Humanoid Robot Affects Human Rationality: A Case Study in a Competitive Game

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

7th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

8th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—As robots are increasingly endowed with social and communicative capabilities, they will interact with humans in more settings, both collaborative and competitive. We explore human-robot relationships in the context of a competitive Stackelberg Security Game. We vary humanoid robot affect (via two conditions, positive and negative, in the form of "encouraging" or "discouraging" verbal commentary) in our study and measure the impact on participants' rationality, strategy prioritization, mood, and perceptions of the robot. We learn that a robot opponent that makes discouraging comments causes a human to play a game less rationally and to perceive the robot more negatively. We also discover that humans may consider an opponent robot to be a "distraction" regardless of the its affect. We also contribute a simple open source Natural Language Processing framework for generating affective sentences, which was used to generate sentences for our autonomous social robot to say during the study.**

***Index Terms*—Robot Affect, Social Robot, Competitive Robot, User Study, Humanoid, Rationality, Quantal Response, Stackelberg Security Game**

I. Introduction

The future will bring humans into contact with embodied agents in a variety of unstructured interactions, many of which will involve engaging robots in verbal dialogue. This includes in-store sales [1], education [2], service interactions [3], and rehabilitation and health care [4]. In some of these settings, one can imagine a robot and human may have different or even conflicting goals. For example, in a sales setting, a robot completing a sale may prioritize convincing a customer to buy a product, whereas the customer aims to make the optimal decision to satisfy their needs. In any interaction like this, linguistic nuances and positive or negative valence of social behavior will impact the smoothness of the interaction.

Positive robot affect has been shown to contribute to perceptions of robots as teammates [5]. On the other hand, humans may be more engaged with robots or may perceive robots as more intelligent when they exhibit certain behaviors that are sometimes considered negative, like acting deceptively in social situations [6], [7] or cheating in a game [8], [9]. Threatening behavior from a robot may also increase humans' attentional control [10]. While it is clear that robot affect can influence humans' experiences and that humans are not as rational as machines nor are they as rational as they believe themselves to be, little is known about the role of a robot's affect in humans' *rational* behavior, which is the main focus of this paper.

Moreover, much work has gone into understanding how humans and social robots interact and partner in cooperative settings [11]–[13]. However, humans and robots will not always be perfectly aligned. We thus focus on the competitive setting, which is understudied in literature despite its significance.

In this study, we examine the impact of robot affect on human rationality and strategy prioritization in a competitive Stackelberg security game. We seek to answer the question: how does the perceived affect of a humanoid robot opponent's spoken language impact a human's rationality and strategy in a game theoretic setting? We implement a system to play the game autonomously with the addition of dialogue generated by an algorithm we develop. In a between-subject study, 40 participants played the game with a humanoid robot in two conditions. During the game, the robot showed positive and negative affect by making encouraging or discouraging comments. We analyze the data collected to obtain insights into how the robot's behavior impacts participants' rationality and emotions. We find that discouraging comments from a robot decreased a participants rationality during gameplay and that negative language contributed to negative social attributions to the robot. We also find that despite our success in providing human-like comments through our NLP method, some participants were still actively aware that their game partner was a robot. These results indicate the degree to which humans see even an autonomous humanoid as a non-agent.

In the following sections we first provide an overview of the existing studies in this domain and discuss how our work differs from related work(section II). We then describe the analysis methods (section III) and study procedures (section IV) followed by a summary of results (V). The discussion section highlights our main findings and their implications in the design of human-robot interaction tools and systems (section VI).

II. Preliminaries and Related Work

Humans interpret emotion through nonverbal and verbal cues, and this is well studied in literature [14]. Notably, it is well documented that, following a social interaction, observing another person's mood can have specific consequences on behavior and change the observer's mood [15].

Affect is a general term relating to emotions, moods, feelings and other such states. Affective states vary in their degree of *activation* (intensity) and *valence* (whether they are positive and negative) [16]. In the psychological and cognitive science literature, affect is often represented via axes in a continuous multi-dimensional space [17], [18].

People exchange verbal messages which contain information conveying their mental and emotional states. This includes the use of emotionally colored words and profanity. The expressions and behaviors of human social partners can subliminally or consciously impact a person's performance [10], [19], risk-taking [20], decision-making [21], and mood [22]. We consider to what degree this holds true for a robot companion as well. Given the importance of affect in language, there has been a fairly substantial amount of research in affective statistical language modeling [23]. This includes the development of affective Natural Language Generation (NLG) for generating medical texts [24] and rule based emotive text generation based on sentence patterns [25]. Notably, there has emerged work in the extension of the Long Short Term Model (LSTM) language model for generating affective text [26].

Research has also shown that humans are capable of perceiving robotic affect and that affect influences the way humans interact with robots. Different forms of affective expression have been modeled with humanoid robots. Bodily expression has been used for ROMAN (robot with facial expressions) [27], [28], NAO (whole body expressions) [27]–[29], KOBIAN (body and facial expressions) [30], and Cozmo (non-verbal behaviors) [31] to model emotional behaviors that improve expression functionality. These studies demonstrated that people are generally capable of recognizing robots' affective states. Work has also been done to develop a parameterized behavior model in which behavior parameters control the spatial and temporal extent of a behavior for mood expression [32]. This includes models that enable the continuous display of mood in an interactive game [13]. Work of this nature with interactive robots is less developed than similar work with computer agents [33], [34]. Moreover, research focusing on the impact of affective language models in human-robot interactions has been limited to joint human-robot tasks in the cooperative setting [11]. Thus, we extend affective language models to the competitive human-robot interactions.

