A Taxonomy of Robot Deception and its Benefits in HRI

Jaeeun Shim
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
jaeeun.shim@gatech.edu

Ronald C. Arkin
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
arkin@cc.gatech.edu

***Abstract*—Deception is a common and essential behavior in humans. Since human beings gain many advantages from deceptive capabilities, we can also assume that robotic deception can provide benefits in several ways. Particularly, the use of robotic deception in human-robot interaction contexts is becoming an important and interesting research question. Despite its importance, very little research on robot deception has been conducted. Furthermore, no basic metrics or definitions of robot deception have been proposed yet. In this paper, we review the previous work on deception in various fields including psychology, biology, and robotics and will propose a novel way to define a taxonomy of robot deception. In addition, we will introduce an interesting research question of robot deception in HRI contexts and discuss potential approaches.**

***Index Terms*—Robot Deception; Robot Behavior; Human-Robot Interaction; Robot Ethics**

I. Introduction

Deception is a common behavior not only in humans but also in animals. Various biological research findings demonstrate that animals act deceptively in several ways to enhance their chances of survival. In human interaction, deception is ubiquitous, occurs frequently during peoples development and is present in personal relationships, sports, culture, and even war. Thus, it is fair to assume that, similar to animals and humans, robots can also benefit from the appropriate use of deceptive behavior.

Since the use of deceptive capabilities in robotics features many potential benefits, it is becoming an important and interesting research question. Despite its importance, however, very little research on robot deception has been conducted until recently. Much of the current research on robot deception [1], [2], [3] focuses on applications and not on fundamental theory, such as the definition of a taxonomy for robot deception. For more details, we review previous research on robot deception in Section II

We contend that defining robot deception and establishing a taxonomy for it are important as a foundation for further robotics research on the subject, and herein we present such a taxonomy. To accomplish this, in Section III we carefully review different ways of defining deception from the fields of psychology, biology, and the military, and survey previous research on robot deception.

Since the use of social robots in everyday life is increasing, we particularly concentrate on robot deception in Human-Robot Interaction (HRI) contexts. We hypothesize that a robot, in order to be intelligent and interactive, should have deceptive capabilities that may benefit not only the robot itself but also its deceived human partner, in some cases to the detriment of the deceiver (the robot). In Section IV, we focus on robot deception, specifically in HRI contexts, and present an interesting research question and potential approaches to its answer.

In sum, this paper has two main goals: 1) to introduce a taxonomy of robot deception, and 2) to provide a research question of robot deception in HRI. First, we present a novel way to define robot deception and then characterize a taxonomy of robot deception. Second, we focus on robot deception in HRI contexts, which leads to the following research question:

*Can a robots deceptive behaviors benefit a **deceived** human partner in HRI contexts?*

In Section IV , some initial ideas to answer this research question are also presented. Even though robot deception can provide several advantages to humans, it is arguable whether it is morally acceptable to deceive humans at all in HRI contexts. We consider this ethical issue and introduce some perspectives to approach robot deception problems in Section V.

II. Previous Research

Endowing robots with the capacity for deception has significant potential utility [4], similar to its use in humans and animals. Clearly, deceptive behaviors are useful in the military domain [5], [6]. Sun Tzu stated in The Art of War, “All warfare is based on deception.” Military robots capable of deception could mislead opponents in a variety of ways. As both individual and teams of robots become more prevalent in the militarys future [7], [8], robotic deception can provide new advantages apart from the more traditional one of force multiplication.

In other areas, such as search and rescue or healthcare, deceptive robots might also add value, for example, for calming victims or patients when required for their own protection. Conceivably, even in the field of educational robots, the deceptive behavior of a robot teacher may potentially play a role in improving human learning efficiency.

Despite the ubiquity in nature and the potential benefits of deception, very few studies have been done on robotics to date. One interesting application in robot deception is the camouflage robot, developed at Harvard University [1]. Camouflage is a widely used deception mechanism in animals and militaries. Inspired by these real-world usages, the researchers at Harvard developed this soft robot, which can automatically change the color of body to match its environment.

Motion camouflage has also been studied for robot systems. Unlike the previous type of camouflage, motion camouflage is a behavioral deception capability observed in dragonflies. By following indirect trajectories, dragonflies can deceptively approach as if they were remaining stationary from the perspective of the prey. Carey et al. [9] developed an optimal control mechanism to generate these motion camouflage trajectories and verified it with simulation results. For real robot systems, more recent research in [10] proposed new motion camouflage techniques that are applicable to unicycle robots.

Floreano's research group [2] demonstrated robots evolving deceptive strategies in an evolutionary manner, learning to protect energy sources. Their work illustrates the ties between biology, evolution, and signal communication and does so on a robotic platform. They showed that cooperative communication evolves when robot colonies consist of genetically similar individuals. In contrast, when the robot colonies were dissimilar, some of the robots evolved deceptive communication signals.

