Detecting lies is a child (robot)'s play: gaze-based lie detection in HRI

Dario Pasquali[*,1,2,3], Jonas Gonzalez-Billandon[1,2], Alexander Mois Aroyo[5], Giulio Sandini[2], Alessandra Sciutti[4], Francesco Rea[2]

ABSTRACT

Robots destined to tasks like teaching or caregiving have to build a long-lasting social rapport with their human partners. This requires, from the robot side, to be capable of assessing whether the partner is trustworthy. To this aim a robot should be able to assess whether someone is lying or not, while preserving the pleasantness of the social interaction. We present an approach to promptly detect lies based on the pupil dilation, as intrinsic marker of the lie-associated cognitive load that can be applied in an ecological human-robot interaction, autonomously led by a robot. We demonstrated the validity of the approach with an experiment, in which the iCub humanoid robot engages the human partner by playing the role of a magician in a card game and detects in real-time the partner deceptive behavior. On top of that, we show how the robot can leverage on the gained knowledge about the deceptive behavior of each human partner, to better detect subsequent lies of that individual. Also, we explore whether machine learning models could improve lie detection performances for both known individuals (within-participants) over multiple interaction with the same partner, and with novel partners (between-participant). The proposed setup, interaction and models enable iCub to understand when its partners are lying, which is a fundamental skill for evaluating their trustworthiness and hence improving social human-robot interaction.

1 Introduction

Trust is a fundamental component of social interaction. For an individual, it is crucial to gain the partners' trust and, at the same time, to assess their trustworthiness. One of the main elements normally adopted to evaluate whether someone should be trusted or not is the veridicality of their claims; since, the occurrence of lies naturally undermines the trust given to a partner [1], [2]. Being able to recognize when someone is lying to us plays an important role in shaping our trust toward them and the entire social rapport.

If robots are meant to become autonomous agents active in our society, they should consider the relevance of mutual trust with their human partners. Recently, researchers and social media raised public awareness on how much artificial intelligence and robots can be trusted [3]. On the other hand, it will be necessary also for the social robot to evaluate how much the human partner is trustworthy and consistently adapt its behavior. Several Human-Robot Interaction (HRI) studies explored the factors that influence humans' trust toward robots. For example, robots' shape and performances can affect trust and its development [4]–[6]. Additionally, robot's transparency [7], [8], behavior explanation [9] and perceived reliability [10]–[13] have been shown to affect trust. To measure trust in human-robot collaboration different scale metrics have been developed [14]–[16]. However, little research has focused on the opposite scenario: how a robot should assess human partner's trustworthiness. Vinanzi et al. [17] and Patacchiola et al. [18] worked on a developmental cognitive architecture based on the Theory of Mind. Their architecture exploits episodic memory to feed a Bayesian model of trust, making the iCub and Pepper humanoid robots able to decide whether to trust or not the human partners. Importantly, in these models, trust is assessed based on whether the human has provided a veridical or a false indication to the robot, but this information is not dynamically updated in further interactions. Hence, the ability to detect lies represents for a robot a crucial skill to evaluate whether its partner should be trusted. Indeed, detecting lies has been proved to be an effective way to evaluate partner's trustworthiness in a social interaction [1]. In the context of human robot interaction, a robot capable of detecting lies, could use it as a quantitative measure to understand and predict the human partners' behaviors.

Lie detection has been well explored in the literature. De Paulo et al. [19] and Honts et al. [20] showed how lying can be related to an increment of cognitive load with respect to truth telling. This cognitive effort is due to the creation and maintenance of a credible and coherent story [21]. Therefore, traditional methods of lie detection involve the monitoring of physiological metrics like skin conductance, respiration rate, heartbeat, or blood pressure, all reflecting variations of cognitive load and stress. The polygraph, one of the most used lie detection devices, relies on the aforementioned metrics reaching an accuracy between 81% and 91% [22] (but see [20] about the possibility of bypassing the measure). Other lie detection methods rely on fMRI images [23], skin temperature variations [24], micro-expressions [25], photoplethysmography [26] or acoustic prosody [27]. Most of these methods (i) are invasive or require cumbersome devices, not easily portable to everyday life scenarios; (ii) are expensive; (iii) or require experts to evaluate the measures. These characteristics make these approaches not suitable for porting them to robotic platforms.