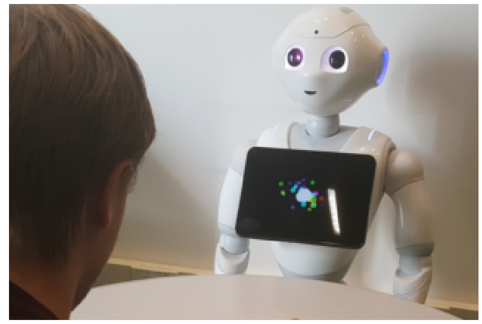


Fig. 1. The study setup from a participant's perspective

Research has shown that affective expression can have positive influence on humans during an interaction [35]. For example, studies with the robot Vikia have demonstrated the capability of an emotionally expressive graphical face for encouraging interactions with a robot [36]. Other works found that humans can identify and respond to a robot's facial or verbal expressions of emotion [37], [38]. This can influence the effectiveness of assistive tasks such as learning and motivation given vocal emotion expression [39] and user behavior during support tasks [40]. Some works incorporate game theoretic notions to codify how humans trust robots during repeated cooperative interactions [12]. However, these cooperative representations fail to extend to competitive settings in which a social robot and human are at odds. Of particular interest to us are studies in the format of interactive, competitive games between social robots and humans.

Mood contagion is a well-researched automatic mechanism whereby the observation of another person's emotional expression induces a congruent mood state in an observer. Several studies also reported consequences of affective virtual agents on humans' task performance. Unlike [41], we use a humanoid instead of a virtual agent. In contrast to the robots in [13], [42], we deal with verbal affect cues instead of gesture and posture. Emotion expression was reported to have effects on users' affective states and behaviors. Reference [33] found that a computer agent's expression of anger and happiness impacted the way humans negotiated with it. Research in HRI has also emphasized the importance of establishing affective common ground between humans and robots in different teamwork-oriented settings [43], [44].

To understand human rationality in a competitive setting, Stackelberg security games are a common choice because they have an optimal strategy [45] and the ability to calculate quantal response, a measure of a participant's rationality, and subjective utility quantal response, a measure of a participant's strategic prioritization [46].

We are among the first to explore the impact of a robot's affective statements on human performance in a competitive game setting. We use a competitive Stackelberg security game to analyze human rationality and strategy via quantal response and human perceptions of the robot from a combination of validated scales and experiment-specific Likert items.

III. Methodology and Contributions

A. Overview of Study

The primary goal of our study was to determine the effect of a robotic opponent's affective state on a human's game-playing strategy in a game theoretic setting. Our hypothesis is that when a human plays a competitive game against a humanoid robot, the human's strategy will differ as the perceived affect of the robot differs.

We conducted a between-subjects experiment in which a human played a repeated Stackelberg security game against a robot. The robot exhibited either "encouraging" or "discouraging" behavior during game play. We measured and calculated the nature of the human's strategy by recording their choices during the game. We administered questionnaires before and after the game to collect data pertaining to the human's perceptions of the task, their performance, and the robot. A detailed description of the experimental setup and procedure can be found in section IV.

Between the two conditions, we manipulated the robot's affect. This is a binary variable, with the two states referred to as "positive" and "negative" or "encouraging" and "discouraging". Giving the robot this affect was accomplished through an NLP model we discuss in section III-B. During all games, the robot's strategy was held constant, regardless of human actions or robot affect. Affect was expressed by the commentary that the robot provided during the game. Robot dialogue and behavior were completely autonomous.

Our primary measures of interest pertained to the participant's strategy. We measured "strategy" in two ways. First, we used a quantal response equilibrium calculation to determine the degree to which the human played rationally. Second, we evaluated the nature of the strategy itself, in terms of which aspects of the game environment the participant prioritized in their decision-making process (assessed via subjective utility quantal response calculation and via self-report). Both of these are discussed in greater detail in section III-C. Other variables of interest were social perceptions, mood, and perceived robot mood, measured in terms of 1) participants' answers to questions about themselves and the robot along Likert scales, and 2) answers to free-response questions about perceptions of the robot and the task after the game is played.

B. NLP Model

To give our robot an affect, we developed an affect-aware bidirectional fill-in-the-blank N-Gram model.

An N-gram model trains on corpora and counts how often each word follows each preceding N-words. From these counts, we construct a probability that any particular word follows a previous sequence of words. This probability is derived by computing the frequency of that particular word's occurrence after a given sequence compared to the frequency of any observed word following the same sequence. After creating counts of sequences of words of length N+1, the probability can be expressed as:

$$P(w_n|w_{n-(N-1)},...,w_{n-1}) = \frac{\text{Count}(w_{n-(N-1)},...,w_{n-1},w_n)+\alpha}{\text{Count}(w_{n-(N-1)},...,w_{n-1},*w)+D\alpha} \quad (1)$$

where $P(w_n|w_{n-(N-1)},...,w_{n-1})$ is the probability that a particular word w_n follows a particular sequence of N other words, and $*w$ is a wildcard meaning "any word observed as completing this sequence". In our usage, as shown, we add $+\alpha$ and $+D\alpha$ terms as Laplacian smoothing to account for situations where a word was not observed. We use $\alpha = 1$ and $D =$ [number of words that could fit $*w$ for the given preceding sequence].

To make the language not just natural but connoting a specific emotion ("affect-aware"), we took advantage of the AFINN Affect Dictionary, which rates the emotional valence of a word on a scale from -5 to 5 [47]. During the game (after the model is trained), we feed in a bank of neutral sentence stems (sentences with fill-in-the-blanks). The model selects words to fill in the blanks that are appropriate given the sentence and that give the sentence appropriate affect. This also allows us to use bidirectional N-grams–training and predicting based on the words preceding *and* following a word to be predicted. We use both bigrams and trigams (N=2, N=3) in the forward and reverse direction.