Wagner and Arkin [4] used interdependence theory and game theory to develop algorithms that allow a robot to determine both when and how it should deceive others. More recent work at Georgia Tech is exploring the role of deception according to Grafen's dishonesty model [11] in the context of bird mobbing behavior [12]. Another study is developing robot's deceptive behavior inspired by biology. It applies squirrel's food protection behavior to robotic systems and shows how a robot successfully uses this deception algorithm for resource protection [3].

Much research on robot deception has also been proposed in HRI contexts. Terada and Ito [13] demonstrated that a robot is able to deceive a human by producing a deceptive behavior contrary to the human subject's prediction. These results illustrated that an unexpected change of the robot's behavior gave rise to an impression in the human of being deceived by the robot.

Other research shows that robot deceptive behavior can increase users' engagement in robotic game domains. Work at Yale University [14] illustrated increased engagement with a cheating robot in the context of a rock-paper-scissors game. They proved greater attributions of mental state to the robot by the human players when participants played against the cheating robots. At Carnegie Mellon University [15] a study showed an increase of users' engagement and enjoyment in a multi-player robotic game in the presence of a deceptive robot referee. By declaring false information to game players about how much players win or lose, they observed whether this behavior affects a human's general motivation and interest based on frequency of winning, duration of playing, and so on. These results indicate that deceptive behaviors are potentially beneficial not only in the military domain but also in a human's everyday context.

Brewer et al. shows that deception can be used in a robotic physical therapy system [16]. By giving deceptive visual feedback on the amount of force patients currently exert, patients can perceive the amount of force lower than the actual amount. As a result, patients can add additional force and get benefits during the rehabilitation.

Recent work from the University of Tsukuba [17] shows that a deceptive robot partner can improve the learning efficiency of children. In this study, the children participated in a learning game with a robot partner, which pretends to learn from children. In other words, the robot partner in this study is a care-receiving robot, which enables children to learn by teaching [18]. The goal of this learning game is for kids to draw the shape of corresponding English words such as circle, square, and so on. The interesting part is that the robot acted as an instructor, but deliberately made mistakes and behaved as if it did not know the answer. According to the results, by showing these unknowing/unsure behaviors, the learning efficiency of the children was significantly increased. Since robots' unsure/dumb behaviors can affect a humans learning efficiency, we assume that these results relate to a robot's deceptive capabilities. As a result, we can conclude that this study provides preliminary results of the positive effects of robots' deceptive behavior in HRI contexts.

III. TAXONOMY OF ROBOT DECEPTION

A. *Taxonomies of Deception from a Human Perspective*

In other disciplines, researchers have developed the definitions and taxonomies of deception drawing from the fields of psychology, biology, military, engineering, etc. In this section, several ways to define and categorize deception in different fields are reviewed followed by a suggested taxonomy of deception from a robotic perspective.

Deception has been studied extensively by observing different human cases. Several ways to define and categorize deception have been proposed already by different psychologists and philosophers. Chisholm and Freehan [19] categorized deception from a logical and philosophical viewpoint. Three dimensions were described for distinguishing among types of deception such as commission-omission (the attitude of the deceiver; the deceiver "contributes causally toward" the mark's changes or the deceiver "allows" the mark's changes with respect to belief states), positive-negative (the belief state of the mark; the deceiver makes the mark believe that false proposition is true vs. true proposition is false), and intended-unintended (whether the deceiver changes the mark's belief state or merely sustains it). From the combination of those three dimensions, they provided eight categories of human deception as shown in Table I(a).

From the results of diary studies and surveys, DePaulo [20] divides deception in four different ways: content, type, referent, and reasons (Table I(b)). Subcategories of these kinds of deception are also observed and defined. Subcategories of content are feelings, achievements, actions, explanations, and facts. In the category of reasons, there are subcategories of self-oriented and other-oriented deception. In type category, outright, exaggerations, and subtle were defined as subcategories. Also, four different referents were suggested such as liar, target, other person, and object/event.

Military is one of the biggest contexts for the use of deceptive behavior. Dunnigan and Nofi [21] proposed a taxonomy of deception based on ways to generate the deceptive behaviors as shown in Table I(d).

Whaley [22], [23] suggested six categories of deception and grouped them into two sets. The six categories of deception are masking, repackaging, dazzling, mimicking, inventing, and decoying. These categories are grouped into dissimulation and simulation (Table I(d)). The first three, masking, repacking and dazzling, are categorized as dissimulation (the concealment of truth) and the others are in the simulation category (the exhibition of false).

Recently, Erat and Gneezy [24] classified four types of deception based on their consequences: selfish black lies, spite black lies, pareto white lies, and altruistic white lies (Table I(c)).