[*] **Corresponding Author**: dario.pasquali@iit.it

[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

[5] Social and Intelligent Robotics Research Laboratory (SIRRL), Department of Electrical and Computer Engineering, University of Waterloo, Canada

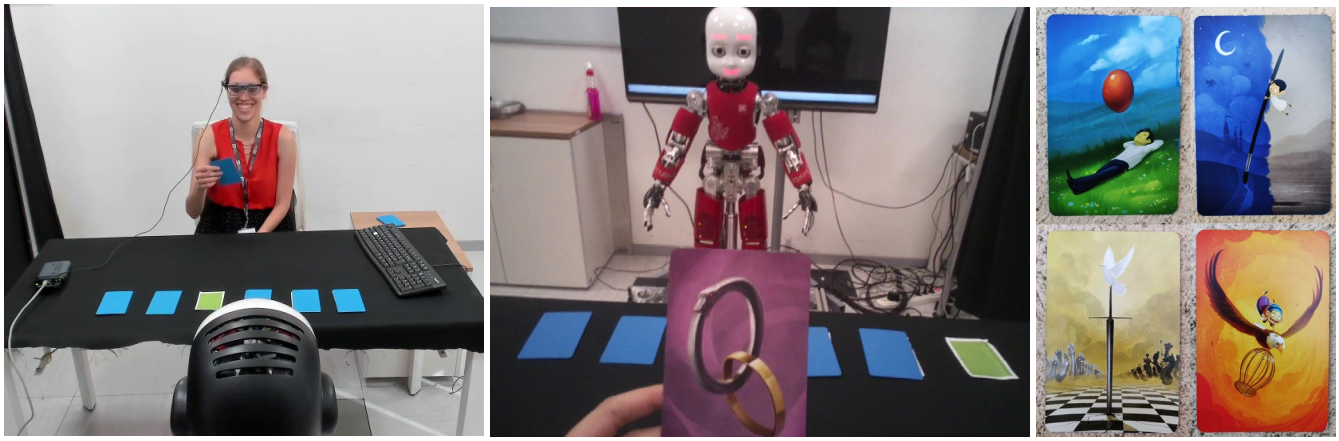

Figure 1: (Left) Participant describing a card to iCub, while wearing the Tobii Pro Glasses 2 eye tracker (Logitech Brio 4k webcam point of view); (Center) Point of view of the participant during the interaction collected through the Tobii glasses; (Right) Examples of Dixit Journey gaming cards (authored by Jean-Louis Roubira, designed by Xavier Collette and published by Libellud).

Recent findings [28]–[33] proved how pupillometry measurements [34] and, in particular, Task Evoked Pupillary Responses (TEPRs) [35], can be used to evaluate the task-evoked cognitive load. Beatty et al. [35] identified mean pupil dilation, peak dilation and latency to peak as useful task-evoked pupillary responses. Dionisio et al. [36] studied the task-evoked pupil dilation related to lie telling. They asked students to lie or tell the truth, answering questions about episodic memory. They reported a significant greater pupil dilation during lie production with respect to truth telling. Gonzalez-Billandon et al. [37] and Aroyo et al. [38] found that participants had a higher mean pupil dilation when lying with respect to telling the truth both in human-human and human-robot interaction. Both, mobile head mounted [39], [40], and remote eye tracker [41], [42] devices have been used as minimally invasive methods to measure pupillometric features, more appropriate for real-world scenarios. Recent research showed the possibility to measure TEPRs from RGB cameras, suitable for robotic platforms, making pupillometry a promising candidate to detect lies in real-life human-robot interactions [43]–[45].

Beyond minimizing the invasiveness of the sensors used, the social robot should perform this evaluation while preserving the pleasantness of the interaction. This is particularly important for humanoid robots that aim to act as teachers, caregivers, or just friendly companions. Conversely, state-of-the-art setups and scenarios for lie detection are long, strict, and interrogatory-like [26], [27], [37].

In this paper, we propose a method to detect lies in real-time via pupillometry-driven cognitive load assessment, by learning how each individual partners' pupil dilation changes while lying. We validate the approach in a quick and entertaining interaction autonomously led by the iCub humanoid robot. The iCub asked participants to describe 12 gaming cards and to lie about a few of them. iCub autonomously processed in real-time participants' pupil dilation to detect the deceptive card description based on our proposed method. During a first phase of the game (*Calibration Phase*) participants had to lie about one predefined card among six. Afterwards, participants could freely decide whether to lie or not for each of the next 6 cards in the game (*Testing Phase*). In this second phase, iCub exploited the knowledge about pupil dilation acquired in the *Calibration phase* to detect the player's lies, without knowing in advance the number of true or false descriptions. The robot obtained an average accuracy of 70.8%, during the game, among the two phases, where the number of lies was either fixed (1 over 6 cards, *Calibration*) or it was arbitrarily chosen by each participant (*Testing*). To improve the robustness of the approach, we designed novel classification methods to adapt iCub's knowledge over multiple interactions with the same individual. Last, we propose an attempt to train a generic machine learning model, able to detect lies without any previous information about the specific human partner.

