Your Robot is Watching: Using Surface Cues to Evaluate the Trustworthiness of Human Actions

Vidullan Surendran[1] and Alan R. Wagner[2]

***Abstract*— A number of important human-robot applications demand trust. Although a great deal of research has examined how and why people trust robots, less work has explored how robots might decide whether to trust humans. Surface cues are perceptual clues that provide hints as to a person's intent and are predictive of behavior. This paper proposes and evaluates a model for recognizing trust surface cues by a robot and predicting if a person's behavior is deceitful in the context of a trust game. The model was tested in simulation and on a physical robot that plays an interactive card game. A human study was conducted where subjects played the game against a simulation, the robot, and a human opponent. Video data was hand coded by two coders with an inter-rater reliability of 0.41 based on Levenshtein distance. It was found that the model outperformed/matched the human coders on 50% of the subjects. Overall, this paper contributes a method that may begin to allow robots to evaluate the surface cues generated by a person to determine whether or not it should trust them.**

I. Introduction

A number of important human-robot applications, such as autonomous driving, demand trust. A great deal of research has examined how and why people trust robots [1]. Some work has also explored how robots might decide whether to trust humans [2]. For humans deciding whether or not to trust another human, surface cues are important [3]. Surface cues are perceptual clues that reflect intent. The interaction between surface cues and human-robot trust is generally understudied, especially when one considers the broad variety of different cues that exist. Nevertheless, for applications ranging from autonomous driving to search and rescue, it would be valuable if a robot could recognize the surface cues that signal a person's intention to trust the robot.

Yet, recognizing the surface cues that signal one's intention to trust is a difficult problem. Related research has shown that culture, appearance, and the task all play an important role in the decision to trust a robot and in the cues a person provides [4], [5], [6], [7]. Individual differences, one's prior history, and even cultural similarity can impact the decision to trust [8], [9], [10]. Although some of the factors that influence trust have been identified, significant gaps in our knowledge of how surface cues foster trust remain [11]. Primary among these are questions related to how machines trigger trustworthiness perceptions, how characteristics of communication such as voice and embodiment impact trust, and what cues are most influential over time. This paper presents a model capable of making predictions about the trustworthiness of an action based on perceived cues.

Trust researchers generally agree that risk is a prerequisite for trust (for example see [12], [13]). Yet the type of risk faced by the trustor may influence the use of trust surface cues and the decision to trust [14], [15]. The purpose of this paper is to create a model that allows a robot to capture and translate these surface cues into actionable information, in particular, recognition that a person's actions are trustworthy.

Games have long been used to evaluate trust in the cognitive science and behavioral economics literature [16], [17]. We propose an interactive card game that allows the robot to iteratively build a model of the person through successive interactions that are structured within the rules of the game. Our long-term goal is to create the underpinnings that will allow a robot to use percievable overt cues to estimate the trustworthiness of an action. We postulate that this information could be used to detect or signal trust.

The remainder of this paper begins by discussing related work. We then describe the game used to evaluate our method and the architecture used. Next the computational model that we have developed to estimate trustworthiness is presented followed by experiments demonstrating the model's performance. We then detail a within person experiment conducted to ascertain if overt cues are observable, and to study if there is a discernible pattern to them. We conclude with a discussion of the impact, assumptions, and directions for future work.

II. Related Work

This research touches on a number of broad robotics, artificial intelligence, and human-robotic interaction topics. With respect to human-robot trust, one avenue of research attempts to model trust as a probabilistic variable indicating system performance [18], [19], [20]. These approaches tends to ignore well established cognitive science research which suggests that trust is also contextual and relates to one's emotional state, personality, and experiences [21], [22]. Recent human-robot interaction experiments in ecologically valid environments also suggest that trust is not a simple function of robot reliability [23], [24].

A large body of literature has demonstrated that appearance cues play a vital role in person perception [25], prediction [26], and trust [27], [28]. Appearance cues are commonly defined as appearance-related perceptual features, which signal underlying behavior, emotions, or motives

*This work was supported by Air Force Office of Sponsored Research contract FA9550-17-1-0017

[1]Prof. Alan R. Wagner is with Faculty of Aerospace Engineering, Pennsylvania State University, PA, USA, 16802 alan.r.wagner@psu.edu

[2]Vidullan Surendran is a PhD candidate in the Department of Aerospace Engineering, Pennsylvania State University, PA, USA, 16802 vus133@psu.edu

[29]. Behavior cues, which are the focus of this paper, are action-related perceptual features [30]. As such, they signal an individual's goals, purpose, and abilities. Behavior cues provide evidence allowing the trustor to predict the trustee's behavior in a situation in which the trustor's utility or reward depends on the trustee. Behavior cues provide insight about the actions that the other individual will select. A variety of surface cues influence a person's decision to trust a robot [31], [32], [33]. This paper presents a model that allows a robot to learn whether the presence or absence of surface cues indicates that a person's behavior in the game is trustworthy.