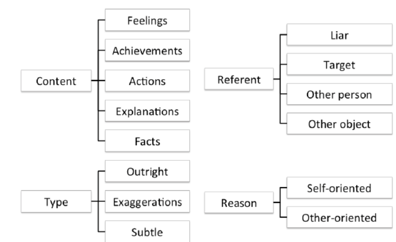
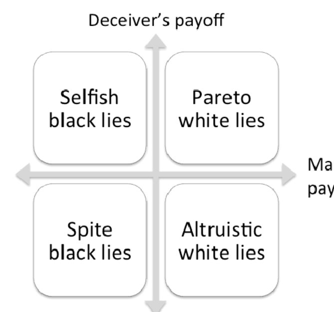
In cyberspace, deception happens frequently and a taxonomy of deception has been proposed by Rowe et al. [25] for this domain. They defined seven categories of cyberspace deception based on linguistic case theory, including: space, time, participant, causality, quality, essence, and speech-act. By exploring subcategories on each case, they proposed 32 types for a taxonomy of cyberspace deception (Table I(e)).

Many deceptive behaviors are also observed in nonhuman cases. Animal deception can be categorized into depending on its cognitive complexity [26], specifically the two categories of unintentional and intentional animal deception (Table I(f)). Unintentional animal deception includes mimicry and camouflage. In contrast, intentional deception requires more sophisticated behavioral capacities such as broken-wing displays or in many non-human primate examples such as chimpanzee communication [27].

Recently, researchers in Human-Computer Interaction (HCI) defined the notion of *benevolent deception*, which aims to benefit not only the developers but also the users [28]. They have not proposed a taxonomy of deception, but provided new design principles regarding deception in HCI.

B. A Taxonomy of Robot Deception

Based on previous efforts in this area, a taxonomy of robot deception was developed. Similar to human and animal deception, robot deception occurs during the interactions among robots or between humans and robots. Therefore, analyzing these interactions can identify the key factors to categorize

<table>
<tr><th>Field</th><th>Method</th><th>Taxonomy</th></tr>
<tr><td>(a)
Philosophy</td><td>Logical and philosophical view points with a proposition</td><td>
<table>
<tr><td></td><td></td><td>Omission (O)</td><td>Commis-sion(C)</td></tr>
<tr><td rowspan="2">Intend -ed (I)</td><td>Pos (P)</td><td>O-P-I</td><td>C-P-I</td></tr>
<tr><td>Neg (N)</td><td>O-N-I</td><td>C-N-I</td></tr>
<tr><td rowspan="2">Unintend -ed (U)</td><td>Pos (P)</td><td>O-P-U</td><td>C-P-U</td></tr>
<tr><td>Neg (N)</td><td>O-N-U</td><td>C-N-U</td></tr>
</table>
</td></tr>
<tr><td>(b)
Psychology</td><td>Analysis results of diary studies and surveys</td><td>
</td></tr>
<tr><td>(c)
Economics[1]</td><td>Deceiver's and mark's consequences</td><td>
</td></tr>
<tr><td>(d)
Military</td><td>Representing deception</td><td>

Dissimulation
Masking: hiding in background
Repacking: hiding as something else
Dazzling: hiding by confusion
Simulation
Mimicking: deceiving by imitation
Inventing: displaying a different reality
Decoying: diverting attention

</td></tr>
<tr><td>(e)
Cyberspace</td><td>Semantic cases (Linguistic case theory)</td><td>
<table>
<tr><td>Space</td><td>Direction, location-at, loc-from, loc-to, loc-through, orientation</td></tr>
<tr><td>Time</td><td>Frequency, time-at, time-from, time-to, time-through</td></tr>
<tr><td>Participant</td><td>Agent, beneficiary, experiences, instrument, object, recipient</td></tr>
<tr><td>Causality</td><td>Cause, contradiction, effect, purpose</td></tr>
<tr><td>Quality</td><td>Accompaniment, content, manner, material, measure, order, value</td></tr>
<tr><td>Essence</td><td>Super type, whole</td></tr>
<tr><td>Speech-act</td><td>External precondition, internal precondition</td></tr>
</table>
</td></tr>
<tr><td>(f)
Biology</td><td>Cognitive complexity</td><td>Intentional vs. Unintentional Deception</td></tr>
</table>

TABLE I: Taxonomies of Deception in Various Fields

[1]The chart is reproduced with permission from [24]'s Figure 1. Taxonomy of Lies on Change in Payoff.

Dimensions	Categories	Specifications
Interaction Object	Robot-human deception (H)	Robot deceives human partners
	Robot-nonhuman deception (N)	Robot deceives nonhuman objects such as other robots, animals, and so on.
Interaction Goal (reason)	Self-oriented deception (S)	Deception for robot's own benefit
	Other-oriented deception (O)	Deception for the deceived other's benefit
Interaction Type	Physical/ unintentional deception (P)	Deception through robot's embodiments, low cognitive / behavioral complexity
	Behavioral/ intentional deception (B)	Deception through robot's mental representations and behaviors, higher cognitive complexity

TABLE II: Three Dimensions for Robot Deception Taxonomy

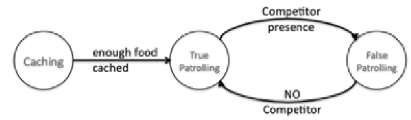
robot deception. Similar to Chisholm and Freehans approach [19], we specify the salient dimensions of robot deception first, and then define a taxonomy of robot deception based on these characteristics.