The final equation our model uses to select the words to fill out the sentences the robot will say is given by

$$\begin{aligned} P(w_n|w_{n+2},w_{n+1},w_{n-1},w_{n-2}) &= z_5 * V(w_n) * A \\ &+ z_1 * P(w_n|w_{n+2},w_{n+1}) + z_2 * P(w_n|w_{n+1}) \\ &+ z_3 * P(w_n|w_{n-2},w_{n-1}) + z_4 * P(w_n|w_{n-1}) \quad (2) \\ A \in \{-1,1\}, \quad \sum_{i=1}^{5} z_i \leq 1 \end{aligned}$$

where the probabilities on the right-hand side are calculated according to (1), V gives the AFINN affective valence of a word (or 0 if not in the dictionary), A indicates whether the affect is encouraging (+1) or discouraging (-1), and the z_i values are weights.

We train our model on transcripts of popular films from the IMSDb archive [48], [49]. During training and prediction, a set of "stop words" (such as "is" and "and") is ignored for the purposes of defining sequences of words. Numerals and punctuation are also filtered out before training. Additionally, during prediction, certain words are ignored (blacklisted), such as "kill". This is to prevent the robot from being too vulgar, saying things that do not make sense, hoping for a participant's death, or making comments that are otherwise offensive or uninterpretable.

The code to generate the model from any corpora and make predictions based on arbitrary sentence stems can be found on Github.[1]

C. Quantal Response

1) Measure of Degree of Rationality: Each participant played several rounds of a Stackelberg security game against the robot [45], [50]. For the purposes of the rationality calculation described here, note that during each round, the participant chose a single action from a set of N options, based on provided information, in an attempt to maximize a numerical reward. Quantal response (q_{c_r}) represents the probability of the participant rationally selecting choice c in round r and is defined in (3), where λ represents the amount of noise in the participant's response (that is, λ captures how rational the participant's decision is). For a given set of rounds Υ, we fit λ to the game play data by estimating the maximum likelihood of the quantal response q_{c_r}. This is also shown in (3) where λ is fit to a subset of rounds, Υ.

$$\lambda = \underset{\lambda}{\operatorname{argmax}} \sum_{r \in \Upsilon} \log(q_{c_r}), \quad q_{c_r} = \frac{\exp(\lambda U_{c_r,r})}{\sum_{j=1}^{N} \exp(\lambda U_{j,r})} \tag{3}$$

where $U_{i,r}$ is the known real utility of choice i in round r (see equation 5), c_r is the number of the choice chosen by the participant in round r, and Υ is the subset of rounds to be used in the calculation.

2) Measure of Prioritization in Strategy: We can also find the Subjective Utility Quantal Response, s_{c_r}, in (4) to determine the probability that a participant selects choice c in round r based on the strategic priority of the participant, W [51]. A participant's **strategic prioritization**, W, denotes the importance to the participant of different attributes of each of the options. We use MLE as shown in (4) to determine this strategic prioritization, W.

$$W = \underset{W}{\operatorname{argmax}} \sum_{r \in \Upsilon} \log(s_{c_r}), \quad s_{c_r} = \frac{\exp(W^T X_{c_r,r})}{\sum_{j=1}^{8} \exp(W^T X_{j,r})}) \tag{4}$$

$$W^T = [w_1 \quad w_2 \quad ... \quad w_n] \qquad X_{i,r}^T \in \mathcal{R}^3$$

where X represents a vector of values for attributes of the choices, and W is the strategic prioritization showing how much weight a participant gives to each attribute in their decision-making.

[1]Note: Github link supplied after review.

IV. Experimental Setup and Protocol

The following sections describe our setup and procedure. We also provide a Github link with code and instructions so anyone with the appropriate robot can replicate this study.[2]

A. Participants

Participants were recruited from the local community. The average age was 27.2 with standard deviation of 11.2. 15 males, 24 females, and 1 nonbinary individual participated in our study. Every participant played at least one game session against the robot. A11 of the participants were selected to play an additional game session. For these participants, the robot exhibited a different affect during each of the two sessions.

B. Robot

We used the Pepper Robot by Softbank Robotics [52]. Pepper is a humanoid robot with arms, a head with cameras and microphones, mobility, and voice abilities. See it pictured in Fig. 1. During the course of a participant‘s interaction with Pepper, Pepper acted autonomously, reciting dialogue per the NLP model and chosen affect, which was set before each game session.

C. Procedure

The experimental procedure was as follows:

1) Consent: The experimenter obtained written consent to participate in the study and verbally informed the participant that video and audio recordings of the session would be made.

2) Pre-Game Survey: Before the game, the experimenter administered a questionnaire to collect demographic information and and measures of pre-task emotional state (more details about the pre- and post-game questionnaire are in section IV-E).

3) Practice Rounds: The participant used a laptop (a "convertible" combination laptop/tablet) to play 2 practice rounds of the Stackelberg game "against the computer". This has the purpose of countering learning effects in participants' performance (they learn to play the game before encountering the robot).

4) Game Session I: After the practice rounds, the experimenter led the participant into a room where the robot sat behind a table. The participant sat across from the robot with a tablet computer (the same as from the practice rounds, but now plugged in) face up between them. While the participant played the game, experimenters sat in a different part of the room and were hidden from view by a screen to prevent the participant from feeling like they were being observed. The participant then played a several rounds of the game "against" the Pepper robot. The robot made periodic comments about the game and the participant. The comments exhibited either positive (**encouraging**) or negative (**discouraging**) affect. Although the commentary was sometimes complimentary or critical in nature, in reality it had nothing to do with the participant's actual performance.

[2]Code to be supplied after review.

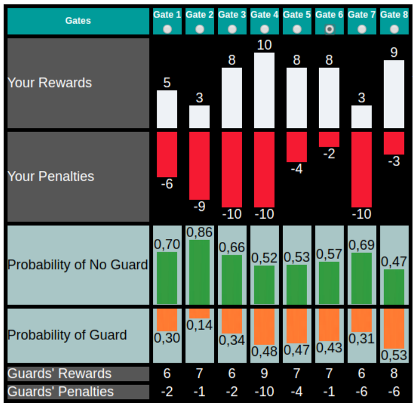


Fig. 2. An example screen presented to the user during a round of the Guards and Treasures game, adapted from [50].

