Human vs Robot Lie Detector: Better working as a team?

Dario Pasquali[*,1,2,3], Davide Gaggero[1], Gualtiero Volpe[1], Francesco Rea[2], Alessandra Sciutti[4]

[1] *Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS), Università di Genova, Opera Pia 13, 16145, Genova, Italy*

[2] *Robotics Brain and Cognitive Sciences (RBCS), Istituto Italiano di Tecnologia, Enrico Melen 83, Bldg B, 16152, Genova, Italy*

[3] *Information and Communication Technologies Directorate (ICT), Istituto Italiano di Tecnologia, Enrico Melen 83, Bldg B, 16152, Genova, Italy*

[4] *COgNiTive Architecture for Collaborative Technologies (CONTACT), Istituto Italiano di Tecnologia, Enrico Melen 83, Bldg B, 16152, Genova, Italy*

ABSTRACT

Human interaction often entails lies. Understanding when a partner is being deceitful is an important social skill, that also robot will need to properly navigate social exchanges. In this work, we investigate how good are human observers at detecting false claims and which features they base their judgment on. Moreover, we compare their performance with that of an algorithm for lie detection developed for the robot iCub and based uniquely on pupillometry. We ran an online survey asking participants to classify as truthful or deceptive 20 videos of individuals describing, either correctly or untruly, complex drawings to iCub. They also had to rate their confidence and provide a written motivation for each classification. Responders achieved an average accuracy of 53.9% with a higher score on detecting lies (55.4%) with respect to true statements (52.8%). Also, they performed better and more confidently on the videos iCub failed to classify than on the ones iCub correctly detected. Interestingly, the human observers listed a wide range of behavioral features as means to decide whether a speaker was lying, while the robot's judgment was driven by pupil size only. This suggests that an avenue for improving lie detection could be a joint effort between humans and robots, where human sensitivity to subtle behavioral cues could complement the quantitative assessment of physiological signals feasible to the robot. Finally, based on the reported motivations, we speculate and give hints on how the lie detection fields should evolve in the future, aiming to portability to real-world interactions.

1 Introduction

Lying is a consistent part of human's social interactions [1], [2], learned since younger age [3], [4]. Feldman et al. found that on a population of students, 60% of the participants lied at least once in a 10-minutes conversation [5] while, in general, people lies at least two times each day [6]. Other than for deceptive and malicious activities, everyone exploits a large amount of "white lies" both to help others and to help ourselves. For instance, we lie to present ourselves better than we are [5], to avoid awkward conversations [1], or to persuade others [7].

Robots will be soon part of our everyday life. Like humans, they will need to be able to detect deception during common human-robot interactions, for instance, to assess partners' trustworthiness [8], to present more efficient support to humans (i.e., in teaching or caregiving) and to maintain a solid social interaction with other individuals in the society. Multiple technical solutions have been developed to detect lies. Traditional methods of lie detection rely on monitoring physiological metrics related to cognitive load and stress, such as

skin conductance, respiration rate, and heartbeat of blood pressure. The polygraph achieves an accuracy between 81% and 91%, making it one of the most used lie detectors [9]. However, literature proves it is possible to bypass its measure [10]. Other state-of-the-art methods rely on fMRI images [11], skin temperature variations [12], micro-expressions [13], photoplethysmography [14], or acoustic prosody [15]. However, most of those methods are either expensive, depend on invasive or cumbersome devices, or require the presence of experts, which limits their portability on robotic platforms and real-life human-robot interactions.

In previous works, we enabled the humanoid robot iCub to detect lies in real-time during an informal and entertaining card game (the Magic Trick, [8], [16]): we asked participants to describe to iCub a set of cards characterized by complex drawings, lying about a few of them while wearing a Tobii Pro Glasses 2 eyetracker; iCub used the pupillometry features collected in real-time from the eyetracker to classify players' lies with an accuracy of 88.2%. To do so, we exploited a well-known effect: lying requires a cognitive effort due to the fabrication and maintenance of a consistent deception [6], [10], and this reflects on measurable Task Evoked Pupillary Responses, like mean pupil dilation and latency to peak [17], which can be used to detect lies [18].

