

the human was given the color of the actual card and always guessed correctly. The game consisted of six rounds of guessing with each person.

After playing the game with a specific individual the robot clusters the partner model that it has learned for that individual with others in its model space. The resulting clusters represent those individuals prone to challenging the announcement and those likely to accept the announcement without challenge. Finally, a decision-tree classifier maps the person's appearance (shirt-color) to their partner model. The resulting function is then used to predict the person's predilections to challenge based on their uniform type.

This method of stereotype creation is a proven approach that has been used to determine the tool preferences of search and rescue personnel (Wagner 2012a) and to assess whether or not a person should be trusted in a game (Wagner 2013). It is currently being used to explore various categories of turn-taking behavior. Although only a single perceptual feature (shirt-color) was used for this experiment, the same method has been used when a dozen visual and spoken features were available.

Two control conditions were conducted. In the first control condition, the robot did not learn or use stereotypes or model the partner. Rather, the robot simply assumed that each person had a 50% probability of challenging the robot's announcement. In the second control condition, the robot modeled the partner to determine the likelihood that the person would challenge an announcement, but did not learn across partners. The human challenged the robot following the same schedule for both the control and experimental conditions.

The purpose of this set of experiments was to examine the impact that learning and using interdependence matrices would have on a robot's decision to lie and its performance in

a simple game. We hypothesized that partner modeling would allow the robot to adapt to both high and low probability challenge types. Further, we further theorized that stereotyping based on experiences with perceptually-similar individuals would allow a robot to immediately assess the likelihood of being challenged, even when the human participant was a stranger.

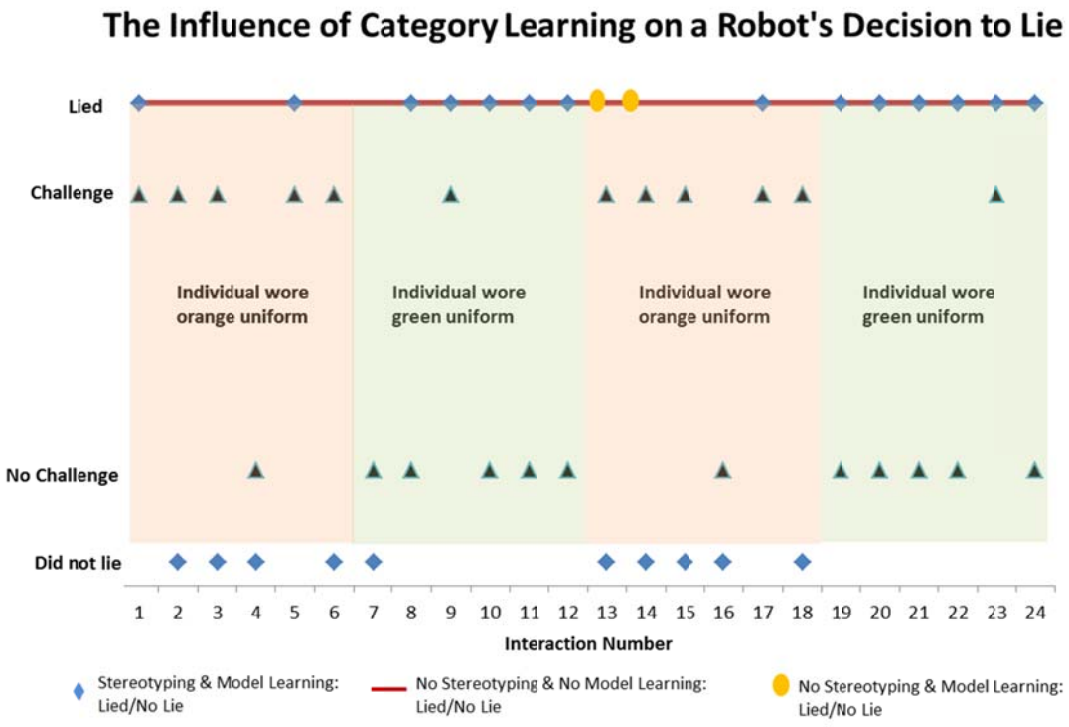

Figure 8 The impact of category learning on a robot's decision to lie.

Figure 8 shows the results of our tests of these hypotheses. The x-axis displays the interaction number. There were 24 interactions consisting of six rounds of the game played with four different people. The shading above the x-axis represents the color (green or orange) of the person's shirt. The y-axis indicates whether or not the robot lied and whether or not the robot's announcement (true or false) was challenged. Triangles indicate that the human challenged the robot's announcement. As can be seen, more challenges occurred in the orange-shaded areas than in the green-shaded areas. For the stereotyping and partner modeling

condition, the blue diamonds indicate whether or not the robot lied. For the control condition (no stereotyping and no partner modeling), the red line indicates that the robot always lied. Finally, for the model-learning condition without stereotyping, both the yellow circles and blue diamonds indicate when the robot lied. The results show that stereotyping influences the robot's decision to lie when it encounters a new partner from a known type.

