# A Robot's Expressive Language Affects Human Strategy and Perceptions in a Competitive Game

Aaron M. Roth[1] Samantha Reig[2] Umang Bhatt[3] Jonathan Shulgach[3]
Tamara Amin[3] Afsaneh Doryab[2] Fei Fang[4] Manuela Veloso[5]
[1]Robotics Institute [2]Human-Computer Interaction Institute
[3]College of Engineering [4]Institute for Software Research [5]Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
Corresponding author email: amroth@umd.edu

***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 expressive language (in the form of "encouraging" or "discouraging" verbal commentary) 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 contribute a simple open source Natural Language Processing framework for generating expressive sentences, which was used to generate the speech of our autonomous social robot.**

## I. Introduction

The future will bring humans into contact with robots 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 [4]. In any interaction like this, linguistic nuances and positive or negative valence of social behavior will impact the result of the interaction. 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. The humans and the robots have to behave strategically in such settings to gain advantage in the interaction. Much work has gone into understanding how humans and social robots interact and partner in cooperative settings [5]–[7]. For example, positive robot affect has been shown to contribute to perceptions of robots as teammates [8]. However, less research has been done to understand how affect impacts interactions when the interests of the humans and robots are not perfectly aligned. In this study, we focus on a competitive setting and study the impact of a robot's affect on humans' *rational* behavior, which is understudied in HRI despite its significance. Acknowledging that *affect* takes many forms, we focus on affect exhibited through **encouraging** and **discouraging** language. This **expressive language** is one manifestation of how positive/negative affects could emerge in a competitive interaction.

In this study, we examine the impact of expressive language from a robot on human rationality and strategy prioritization in a representative general-sum competitive game, the Guards and Treasures game. (This paper considers rationality in the context of maximizing expected utility.) This game has been used extensively in the literature on Stackelberg Security Games to collect human play data and analyze bounded rationality of humans [9], [10]. We adopt this game as it provides a simple environment in which players' interests are not fully aligned. We seek to answer the question: how does encouraging or discouraging language from a humanoid robot opponent impact a human's rationality and strategy in this example strategic game? We expect the results from this study to shed light on more general settings of competitive or semi-competitive interaction between robots and humans. We implement a system to play the game autonomously with dialogue generated by our expressive language algorithm. In a between-subjects study, 40 participants played the game with a humanoid robot. Each participant was exposed to one of two conditions in which the robot made either encouraging or discouraging comments. We analyze the collected data to obtain insights into how the robot's behavior impacts participants' rationality and emotions. Some existing work shows that threatening behavior from a robot may increase humans' attentional control [11]. In contrast, in our study, discouraging comments from a robot decreased a participants rationality during gameplay. In addition, negative language contributed to negative social attributions to the robot.

In addition to investigating the impact of a robot's use of expressive language on a human opponent's strategy, risk-taking, and performance in a competitive game setting, we contribute an open-source Natural Language Processing (NLP) model that is affect-aware. We discuss these findings and others in section V.

## II. Background and Related Work

Observing others' moods can have specific consequences for the observer [12], impacting their performance [11],

[13], risk taking [14], decision making [15], and mood [16]. Mood-contagion is a well-researched automatic mechanism whereby the observation of another person's emotional expression induces a congruent state of mood in the observer.

Affect is a general term relating to emotions, moods, feelings and desires that may influence behavior. Affective states vary in their degree of *activation* (intensity) and *valence* (whether they are positive and negative) [17]. Research has shown that humans' perception of robots' affect and/or expressive language can influence interactions. Various studies have used the ROMAN robot for facial expressions [18], [19], NAO for body expressions [18]–[20], KOBIAN for body and facial expressions [21], and Cozmo for nonverbal behaviors [22]. One study found that humans can identify and respond to a robot's expressive language [23]. Robot language can influence the effectiveness of assistive tasks including learning or receiving vocal encouragement expressions [24]. Another study found that a computer agent's expression (anger and happiness) impacted the way humans negotiated with it [25]. Research on the impact of expressive language between language models in human-robot interaction has been limited to joint human-robot tasks in a cooperative setting; therefore we extend an expressive language model to a competitive human-robot interaction situation.