Humans however cannot have access to precise information about the pupillometry of the partner, but still can sometimes detect lies. On average human performance in lie, detection is 54% [19], with an accuracy of 47% on detecting false statements and of 61% on detecting true ones. With training, experts, such as law enforcement or secret service officers could reach and accuracy of 65%; however, they report their detection is based more on a gut feeling and past experiences. Indeed, one of the main reasons detecting lies is a hard problem is the absence of a finite and objective set of behavioral cues that can be directly related to deception [20]. As reported by De Paulo et al. [6] and Vrij et al. [21] what usually happens is a combination of multimodal and context-based cues related to the control of body reactions or to hiding an internal feeling. Some of those cues are the increase of body movements, impossibility to stay still, speech hesitation, complexity of the speech, mutual gaze avoidance, hand movements, the covering of face and mouth, and increased number of stopwords. However, recent research started questioning the reliability of behavioral cues to detect deception [21], [22]. For a robot, it could be relevant to understand which features enable human observers to tell a partner is lying. Such intuition, paired with the technical solutions potentially portable on robots, could help them to better understand human partners' behaviors.

In this paper, we propose an online study meant to evaluate how humans perform at detecting lies in the same game scenario on which we developed our above-mentioned solution [16]. We asked participants to take the role of iCub in the Magic Trick card game, classifying 20 videos as truthful or deceptive. A similar lie catcher study has been done recently in [15], [23] even if the focus there was on acoustic and prosodic features We compare participants' performances with those of the purely pupillometry-based method we endowed iCub with and we analyze which other features participants based their judgments on. Results provide useful hints on how improving our system and how the lie detection field in human-robot interaction should evolve in the future.

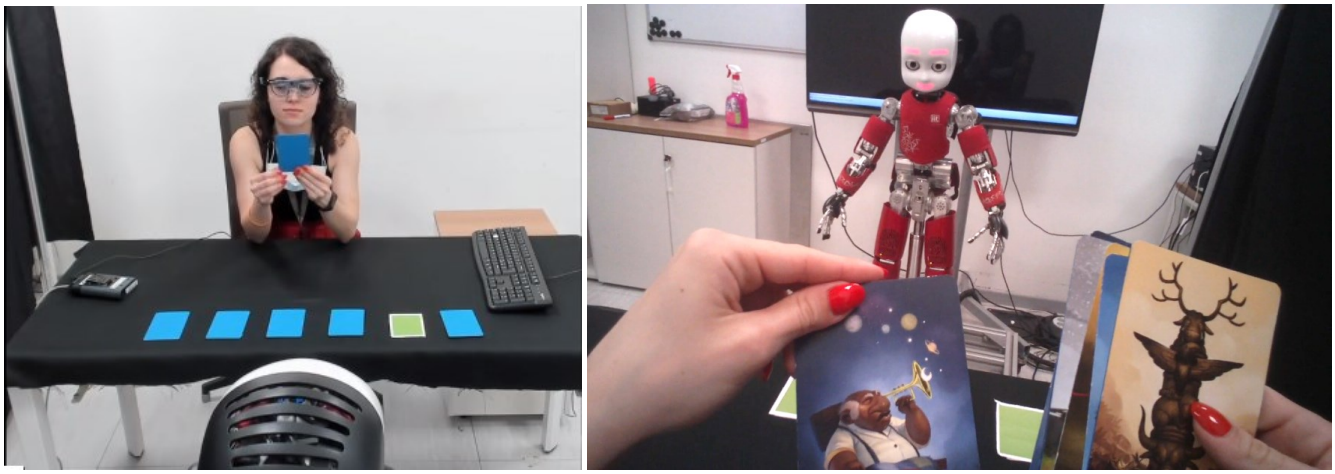

Figure 1: (Left) Participant describing a card to iCub, while wearing the Tobii Pro Glasses 2 eyetracker (Logitech Brio 4k webcam point of view); (Right) Point of view of the participant during the interaction collected through the Tobii glasses with an example of the described cards.

2 Methods

For the online survey, we used the videos collected during the Magic Trick card game presented in [8], [16].

2.1 The Magic Trick Card Game

The Magic Trick is a game-like human-robot interaction where players describe 6 gaming cards from the Dixit card game [24] to the humanoid robot iCub (see Fig. 1). Players were asked to describe some of the cards creatively and deceitfully, while describing the others truthfully. No limitation was provided on the number of cards described falsely neither on the length of the descriptions. After each card, iCub tried to classify its description as true or false. iCub's classifications were based on the real-time reading of players' pupil dilation via the Tobii Pro Glasses 2 eyetracker they wore (see Fig 1, left). During a previous interaction [16], iCub learned on a similar task how players' pupils dilate in response to a lie. Then, it exploited this information to classify each card description: pupils are known to dilate in response to an increase of cognitive load, like the one generated by the fabrication of a false description; in the first interaction iCub learns the mean pupil dilation of players for truthful and deceptive descriptions, then those values are compared with the mean pupil dilation of new card descriptions; the closer score is the assigned class. N=34 participants played the Magic Trick Card Game and iCub could correctly classify players' descriptions with accuracy = 70.8%, precision = 73.6%, recall = 57%, and F1 score = 64.2% (N=34). For a deeper analysis of both interactions see [16] and [8].