III. Card Game

Verish Ne Verish is a 2-6 player card game where players take turns selecting some number of cards to play face down onto a discard pile and then name the rank and number of the cards they discarded. The player may be truthful or not about the rank of the cards they have just discarded. For example, stating that they have discarded three Ace's when, in fact, they have discarded one Ace and two Kings. The other players then have opportunity to state "I don't trust you" and expose the cards that the player has just discarded. If the discarded cards are not all of the rank that the player claimed, then they must pick up the entire discard pile and add it to their hand. If, on the other hand, all of the cards are of the claimed rank (the player was honest) then the player stating "I don't trust you" must add the discard pile to their hand. The first player to get rid of all of their cards wins.

The crux of this game involves judging the trustworthiness of a person's statement about their discarded cards. For the model discussed in this paper we focus on the evaluation of the trustworthiness of the discarding play. This reduced form of the game consisted of the subject discarding a card face down and stating the suit. The opponents goal was to determine if the subject was to be trusted and relaying this information by stating either, "I think you are telling the truth", or "I think you are lying".

IV. Computational Model

Research from cognitive science indicates that humans tend to form behavioral habits that are associated with an environment and that are linked to the person's goals [34]. In games like poker individuals may display surface cues indicating that they are bluffing mixed with surface cues indicating that they are not [35], [36]. For a robot interacting with a person, detecting and understanding what these cues mean in terms of trust may allow the robot to better collaborate and assist a person [37]. In our application, we assume that the game, not unlike poker, will foster the development of observable and stable surface cues depicting whether or not the person's statement should be trusted. By stable we mean that the sequence of surface cues accompanying a bluff would be constant for many rounds or even games. Further, we hope that this model will be used in more general settings as means for recognizing and evaluating a person's underlying needs. In a healthcare setting, for example, patients may make overt statements that contradict their surface cues.

Our goal is to develop a model that is able to predict whether to trust based on the observation of a sequence of cues. We do not assume any prior knowledge about the players, nor the likelihood that a particular sequence of cues reflects a signal to trust or distrust the player. We also do not assume any prior information about the sequence of the cues themselves. Because the application domain is a game, the model should adapt to the play of the human player, without a lengthy training period or long history of data.

The developed model classifies the observation of a sequence of cues witnessed by the the robot during a round of the game into a binary variable, $X = \{lie, truth\}$. The set of all cues that can be detected by the system is denoted as G, which in this game was $G = \{1, 2, 3, 4, 5\}$ where the number refers to the description in Section V. During each round of play, a sequence of cues, S_g, is generated by the player. S_l represents a sequence of cues that indicates an untrustworthy statement called the 'lie sequence'. The length of the cue sequence was limited to between 1 and 10 cues in order for the pace of the game to be reasonable. We note however that in reality, cue sequences can be arbitrarily long and that the model can still be used with arbitrarily long sequences.

Our goal is to determine the probability $P(X = lie \mid S_g)$. To do this we use Bayes theorem to estimate the posterior probability that the statement is a lie given the observed sequence of cues. In other words,

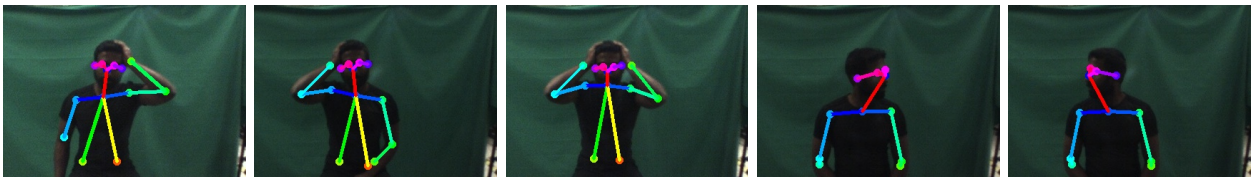
$$P(X = lie \mid S_g) = \frac{P(S_g \mid X = lie)\ P(X)}{P(S_g)} \tag{1}$$

Given a maximum sequence length of 10, the probability of observing any particular sequence is on the order of 10^{-8} if all possibilities are considered. Since, in practice, the whole search space is unlikely to be seen, we restrict the search space to the cues observed during game-play. Using data collected during the game, the priors are defined as follows:

$$P(X) = \frac{number\ of\ times\ class\ was\ observed}{total\ number\ of\ observations} \tag{2}$$

$$P(S_g) = \frac{number\ of\ times\ S_g\ was\ observed}{total\ number\ of\ observations} \tag{3}$$

The conditional probability of observing a sequence when it belongs to a class is empirically calculated by counting the number of times it has been observed when the round resulted in an outcome of that class. The posterior probability allows us to predict if the encountered sequence S_g belongs to the outcome class lie or $truth$. Yet, we also need to detect when a player masks the lie sequence by performing other, unrelated cues. In other words, S_l may be a sub-sequence of S_g. To overcome this, the model needs to predict the sequence S_l correctly to aid in identifying any sequence that contains a lie regardless of such masking. To address



(a) Left hand on head (b) Right hand on head (c) Both hands on head (d) Looking left (e) Looking right

Fig. 1: List of recognized cues overlaid with pose estimates

this problem we calculate not just the probability of the observed sequence S_g being the lie sequence, but also the probability of each subsequence of S_g being the lie sequence. Intuitively this corresponds to hypothesizing that all possible subsequences observed are equally probable to be a lie and then discounting based on future observations. A drawback of using observations of sequences to inform the probability of the subsequences is that many of the subsequences are unlikely to have been observed earlier in the game resulting in a probability of zero, but may nevertheless be probable.