One experiment showed that human strategy is different when facing a text-based mediator vs. a mediator with an avatar [26], which indicates that the form of interaction between a robot and a human matters. Several other studies have shown the impact of affective virtual agent behavior on human task performance [7], [27], [28]. We consider a competitive scenario and analyze human performance and strategy when facing a humanoid robot.

To our knowledge, we are the first to explore the impact of a robot's expressive language statements on human performance in a competitive game setting. We use a game which can be mathematically modeled as a Stackelberg security game [29] to analyze human rationality and strategy as well as human perception of the robot from a combination of validated scales and experiment-specific Likert items. Quantal response [30] and its variant—subjective utility quantal response [31]—have been proposed and used in game theory literature to quantitatively model the bounded rationality of human players and their prioritization of different factors impacting their decision making. We leverage these models for strategic human-robot-interaction. We use a humanoid robot in this study. In contrast to [7] and [28], we deal with verbal affect cues instead of gesture and posture.

## III. Methodology and Contributions

### A. Overview of Study

The primary goal of our study was to determine the effect of a robotic opponent's expressive language on a human's game-playing strategy in a competitive and strategic game. We hypothesized that (1) when playing a competitive game against a humanoid robot, a human's strategy will be influenced by the expressive language of the robot, and (2) encouraging expressive language will positively impact participants' social perceptions of the robot. We conducted a between-subjects experiment in which a human played a repeated Stackelberg security game against a robot. (Specifically, the human plays two practice rounds without the robot and then 35 rounds of the game "against" the robot.) The robot made either *encouraging* or *discouraging* comments during game play. We recorded participants' actions to understand the nature of their game play strategy. We also measured their perceptions of the task, of their performance, and of the robot. A detailed description of the experimental setup and procedure can be found in section IV.

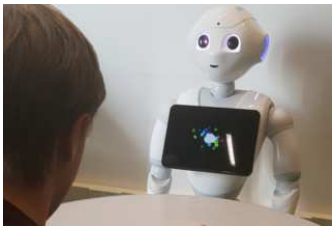


Fig. 1. The study setup from a participant's perspective

We manipulated robot expressive language in the form of periodic utterances that were either *encouraging* or *discouraging*. Utterances were generated via an NLP model we discuss in section III-B. The robot exhibited optimal strategy in all games, regardless of the condition or the human player's actions. The robot moved and spoke autonomously according to a script that our framework generated ahead of time.

Our primary measures of interest pertained to the participant's strategy. We analyzed their "strategy" in two ways. First, we used a quantal response equilibrium model 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 the parameter values of the subjective utility quantal response model that can best fit the human play data 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 expressive language, we developed an affect-aware bidirectional fill-in-the-blank N-Gram model.

We construct a probability that a particular word follows a previous sequence of words using an N-gram [32] method:

$$P(w_n|w_S,...,w_{n-1}) = \frac{\mathrm{C}(w_S,...,w_{n-1},w_n)+\alpha}{\mathrm{C}(w_S,...,w_{n-1},{*w})+D\alpha} \quad (1)$$

where $S = n-(N-1)$, C() means "count of", and $*w$ is a wildcard meaning "any word observed as completing this sequence". Thus, $P(w_n|w_S,...,w_{n-1})$ is the probability

that a particular word $w_n$ follows a particular sequence of $N$ other words. 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 affect-aware, we used the AFINN Affect Dictionary, which rates the emotional valence of a word [33]. We constructed sentence stems (sentences with fill-in-the-blanks) such that positive or negative fill-in words result in encouraging or discouraging sentences. (These sentence stems can be fed into our model.) We use both bigrams and trigams ($N = 2$, $N = 3$), and train our model in both forward and reverse direction. The final equation used to select the words to complete neutral sentence stems 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 [34], [35]. 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 Guards and Treasures, a Stackelberg security game [29], [36], against the robot. 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, in an attempt to maximize the expected numerical reward.

Quantal response model assumes that a human player is more likely to choose more promising options. Mathematically, let $q_{c_r}$ represent the probability of the participant selecting choice $c$ in round $r$. It is defined in (3), where $\lambda$ is a parameter that can control or represent how rational the participant's decision is.

$$q_{c_r} = \frac{\exp\left(\lambda U_{c_r,r}\right)}{\sum_{j=1}^{N} \exp\left(\lambda U_{j,r}\right)} \quad (3)$$

Assuming that participants follow the quantal response model, given the game play data of a given set of rounds $\Upsilon$, we can learn the value of parameter $\lambda$ that can best fit the data via maximum likelihood estimation. This is shown in (4) where $\lambda$ is fit to a subset of rounds $\Upsilon$.