5) Post-Game Survey and Video: After the games concludes, a researcher was notified to re-enter and gave the participant a written survey to fill out. Then, the research verbally asked a series of questions to the participant, who responded on video.

6) Game Session II: A selected subset of participants played a second game session against the robot, which exhibited the opposite affect as from the first game.

7) Post-Game Survey and Video II: If a participant played a second game session, they were given a second post-game survey and asked the same set of verbal questions again.

8) Debriefing: Initially, participants were told they would play against a robot but were not informed of the true purpose of the study. Participants were debriefed after the study ended.

D. *The Stackelberg Security Game Interface*

The "Guards and Treasures" game[3] that the participants played consists of 35 rounds (following 2 practice rounds). It is a modified version of the game from [50].[4] Each round, the participant is shown a screen as in Fig. 2. The idea is that the player can choose to "attack" (select) a gate. If the defending player (the robot) places a guard at the gate, the human player incurs the penalty for that gate. If the chosen gate is not defended, the human player receives the reward instead. The probability that a guard is behind a particular gate is also displayed. The player selects one gate each round and only sees their results after all the rounds are complete.

The "choices" referred to in section III-C are the different gates than can be chosen. The expected utility $U_{i,r}$ of a particular gate i for a given round r (referenced in (3)) can be found by

$$U_{i,r} = R_{i,r}(1 - g_{i,r}) - g_{i,r}Y_{i,r} \quad (5)$$

where R is the reward, g is the probability a guard will defend a gate, and Y is the penalty for a particular gate i in round r. ($R \in \mathbb{Z}, 1 \leq R \leq 10$, Y $\in \mathbb{Z}$,-10 $\leq Y \leq -1$, and $g \in [0, 1]$) $N = 8$ because there are 8 gates, each gate is one of the choices a participant can make.

The $X_{i,r}$ from (4) is $X_{i,r}^T = [R_{i,r} \quad Y_{i,r} \quad g_{i,r}]$. W has three components, w_1, w_2, w_3, which refer to how much weight a participant gives reward, penalty, and probability of seeing a guard, respectively, in determining their strategy.

E. *Surveys and Data Collection*

Our data collection included i) a pre-task questionnaire, ii) records of actions taken during the game, iii) a post-task questionnaire, iv) a post-game verbal interview (recorded on camera), and (for some participants) v) video of the participant playing the game against the robot. [5] Participant data was encrypted and stored according to assigned ID numbers (and never paired with names).

The pre-task questionnaire included demographic information and numerical ratings of familiarity with robots and with technology. The post-task questionnaire asked the participant to assess their own performance and their experience with the game and the robot. The aforementioned questions are on yes/no or numeric scales. Free-response questions related to a person's strategy and their perceptions of the robot were also asked. Some questions we originated and others we drew from [53] or [54]. The pre-task and post-task questionnaires also both made use of the Self Assessment Manikins [55], which measure affect via three dimensions: valence, arousal, and dominance. In the first questionnaire, participants assessed themselves on these scales, and in the second, they assessed both themselves and the robot.

In a post-task semi-structured interview, we asked 9 questions. The questions pertained to participants' overall perception of the robot, their overall thoughts about the experience, self-assessment of performance, perceptions of the robot's goal, and their game-playing strategy. Participants who played the additional session answered the pre-task questionnaire once and the post-task questionnaire and interview questions after each game session.

V. RESULTS

A. *Analysis on Gameplay*

We used Maximum Likelihood Estimation (MLE) to solve (3) and (4) over data from aggregate groups of participants. In table I, find λ and W values corresponding to various populations (Υs). These Υs contain multiple participants, divided by affect, and include all rounds for each participant in each subset. The Quantal Response parameter specifies that $\lambda = 0$ indicates random behavior (i.e., no rationality) and $\lambda = \infty$ indicates perfect rationality (i.e. optimal strategy). W values (the strategic prioritization, our variant of the Subjective Utility Quantal Response) describes a participant's strategy prioritization (w_1 is the reward component, w_2 is the penalty component, and w_3 is the probability of guard component). For reference, previous work [51] obtained $\lambda = 0.77$ via a

[3]The code we used to run the game can be found at [supplied after review] and is open source.

[4]This original game can be found at http://teamcore.usc.edu/Software.htm

[5]Exact pre- and post-surveys used are available for download at [supplied after review].

TABLE I
λ AND W FOR VARIOUS POPULATIONS Υ

	λ			W								
Affect	**Both**	**Positive**	**Negative**	**Both**			**Positive**			**Negative**		
				w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
Basic Games For All	0.5432	0.5828	0.5064	0.3261	0.1697	-10.4838	0.3586	0.1573	-11.1006	0.2965	0.1819	-9.939
Basic Games for Selected Group	0.3269	0.2256	0.3929	0.1649	0.1061	-8.007	0.0893	0.0818	-5.6158	0.2190	0.1254	-9.7572
Additional Games for Selected Group	0.4128	0.4892	0.3015	0.1907	0.2021	-10.2328	0.0742	0.2028	-9.7686	0.2624	0.2041	-10.6888
Basic and Additional Games for All	0.5121	0.5568	0.4660	0.2947	0.1761	-10.3758	0.3318	0.1692	-10.9512	0.2564	0.1840	-9.8081

group of Amazon Mechanical Turk (AMT) workers playing this game online without a lab environment or the presence of a robot. In addition, [51] reports $W = [0.37, 0.15, -9.85]$ (when converted to our representation) for the Subjective Utility Quantal Response model for a general population playing this game.

The first row in the table has basic, first round games for all participants. The "selected group" of participants refers to those who played an additional game with the reverse affect (mentioned in sections IV-C6 and IV-C7).