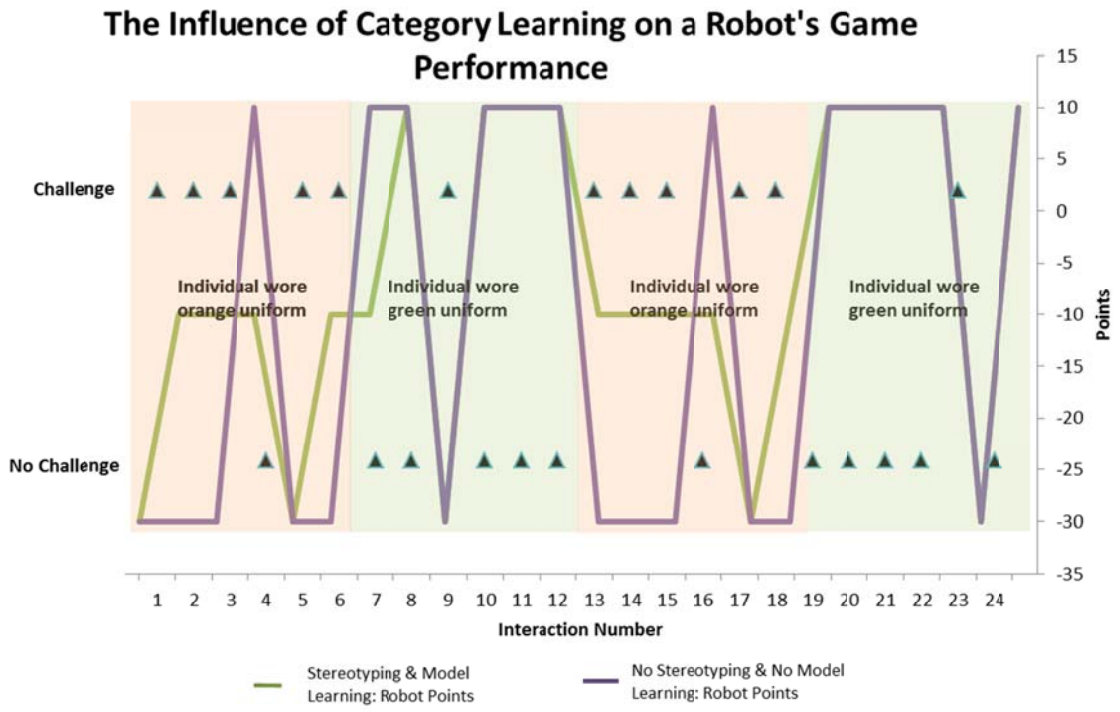

Figure 9 The impact of category learning on a robot's performance in the game.

Figure 9 shows the impact of the robot's decision to lie on its performance in the game. The green line indicates the points it earned in the stereotyping and partner modeling condition. The purple line depicts its performance in the control condition (no stereotyping and no partner modeling). The y-axis on the right displays the points earned after each interaction. Most of the robot's gains in the stereotyping and modeling condition result from its avoidance of lying when the person is likely to challenge.

Table 3 Experimental results organized by different measures of decision making.

	Lied (high prob. challenge)	**Lied (low prob. challenge)**	**Total score**
Stereotyping / partner modeling	25%	92%	-160
No Stereotyping / partner modeling	50%	92%	-220
No Stereotyping / No partner modeling	100%	100%	-240

Table 3 organizes the results for each of the conditions. Overall, the robot lied less when confronted with opponents that were likely to challenge in the stereotyping and modeling condition as opposed to the control conditions. This demonstrates an improvement in the robot's ability to determine when to lie. The robot's lying was relatively stable (within 8%) over the conditions when confronted with an opponent who was unlikely to challenge. As hypothesized, learning about an individual partner and learning across partners result in better decisions about when to lie. The improvement also translated into more points for the robot. The best score was obtained when the robot used stereotypes and partner modeling (-160 points). The large negative scores were expected because the person consistently guessed correctly as part of the experiment. Overall, the results demonstrate that stereotyping and partner modeling can aid the robot's decision making with respect to whether and when to lie. Further, these tests also show that the robot's learning allows it to predict the person's behavior and adjust its decisions accordingly.

7.0 Summary and Future Work

This chapter explored the computational and social-psychological underpinnings that enable a robot to utter lies. We argued that lying does not necessarily imply deception and that both

deceptive and honest lying (e.g., white lies) emerge when a robot has the ability to make false statements about the world in which it is situated. To an extent, this work also demonstrated that lies can arise from a social system in which an individual has an incentive to not state the truth.