The three dimensions of robotic deception are defined as deception object, deception goal, and deception method (Table II). First, interaction object indicates with whom the robot interacts with and thus tries to deceive. In this dimension, deception can be classified into the two categories of robot-human deception, and in robot-nonhuman deception.

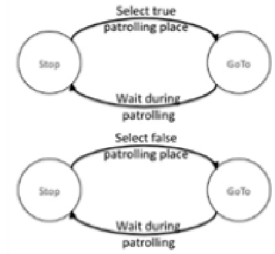
The second dimension is deception goal. This approach is similar to the distinctions in DePaulo's taxonomy, especially the reason category [20], by categorizing robot deception based on the reason why a robot tries to deceive others: self-oriented deception and other-oriented deception. Self-oriented deception means that the robot's deceptive behaviors benefit the robot itself. In contrast, other-oriented deception occurs when the goal of robot deception is to provide advantage to the deceived robots or human partners, even at the robot's own expense.

The final dimension is deception method, which is the way by which the robot generates deception. This dimension is similar to the taxonomy of animal deception: intentional and unintentional deception. It includes embodiment/physical deception and mental/behavioral deception. Embodiment deception indicates deception resulting from the robot's morphology such as camouflage. In mental/behavioral deception, a robot deliberately generates intentional deceptive behaviors.

From the combinations of those three dimensions, we can define a taxonomy of robot deception as shown in Table III. Each element of the taxonomy (type) consists of a combination of three categories, one from each dimension, providing eight different types of robot deception. The table also includes examples of each type of robot deception. As shown in this table, N-S-P and N-O-P types do not have specific examples in robot contexts yet. Therefore, we exclude these two types from further consideration. Thus, based on the characteristics of interactions in current robot systems, six different types of robot deception are defined to constitute our taxonomy of robot



(a)



(b)

Fig. 1: FSA for robot deception algorithms: (a) High-level FSA: caching and patrolling behaviors of squirrel robot, (b) Sub-FSA: food patrolling (top: normal, bottom: deception)

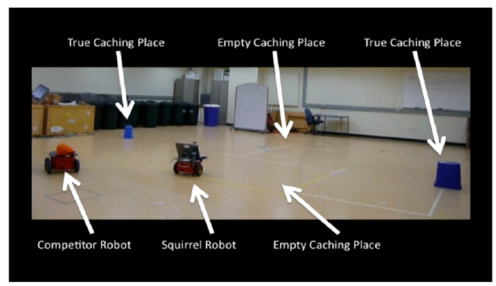


Fig. 2: Robot Experiment Layout with Two Pioneer Robots

deception.

C. Robot Deception: A Case Study

A taxonomy for robot deception was presented above by defining salient dimensions and categories. In our previous research [3], we developed and evaluated a robot's deceptive behaviors for resource protecting strategies, which is potentially applicable in military context and inspired by biology. The patrolling strategy used by Eastern Grey Squirrels is one interesting example in nature regarding the possible role of deception [30], where they use deception to protect their food caches from other predators.

Briefly, the squirrel spends time visiting stocked food caches. It was observed, however, that when a predator is present, the squirrel changes its patrolling behavior to spend time visiting empty cache sites, with the apparent intent to mislead the raider into the belief that those sources are where the valuables are located, a diversionary tactic of sorts.

Inspired by these deceptive behaviors of squirrels, we developed and implemented deception algorithms for a robot. Figure 1a illustrates the high-level model of algorithms using a finite state acceptor (FSA). After caching is complete, the robot then begins to move between the caching locations in order to patrol its resources. The behaviors of the robot include goal-oriented movement, selecting places, and waiting behavior as shown

Taxonomy	Definition	Examples
H-S-P	Deceiving human for deceiver robot's own benefit using physical interactions	Camouflage robots - DARPA's soft robot [1]
N-S-P	Deceiving other robot or nonhuman for deceiver robot's own benefit using physical interactions	N/A
H-O-P	Deceiving human for deceived human's benefit using physical interactions	Android Robots
N-O-P	Deceiving other robot or nonhuman for deceived other's benefit using physical interactions	N/A
H-S-B	Deceiving humans for deceiver robot's own benefit using behavioral interactions	Robot deception in HRI [13]
N-S-B	Deceiving other robots or nonhumans for deceiver robot's self benefit using behavioral interactions	Mobbing robot [12], Robot deception using interdependence theory [4], Squirrel-like resource protection robot [3]
H-O-B	Deceiving humans for deceived human's benefit using behavioral interactions	Robot deception in entertainment [14], Deceptive robot learner for children [17], Robot referees for human game players [15]
N-O-B	Deceiving other robots or nonhumans for deceived other's benefit using behavioral interactions	Robot Sheepdog [29]

TABLE III: Robot Deception Taxonomy

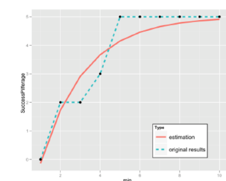
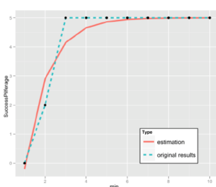
	With Deception	**Without Deception**
Result		
Convergence Rate	-2.2188	-1.1035

TABLE IV: Robot Experiment Results with Convergence Rate

in Figure 1b. Initially, the robot employs the true patrolling strategy when the select-true-location trigger is activated. This trigger calculates which of the many caching locations the robot should patrol in the current step. The calculation is made by random selection based on the transition probabilities among the places. Transition probabilities are determined by the number of cached items. In other words, if a place has more items, the probability of visiting that place is higher. When a competitor robot is detected, the squirrel robot starts the false patrolling strategy and selects/goes to false-patrolling-locations as shown in Figure 1b.