[1]Find the NLP model code here: https://github.com/AMR-/fill_in_the_blank_word_prediction

$$\lambda = \underset{\lambda}{\operatorname{argmax}} \sum_{r \in \Upsilon} \log(q_{c_r}) \quad (4)$$

where $U_{i,r}$ is the known real utility of choice $i$ in round $r$ (see (7)), $c_r$ is the number of the choice chosen by the participant in round $r$, and $\Upsilon$ is the subset of rounds to be used in the calculation.

*2) Measure of Prioritization in Strategy:* We can also follow the Subjective Utility Quantal Response (SUQR) model, defined in (5), to determine the probability $s_{c_r}$ that a participant selects choice $c$ in round $r$ based on the parameters $W$ representing the **strategic priority** of the participant [31]. $W$ denotes the importance to the participant of different attributes of each of the options.

$$\begin{aligned} s_{c_r} &= \frac{\exp\left(W^T X_{c_r,r}\right)}{\sum_{j=1}^{8} \exp\left(W^T X_{j,r}\right)} \\ W^T &= [w_1 \quad w_2 \quad ... \quad w_n] \qquad X_{i,r}^T \in \mathcal{R}^3 \end{aligned} \quad (5)$$

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.

$$W = \underset{W}{\operatorname{argmax}} \sum_{r \in \Upsilon} \log(s_{c_r}) \quad (6)$$

We use Maximum Likelihood Estimation as shown in (6) to determine the values of strategic prioritization $W$ that best fit the data.

## IV. Experimental Setup and Protocol

### A. Participants

We recruited 40 participants from the local community (15 M, 24 F, 1 nonbinary, $M_{age} = 27.2, SD_{age} = 11.2$). All participants played Game Session I consisting of 35 rounds of the Guards and Treasures game against the robot, referred to as the "basic game". A selected group of the participants, referred to as the "two-session group", also played Game Session II, consisting of another 35 rounds, referred to as the "additional games", in which the robot exhibited the opposite language behavior as from the basic games.

### B. Robot

We used the Pepper Robot by Softbank Robotics [37] (a research robot provided for participation in the RoboCup Social Standard Platform League).[2] Pepper is a humanoid robot with arms, a head with cameras and microphones, mobility, and voice abilities. See it pictured in Fig. 1. HRI research suggests that physically present, embodied robots may make interactions more engaging and enjoyable and increase social presence [38], [39]. We used a humanoid robot for these reasons, as well as to give the participants a visual, physical reference for interacting with their opponent (for example, Pepper made eye contact with participant). We

[2]http://www.robocupathome.org/athome-spl

hoped that using an embodied robot would maximize the opportunity for participants to attribute social characteristics to the opponent and exhibit task-relevant and affective responses to its behaviors. Pepper acted autonomously according to a script that 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 measures of pre-task emotional state.

*3) Practice Rounds:* The participant played on a "convertible" combination laptop/tablet. In order to counter learning effects, we had participants play two practice rounds of the game "against the computer".

*4) Game Session I (Basic Games):* After the practice rounds, the participant was led into a room where the robot sat behind a table. The participant sat across from the robot with the tablet face-up between them. The participant then played several rounds of the game "against" the Pepper robot. The robot made periodic comments about the game and the participant. The comments exhibited either **encouraging** or **discouraging** expressive affect. Although the commentary was sometimes complimentary and sometimes critical in nature, in reality it had nothing to do with the participant's actual performance.

*5) Post-Game Survey and Video:* Upon completion of the games, the participant notified the researcher that they had completed the task. The researcher then administered a written survey and a verbal semi-structured interview, which was video-recorded.