We can smooth the model by preventing an element from being assigned a zero probability by assuming it has been observed non-zero times. The sequence of cues, S_g, is similar in many ways to N-grams from language modelling problems. An N-gram being a contiguous sequence of n items from a given sample. Empirical studies have shown that the Kneser-Ney smoothing algorithm, and its variants consistently outperform most algorithms evaluated for language modelling [38]. However, for our problem the length of the N-gram is unknown as the number of cues exhibited by the human player varies from round to round. Moreover, absolute discounting such as Kneser-Ney might not be appropriate in our situation because we have no evidence to show that N-grams of a particular length are more likely in our game unlike in language processing. Add-one smoothing [39], however, performs well in our situation because this approach assumes N-grams of different lengths are equally likely. Using Add-one smoothing, we assume that all sequences have been observed at least once resulting in non-zero probabilities.

Let U be a set of ordered pairs where each pair represents a subsequence and its posterior probability respectively. This set contains pairs that represent all subsequences generated during the game and the posterior probabilities are updated every round. This information is used to determine the most probable sequence denoted by S_p which is the element with the highest probability.

$$S_p = \{(g,p) \in U : p = max(X) \\ where\ X = \{p : (g,p) \in U\}\} \quad (4)$$

It is possible that multiple sequences have equal probabilities of being the lie sequence especially when minimal data is available at the start of the game. In cases where $|S_p| > 1$, a random element $x \in S_p$ is chosen to be the lie sequence. The model then infers that if $x \not\subseteq S_g$ the player is being trustworthy. With additional data we expect the cardinality of the set S_p to approach 1 representing that a single sequence has been identified. In Section VI we demonstrate the use of this model to learn and evaluate patterns of cues.

V. Cue Recognition

We require a sequence of cues to be identified during each player's turn. To recognize these cues we used OpenPose to determine the upper body pose of the player from camera images [40], [41], [42]. OpenPose is a real time pose detection library that creates 2D pose estimates using keypoint detection. OpenPose generates a feature vector identifying the pixel position of the person's joints. In order to reduce the space of all possible cues and the perceptual demands, the location of the upper body joints (wrists, elbows, shoulders, neck, nose, and eyes) were extracted and used to classify the following five nominal cues: (1) left hand touching the head, (2) right hand touching the head, (3) both hands on the head, (4) turning to the left, and (5) turning to the right. These cues were chosen due to the ease of their detection, being fully described by the location of the joints, and not depending on the path taken or velocities of the joints. Figure 1 depicts the cues. For the remainder of this paper these cues will be referred to by their numbers.

Pose estimation was first performed on videos of the participant and the location of the joints was then used to classify the gesture. A low pass filter was used to reduce noise and the variance between different cues. The output was coded as 1 to 5 representing each of the nominal cues to generate a sequence of cues. As mentioned earlier, this sequence is denoted the 'lie sequence'.

As these cues occurred within the context of the game, we could exploit changes in the game state to inform the system that a sequence begins and ends. Conceptually, the change in game state offers contextual clues signifying situations that may demand trust [43]. Not all situations in a game demand trust. The primary factor for evaluating whether or not a situation demands trust is the risk entailed by relying on the interactive partner's predicted behavior. For this game, the act of playing a card effectively ends the player's turn. We assume that this act also signifies the end of cues signalling the person's underlying behavior. This sequence has the potential to be infinitely long. We required that the sequence be performed within 30 seconds to reduce computational time required to post process the camera data. Empirically it was observed that it took approximately 2-3 seconds to perform each cue at a gentle pace. This resulted

TABLE I: Accuracy with respect to lie sequence length

		Game Number				
		1	2	3	4	5
Sequence Length	2	0.98	0.98	0.90	0.96	0.96
	4	0.94	0.94	0.96	0.90	1.00
	8	0.96	0.88	0.96	1.00	0.98

in a constraint that the maximum number of cues performed before each player's move of playing a face down card is limited to a maximum of 10. Consequently, the lie sequence was also constrained to the same limit.

VI. Experiment One: Quantifying The Model's Behavior in Simulation

The purpose of this initial experiment was to study how the model behaved with respect to the length of the input cue sequence, the change in prediction accuracy with the number of rounds played, and accuracy when a lie sequence changed midway through a game.

The changing length of lie sequence was studied by playing a simulated game consisting of 50 rounds. We tested the model on lie sequences consisting of 2, 4, and 8 cues. The cues making up the lie sequence were chosen at random and the sequence was kept constant throughout each game. Five games of each condition were played and the results are shown in Table I. For a lie sequence length of 2 the accuracy ranged from 0.90 to 0.98 with a mean accuracy of 0.96, a length of 4 yielded a mean accuracy of 0.94 with a range of 0.90 to 0.98, and a length of 8 had a range of 0.88 to 1 with a mean accuracy of 0.96. The variation in accuracy occurs because in the very first round the model must guess whether or not a sequence is a lie due to a lack of prior data. Once sufficient data is obtained, the predictions have a higher confidence score associated with them. The model's accuracy is due, in part, to the fact that the lie sequence remains constant throughout a game. Hence, once the lie sequence is discovered by the model, all further rounds are correctly classified.