This is a form of misdirection, where communication is done implicitly through a behavioral change by the deceiver. We implemented this strategy in simulation [3], and showed that these deceptive behaviors worked effectively, enabling robots to perform better using deception than without with respect to delaying the time of the discovery of the cache. We also evaluated this algorithm by applying it to real robot system using the experimental layout in Figure 2. Table IV illustrates the results of experiments. As the graphs show, experimental results with deception converges to the predator's maximum successful pilferages more slowly than without deception. Therefore, it can be concluded that the deception algorithm leads to a robot's better resource protection performance.

This deception capability for a robot can be categorized in terms of our taxonomy. First, the object that a robot tries to deceive is other competitor robot, which means nonhuman objects (N). The deception happens through the robot's behaviors by intentionally misleading the competitor robot (B). In deception goal dimension, the benefits of this deception capability are protecting the deceiver's resources longer, so the squirrel robot obtains advantage: i.e., self-oriented deception (S). As a result, this squirrel robot deception is classified as N-S-B type in our taxonomy.

Other applications of robot deception can be similarly categorized in the taxonomy. Examples of how to categorize various forms of robot deception in the taxonomy are shown in Table III.

IV. OTHER-ORIENTED ROBOT DECEPTION IN HRI

Many researchers aim to build social robots that feature intentionality. According to Dennett [31], a high-order intentionality can be achieved by adding several different features, notably deception capability. In other words, more intentional and autonomous social robots are possible when deception capabilities are added. Therefore, research on robot deception is arguably an important topic in HRI studies.

HRI studies generally aim to evaluate how a robot affects human partners, and the goal of such studies is usually to achieve effective and positive interactions between robots and humans. Similarly, it is necessary to consider the potential benefits to human partners when we are dealing with robot deception in HRI contexts. In other words, one goal of a robot's deceptive behaviors in HRI should be to provide advantage to the deceived human.

Earlier we saw several studies in the field of psychology that defined deception in different ways. In particular, DePaulo [20] defined and characterized a taxonomy for human deception based on motivations such as self-oriented and other-oriented deception. We also capture this category in the taxonomy of robot deception presented in the previous section and defined one dimensions based on the deception's goal.

Other-oriented robot deception should be developed and evaluated when applying deception capabilities to robots that interact with human partners in different HRI contexts for their benefit. As a result, we ask:

*Can a robots deceptive behaviors benefit a **deceived** human partner in HRI contexts?*

The notion of other-oriented deception has been also proposed in HCI as *benevolent deception* [28]. However, the embodied nature of a robotic agent distinguishes it from traditional HCI with respect to the effect on a human actor. Therefore, it is required to develop principles of other-oriented robot deception in HRI separately.

To answer this research question, the following research steps are being undertaken. First, a computational model and associated algorithms need to be developed specifically for other-oriented robot deception. The computational model must then be implemented and embedded on a robotic platform. The effects of this type of deception should be evaluated using HRI studies that are carefully constructed and conducted to observe whether the deceived human partners could truly obtain benefit from a robot's other-oriented deception capabilities.

We also characterize the situational conditions pertaining to the application and utility of other-oriented deception, by grouping and characterizing relevant situations of other-oriented deception. From reviewing various situations when other-oriented deception occurs between humans, they can be grouped along two dimensions: 1) the time duration of the deception; and 2) the payoff of the mark (the deceived person). The time dimension ranges from one-shot to short-term to long-term, referring to the length of time deception is maintained by the deceiver's actions. The mark's payoff is categorized by the effect on the mark's outcome (ranging from high to low payoff).

As shown in Figure 3, representative other-oriented examples in these dimensions are illustrated by their location in this two dimensional space. They include:

1. *Crisis management* is a situation where the deceiver's deceptive behaviors or lies must have a rapid effect on the mark (short-term) perhaps in a life-threatening high payoff situation. For example, other-oriented deception in a search-and-rescue situation may involve immediate emotional or physiological remediation for a victim. Lying to a mark regarding the direness of their situation in order to calm him/her down or to increase their confidence may increase their likelihood of survival in this life critical situation.

2. When someone faces highly stressful situation such as a big presentation in front of huge crowd or an athletic trial, people sometimes lie to *cheer up / increase their confidence* to let the speaker calm down in short-term such as "Don't worry! You're perfectly prepared" or "I know you can successfully do this."