2.2 Materials

A Logitech Brio 4k webcam, fixed on a television behind iCub, recorded the interaction from iCub's point of view at a resolution of 1080p (Fig. 1, left). We segmented in 6 card descriptions the videos of the 34 participants who took part in the experiment, resulting in 204 videos. From these videos we discarded: (i) the players who did not give the consent to share the videos recorded during the experiment (N=3); (ii) the players who wore a surgical mask or other accessories which prevent a complete vision of players' face (N=4); (iii) the videos affected by recording technical issues (N=7). Then, we picked a balanced set of 20 videos following a 2 x 2 set of conditions: (i) *Card Label*: 10 videos present a truthful description (*True* videos) and 10 a deceitful description (*False* videos); (ii) *Difficulty*: among each sub-group, 5 videos have been successfully classified by iCub during the game (*robot-easy* videos) while for the other 5 iCub's classification failed (*robot-difficult* videos). Moreover, we ensured each video involved a different actor and a different card, even if described falsely. The resulting set of videos lasted on average 27 seconds (SD=15 seconds). We uploaded the 20 selected videos on Vimeo [25], and linked them on SurveyMonkey [26], the platform used to administrate the online survey.

2.3 Procedure

We designed the online survey as a game in which responders compete on detecting the highest number of deceptive card descriptions. Before starting the survey, responders were asked to accept an informed consent, they had to select a nickname for anonymization purposes and were asked to wear headphones and carefully listen. The survey consisted of three phases:

2.2.1 Pre-questionnaire

Responders answered questions about their sex and age and filled in the Italian version of the Ten-Items Personality Inventory (TIPI) (extroversion, agreeableness, conscientiousness, emotional stability, openness to experiences) [27]. Then, they were informed they were going to see 20 videos of players describing gaming cards in front of iCub and that they had to judge each description as real or deceptive. After that, they saw an example of a video in which the falsely described card was presented in the top right corner.

2.2.2 Lie Detection Survey

After that, responders saw the 20 videos of card descriptions selected from the original Magic Trick card game. For each video, responders had to answer three questions: (i) whether the person in the video was lying or not (Yes or No answer); (ii) their confidence in this answer (slider from 0 to 100) and (iii) the reason why they provided such judgement. Responders could see the videos any time they wanted, but they could not go back after providing a judgment for a video. SurveyMonkey platform shuffled the videos for each responder to compensate for any order effect.

2.2.3 Post-questionnaire

Responders were presented with a list of common deceptive behaviors extracted from the literature [6]: uncertainty, an increasing number of stopwords, delay in providing an answer, repetitions and autocorrection, complexity of the answer, negativity, voice tone, eyebrows movements, touching the face, covering the mouth, avoiding mutual gaze, head wandering, fast body movements/breathing, eyes wide-opened, and fake smile. Responders had to rate on a 7-points Likert scale how much they relied on each of them. Finally, responders could report any other method or cue they used in the survey.

2.4 Participants

163 responders (82 males, 78 females, 3 preferred to not answer), with an average age of 40 years (SD=16) took part in the online survey. Responders were recruited among authors' colleagues and friends through word-to-mouth sharing, and they received no monetary compensation. They all accepted an informed consent form approved by the ethical committee of the Regione Liguria (Italy). They all agreed on using their data for scientific purposes. Among the 163 responders, only 117 completed the survey entirely. They were 54 males and 63 females (1 preferred to not answer) with an average age of 39 years (SD=14).

3 Results

Considering both truthful and deceptive descriptions, responders correctly guessed them with an accuracy score of 53.9% (SD=10.7%). Interestingly, nobody correctly guessed all the card descriptions, but the best performer reached an accuracy of 95%, missing the classification of a single video. Regarding confidence, responders reported an average confidence of 67.1% (SD=13.8%). A Shapiro-Wilk normality test showed that the confidence score is normally distributed, whilst the accuracy score is not. Therefore, in the following, a non-parametric analysis was conducted on the accuracy score and a parametric one on the confidence score.