If we aggregate all 15 trials (five games times 3 different lie lengths), for an arbitrary lie sequence length and a total sequence length of 10, we can estimate the overall mean accuracy of the model to be 0.95 with a sample standard deviation of 0.03. A model randomly guessing lie or no lie would have an accuracy of 0.5. Data was collected for a model that randomly guessed the outcome class. The mean accuracy for this model was 0.49 with a standard deviation of 0.04. A unpaired t test was conducted to obtain $t = 38.01$ and $p < 0.01$ confirming that our model is statistically better than chance at predicting the outcome class. Although comparison to random guessing is a weak control, to the best of our knowledge no other models exist to predict trustworthiness from surface cues.

Next we examined how the model's accuracy changes as the game goes on and the number of rounds increases. We hypothesized that the accuracy would increase proportional to the number of observations. For a gesture length of 10,

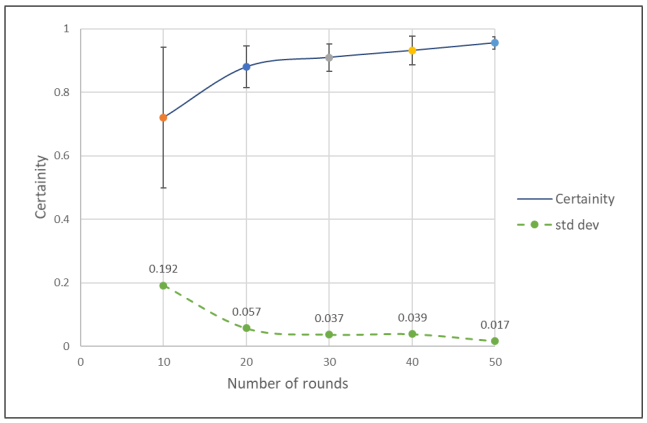


Fig. 2: Improvement in model's prediction of lie sequence.

the change in accuracy was difficult to observe because the model obtained an accuracy of 0.9 within 10-15 rounds and changed little (~10%) thereafter. We therefore increased the maximum sequence length to 20 in order to increase the search space and the number of rounds necessary to obtain an accuracy of 0.9. The length of the arbitrarily chosen lie sequence was 5. This experiment was repeated five times and the mean accuracy of these 5 trials in shown in Fig. 2. After 10 rounds the mean of the system's accuracy was estimated to be about 70% eventually increasing to about 95% at the end of 50 rounds. Confidence intervals are shown at rounds 10, 20, 30, 40, and 50 to illustrate that on average there is a strong trend of increasing confidence in the system's identification of the unknown lie sequence. The dotted line (green) represents the standard deviation of the trials further confirming that the average spread of the mean reduced over the trials. This confirms our hypothesis that the system's accuracy increases proportional to the number of rounds.

It is possible that the accuracy rates of the model are not indicative of its success in identifying the correct lie sequence but rather a reflection of a player that doesn't lie often. For example, if the lie sequence was $\{1, 2, 3\}$ and the system identified it as $\{2, 2\}$, then player moves such as $\{\{1, 2, 1\}, \{2, 4, 5\}, \{3, 4, 5\}, \ldots\}$ would be deemed

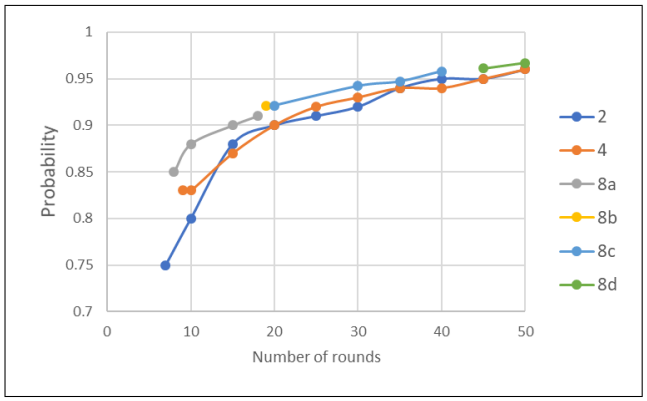


Fig. 3: Effect of sequence length on number of rounds required to correctly identify the lie sequence

TABLE II: Detection of a changing lie sequence

		Trial Number				
		1	2	3	4	5
Sequence Length	2	6/- (0.88)	5/- (0.76)	5/33 (0.7)	13/41 (0.76)	7/- (0.66)
	4	7/- (0.76)	5/- (0.72)	45/- (0.68)	10/- (0.66)	6/- (0.68)

correctly as truths improving the model accuracy but masking the failure of the system. To investigate if this was an issue, the average number of moves to correctly identify the player's lie sequence was analyzed along with the accuracy at each step. A gesture length of 10 was used along with a lie sequence length of 2,4 and 8. Once again each trial consisted of 50 rounds.