3. *Quality of Life Management (QoLM)* involves maintaining deception over long periods of time, again for potential life-critical (health) situations in therapeutic treatment of serious or generative illness, or regarding status of long-term economic well-being. For example, *placebos* may be persistently used for a deceived patient's long-term benefit [32]. Long-term lying can also be used in a similar manner with the hopes of benefitting the patient.

4. Sometimes, teachers also behave deceptively or lie for

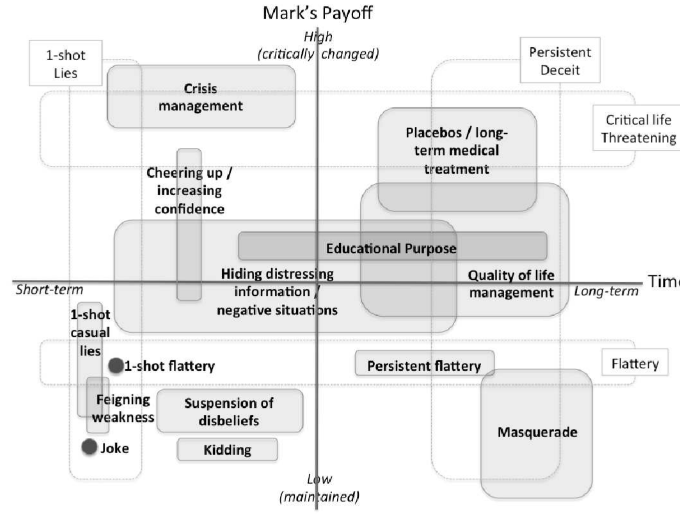


Fig. 3: Situational Conditions for Other-oriented Deception with Examples

educational purposes, perhaps playing dumb for example [17]. This deception can increase the student's learning efficiency, and it produces long-term benefits to the mark, although the deceit may be either short or long-term.

5. One-Shot *Casual Lies* are common in general conversation [20]. Generally, deceivers act deceptively or tell a lie to maintain the mark's emotional state for good. For example, general lies such as "you look nice today" or "I like your clothes" are obvious examples of 1-shot casual lies. These are not life-critical situations. "That was a great presentation" can also be another such example.

6. *Flattery* also ranges from the short- to long-term to make the mark's emotional state beneficial to their performance. *Persistent flattery* is an example (e.g., "a**-kissing") that makes the mark feel undeservedly better about themselves for a relatively long-term period. In this long-term case, benefit (payoff) accrues for both the mark and the deceiver, but we focus for now solely on the benefits to the mark.

7. Peoples sometimes *feign weakness* to make marks feel better by helping deceivers in short-term periods. Marks can maintain emotionally good state or feel better and confident from this deception. For example, a woman might pretend not to be able to open a jar just to make the man feel better and more confident about himself.

8. One-shot *Jokes* or more persistent *kidding* using deception is also an example of short-term lies, since they aim to maintain a good atmosphere of social community by making marks feel at ease perhaps by stating falsehoods about themselves, others, or a situation in a humorous and non-truthful way.

9. Promotion of *suspension of disbelief* uses deception to provide the mark with fun and enjoyment. For example, movies, magic, or other fictional works use illusion to deceive marks. This differs from other examples of deception, since the mark voluntarily allows herself to be deceived.

10. A *masquerade* is characterized by deception that persists

for extended periods of time to create an illusion regarding something that does not exist, but may make the mark feel better about themselves.

11. Sometimes, people *hide distressing information or negative situations* from others, assuming they may be able to resolve it on their own without additional help from the mark and so not induce anxiety in the deceived.

HRI studies are being developed that will address various sectors of the other-oriented situational space (Fig. 3). A computational model will be embedded to capture these relationships and empirically test the value of robotic other-deception in various contexts.

We aim to approach this problem inspired by Gneezy's deception study in economics [33]. He proposed the formulation of humans deceptive behaviors in pair interactions based on the role of consequences. He modified the utility equation and proposed a framework for when a deceiver determines to perform deceptive behaviors.

$$U = \alpha(D^L - D^T) + (1 - \alpha)(M^L - M^T)/(D^L - D^T)$$

This formulation calculates the utility of situation. Here, U is the utility, D is the deceiver, and M is the mark. X^L indicates the payoffs to X when acting deceptively (or lying). X^T is the payoffs to X from acting truly. α is the relative weight on deceiver's and mark's payoffs. Deceiver chooses deceptive behaviors when the value U from this formulation is larger than 0 in certain situation.

According to this formulation, the deceiver chooses deceptive behaviors based not only on the deceiver's payoff but also on the mark's payoff. Since it is the intention that a robot's deceptive behaviors increase the payoff of the deceived human partner, we extend Gneezy's formulation to apply it to robot systems. To do so, the critical part is how to define and quantify the payoff for the mark. It is necessary to define the payoff matrix for each situation. We can use situational conditions as illustrated earlier to group and generalize different payoff matrices across entire cases.