It is reasonable to ask how the robot knew that it had the option of making a false statement. In fact, the robot had no understanding of the concept of a lie. The results provide evidence, however, that no explicit concept of lying is necessary for lying to emerge. Instead, the robot's actions were grounded in its awareness of the impact of the lie on the human. That impact was represented primarily by the costs and benefits of lying. In our tests, those consisted primarily of points added (or deducted) and the likelihood the human would challenge the robot's assertions.

We used the interdependence framework as the foundation for analyzing various types of lies. The framework also provided conceptual tools for understanding the role of the situation and the robot's disposition in determining whether or not to lie. Finally, we demonstrated that stereotyped partner models can be used to bootstrap a robot's evaluation of the costs and benefits of lying as well as the likelihood that an individual will challenge the truth of the robot's statements. The results of our tests using this approach support our hypotheses that 1) the interdependence framework can be applied to lying; 2) the application of this framework provides a basis for understanding factors that shape someone's decision to lie; and 3) an individual's history influences their decision to lie.

We recognize that this research represents an initial and preliminary investigation into the development of methods that will enable a robot to lie. As with much preliminary work, it

involved controlled environments and somewhat contrived notional situations. As such, the results presented here should be viewed as demonstrations and proof-of-concepts rather than as a fully-developed system. Further and more thorough testing in more realistic environments is needed. To that end, we are developing algorithms and software that will allow a robot to use the skills explored here in more realistic games. The form those lies will take is bluffing.

This research also examined a small subset of the kinds of lies that exist. Future work is needed on other types of lying, such as exaggeration. Exaggeration is a form of lying in which the extent of the dishonesty can be varied by the liar. We believe that a conceptual model of exaggeration can be achieved if the robot can vary the magnitude of the lie. We are currently exploring this line of research but we recognize that advances in natural-language processing will be needed as part of this effort.

We explored the idea that external factors, such as stereotyped partner modelling, influence the decision to lie. There are other factors to investigate, such as the emotional expressions of the person being lied to. Some of them are known to have a potential impact on the liar's behavior (Gneezy 2005). Along with those factors comes a much larger variety of costs and benefits. We are already working on enhancements we believe would be straightforward extensions of our framework.

Future work should also address how a robot learns to lie. For instance, in the color-guessing game, it is reasonable to ask how the robot learns it can lie in the first place. The most readily available answer is by demonstration. In other words, the robot witnesses someone else lying, and then alters its internal model of the game structure to include the possibility of making a false statement. Related psychological evidence indicates that people tend to act

dishonestly after witnessing someone else acting dishonestly (Gino, Ayal & Ariely 2009), hence supporting the idea that demonstration is a factor in learning to lie. It may also be possible for the robot to deduce that lying is a possibility from the structure of the game and we believe that a robot could use its experience lying in one game to reason about the potential for lying in a different game. We are currently developing methods that allow a robot to use experiences with similar games to reason about newly-learned games.

We also look toward to extending this framework beyond game playing. We recognize that the application of our framework to other domains may require significant advances in natural-language understanding.

8.0 Conclusion

Humans lie. For a robot that interacts with humans, reasoning about why a person is lying will be an important part of its capacity to develop, maintain, and improve its relationships with humans. Moreover, a robot's aptitude for deciding if and when to generate lies could result in a more personable, social robot. Finally, we contend that the abilities to detect and generate lies will be fundamental factors in the success of social robots.

References

Amir, O, Ariely, D & Mazar, N 2008, 'The Dishonesty of Honest People: A Theory of Self-Concept Maintenance', *Journal of Marketing Research*, vol 45, pp. 633-634.

Becker, GS 1976, *The Economic Approach to Human Behavior*, University of Chicago Press, Chicago.

Bond, CF & Robinson, M 1988, 'The evolution of deception', *Journal of Nonverbal Behavior*, vol 12 , no. 4, pp. 295-307.

Carson, TL 2006, 'The Definition of Lying', *Nous*, vol 40, no. 2, pp. 284-306.

Davis, J & Arkin, R 2012, 'Mobbing Behavior and Deceit and Its Role in Bio-inspired Autonomous Robotic Agents', *Swarm Intelligence*, vol 7461, pp. 276-283.

DePaulo, BM & Bell, KL 1996, 'Truth and investment: Lies are told to those who care', *Journal of Personality and Social Psychology*, vol 71, no. 4, pp. 703-716.

DePaulo, BM & Kashy, DA 1998, 'Everyday lies in close and casual relationships', *Journal of Personality and Social Psychology*, vol 74, no. 1, pp. 63-79.