From the data shown in Figure 3, we observe that for a lie sequence of length 2, the system took 7 rounds to identify the correct sequence with a final certainty of 96%. Similarly, we see that when the length of the sequence was 4, it took 9 rounds to identify the sequence with a final accuracy of 95%. The case of a sequence length of 8 illustrates the problem discussed. The data is split into lines 8a, 8b, 8c, and 8d to denote where changes in the system's prediction of the lie sequence occurred. In this case the lie sequence chosen was $\{1, 4, 3, 2, 2, 4, 5, 1\}$. The system correctly guessed the chosen lie sequence after 8 rounds with a probability of 0.85 (Line 8a). By round 18 the probability rose to 0.91. But then the system erroneously switched its guess to the sequence $\{1, 4, 3\}$ at step 19 (Line 8b) and then sequence $\{1, 4, 3, 2, 2\}$ (Line 8c). At step 45, it then switched to sequence $\{2, 4, 5, 1\}$ achieving an overall accuracy of 0.97 at round 50. Strictly speaking this could be considered a failure of the system, but it should also be noted that all of the sequences considered were non-trivial subsequences of the actual lie sequence. If these erroneous sequences were encountered and reinforced as being incorrect guesses, we would expect the system to return to the originally correct guess. With more data this is the likely outcome.

Finally we examined the ability of the model to respond to a changing lie sequence. We hypothesized that it would take longer to determine the second sequence as the model would have to 'forget' its initial guess. This would mean that wrong predictions have to be made before the model's most probable lie sequence is deemed to be erroneous and another sequence can replace it. Five games of 50 rounds each were run. In each of the 5 games, after the 25^{th} round the lie sequence was changed. Table II depicts the results. Each entry indicates the round that the system identified the lie sequences and the overall accuracy for that trial. A dash (-) indicates that the sequence was never correctly identified. The accuracy is shown in parentheses. For example, the entry 5/33 (0.7) indicates that the first lie sequence was identified in the 5^{rd} round, the second lie sequence was identified in the 33^{th} round, and that the overall accuracy for that trial was 70%.

From the data we see that while the system can determine the first sequence correctly in under 15 rounds in 9 out of 10 trials, when the lie sequence changes the model is rarely able to deduce the new sequence within the 50 round limit. With additional observations we believe that the new lie sequence would be found because the model's accuracy increases with data. It should also be noted that for some of the trials and sequence lengths even though the system does not detect the lie sequence its accuracy is nevertheless 0.75 percent or greater. As alluded to previously, this result arises from either the model correctly determining a subset of the lie sequence or because of the random nature of the moves made by the player.

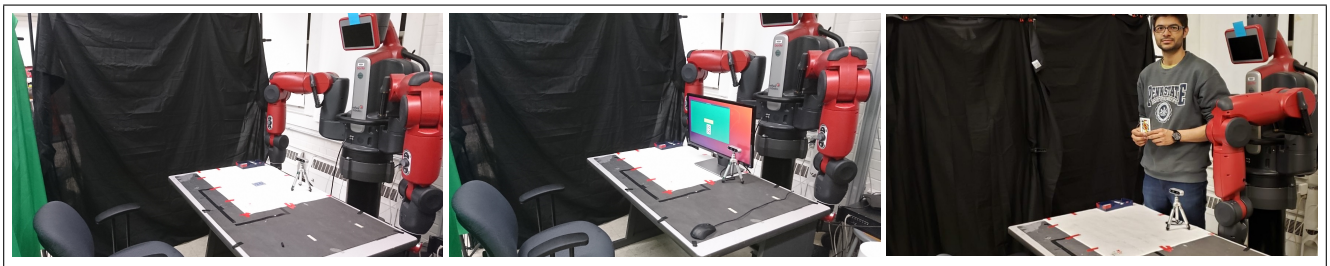
VII. Experiment Two: Testing the Model on Human Data

The previous experiment tested aspects of the model under the simplifying assumption that overt, easily detectable cues were available. This second experiment attempts to identify whether or not such cues exist while the person is discarding (as described in Section III) and if any identifiable cues can be fed into the model described in Section IV to predict truth-telling versus lying.

All subjects played against a computer, a human, and a robot opponent. The order of the type of opponent was counterbalanced. In all three cases, the amount of time taken to play each round, the phrasing of the opponent's questions, and the movement speeds were matched as closely as possible in order to maximize the similarity in play styles of the three types of opponents. Subjects were instructed that they would be playing against an opponent that would analyze visual data collected from a stereo camera to detect the trustworthiness of their move, but in reality all three opponents played randomly. This was necessary so that the capability of the opponent did not affect the subject's game play. Human subjects played as many rounds as possible within 7 minutes against each type of opponent. The average number of rounds played was found to be 14.5 against the computer, 20.2 against the human, and 11.9 against the robot. In total 466 rounds were played by the subjects which was then used to quantify system performance.

For this experiment the subject and opponent sat across a table from one another as shown in Fig. 4. An Intel D435 stereo camera was placed across from the subject in order to capture most of the upper body of a person seated in the chair. A sense of privacy was provided using black cloth screens to isolate the subject. The screen behind the participant was green to enhance contrast to aid in image post processing. A microphone was placed on the table to capture the subject's voice commands.