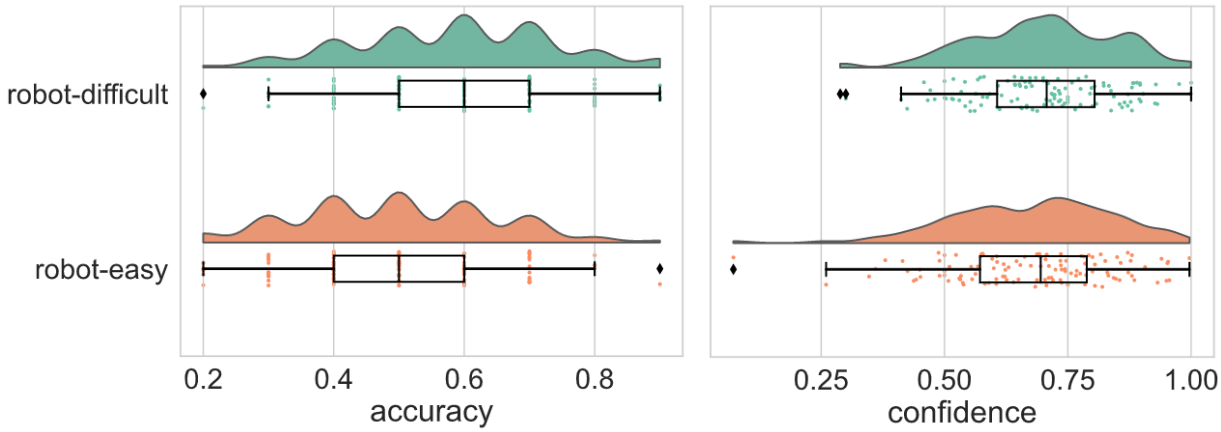

Figure 2: Average accuracy (Left) and confidence score (Right) for *robot-easy* and *robot-difficult* card descriptions.

3.1 Comparison of the conditions

Assuming detecting deception is a tougher task, we compared the accuracy score and the confidence of responders among truthful and deceptive card descriptions. Responders classified truthful descriptions with an accuracy score of M=52.8% (SD=16.1%) and deceptive descriptions with an accuracy of M=55.4% (SD=13.7%). Even if the score for false card descriptions is higher, a Wilcoxon signed-rank test did not reveal a significant difference ($W(115)=1940$, $p=0.343$). Also, the reported confidence between truthful and deceptive descriptions is not statistically different ($t(115)=-1.59$, $p=0.115$) with an average confidence of M=68.1% (SD=16.6%) for truthful descriptions against an average confidence of M=69.7% (SD=13.8%) for deceptive ones.

More interesting is the comparison between *robot-easy* and *robot-difficult* card descriptions. As a remark, this concept is defined from iCub's perspective: we selected the *robot-easy* descriptions among the ones iCub correctly classified, while the *robot-difficult* ones where

chosen among the ones for which iCub failed the classification. Interestingly, responders achieved a statistically higher score on *robot-difficult* card descriptions (M=58.4%, SD=15.1%) with respect to the *robot-easy* ones (M=49.6%, SD=14.9%), as proved by a Wilcoxon signed-rank test (W(115)=3373, p<0.001) (see Fig. 2, Left). Moreover, the reported confidence also follows a similar pattern, with statistically higher confidence for *robot-difficult* card descriptions (M=70.2%, SD=14.5) with respect to *robot-easy* ones (M=67.6%, SD=16.3). We confirmed it with a paired t-test (t(115)=2.42, p=0.017) (see Fig. 2, Right).

Also, we compared the accuracy score and the confidence score within each condition. A Wilcoxon signed-rank test showed a statistically higher score for false-*robot-difficult* descriptions with respect to false-*robot-easy* ones (W(103)=952, p<0.001). Conversely, there is no significant statistical difference among true-*robot-difficult* and true-*robot-easy* card descriptions (W(103)=1103, p=0.117). Regarding the reported confidence, responders were more confident for true-*robot-difficult* descriptions with respect to true-*robot-easy* ones with a statistically significant difference (t(103)=3.485, p<0.001); however, we did not find any statistical difference among false-*robot-difficult* and false-*robot-easy* card descriptions (t(103)=0.553, p=0.581).

Finally, we explored the correlation between the average confidence and the average accuracy score for each video. We fit a linear regression model with the average accuracy score as the dependent variable and the average confidence as the independent variable. Results show that the average confidence score inversely correlates with the average accuracy score (t(19)=-0.084, p=0.024, Adj R^2=0.21). We also tested whether the videos' duration correlated with their accuracy score or the average confidence, but we did not find any significant result.