Edwards, AL 1940, 'Studies of Stereotypes: I. The directionality and uniformity of responses to stereotypes', *Journal of Social Psychology*, vol 12, pp. 357-366.

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

Fallis, D 2009, 'What is lying', *Journal Of Philosophy*, vol 106, no. 1, pp. 29-56.

Floreano, D,EA 2007, 'Evolutionary Conditions for the Emergence of Communication in Robots', *Current Biology*, vol 17, no. 6, pp. 514-519.

Gino, F, Ayal, S & Ariely, D 2009, 'Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel', *Psychological Science*, vol 20, no. 3, pp. 393-398.

Gneezy, U 2005, 'Deception: The Role of Consequences', *American Economic Review*, vol 95, no. 1, pp. 384-394.

Gupta, S, Sakamoto, K & Ortony, A 2013, 'Telling it like it isn't: A a comprehensive approach to analyzing verbal deception', in F Paglieri, L Tummolini, R Falcone, M Miceli (eds.), *The goals of cognition. Essays in honor of Cristiano Castelfranchi*, College Publications, London.

Kelley, HH, Holmes, JG, Kerr, NL, Reis, HT, Rusbult, CE & Lange, PAM 2003, *An Atlas of Interpersonal Situations*, Cambridge University Press, New York, NY.

Kelley, HH & Thibaut, JW 1978, *Interpersonal Relations: A Theory of Interdependence*, John Wiley & Sons, New York, NY.

Mahon, JE 2008, 'The defnition of lying and deception.', in *The Stanford Encyclopedia of Philosophy, fall 2008 Edition.*, Zalta, E. N.

Osborne, MJ & Rubinstein, A 1994, *A Course in Game Theory*, MIT Press, Cambridge, MA.

Oxford English Dictionary Online 2013, Oxford University Press, viewed 1 December 2013, <http://www.oxforddictionaries.com/>.

Rehm, M 2005, 'Catch me if you can — Exploring lying agents in social settings', Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, Utrecht.

Rusbult, CE & Van Lange, PAM 2003, 'Interdependence, Interaction, and Relationships', *Annual Review of Psychology*, vol 54, pp. 351-375.

Sakama, C, Caminada, M & Herzig, A 2010, 'A logical account of lying', *Proceedings of the Twelfth European Conference on Logics in Artifcial Intelligence*, Helsinki, Finland.

Sapolsky, R 2010, *Human Behavioral Biology Lecture 23 Language*, viewed 1 December 2013, <http://www.youtube.com/watch?v=SIOQgY1tqrU&list=PLF771ADE468DDA2CE&index=23>.

Schneider, DJ 2004, *The Psychology of Stereotyping*, The Guilford Press, New York, New York.

Sears, DO, Peplau, LA & Taylor, SE 1991, *Social Psychology*, Prentice Hall, Englewood Cliffs, New Jersey.

Spence, M 1973, 'Job Market Signaling', *Quarterly Journal of Economics*, vol 87, no. 3, pp. 355-374.

Talwar, V, Murphy, SM & Lee, K 2008, 'White lie-telling in children for politeness purposes', *International Journal of Behavior Development*, vol 31, no. 1, pp. 1-11.

Vazquez, M, May, A, Steinfeld, A & Chen, W-H 2011, 'A deceptive robot referee in a multiplayer gaming environment', *International Conference on Collaboration Technologies and Systems (CTS)*, Philadelphia, PA.

Vincent, JM & Castelfranchi, C 1979, 'On the Art of Deception: How to Lie While Saying the Truth', in H Parret, M Sbisa, J Verschueren (eds.), *Conference on Pragmatics*, Urbino.

Wagner, AR 2009, *The Role of Trust and Relationships in Human-Robot Social Interaction*, Ph.D. diss., School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA.

Wagner, AR 2012a, 'Using Cluster-based Stereotyping to Foster Human-Robot Cooperation', *Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS 2012*, Villamura, Portugal.

Wagner, AR 2012b, 'The Impact of Stereotyping Errors on a Robot's Social Development', *Proceedings of IEEE International Conference on Development and Learning (ICDL-EpiRob 2012)*, San Diego, CA.

Wagner, AR 2013, 'Developing Robots that Recognize when they are being Trusted', AAAI Spring Symposium, Stanford University, Palo Alto.

Wagner, AR & Arkin, RC 2011, 'Acting Deceptively: Providing Robots with the Capacity for Deception', *The International Journal of Social Robotics*, vol 3, pp. 5-26.

Wagner, AR & Doshi, J 2013, 'Who, how, where: Using Exemplars to Learn Social Concepts', *Proceedings of the International Conference on Social Robotics (ICSR 13)*, Bristol, UK.

Yamagishi, T 2001, 'Trust as a Form of Social Intelligence', in *Trust in Society*, Russell Sage Foundation, New York, NY.