A Baxter robot was used for the robot portion of the experiment. The Baxter robot (Fig. 4 (a)) is a torso with two 7 DOF arms. A suction gripper end effector was used to perform card manipulations such as picking up a card and discarding. When playing against the robot, the subject's were asked to discard their card in a specified area of the table that was within reach of the robot's arm. The human player's discard was detected using a camera in the robot's



(a) Robot Opponent (b) Computer Opponent (c) Human Opponent

Fig. 4: The experiment was designed to maintain uniformity across all three conditions to reduce noise. In all cases an Intel D435 stereo camera was used to capture subject video and was placed in the same location along with a common 3.5mm computer microphone to capture subject suit declarations.

head. Next the robot would ask the player, "What suit did you play?" The subject's reply was captured using the microphone. Google cloud speech recognition was used to detect the suit type verbally stated by the person. The robot then picked up the card using its suction gripper and moved it to a pre-specified drop location at the rear right corner of the table. It would then either drop the card face down and say "I think you are telling the truth" or flip the card over and say "I think you are lying." It would then return its arm to the initial position and start the next round by saying, "It is your turn to play."

Play against the computer (Fig. 4 (b)) followed a similar procedure and used the exact same voice messages. Against the computer the subject was shown a virtual card on the monitor. The participant then used a mouse to click on a button titled 'play'. Next the computer would then ask them what suit was played. Speech recognition identified their stated suit and the card would slowly spin on the screen for a time period similar to the time taken for the robot to move the card to its drop off location. Finally, the computer's decision was revealed with the same statements ("I think you are telling the truth" or "I think you are lying").

Against the human opponent (Fig. 4 (c)), a similar procedure was followed with the researcher repeating the same phrases. For this condition, however, the human opponent stood in front of the Baxter robot so that the environment remained as similar as possible.

Ten participants were recruited with an equal number of male and female participants. Internal Review Board approval was obtained. This research was primarily exploratory in nature in that our goal was to determine if people do generate recognizable cues while playing this game. We believed that the participants would exhibit a range of cues related to the trustworthiness of their actions in the game. We further believed that these cues would be individualistic and that any patterns, if found, would remain consistent throughout a particular experimental condition. We do not have enough evidence to say for certain if the patterns remain consistent across different opponents such as the robot, another human, or an online version of the game.

Twelve types of cues were chosen based on a quick analysis of the video. The classes were: no gesture, eyebrow motion, motion of the lips, gaze to the left, gaze to the right, gaze ahead/up/down, smiling, laughing, arbitrary facial expression, arbitrary body motion, head to the left, and head to the right. The cue detection system as presented in Section V was limited to five cues and while able to detect each cue when presented individually, was unable to accurately demarcate multiple cues as would be observed in any human video. The problem of autonomously identifying cues/gestures from a video where the human might perform sequential and simultaneous cues is out of the scope of this paper. We validated our model by using two independent coders to hand annotate all the videos.

Since two different coders were used and they could detect an arbitrary number of cues, their codes could have unequal lengths making it unsuitable to apply Cohen's kappa or Krippendroff's alpha as an inter-rater reliability metric. If coder 1 detected a cue sequence of length L_1 and coder 2 detected a sequence of legnth L_2, the inter-rater reliability, R, was calculated as

$$R = max(L_1, L_2) - D/max(L_1, L_2)$$
$$where\ D = Levenshtein\ Distance \quad (5)$$

The inter rater reliability was calculated to be 0.41 meaning that on average at least 41% of the codes had an overlap. We realize that even for human coders, recognizing cues is challenging and subjective. If we employ a sliding window to equalize the length of the codes, that is we only consider the most similar code of length $min(L_1, L_2)$ from both codes, we calculate the inter-reliability to be 67%. This tells us that on average even if one of the coders annotates more cues than the other, at least 67% of the sequence is in agreement.

The coders were also asked to judge if the subject in the video was bluffing or not to quantify how accurately they were able to predict lying versus truthful play as a control for comparision. Once again with two coders there is a variation in performance. The coder accuracy reported in Table III is the mean of both these values for each participant (columns) and opponent type (rows). Over all the cases, the average difference in prediction accuracy between the two coders was 0.13 (std-dev = 0.07) with a minimum and maximum difference of 0 and 0.33 respectively.

[b]

TABLE III: Prediction accuracy w.r.t. computer (S), human (H), and robotic (B) opponent. The model accuracy is reported first followed by the coder accuracy. The average represents performance for that particular subject over all opponent types.

	Participant Number									
	1	2	3	4	5	6	7	8	9	10
S	0.57 | 0.56	0.56 | 0.44	0.46 | 0.40	0.36 | 0.58	0.35 | 0.71	0.59 | 0.50	0.55 | 0.62	0.30 | 0.67	0.61 | 0.39	0.80 | 0.54
H	0.59 | 0.72	0.36 | 0.45	0.64 | 0.38	0.55 | 0.55	0.45 | 0.55	0.57 | 0.49	0.43 | 0.62	0.43 | 0.60	0.55 | 0.50	0.50 | 0.53
B	0.54 | 0.54	0.84 | 0.59	0.66 | 0.53	0.53 | 0.60	0.63 | 0.38	0.77 | 0.46	0.59 | 0.50	0.41 | 0.60	0.65 | 0.65	0.24 | 0.46
Avg	0.57 | 0.61	0.59 | 0.49	0.59 | 0.43	0.48 | 0.58	0.48 | 0.54	0.64 | 0.48	0.52 | 0.58	0.38 | 0.62	0.60 | 0.51	0.51 | 0.51

In order to give the coders the same information as the model, the coders were told the ground truth after each video so that they could use this information to judge future rounds. A qualitative analysis of the videos shows that participant's are much more likely to be animated and express themselves when playing against the physical robot. When playing against a human, participants seemed hesitant to make prolonged eye contact. On the other hand when playing against a computer, participants are extremely somber making it hard to judge their intentions.