Developing computational models and algorithms for deceptive capabilities is not an easy task since these capabilities are also highly related to understanding states of mind. Previous research in our laboratory has already proposed a novel algorithm for capturing another robot's internal states during the decision-making process for deception [4]. The development of this algorithm was based on human interdependence theory. By expanding upon this previous research, we expect to develop computational models for robots' other-oriented deception.

In sum, we are in the process of building these computational models for a robot's other-oriented deception, where we will extend this framework and apply it to robot systems in specific situations. It will then be necessary to conduct HRI studies to evaluate our methods quantitatively.

V. ETHICAL ISSUES

Despite the potential benefits of robot deception, relatively little research has been conducted to date on this topic, perhaps due to ethical considerations involving this somewhat controversial topic. Robot ethics is a rapidly expanding area [34], [35], [36], [37]. Is deception acceptable even in humans? Should a robot be allowed to lie?

According to Kantian theory (a deontological perspective), deception or lies should always be prohibited, a standard outcome of any ethics classroom in the application of the Categorical Imperative. By this standard, any deceptive behaviors or lies are morally incorrect, human or robot. The utilitarian perspective on the other hand argues that an action is morally right and acceptable if it leads to increasing total happiness over all relevant stakeholders. Therefore, we can also argue that if deception increases the total benefits among the involved relationships, it is ethically correct [38], [39].

We do not intend to resolve this argument here. In robotics, it is even more complicated to state the ethical issues related to deception, as to whether machines should be given the authority to deceive a human being. It should be carefully and thoughtfully considered [40] before being developed and applied, and it is an integral part of our research.

First, we recommend that the use of robot deception should be in *appropriate* domains. In this perspective, military robot deception would be acceptable as long as it is in accord with the Laws of War. However, we are considering the use of robot deception in HRI contexts, in which it is currently difficult to state what if any situations constitute appropriate use. We intend to evaluate this with our study instruments (e.g., surveys and post-experiment comments).

The objective or goal is an important criterion in which to argue for or against robot deception as well as the ethical framework–i.e, is consequentialism, where the ends justifies the means, appropriate for a particular situation. The aim of this paper is not to justify whether robot deception is morally right or wrong. Rather the main point of this section is to illustrate the applicability of various ethical frameworks in considering robotic deception and to make clear that further discussions of these ethical issues are required by all affected parties.

VI. SUMMARY AND CONCLUSION

Deception is one of the capabilities that is needed to achieve high-order intentionality. Therefore, we can assume that adding deceptive capabilities to robot systems is needed for achieving social robots; however, few studies have been conducted on deception in robotics. Furthermore, there is a lack of studies on fundamental theory, such as the definition of a taxonomy for robot deception. In this paper, we reviewed previous research on robot deception and developed a novel taxonomy for its classification.

Our focus is on the use of other-oriented deception in HRI contexts. We expect that the use of deception in HRI will necessarily grow to achieve more intelligent, effective and natural social robots, and deception, for better or worse, is an inherent part of the human condition.

A research question has been posed which frames our future research: what if any are the *benefits* of robot deception on their human partners. The situational conditions of other-oriented deception were characterized serving as a preliminary stage for the computational models and algorithms being developed for implementation and testing.

We acknowledge that robot deception is a controversial research topic from an ethical perspective. The implications of this and related research should be thoughtfully and carefully established and discussed.

Acknowledgment

This research was supported by the Office of Naval Research under MURI Grant #N00014-08-1-0696.

References

[1] S. A. Morin, R. F. Shepherd, S. W. Kwok, A. A. Stokes, A. Nemiroski, and G. M. Whitesides, "Camouflage and Display for Soft Machines," *Science*, vol. 337, no. 6096, pp. 828–832, Aug. 2012.

[2] D. Floreano, S. Mitri, S. Magnenat, and L. Keller, "Evolutionary conditions for the emergence of communication in robots." *Current Biology*, vol. 17, no. 6, pp. 514–519, Mar 2007.

[3] J. Shim and R. C. Arkin, "Biologically-inspired deceptive behavior for a robot," *12th International Conference on Simulation of Adaptive Behavior*, pp. 401–411, 2012.

[4] A. R. Wagner and R. C. Arkin, "Acting deceptively: Providing robots with the capacity for deception," *I. J. Social Robotics*, vol. 3, no. 1, pp. 5–26, 2011.

[5] L. Hawthorne, "Military deception," *Joint Publication, JP 3-13.4*, 2006.

[6] W. J. Meehan, "Fm 90-2 battlefield deception," *Army Field Manuals*, 1988.

[7] U. D. of Defense, "Unmanned systems integrated roadmap," *FY 2009-2034*, 2009.

[8] D. J. Sexton, "The theory and psychology of military deception," *SUNY press*, 1986.

[9] N. Carey, J. Ford, and J. Chahl, "Biologically inspired guidance for motion camouflage," in *Control Conference, 2004. 5th Asian*, vol. 3, 2004, pp. 1793–1799 Vol.3.