Using the procedure in [56], we can analyze changes in λ and W between the selected group participant's basic and additional games. We found that those who played a negative session first had a 21% increase in λ and a 104% in the 1-norm of the strategic prioritization vector, W. On the other hand, those who played the positive session first had a 28% increase in λ and a 110% in the 1-norm of W.

Fig. 3(a) shows the trend for λ for positive and negative affect over time for participants' basic games.

We notice that, though participants place a higher priority on reward than on penalty, all participants place more emphasis on penalty over time in Fig. 3(b). Just as weight of reward (w_1) is relatively steady over multiple intervals of five rounds, we found that the weight participants placed on the probability of a guard being present (w_3) was steady over time, though the weight placed on the guard's presence (w_3) was two orders of magnitude larger than the other two weights.

In the positive affect condition, 15 participants won the game and 5 lost. In the negative affect condition, 16 participants won and 4 lost. This difference in performance was not significant. We also did not find any effects of robot affect on prioritization of the guard, reward, and penalty factors.

B. Perception of Robot

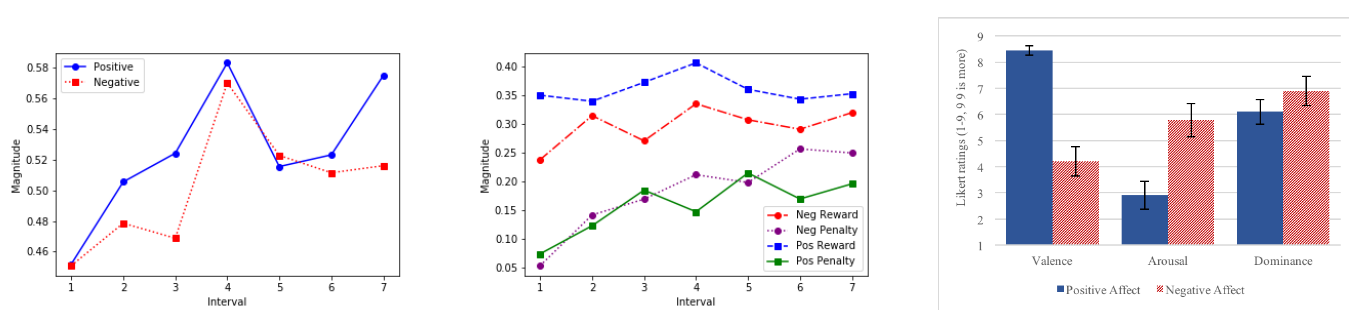
Age was relevant in participants' perceptions of the robot. Negative affect and low valence prior to the start of the experiment led to a lower perception of humanlike-ness. However, negative affect mattered less to younger people.

C. Analysis of Self Assessment Manikin

Responses to the questionnaire about perceptions of the robot were also analyzed. Because the data were not normally distributed and the sample was small ($n < 50$), the non-parametric Wilcoxon/Kruskal-Wallis rank sum test for two samples was used to compare the data from the two conditions for all variables.

Robot affect significantly impacted several measures of positive social assessments of the robot, including perceptions that the robot was encouraging, $\chi^2(1, N = 40) = 31.55$, $p = 0.008$, optimistic, $\chi^2(1, N = 40) = 23.48$, $p < 0.0001$, and cheerful, $\chi^2(1, N = 40) = 28.33$, $p < 0.0001$. For all of these variables, positive affect led to higher ratings. Robot affect did not impact perceptions of humanlikeness or cuteness. These results confirm the validity of our affect manipulation.

We found a significant main effect of robot affect on participants' liking of the activity, $\chi^2(1, N = 40) = 6.97$, $p = 0.008$. We used the Self Assessment Manikin (SAM) scale [55] to measure the participants' mood and before and after the game in terms of emotional valence (i.e. how happy or unhappy they felt), arousal (i.e. how excited or unexcited they felt), and dominance (i.e. how in-control they felt). We also



(a) Value of λ_P (positive) and λ_N (negative) over time (captured at seven 5-round negative across an affect class)

(b) Values of w_1 and w_2 (Reward and Penalty components of w over time. Each interval is 5 rounds.)

(c) Mean scores for Self Assessment Manikin For Robot by affect class. Error bars represent ± 1 standard error of the mean.

Fig. 3. Results: Trends and Comparisons

used this scale to assess participants' perceptions of the robot's mood after the game. There was a significant main effect of robot affect on post-task participant valence, $\chi^2(1, N = 40) = 4.36$, $p = 0.037$, and perceived robot valence, $\chi^2(1, N = 40) = 20.87$, $p < 0.0001$, Fig. 3(c). For these variables, participants in the Positive affect condition had higher ratings. We also found an effect wherein negative affect positively impacted ratings of perceived robot arousal, $\chi^2(1, N = 40) = 10.07$, $p = 0.002$. There were no main effects of affect on post-task participant arousal, participant dominance, or perceived robot dominance. Collectively, these findings corroborate previous work [12], [33] suggesting that robot affect has a strong impact on participants' feelings in a dyadic interaction, though in our case, the setting is competitive rather than cooperative.

We suspected that other independent variables such as age, gender, preconceived notions about robots, and mood prior to the experiment may also play a role in participants' strategy and evaluations of the robot. We looked for correlations among these variables and ran our analyses again with correlated variables as covariates. We found a main effect of age on participants' belief that the robot was humanlike, $p = 0.003$, in which younger participants thought it was more humanlike. We also found a significant interaction effect of age and robot affect condition on perceptions that the robot was humanlike, $p = 0.012$, and ratings of the robot's *dominance*, $p = 0.004$: negative affect mattered less for younger participants in assessments of human likeness and dominance. We found an interaction effect of participant valence prior to the start of the study and robot affect on perceptions that the robot was humanlike, $p = 0.015$, in that Negative robot affect and low valence prior to the start of the experiment led to lower perceptions of humanlikeness.