*6) Game Session II (Additional Games):* A selected subset (the "two-session group") 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 Guards and Treasures Game Interface

One round of the "Guards and Treasures" game is a modified version of the game from [29].[3] This specific game is useful for studying bounded rationality. It features a virtual scenario and provides a limited set of options in each round. In each round, the participant is shown a screen as in Fig. 2. The central idea of the game is that the player can choose to "attack" (select) each of several gates. If the defending player (the robot) places a guard at the gate, the human player incurs

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

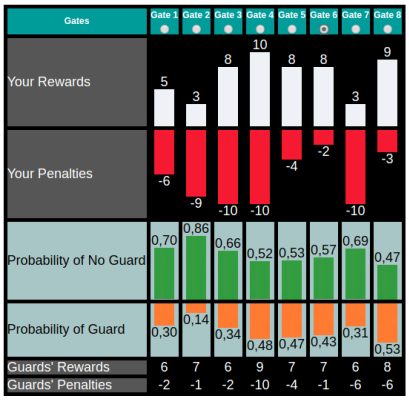



Fig. 2. A round of the Guards and Treasures game, adapted from [29].

the penalty for that gate. If the chosen gate is not guarded, 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 (whether the gate they chose is guarded and whether they got the reward or penalty) after all the rounds are complete.

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

$$U_{i,r} = R_{i,r}(1 - g_{i,r}) - g_{i,r}Y_{i,r} \tag{7}$$

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 $r$. ($R \in \mathbb{Z}, R \in [1, 10], Y \in \mathbb{Z}, Y \in [1, 10]$, and $g \in [0, 1]$) $N = 8$ as there are 8 gates, each gate is one of the choices a participant can make. The $X_{i,r}$ from (6) 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 to reward, penalty, and probability of seeing a guard, respectively.

### E. Surveys and Data Collection

Our measures consisted of i) a pre-task questionnaire, ii) records of actions taken during the game, iii) a post-task questionnaire, iv) a post-game verbal semi-structured interview (recorded on camera), and (for some participants) v) video of the participant playing the game against the robot.

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. We originated some of the questions, and other questions we drew from [40] and [41]. The pre-task and post-task questionnaires both made use of the Self Assessment Manikins [42], 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 pertaining to participants' overall perception of the

TABLE I
$\lambda$ AND $W$ FOR VARIOUS POPULATIONS $\Upsilon$

| | $\lambda$ | | | $W$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Affect:** | **Both** | **Positive** | **Negative** | **Both** | | | **Positive** | | | **Negative** | | |
| **Rounds $\Upsilon$ from Population:** | | | | $w_1$ | $w_2$ | $w_3$ | $w_1$ | $w_2$ | $w_3$ | $w_1$ | $w_2$ | $w_3$ |
| 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 Two-Session 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 Two-Session 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 |

robot, overall thoughts about the experience, self-assessment of performance, perceptions of the robot's goal, and 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

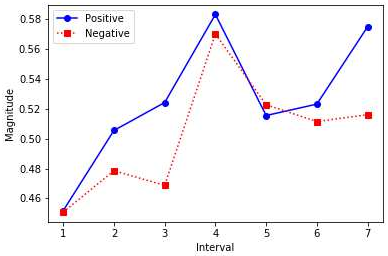
We solve (4) and (6) over data from rounds from aggregated groups of participants. In table I, we show $\lambda$ and $W$ values corresponding to rounds played by various populations ($\Upsilon$s). These $\Upsilon$s span multiple participants, divided by affect, and include all rounds for each participant in each subset. In Quantal Response mode, $\lambda = 0$ indicates uniform random behavior and $\lambda = \infty$ indicates perfect rationality (i.e., best response). The parameter values of $W$ in our variant of the SUQR model describes a participant's strategic prioritization over different factors influencing their decision making ($w_1$: reward component, $w_2$: penalty component, $w_3$: probability-of-guard component). For reference, previous work [31] obtained $\lambda = 0.77$ via a group of Amazon Mechanical Turk (AMT) workers playing this game online. In addition, [31] reports $W = [0.37, 0.15, -9.85]$ (converted to our representation) for the SUQR model for a general population playing this game.

The first row in the table has basic, first round games for all participants. The "two-session group" of participants are those who played an additional game with the reverse affect (mentioned in sections IV-C6 and IV-C7). In addition, using the procedure in [43], we analyze changes in $\lambda$ and $W$ between the basic and additional session for each individual participant in the two-session group. We found that those who played a negative session first had a 21% increase in $\lambda$ and a 104% increase in the 1-norm of the strategic prioritization vector, $W$, on average. On the other hand, those who played the positive session first had a 28% increase in $\lambda$ and a 110% increase in the 1-norm of $W$. Further, we divide the 35 rounds of game into seven 5-round intervals and analyze the best parameter values in each stage. Fig. 3(a) shows the trend for $\lambda$ for positive and negative affect over time for participants' basic games.