3.3 Responders vs iCub performances

As specified in section 2.2, the videos were selected to be half among the ones that the iCub correctly classified during the game (*robot-easy*) and a half among the ones the robot misclassified (*robot-difficult*). However, after the full data collection of the game was completed, we could post-hoc train an algorithm, based on a larger dataset, which led to improved performances in iCub's lie detection. To better compare iCub's and responders' performance we assessed what performance would iCub have had, based on the new algorithm. . We exploited the pupillometry data collected for the N=34 players of the original Magic Trick card game [8], [16]. During the game, iCub asked players to describe 6 cards with a pointing gesture. The player was instructed to take the card as iCub pointed it, describe it while keeping it in the hands, and finally place it back on a marker on the table. During the whole interaction, participants wore a Tobii Pro Glasses 2 eyetracker, recording their pupillometry at 100 Hz. Post-hoc, we cleaned and segmented the pupillometry data for each pointing and card description and applied a baseline correction, subtracting to each segment the average pupil dilation during the 5 seconds before the relative iCub's pointing. Then, for each card, we computed the mean, max, min and standard deviation of the pupil dilation, along with the duration during the pointing, the card description, and the whole interval. The result is a dataset of 15 features for 228 cards. We split this dataset considering the 20 card descriptions presented in the survey as test set and the remaining as training set. We then trained a random forest classifier with the best hyperparameters selected in [8]. If iCub had embedded the model during the Magic Trick card game, it would have correctly classified the 20 card descriptions with an accuracy, precision, recall and F1 score of 70%. We statistically compared this 70% accuracy score with respect to the 53.9% average accuracy of the responders; results show the accuracy score of the random forest is higher, however, this difference is not statistically significant (z=1.43, p=0.07). Also, we tested the new model on *robot*-easy and *robot*-difficult card descriptions: results show it can classify *robot-easy* card descriptions with an accuracy of 90%, a performance consistent with the in-game results and statistically higher than humans' performance (49.6%) on those videos (z=2.57, p=0.005). However, on *robot-difficult* videos it still performs worst than humans (50% for iCub against 58.4% for humans), even if the difference is not statistically significant (z=0.55, p=0.29).

3.2 Pre-questionnaire analysis

We then explored whether responders' personality traits influenced their performance or confidence in the online survey. From the Ten-Items Personality Inventory (TIPI), filled in before the survey, participants average scores were: Agreeableness: M=5.11, SD=1.08; Conscientiousness: M=5.12, SD=1.59; Emotional Stability: M=4.52, SD=1.39; Openness to experiences: M=4.66, SD=1.05 and Extraversion: M=4.0, SD=1.41. We fit two multiple linear regression models with the personality traits as independent variables and the average accuracy score or the average confidence for each responder as the dependent variable. Results show that only emotional stability correlates significantly with the average accuracy score (t=0.022, p=0.004, Adj R^2=0.046). Also, a comparison of the confidence and accuracy score among male and female responders showed no relevant results. Finally, we fit two linear regression models with responders' age as the independent variable and the confidence or accuracy score as the dependent variable but we did not find any significant effect.

3.4 Motivations and Post-Questionnaire analysis

Other than classifying each card description as truthful or deceptive, responders were asked to report the motivation which drove their decisions. We applied a stopword filter and a lemmatization to clean the reported motivations. From a qualitative analysis, responders focused more on how the actor described the card, reporting words like "precise", "details", "confident", "sincere", "thinking", "quick", "pauses", "short", "fluid", "time", "(un)decided". Also, responders reported elements related to what they were looking at with words like: "looking", "gaze", "voice", "hands", "touch", "smiling", "laughing", "face", "eye", "leg". Comparing the motivations of truthful and deceptive videos or *robot-easy* and r*obot-difficult* ones did not reveal any clear difference.

Also, we run a deeper analysis on the motivations reported by the responder which achieved an accuracy score of 95%. We did not assess the profession of the responder; hence we could not know if he is an expert or a professional on lie detection, still, he was the best on the task. Looking at his motivations we found he focused on three main features: (i) the fluidity of the communication (i.e., the complexity of the speech, the rephrasing, or the presence of "hmmm"s); (ii) the consistency between verbal communication and body movements (i.e., moving the body from right to left); (iii) the injection of emotional or personal thought on the card description. Interestingly, he used the presence of reflection pauses as a criterion to classify card descriptions as truthful – he reported it on 8 cards over 10. Lastly, he classified all the deceptive card descriptions as so, but he misclassified one of the true cards: he has been fooled by a leg movement, a potential sign of stress.