The effect of pre-task valence on overall λ was significant ($p = 0.035$). Participants with lower valence at the start trended towards higher overall λ scores. Higher familiarity with robots also led to significantly higher ratings of liking the activity, $F(1, 1) = 11.85$, $p = 0.039$.

We found that negative robot affect significantly lowered participants' beliefs that the robot was encouraging, $F(1, 9) = 262.09$, $p < 0.0001$, optimistic, $F(1, 9) = 65.05$, $p < 0.0001$, cheerful, $F(1, 9) = 45.64$, $p < 0.0001$, and cooperative, $F(1, 9) = 24.77$, $p = 0.008$. This was similar to the findings from our between-subjects analysis. Here, we also found an effect of affect on perceptions that the robot was cute, $F(1, 9) = 6.92$, $p = 0.027$, whereby positive affect increased perceptions that the robot was cute.

D. Feedback from Participants Through Oral Interviews

12 participants (4 in the positive condition and 8 in the negative condition) explicitly stated that they believed the robot's goal involved distracting them. A participants in the negative affect condition said, "it would have helped my concentration if it wasn't talking as much" (P102). Participants in the positive condition said, "When I was trying to determine what move to make, it took me out of that zone fora bit" (P220), and, "It felt like I was doing homework and my friend kept talking to me" (P201). An additional 4 individuals (2 negative and 2 positive) used language that indicated a perception that the robot was perceived as distracting without saying so explicitly ("disrupt my concentration" (P112)). Altogether, 30% of participants explicitly classified the robot's goal as "distraction" and 43% used language that expressed a similar perception.

Another theme was the dehumanization of the robot. Participants noted that the robot was a robot in explaining why they did not believe they had been influenced by it. One participant said, "It kept on insulting me the entire time. But like, my friends also tease me on a regular basis so it was funny... it doesn't hurt" (P110). Another said, "I don't like some of the stuff it was saying. But that's the way it was programmed so I can't blame it" (P104).

Finally, interviews revealed that participants were encouraged by the robot in the positive affect condition and were especially discouraged in the negative affect condition. When asked about the robot's goal, a participant in the positive condition answered, "To encourage me to do well... it seemed to [succeed in that goal]" (P117), while a participant exposed to the negative affect said "It kept making me doubt myself" (P214).

VI. Discussion

A. Validation of NLP Model

Participants perceived an encouraging robot as encouraging, cheerful, and optimistic and a discouraging robot as discouraging and pessimistic. Interviews confirmed the quantitative ratings. Example sentences generated by the model can be found in Fig. 4. This validates the affect-aware bidirectional fill-in-the-blank N-gram NLP model we developed, and demonstrates that our simple word choice model achieved the desired result.

B. Discouraging Affect Reduces Both Rationality & Learning

Overall, we found that negative affect causes worse (less rational) performance ($\lambda = 0.51$ for negative vs. $\lambda = 0.58$ for positive). This is in line with what might be expected and with previous work [51], [57], [58]. A participant will believe they will make better choices when encouraged, whereas a discouraged individual will make more mistakes. Pepper's form is particularly similar to that of a human (two arms, fingers, head, torso), so certain aspects of the interaction may more closely mimic human-human game play than they would have if our participants had played with a less humanoid robot robot.

Participants in the selected group (who played an additional game) performed more rationally and more strategically (as noted by the increase in the 1-norm of W) in the additional session compared to the basic session. Those who played the positive affect session first had a higher increase in these metrics than those who played a negative affect session first. One possible explanation is that in the first case, residual encouragement from the initial positive session continued to buoy the participant in the second session.

Positive Sentences	Negative Sentences
You seem to be considering your moves in a practiced manner.	You seem to be considering your moves in a bizarre manner.
Honestly this game is a wonderful experience.	Honestly this game is a bad experience.
I have to say you are a great player.	I have to say you are a terrible player.
Over the course of the game your playing has become brilliant.	Over the course of the game your playing has become confused.

Fig. 4. Example sentences built by our affective NLP algorithm

C. Overall Rationality of Population Was Lower Than Expected

While there were standout individuals, the overall rationality for our populations was below that found among the crowd-sourced AMT population from [51]. While the discrepancy may be attributable to differences in the game framing, timing, population, or noise, we considered whether factors in our study may have further contributed to an overall degradation of rationality beyond the influence of affect. One possible explanation is that the amount of money our participants received was fixed as opposed to dependent on their performance, like AMT workers. In other words, they themselves derived no benefit from how much effort they actually put in.

The other major difference between that study and our own is that AMT workers were competing in the game against a computer, while physically located in (presumably) a location of their choice. Our participants were face to face with a robot "opponent" and in an unfamiliar room. An unfamiliar setting can influence a participant's decision rationale and may have been an additional factor hampering the competitive abilities of many participants. [59] Dialogue can also be a distraction, regardless of content [60].

Another reason for decreased rationality may be the competitive nature of the task. While emotion is contagious in a cooperative setting (robot encouragement would be expected to help a human), that may not apply in a competitive setting. [61] Indeed, participants in both affect classes sometimes saw the robot as a "distraction".

D. Robot Opponent as a Distraction, Regardless of Affect

This was one of the most unexpected results of our study. Although we had no intent to measure how distracting the robot was perceived, over a quarter of participants **explicitly** answered that "distracting" the participant was one of the goals or even the goal of the robot. This suggests that, given the competitive setting, some participants were focused more on winning the game than on interacting with the robot. Also possible is that participants had less empathy for the robot due to the competitive nature of the scenario [61].

E. When a Robot is Clearly a Robot

One of the limitations of our technique may be that the robot is not subtle. It is possible that the robot is so discouraging that it veers into the absurd as opposed to the legitimately discouraging. While some participants, as noted, legitimately were made to feel that they were doing something wrong, others more than took it in stride, finding it humorous.