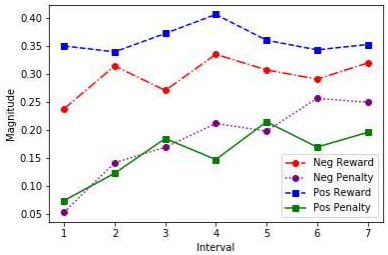
We notice that although participants place a higher priority on reward than on penalty, all participants place more emphasis on penalty over time in Fig. 3(b). Just as weight of reward ($w_1$) is relatively steady over multiple intervals of five rounds, we found that the weight participants placed on 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 (i.e., the total utility the participant achieved was higher than their opponent) and 5 lost. In the negative affect condition, 16 participants won and 4 lost. This difference in performance was not significant. We found no effect of affect on prioritization between W components.

### B. Analysis of Self Assessment Manikin

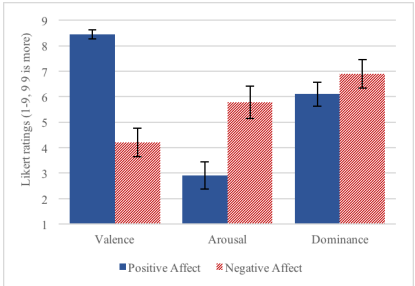
The questionnaire about perceptions of the robot was also analyzed. Because the data were not normally distributed and the sample was small ($n < 50$), the non-parametric




(a) Value of $\lambda_P$ (positive) and $\lambda_N$ (negative) over time (captured at seven 5-round intervals across an affect class) Both increase, but $\lambda_P$ more so




(b) Values of $w_1$ and $w_2$ (Reward and Penalty components of $w$ over time. Each interval is 5 rounds.) Participants place more value on penalty over time.




(c) Mean scores for Self Assessment Manikin for perception of robot by affect class. Error bars represent $\pm 1$ standard error of the mean.

Fig. 3. Results: Trends and Comparisons

| 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

Wilcoxon/Kruskal-Wallis rank sum test for two samples was used to compare the two conditions for all variables.

The robot's expressive language significantly affected 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, encouraging language led to higher ratings. There was no effect of the language manipulation on perceptions of humanlikeness or cuteness. These results serve as a validity check, suggesting that the encouraging commentary was perceived as positive and and the discouraging commentary was perceived as negative.

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 [42] to measure the participants' mood and before and after the game in terms of emotional valence (how happy or unhappy they felt), arousal (how excited or unexcited they felt), and dominance (how in-control they felt). We also 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, $\chi^2(1, N = 40) = 20.87$, $p < 0.0001$.For these variables, participants in the encouraging language condition had higher ratings. We also found an effect wherein discouraging language positively impacted ratings of perceived robot arousal, $\chi^2(1, N = 40) = 10.07$, $p = 0.002$. There were no main effects of language on post-task participant arousal, participant dominance, or perceived robot dominance. The effect of condition on perceptions of robot affect can be seen in Fig. 3(c). Collectively, these findings corroborate previous work [6], [25] suggesting that affective robot behavior is able to strongly influence people's feelings in a dyadic interaction. Our case differs from this work in that 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 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 expressive language 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 humanlikeness and dominance. We found an interaction effect of pre-task participant valence and robot expressive language on perceptions that the robot was humanlike, $p = 0.015$, in that discouraging language and low valence prior to the start of the experiment led to lower perceptions of humanlikeness.

We found that discouraging language significantly lowered participants' beliefs that the robot was 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 that encouraging language increased perceptions that the robot was cute, $F(1, 9) = 6.92$, $p = 0.027$.