Once coded, the sequence was fed into the computational model described in Section IV. Table III depicts the model's accuracy for each participant and each type of opponent. Over all opponent types and subjects, the computational model has an average prediction accuracy of 0.53 which is equivalent to the human average of 0.53. For each opponent type the accuracy of the model compared to the coders is 0.51 vs. 0.54 (computer), 0.51 vs. 0.54 (human), and 0.58 vs 0.53 (robot). Most importantly, the model outperforms or matches coder accuracy on five of the ten participants.

We suspect that the number of cues the participant displays may be the source of this difference. The more cues the better the model performances when compared to the human coders. This may reflect the model's ability to record and use a large number of cues whereas human's memory is limited. On the other hand, in participants that are less animated, the human coder performs better as they are able to draw on prior knowledge, stereotypes, and reasoning about patterns of deception observed in the previous rounds whereas the model purely relies on the observed surface cues.

VIII. Conclusions

A model for evaluating an individual's trustworthiness based on surface cues is presented. This paper develops the model and demonstrates the use of the model in a trust-deception card game against a robot. We have shown that as the model gains data its predictions improve. We have also presented preliminary experiments comparing the model's performance to a human's showing that the model is slightly better.

Still, the model does make assumptions and has limitations that will need to be addressed in future work. We have shown that the model's performance drops if the sequence of cues indicating a lie changes. Yet the ability of the model to classify a player's move correctly suggests that the sequence of cues displayed by the subjects during the bluffing play has a high degree of repeatability and at the very least, is stable over a short duration of time. The results from our second experiment are based on a small population of subjects. Future work will need to include more subjects. Finally, our current system relies on human coders to generate the sequence of cues from video data. Clearly a traditional machine classifier could be used to automate this process and, if accurate, result in a system pipeline in which cues are automatically recognized used to influence the robot's decision-making.

Our hope is that the model can one day be used to make predictions about whether or not a robot should trust the behavior of a human based on prior experiences and recognized surface cues. Accurate cue/gesture recognition is essential. Recognition of a greater variety of simultaneously occurring cues could improve the value and performance of our model. By linking cues to a person's underlying state in the proposed manner, we seek to develop a method that can determine one's motivation. The results of this work may be broadly applicable to situations in which a person intentionally or unintentionally disguises their motives. In health care, for example, it may be critical for a robot to recognize when a person's cues indicate an internal state that differs from what they are saying in order manage how to best help them.

ACKNOWLEDGMENT

This work was supported by Air Force Office of Sponsored Research contract FA9550-17-1-0017.

References

[1] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011.

[2] A. R. Wagner, "Developing robots that recognize when they are being trusted," *AAAI Spring Symposium*, pp. 84–89, 2013.

[3] T. R. Campellone and A. M. Kring, "Who Do You Trust? The Impact of Facial Emotion and Behaviour on Decision Making," *Cognition and Emotion*, vol. 27, no. 4, pp. 603–620, 2013.

[4] D. Li, P. L. P. Rau, and Y. Li, "A Cross-cultural Study: Effect of Robot Appearance and Task," *International Journal of Social Robotics*, vol. 2, no. 2, pp. 175–186, June 2010. [Online]. Available: https://doi.org/10.1007/s12369-010-0056-9

[5] P. L. P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior*, vol. 25, no. 2, pp. 587–595, Mar. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563208002367

[6] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *Proceedings of the 12th IEEE international workshop on robot and human interactive communication*. IEEE Press Piscataway, NJ, 2003, pp. 55–60.

[7] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Cultural differences in attitudes towards robots," in *Robot companions : hard problems and open challenges in robot-human interaction : AISB'05 convention, 12-15 April 2005, Hatfield, UK*. Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), 2005, pp. 1–4. [Online]. Available: https://research.tue.nl/en/publications/cultural-differences-in-attitudes-towards-robots

[8] S. You and L. P. Robert Jr., "Human-Robot Similarity and Willingness to Work with a Robotic Co-worker," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 251–260. [Online]. Available: http://doi.acm.org/10.1145/3171221.3171281

[9] M. Lohani, C. Stokes, M. McCoy, C. A. Bailey, and S. E. Rivers, "Social interaction moderates human-robot trust-reliance relationship and improves stress coping," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2016, pp. 471–472.

[10] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2013, pp. 251–258.

[11] D. Atkinson, P. Friedland, and J. Lyons, "Human-machine trust for robust autonomous systems," in *Proc. of the 4th IEEE Workshop on Human-Agent-Robot Teamwork. In conjunction with the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012) Boston, USA*, 2012.

[12] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of management review*, vol. 23, no. 3, pp. 393–404, 1998.