While many individuals anthropomorphized the robot, multiple participants described the robot in ways that dehumanized it. This awareness or assumption of the robot's lack of agency (despite the fact that it was autonomous) could also have contributed to a participants being less impacted by it overall. It is possible that a participant might have been more heavily influenced by a human who spoke in a discouraging matter but believed (consciously or unconsciously) that since a robot can't be "blamed" for its actions, there was no reason to take the degrading speech personally.

F. Less Impact on Young People

A possible explanation for why younger participants were less influenced by affect could be because younger individuals are more used to thinking of robots as machines. Robot developers will need to keep this in mind when designing robotic systems for older users, who may be more likely to perceive or value emotional connection as conducive to pleasant interaction.

VII. Conclusion and Future Work

A humanoid robot that encourages or discourages a human opponent can impact that human's rationality. A discouraging robot leads to lower rationality while an encouraging robot is associated with higher rationality.

The insights documented here will be useful for future designers of robots. Game developers can also use this knowledge to create more interactive opponents to increase the sense of engagement and enjoyment. In the field of education, we can be aware that were a humanoid robot exam proctor to express affect in its language while administering an exam to students, the students' performance could be influenced, for better or for worse. Creators of social robots can more precisely understand the difference between how a humanoid robot and human are perceived in non-cooperative interactions.

We have developed open source tools, such as the program to generate and run the affect-aware bidirectional fill-in-the-blank N-gram NLP model, as well as code for running it and the whole experiment on the Pepper robot. These resources are validated as serving their intended purpose and we invite other researchers to take advantage of and use them.

Useful future work would be to further investigate consequences of humanoid robot affect by incorporating nonverbal modes of affect expression (for example, body movement and gestures). It would also be worthwhile to investigate whether rationality and other measures discussed here are influenced in a similar manner in additional types of competitive settings.

REFERENCES

[1] H.-M. Gross, H. Boehme, C. Schroeter, S. Müller, A. König, E. Einhorn, C. Martin, M. Merten, and A. Bley, "Toomas: interactive shopping guide robots in everyday use-final implementation and experiences from long-term field trials," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 2005–2012.

[2] E. Hyun, H. Yoon, and S. Son, "Relationships between user experiences and children's perceptions of the education robot," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 199–200.

[3] M. de Jong, K. Zhang, A. M. Roth, T. Rhodes, R. Schmucker, C. Zhou, S. Ferreira, J. Cartucho, and M. Veloso, "Towards a robust interactive and learning social robot," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 883–891.

[4] E. Broadbent, I. H. Kuo, Y. I. Lee, J. Rabindran, N. Kerse, R. Stafford, and B. A. MacDonald, "Attitudes and reactions to a healthcare robot," *Telemedicine and e-Health*, vol. 16, no. 5, pp. 608–613, 2010.

[5] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "It's all in the game: Towards an affect sensitive and context aware game companion," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–8.

[6] J. Shim and R. C. Arkin, "A taxonomy of robot deception and its benefits in hri," in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Oct 2013, pp. 2328–2335.

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

[8] E. Short, J. Hart, M. Vu, and B. Scassellati, "No fair!! an interaction with a cheating robot," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 219–226.

[9] A. Litoiu, D. Ullman, J. Kim, and B. Scassellati, "Evidence that robots trigger a cheating detector in humans," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 165–172. [Online]. Available: http://doi.acm.org.proxy.library.cmu.edu/10.1145/2696454.2696456

[10] N. Spatola, C. Belletier, A. Normand, P. Chausse, S. Monceau, M. Augustinova, V. Barra, P. Huguet, and L. Ferrand, "Not as bad as it seems: When the presence of a threatening humanoid robot improves human performance," *Science Robotics*, vol. 3, no. 21, 2018. [Online]. Available: http://robotics.sciencemag.org/content/3/21/eaat5843

[11] M. Scheutz, P. Schermerhorn, and J. Kramer, "The utility of affect expression in natural language interactions in joint human-robot tasks," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, ser. HRI '06. New York, NY, USA: ACM, 2006, pp. 226–233. [Online]. Available: http://doi.acm.org/10.1145/1121241.1121281

[12] E. Paeng, J. Wu, and J. Boerkoel, "Human-robot trust and cooperation through a game theoretic framework," in *AAAI Conference on Artificial Intelligence*, 2016.

[13] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Robot mood is contagious: effects of robot body language in the imitation game," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 973–980.

[14] N. Frijda, "Emotion experience," *Cognition & Emotion*, vol. 19, no. 4, pp. 473–497, 2005.

[15] B. Wild, M. Erb, and M. Bartels, "Are emotions contagious? evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences," *Psychiatry research*, vol. 102, no. 2, pp. 109–124, 2001.

[16] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322–332, 2010.

[17] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[18] G. Loewenstein, "Emotions in economic theory and economic behavior," *American Economic Review*, vol. 90, no. 2, pp. 426–432, May 2000. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.90.2.426

[19] O. B. Conmy, *Trash talk in a competitive setting: Impact on self-efficacy, affect, and performance*. The Florida State University, 2008.

[20] L. Steinberg, "A social neuroscience perspective on adolescent risk-taking," *Developmental review*, vol. 28, no. 1, pp. 78–106, 2008.

[21] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[22] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[23] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," 2014.

[24] S. Mahamood and E. Reiter, "Generating affective natural language for parents of neonatal infants," in *Proceedings of the 13th European Workshop on Natural Language Generation*. Association for Computational Linguistics, 2011, pp. 12–21.

[25] F. Keshtkar and D. Inkpen, "A pattern-based model for generating text to express emotion," in *Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 11–21.

[26] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, "Affect-lm: A neural language model for customizable affective text generation," *arXiv preprint arXiv:1704.06851*, 2017.

[27] J. Hirth, N. Schmitz, and K. Berns, "Towards social robots: Designing an emotion-based architecture," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 273–290, 2011.

[28] M. Häring, N. Bee, and E. André, "Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *Ro-Man, 2011 Ieee*. IEEE, 2011, pp. 204–209.