After the survey, we asked responders to rate on a 7-points Likert scale how much they relied on the state-of-the-art methods used to detect a liar; also, we asked them to report any other method they rely on. The complexity of the description (M=4.89, SD=1.62), presence of stopwords (M=4.68, SD=1.61), the uncertainty of the description (M=4.67, SD=1.69), fake smiling (M=4.65, SD=1.79), voice tone (M=4.54, SD=1.77), absence of mutual gaze (M=4.27, SD=2.01), fast movements and breathing (M=4.07, SD=1.83), touching nose or face (M=4.01, SD=2.02) were the most used ones. Then head movements (M=3.78, SD=1.84), repetitions and autocorrections (M=3.78, SD=1.84), description time (M=3.72, SD=1.05), eyebrow movements (M=3.4, SD=1.69), covering the mouth (M=3.28, SD=1.98), eye movements (M=3.14, SD=1.89), and negative words in the description (M=2.73, SD=1.55) follow. Finally, a few responders reported other features used to detect liars: 9 responders took into account the amount of body movement, the impossibility to stay still, or the position of leg and hands; also 8 responders focused more on the content of the descriptions rather than on the visual appearance like too creative descriptions, a high number of details or adjectives, or a feeling of premeditation of the description.

4 Discussion

In this study, we compared human and robot performances on detecting lies during an informal interaction and explored which behavioral cues are used with the purpose to improve our system. Being able to detect lies in a real-world informal scenario is a mandatory requirement to port lie detection methods out of laboratory scenarios. Even if state-of-the-art methods work on constrained and formal setups, they usually depend on cumbersome devices and lack the intuition and experience that makes humans able to detect liars. In this manuscript, we explored what robots should look at to overcome that limitation. To do so, we ran an online survey where responders had to classify a set of videos, recorded during an informal game-like human-robot interaction from iCub humanoid robot point of view, as truthful or deceptive. We also asked for each video the confidence on the classification and an open-ended motivation of what led the decision.

Responders achieved an accuracy score of 53.9% on classifying deceptive and truthful card descriptions, which is consistent with the average 54% from the literature [19]. Also, they outperformed iCub achieving better performance on *robot-difficult* than on *robot-easy* card descriptions. To run a fairer comparison between iCub and responders' performances, we trained a random forest classifier on the pupillometry data collected during the original Magic Trick card game. Testing the model on the 20 card descriptions of the survey (excluded from the training set) revealed an accuracy score of 70%, higher than the average score of humans (53.9%) even if not statistically higher. As a remark, each player of the magic trick described 6 cards to iCub, but we excluded from the training set only the card descriptions used in the survey, not the whole participants. Hence, the random forest classifier embeds a little information on the actors it classifies in the test set. We took this decision to replicate the population of actors and responders of the survey. Indeed, most of the actors and most of the responders were internal confederates and we cannot exclude they know each other; hence it is possible that a subset of the responders had some prior knowledge on how the actors lie or tell the truth, even if we cannot spot those connections due to the anonymization of the data.

Looking at the reported motivations for each video and at the end of the survey, we have an insight into what a social robot should look at to improve its lie detection abilities. Responders mainly pointed out two major aspects to take into account: (i) *how* the actor described the card (i.e., "quick", "(un)decided", "precise", "fluid"); and (ii) what to look at (i.e., "face", "gaze", "hand", "leg", "smile"). Those motivations

are supported and extended by the ratings at the end of the survey: responders focused mainly on (i) the content, fluidity, and complexity of the descriptions; and (ii) on the body movements of the actors. Interestingly, responders focused less on facial features postural features than what was expected from the literature. We speculate this depends on the setup in which the videos were acquired: participants wore a Tobii eyetracker which, partially cover their face, and sat behind a table covering their lower bodies. Also, actors mostly looked to the cards they were holding in their hands rather than looking to iCub. Still, motivations and final ratings suggest a combination of visual and prosodic features could be a good candidate to improve iCub's lie detection performances on real-life informal scenarios, as also supported by the literature [28]. Moreover, those features could be extracted from the devices (i.e., RGB cameras and stereo microphones) already equipped on the iCub humanoid robot. Overall, the reported motivations suggest that both the behaviors of the actors and the qualities of such behaviors (including expressive, emotional facets) have an important role in detecting lies so that the robot should also be endowed with techniques for detecting behavior and for analyzing their expressive qualities.