[13] J. A. Simpson, "Psychological foundations of trust," *Current directions in psychological science*, vol. 16, no. 5, pp. 264–268, 2007.

[14] P. Robinette, A. R. Wagner, and A. M. Howard, "Building and maintaining trust between humans and guidance robots in an emergency." Georgia Institute of Technology, 2013.

[15] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing Is Key for Robot Trust Repair," in *Social Robotics*, ser. Lecture Notes in Computer Science, A. Tapus, E. Andr, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Springer International Publishing, 2015, pp. 574–583.

[16] J. Rilling, D. Gutman, T. Zeh, G. Pagnoni, G. Berns, and C. Kilts, "A neural basis for social cooperation," *Neuron*, vol. 35, no. 2, pp. 395–405, July 2002.

[17] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz, and P. R. Montague, "Getting to know you: reputation and trust in a two-person economic exchange," *Science (New York, N.Y.)*, vol. 308, no. 5718, pp. 78–83, Apr. 2005.

[18] D. Gambetta, "Can We Trust Trust," in *Trust: Making and Breaking Cooperative Relations,*, electronic edition ed. Oxford, UK: University of Oxford, 2000, vol. 13, pp. 213–237. [Online]. Available: ¡http://www.sociology.ox.ac.uk/papers/ gambetta213-237.pdf¿

[19] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM.

[20] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with Trust for Human-Robot Collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 307–315. [Online]. Available: http://doi.acm.org/10.1145/3171221.3171264

[21] J. R. Dunn and M. E. Schweitzer, "Feeling and believing: the influence of emotion on trust," *Journal of Personality and Social Psychology*, vol. 88, no. 5, pp. 736–748, May 2005.

[22] M. Freitag and P. C. Bauer, "Personality traits and the propensity to trust friends and strangers," *The Social Science Journal*, vol. 53, no. 4, pp. 467–476, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0362331915001123

[23] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2016, pp. 101–108.

[24] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[25] H. Aarts, T. L. Chartrand, R. Custers, U. Danner, G. Dik, V. E. Jefferis, and C. M. Cheng, "Social stereotypes and automatic goal pursuit," *Social Cognition*, vol. 23, no. 6, pp. 465–490, 2005.

[26] N. O. Rule and N. Ambady, "The face of success: Inferences from chief executive officers' appearance predict company profits," *Psychological science*, vol. 19, no. 2, pp. 109–111, 2008.

[27] F. Ma, F. Xu, and X. Luo, "Children's and adults' judgments of facial trustworthiness: the relationship to facial attractiveness," *Perceptual and Motor Skills*, vol. 121, no. 1, pp. 179–198, 2015.

[28] C. Sofer, R. Dotsch, D. H. Wigboldus, and A. Todorov, "What is typical is good: The influence of face typicality on perceived trustworthiness," *Psychological Science*, vol. 26, no. 1, pp. 39–47, 2015.

[29] L. A. Zebrowitz and J. M. Montepare, "Social psychological face perception: Why appearance matters," *Social and personality psychology compass*, vol. 2, no. 3, pp. 1497–1517, 2008.

[30] G. Dik and H. Aarts, "Behavioral cues to others' motivation and goal pursuits: The perception of effort facilitates goal inference and contagion," *Journal of Experimental Social Psychology*, vol. 43, no. 5, pp. 727–737, 2007.

[31] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, "Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 493–500.

[32] S.-Y. Lee, "Impact of Human like Cues on Human Trust in Machines: Brain Imaging and Modeling Studies for Human-Machine Interactions," Korea Advanced Institute of Science and Technology Taejon Korea, South, Tech. Rep., 2018.

[33] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams," in *International Conference on Persuasive Technology*. Springer, 2018, pp. 56–69.

[34] D. T. Neal, W. Wood, J. S. Labrecque, and P. Lally, "How do habits guide behavior? perceived and actual triggers of habits in daily life," *Journal of Experimental Social Psychology*, vol. 48, no. 2, pp. 492–498, 2012.

[35] M. Caro, *Caro's book of poker tells*. Cardoza Publishing, 2003.

[36] P. Ekman, "Lie catching and microexpressions," *The philosophy of deception*, pp. 118–133, 2009.

[37] D. Atkinson, P. Friedland, and J. B Lyons, "Human-Machine Trust for Robust Autonomous Systems," Mar. 2012.

[38] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[39] B. MacCartney, "Nlp lunch tutorial: Smoothing, 2005," *URL http://nlp. stanford. edu/~ wcmac/papers/20050421-smoothing-tutorial. pdf*.

[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields," *CoRR*, vol. abs/1611.08050, 2016. [Online]. Available: http://arxiv.org/abs/1611.08050

[41] T. Simon, H. Joo, I. A. Matthews, and Y. Sheikh, "Hand Keypoint Detection in Single Images using Multiview Bootstrapping," *CoRR*, vol. abs/1704.07809, 2017. [Online]. Available: http://arxiv.org/abs/1704.07809

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," *CoRR*, vol. abs/1602.00134, 2016. [Online]. Available: http://arxiv.org/abs/1602.00134

[43] A. R. Wagner and R. C. Arkin, "Recognizing situations that demand trust," in *RO-MAN, 2011 IEEE*. IEEE, 2011, pp. 7–14.