### C. Participant Interviews

To gain further insight into participants' impressions of the robot, we conducted semi-structured interviews with our participants. Twelve participants (four in the encouraging condition and eight in the discouraging condition) reported a belief that the robot's goal involved distracting them. Participants in the encouraging condition said, "When I was trying to determine what move to make, it took me out of that zone for a bit" (P220), and, "It felt like I was doing homework and my friend kept talking to me" (P201). Altogether, 30% of participants explicitly classified the robot's goal as "distraction". Participants also spoke about the robot's behavior as a result of its programming. For example, P104 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". Interviews also further confirmed participants were encouraged by the robot in the encouraging condition and were especially discouraged in the discouraging condition. When asked about the robot's goal, a participant in the encouraging condition answered, "To encourage me to do well... it seemed to [succeed in that goal]" (P117), while a participant exposed to the discouraging language 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 supported the quantitative results. 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. Population Rationality

Overall, we found that discouraging expressive language caused 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 [31], [44], [45]. 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.

Participants who played an additional game (players in the "two-session group") 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.

While there were outliers, our participants' overall rationality was below that of the crowd-sourced AMT population from [31]. The discrepancy may be attributable to differences in the game framing, timing, population, or noise. One possible explanation is that the amount of money our participants received was fixed as opposed to dependent on their performance, like AMT workers. 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" 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 [46]. Dialogue can also be a distraction, regardless of content [47].

Over a quarter of all participants **explicitly** expressed a belief that that one of the robot's goals was to distract them. This suggests that, given the competitive setting, some participants were focused more on winning the game than on interacting with the robot. 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), it may not be in a competitive setting.

### C. Perception of Humanoid Robot

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. Younger participants were less influenced by affect, which could be due to a younger generation more used to thinking of robots as machines.

## VII. Conclusion

A humanoid robot that encourages or discourages a human opponent can impact that human's rationality. In our study, a discouraging robot led to lower rationality while an encouraging robot was associated with higher rationality. The insights documented here may 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. Our findings may serve to help future robot designers develop a better understanding of how affect impacts perceptions of a social robot during non-cooperative interactions. Useful future work would be to investigate nonverbal modes of expression, like body movement and gestures, in competitive settings.

## Acknowledgment

This work was supported in part by NSF grant IIS-1850477. We would like to thank Jeffery Cohn (Robotics Institute/Department of Psychology, CMU/Pitt) and Louis Philippe Morency (Language Technologies Institute, CMU) for their guidance and advice. We thank Tianyu Gu and Ashley Liu for help running some of the experiments. We thank Gayatri Shandar who transcribed and annotated some of the videos.

## 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] 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

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

[7] 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.

[8] 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.

[9] J. Pita, M. Jain, F. Ordóñez, M. Tambe, S. Kraus, and R. Magori-Cohen, "Effective solutions for real-world stackelberg games: When agents must deal with human uncertainties," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 369–376.

[10] J. Pita, R. John, R. Maheswaran, M. Tambe, R. Yang, and S. Kraus, "A robust approach to addressing human adversaries in security games," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 1297–1298.

[11] 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

[12] 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.

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

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

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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 *The 18th IEEE International Symposium on Robot and Human Interactive Communication, 2009.* IEEE, 2009, pp. 381–386.

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

[23] 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 Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[24] 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

[25] 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.

[26] R. Lin, Y. Gev, and S. Kraus, "Bridging the gap: Face-to-face negotiations with an automated mediator," *IEEE Intelligent Systems*, no. 6, pp. 40–47, 2011.

[27] 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.

[28] 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.

[29] 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.

[30] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games and economic behavior*, vol. 10, no. 1, pp. 6–38, 1995.

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

[32] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

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

[34] 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.

[35] 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.

[36] 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.

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

[38] S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey, "Anthropomorphic interactions with a robot and robot–like agent," *Social Cognition*, vol. 26, no. 2, pp. 169–181, 2008.

[39] A. Powers, S. Kiesler, S. Fussell, S. Fussell, and C. Torrey, "Comparing a computer agent with a humanoid robot," pp. 145–152, 2007.

[40] 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.

[41] 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.

[42] 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.

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

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

[45] 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

[46] 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.

[47] 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.

[48] A. M. Roth, U. Bhatt, T. Amin, A. Doryab, F. Fang, and M. Veloso, "The impact of humanoid affect expression on human behavior in game-theoretic setting," in *Proceedings of the IJCAI'18 Workshop on Humanizing AI (HAI), the 28th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, July 2018.