[10] I. Rano, "An optimal control strategy for two-dimensional motion camouflage with non-holonimic constraints," *Biological Cybernetics*, vol. 106, no. 4-5, pp. 261–270, 2012.

[11] R. A. Johnstone and A. Grafen, "Dishonesty and the handicap principle," *Animal Behaviour*, vol. 46, no. 4, pp. 759–764, Oct. 1993.

[12] J. Davis and R. Arkin, "Mobbing behavior and deceit and its role in bio-inspired autonomous robotic agents," *International Conference on Swarm Intelligence*, pp. 276–283, 2012.

[13] K. Terada and A. Ito, "Can a robot deceive humans?" in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, ser. HRI '10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 191–192.

[14] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!!: an interaction with a cheating robot," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, ser. HRI '10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 219–226.

[15] M. Vazquez, A. May, A. Steinfeld, and W.-H. Chen, "A deceptive robot referee in a multiplayer gaming environment," in *Collaboration Technologies and Systems (CTS), 2011 International Conference on*, 2011, pp. 204–211.

[16] B. Brewer, R. Klatzky, and Y. Matsuoka, "Visual-feedback distortion in a robotic rehabilitation environment," *Proceedings of the IEEE*, vol. 94, no. 9, pp. 1739–1751, 2006.

[17] S. Matsuzoe and F. Tanaka, "How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot," in *RO-MAN, 2012 IEEE*, 2012, pp. 339–344.

[18] F. Tanaka and T. Kimura, "Care-receiving robot as a tool of teachers in child education," *Interaction Studies*, vol. 11, no. 2, pp. 263–268, 2010.

[19] R. M. Chisholm and T. D. Feehan, "The intent to deceive," *Journal of Philosophy*, vol. 74, no. 3, pp. 143–159, 1977.

[20] B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, and J. A. Epstein, "Lying in everyday life." *Journal of personality and social psychology*, vol. 70, no. 5, pp. 979–995, May 1996.

[21] J. F. Dunnigan and A. A. Nofi, "Victory and deceit, 2nd edition: Deception and trickery in war," *Writers Press Books*, 2001.

[22] B. Whaley, "Toward a general theory of deception," *Journal of Strategic Studies*, vol. 5, no. 1, pp. 178–192, Mar. 1982.

[23] J. Bell and B. Whaley, *Cheating and deception*. TRANSACTION PUBL, 1991.

[24] S. Erat and U. Gneezy, "White lies," *Rady Working paper, Rady School of Management, UC San Diego*, 2009.

[25] N. C. Rowe, "Designing good deceptions in defense of information systems," in *Proceedings of the 20th Annual Computer Security Applications Conference*, ser. ACSAC '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 418–427.

[26] D. C. Dennett, *The Intentional Stance (Bradford Books)*, reprint ed. Cambridge, MA: The MIT Press, Mar. 1987.

[27] F. De Waal, "Deception in the natural communication of chimpanzees," *Mitchell and Tompson*, 1986.

[28] E. Adar, D. S. Tan, and J. Teevan, "Benevolent deception in human computer interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 1863–1872.

[29] R. T. Vaughan, N. Sumpter, J. Henderson, A. Frost, and S. Cameron, "Experiments in automatic flock control," *Robotics and Autonomous Systems*, vol. 31, no. 1-2, pp. 109–117, 2000.

[30] M. A. Steele, S. L. Halkin, P. D. Smallwood, T. J. McKenna, K. Mitsopoulos, and M. Beam, "Cache protection strategies of a scatter-hoarding rodent: do tree squirrels engage in behavioural deception?" *Animal Behaviour*, 2008.

[31] D. C. Dennett, "When hal kills, who's to blame? computer ethics," in *HAL's Legacy: 2001's Computer as Dream and Reality*, D. G. Stork, Ed. Cambridge, MA: MIT Press, 1997.

[32] F. Miller, D. Wendler, and L. Swartzman, "Deception in research on the placebo effect," *PLoS Med*, vol. 2, no. 9, p. e262, 2005.

[33] U. Gneezy, "Deception: The role of consequences," *American Economic Review*, vol. 95, no. 1, pp. 384–394, September 2005.

[34] N. Sharkey, "The ethical frontiers of robotics," *Science*, vol. 322, no. 5909, pp. 1800–1801, 2008.

[35] T. Economist., *Morals and the machine*. The Economist Newspaper Limited, 2012.

[36] P. Lin, K. Abney, and G. Bekey, *Robot Ethics*. MIT Press, 2011.

[37] W. Wallach and C. Allen, *Moral Machines : Teaching Robots Right from Wrong: Teaching Robots Right from Wrong*. Oxford University Press, USA, 2008.

[38] W. Sinnott-Armstrong, "Consequentialism," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., 2012.

[39] K. Christine, "Two arguments against lying." *Argumentation2*, 1988.

[40] R. Arkin, "The ethics of robotics deception," *1st International Conference of International Association for Computing and Philosophy*, pp. 1–3, 2010.