From our results, we could say that what is "difficult" for a robot that embeds a pupillometry-based technical solution is "easy" for humans that use behavioral cues and vice versa. This might happen because when lying some actors rely on special behaviors (e.g. pauses, body motions, slowing down.) to reduce the cognitive load, in turn minimizing the pupillary change associated with the latter. In these cases, a robot focusing on pupillometry alone could never realize that the partner is lying. Conversely, a keen human observer could notice these tell-tale signs, most probably missing instead the cases in which only the pupil variation reveals the deception. Hence, we speculate the cooperation of those two systems will be a key factor for future developments of lie detection in human-robot interaction. To improve, a robot should be able to "look at humans as other humans do" combining our fuzzy evaluation with the rigour of technical and physiological metrics.

In the future, it would be interesting to integrate our pupil-based approach with the processing of visual features (i.e. body posture, body movements, or facial expression) and audio features (i.e., word embedding or prosodic analysis of the descriptions). To validate such multimodal system on our setup, aiming to port it to a real-world scenario, it would be mandatory to overcome the limitation posed by the Tobii Pro Glasses 2 eyetracker since it partially occludes actors' faces, limiting the usage of visual features. Recent findings [29] suggest it will be soon possible to measure pupillometric features with common RGB cameras like the ones embedded on the iCub robotic platform. Finally, it would be necessary to push the research field to more ecological and real-life scenarios. Indeed, most of the state-of-the-art research focus on strict and interrogatory-like setups that for sure happen in the real world; however, they represent a strict subset of the variety of interaction that happens and in which both humans and robot could take advantage from detecting lies. For instance, a more portable lie detector system could help in airports or sensible buildings to prevent dangerous situations; while a social robot could use it to better understand humans, give reason to human behaviors, assess their trustworthiness, and provide better support in professions like teaching, caregiving, or law enforcing.

5 Conclusion

In this work, we assessed humans' performance on detecting lies in an informal scenario and compared them with iCub's performance. Responders had a similar performance as iCub but showed a significantly better performance in those videos which resulted more difficult for the robot, than in those I iCub classified correctly. Integrating iCub's pupillometry-based approach and humans' behavioral-cues-based approach could be the key solution to improve lie detection in human-robot interaction. Robots able to detect lies "from a human point of view" could better support humans in roles professions like teaching, caregiving or law enforcement, other than improve their ability to interact socially with human partners. In our view, these aspects will deserve further investigation e.g., in the framework of emerging research areas such as Human-Centered Artificial Intelligence and hybrid intelligence human-robot communities.

KEYWORDS

Lie detection, machine learning, human-robot interaction.

FUNDING

This work has been supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. G.A. No 804388, wHiSPER.

REFERENCES

[1] C. Tosone, “Living everyday lies: The experience of self,” *Clin. Soc. Work J.*, vol. 34, no. 3, pp. 335–348, 2006, doi: 10.1007/s10615-005-0035-z.

[2] B. M. DePaulo, S. E. Kirkendol, D. A. Kashy, M. M. Wyer, and J. A. Epstein, “Lying in Everyday Life,” *J. Pers. Soc. Psychol.*, vol. 70, no. 5, pp. 979–995, 1996, doi: 10.1037/0022-3514.70.5.979.

[3] C. Stern and W. Stern, “Recollection, testimony, and lying in early childhood.,” *Recollect. testimony, lying early childhood.*, Oct. 2004, doi: 10.1037/10324-000.

[4] V. Talwar and K. Lee, “Development of lying to conceal a transgression: Children’s control of expressive behaviour during verbal deception,” *Int. J. Behav. Dev.*, vol. 26, no. 5, pp. 436–444, 2002, doi: 10.1080/01650250143000373.

[5] R. S. Feldman, J. A. Forrest, and B. R. Happ, “Self-Presentation and Verbal Deception : Do Self-Presenters Lie More ?,” *Basic Appl. Soc. Psych.*, no. January 2014, pp. 37–41, 2010, doi: 10.1207/S15324834BASP2402.

[6] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, “Cues to deception.,” *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003, doi: 10.1037/0033-2909.129.1.74.

[7] C. Hadnagy, *Social engineering : the science of human hacking*. 2018.

[8] D. Pasquali, J. Gonzalez-billandon, A. M. Aroyo, G. Sandini, and F. Rea, “Detectng lies is a child (robot)’s play: gaze-based lie detecton in HRI.”

[9] A. Gaggioli, “Beyond the Truth Machine: Emerging Technologies for Lie Detection,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 21, no. 2, pp. 144–144, Feb. 2018, doi: 10.1089/cyber.2018.29102.csi.