[29] A. Beck, B. Stevens, K. A. Bard, and L. Cañamero, "Emotional body language displayed by artificial agents," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, p. 2, 2012.

[30] M. Zecca, Y. Mizoguchi, K. Endo, F. Iida, Y. Kawabata, N. Endo, K. Itoh, and A. Takanishi, "Whole body emotion expressions for kobian humanoid robotpreliminary experiments with different emotional patterns," in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE, 2009, pp. 381–386.

[31] J. Skågeby, "well-behaved robots rarely make history: Coactive technologies and partner relations," *Design and Culture*, pp. 1–21, 2018.

[32] J. K. Lee and C. Breazeal, "Human social response toward humanoid robot's head and facial features," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '10. New York, NY, USA: ACM, 2010, pp. 4237–4242. [Online]. Available: http://doi.acm.org/10.1145/1753846.1754132

[33] C. M. de Melo, P. Carnevale, and J. Gratch, "The effect of expression of anger and happiness in computer agents on negotiations with humans," in *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, ser. AAMAS '11. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 937–944.

[34] C. Pelachaud, "Studies on gesture expressivity for a virtual agent," *Speech Communication*, vol. 51, no. 7, pp. 630–639, 2009.

[35] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, NY, USA: ACM, 2012, pp. 367–374. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157811

[36] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, vol. 4. IEEE, 2002, pp. 4138–4142.

[37] R. Gockley, J. Forlizzi, and R. Simmons, "Interactions with a moody robot," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, ser. HRI '06. New York, NY, USA: ACM, 2006, pp. 186–193. [Online]. Available: http://doi.acm.org/10.1145/1121241.1121274

[38] R. Gockley, R. Simmons, and J. Forlizzi, "Modeling affect in socially interactive robots," in *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*. IEEE, 2006, pp. 558–563.

[39] J. M. Kessens, M. A. Neerincx, R. Looije, M. Kroes, and G. Bloothooft, "Facial and vocal emotion expression of a personal computer assistant to engage, educate and motivate children," in *Affective Computing and In-*

telligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009, pp. 1–7.

[40] B. Robins, K. Dautenhahn, and P. Dickerson, "From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot," in *Advances in Computer-Human Interactions, 2009. ACHI'09. Second International Conferences on*. IEEE, 2009, pp. 205–211.

[41] D. C. Berry, L. T. Butler, and F. De Rosis, "Evaluating a realistic agent in an advice-giving task," *International Journal of Human-Computer Studies*, vol. 63, no. 3, pp. 304–327, 2005.

[42] I. Leite, A. Pereira, C. Martinho, and A. Paiva, "Are emotional robots more fun to play with?" in *Robot and human interactive communication, 2008. RO-MAN 2008. The 17th IEEE international symposium on*. IEEE, 2008, pp. 77–82.

[43] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 229–236.

[44] M. F. Jung, "Affective grounding in human-robot interaction," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 263–273.

[45] A. Sinha, F. Fang, B. An, C. Kiekintveld, and M. Tambe, "Stackelberg security games: Looking beyond a decade of success." in *IJCAI*, 2018, pp. 5494–5501.

[46] J. K. GOEREE, C. A. HOLT, and T. R. PALFREY, *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press, 2016.

[47] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[48] M. A. Walker, R. Grant, J. Sawyer, G. I. Lin, N. Wardrip-Fruin, and M. Buell, "Perceived or not perceived: Film character models for expressive nlg," in *International Conference on Interactive Digital Storytelling*. Springer, 2011, pp. 109–121.

[49] M. A. Walker, G. I. Lin, and J. Sawyer, "An annotated corpus of film dialogue for learning and characterizing character style." in *LREC*, 2012, pp. 1373–1378.

[50] R. Yang, C. Kiekintveld, F. Ordonez, M. Tambe, and R. John, "Improving resource allocation strategy against human adversaries in security games," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 458.

[51] T. H. Nguyen, R. Yang, A. Azaria, S. Kraus, and M. Tambe, "Analyzing the effectiveness of adversary modeling in security games." in *AAAI*, 2013.

[52] E. Guizzo, "A robot in the family," *IEEE Spectrum*, vol. 52, no. 1, pp. 28–58, 2015.

[53] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *Humanoid robots, 2006 6th IEEE-RAS international conference on*. Citeseer, 2006, pp. 518–523.

[54] H.-J. Suk, "Color and emotion-a study on the affective judgment across media and in relation to visual stimuli," Ph.D. dissertation, Universität Mannheim, 2006.

[55] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[56] J. Pita, R. John, R. Maheswaran, M. Tambe, and S. Kraus, "A robust solution concept for human bounded rationality in security games," 2012.

[57] I. N. Debra Bernstein, Kevin Crowley, "Working with a robot," 2007.

[58] M. Gouko and C. H. Kim, "Fundamental study of robot behavior that encourages human to tidy up table," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser. HRI'15 Extended Abstracts. New York, NY, USA: ACM, 2015, pp. 89–90. [Online]. Available: http://doi.acm.org/10.1145/2701973.2701981

[59] M. Nooraie, "Factors influencing strategic decision-making processes," *International Journal of Academic Research in Business and Social Sciences*, vol. 2, no. 7, p. 405, 2012.

[60] F. A. Drews, M. Pasupathi, and D. L. Strayer, "Passenger and cell phone conversations in simulated driving." *Journal of Experimental Psychology: Applied*, vol. 14, no. 4, p. 392, 2008.

[61] P. Gilbert, F. Catarino, J. Sousa, L. Ceresatto, R. Moore, and J. Basran, "Measuring competitive self-focus perspective taking, submissive compassion and compassion goals," *Journal of Compassionate Health Care*, vol. 4, no. 1, p. 5, 2017.

[62] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[63] S. D. Akshay Agrawal, Robin Verschueren and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.