[10] C. R. Honts, D. C. Raskin, and J. C. Kircher, “Mental and physical countermeasures reduce the accuracy of polygraph tests.,” *J. Appl. Psychol.*, vol. 79, no. 2, pp. 252–9, Apr. 1994, Accessed: Jul. 07, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8206815.

[11] M. Gamer, “Detecting of deception and concealed information using neuroimaging techniques,” in *HRI’20 Human-Robot Interaction*, 2011, pp. 90–113, doi: 10.1017/CBO9780511975196.006.

[12] B. A. Rajoub and R. Zwiggelaar, “Thermal Facial Analysis for Deception Detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 6, pp. 1015–1023, Jun. 2014, doi: 10.1109/TIFS.2014.2317309.

[13] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, “TS-LSTM and Temporal-Inception: Exploiting Spatiotemporal Dynamics for Activity Recognition,” Mar. 2017, Accessed: Jul. 07, 2019. [Online]. Available: http://arxiv.org/abs/1703.10667.

[14] V. Karpova, V. Lyashenko, and O. Perepelkina, “‘ Was It You Who Stole 500 Rubles ?’ — The Multimodal Deception Detection,” in *ICMI ’20 Companion: Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 112–119, doi: ttps://doi.org/10.1145/3395035.3425638.

[15] X. (Leslie) Chen, S. Ita Levitan, M. Levine, M. Mandic, and J. Hirschberg, “Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies,” *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 199–214, 2020, doi: 10.1162/tacl_a_00311.

[16] D. Pasquali, J. Gonzalez-billandon, F. Rea, G. Sandini, and A. Sciutti, “Magic iCub: a humanoid robot autonomously catching your lies in a card game,” 2021.

[17] J. Beatty, “Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources,” 1982.

[18] D. P. Dionisio, E. Granholm, W. A. Hillix, and W. F. Perrine, “Differentiation of deception using pupillary responses as an index of cognitive processing.,” *Psychophysiology*, vol. 38, no. 2, pp. 205–11, Mar. 2001, Accessed: Jul. 07, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11347866.

[19] C. F. Bond and B. M. DePaulo, “Accuracy of deception judgments,” *Personal. Soc. Psychol. Rev.*, vol. 10, no. 3, pp. 214–234, 2006, doi: 10.1207/s15327957pspr1003_2.

[20] A. Vrij, “Why professionals fail to catch liars and how they can improve,” *Leg. Criminol. Psychol.*, vol. 9, no. 2, pp. 159–181, Sep. 2004, doi: 10.1348/1355325041719356.

[21] A. Vrij, M. Hartwig, and P. A. Granhag, “Reading Lies: Nonverbal Communication and Deception Aldert,” 2019.

[22] T. Brennen and S. Magnussen, “Research on Non-verbal Signs of Lies and Deceit: A Blind Alley,” *Front. Psychol.*, vol. 11, no. December, pp. 1–4, 2020, doi: 10.3389/fpsyg.2020.613410.

[23] S. I. Levitan, X. Tan, and J. Hirschberg, “LieCatcher: Game Framework for Collecting Human Judgments of Deceptive Speech,” *ICMI 2020 - Proc. 2020 Int. Conf. Multimodal Interact.*, pp. 762–763, 2020, doi: 10.1145/3382507.3421166.

[24] “Dixit 3: Journey | Board Game | BoardGameGeek.” https://boardgamegeek.com/boardgame/119657/dixit-3-journey (accessed Sep. 27, 2020).

[25] “Vimeo.” https://vimeo.com/.

[26] “SurveyMonkey Audience.” www.surveymonkey.com/mp/audience.

[27] C. Chiorri, F. Bracco, T. Piccinno, C. Modafferi, and V. Battini, “Psychometric properties of a revised version of the ten item personality inventory,” *Eur. J. Psychol. Assess.*, vol. 31, no. 2, pp. 109–119, 2015, doi: 10.1027/1015-5759/a000215.

[28] J. Zhang, S. I. Levitan, and J. Hirschberg, “Multimodal deception detection using automatically extracted acoustic, visual, and lexical features,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-October, pp. 359–363, 2020, doi: 10.21437/Interspeech.2020-2320.

[29] R. Mazziotti *et al.*, “MEYE: Web-app for translational and real-time pupillometry,” *bioRxiv*, p. 2021.03.09.434438, 2021, [Online]. Available: https://doi.org/10.1101/2021.03.09.434438.