In the following sections we will first describe the experimental procedure and the setup used to run the validation experiment (section 2), the collected measures (section 3) and the architecture enabling the robot to conduct the game and detect lies (section 4). Then, we describe the data preparation procedures and the datasets built with the collected data (section 5). Last, we will report the results of the experimental validation with naïve subjects and the results of machine learning methods aimed at improving within-subject detection and lie detection in presence of novel partners (section 6). Results suggest that with the proposed interaction and lie detection models iCub could reliably assess when the human partners were lying.

2 Methods

To prove the effectiveness of our lie detection method, we performed an HRI experiment. The setup and a subset of the procedure have been previously described in [46].

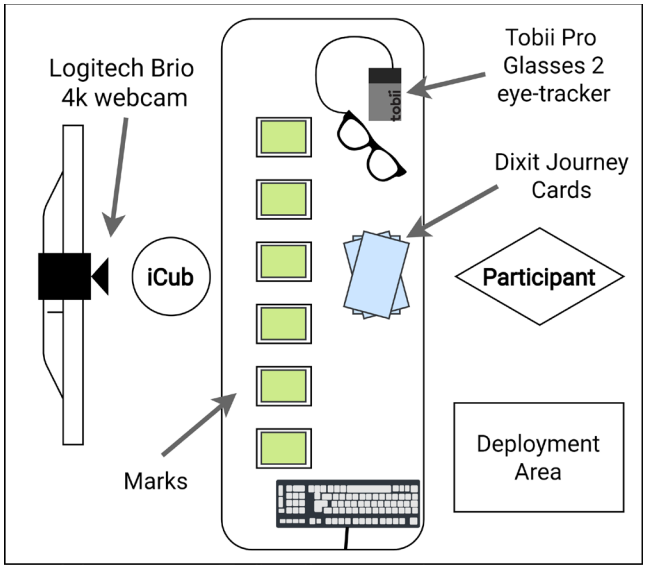

Figure 2: Card game experimental setup with iCub (left) and the participant (right) sitting on a table. The deployment area is the location where the remaining Dixit Cards after each drawing were placed.

2.1 Setup and Materials

The room was arranged to replicate an informal interaction scenario between a human and a robot (Figure 2). The participants sat in front of the iCub humanoid robot separated by a table covered with a black cloth. On the table, the experimenter placed: six green marks (95x70 mm); a deck of 84 cards from Dixit Journey card game with the back painted in blue; a keyboard; and a Tobii Pro Glasses 2 eye-tracker. On participants' left there was a little drawer (deployment area); while on the right, a black curtain hid the experimenter from participants' sight. Behind iCub, a 47 inches television showed iCub's speech during the interaction (to prevent any speech misunderstanding). A Logitech Brio 4k webcam, fixed on the television, recorded the scene from iCub's point of view at a resolution of 1080p (Figure 1, left).

The Dixit Journey card deck is composed by 84 cards (80x120 mm) with different toon-styled drawings meant to stimulate creative thinking [47] (Figure 1, right). Designing the card game, we tried to avoid any cue – other than the wearable eye-tracker – for the participants about the method used by iCub to detect their false card descriptions; in this sense, we avoided any machine-readable mark (i.e., QR codes on cards' back) that iCub could use to recognize the cards. The Tobii Pro Glasses 2 eye tracker recorded participants' pupillometric features at a frequency of 100 Hz and streamed in real-time the participants' pupil dilations at a frequency of 10 Hz (Figure 1, center). The window blinders were closed, and the room was lit with artificial light to ensure a stable light condition for all the participants.

The iCub humanoid robotic platform [48] played the role of a magician. iCub autonomously led the whole interaction thanks to the autonomous end-to-end (E2E) architecture in Figure 3 (see section 4). The experimenter monitored the scene through the iCub's left eye ensuring the safety of the participants and the correct execution of the experiment.

2.2 Procedure

At least a day before the experimental session, the participants filled in a set of questionnaires meant to assess their personality (see Section 3). After signing the informed consent, the experimenter led the participants to the experimental room. They were asked to sit on the chair in front of iCub and informed that the robot would have played a game with them. Then, the experimenter hid himself behind the black curtain and started the experiment.

The human-robot interaction was composed of two phases, *Calibration Phase* and *Testing Phase*, both led autonomously by iCub.

2.4.1 Calibration Phase. As the game started, iCub asked the participants to shuffle the cards deck, extract six cards without looking at them and put the deck on the deployment area. Then, iCub asked them to draw out one of the cards (referred as *secret card*) and memorize it. Afterwards, iCub instructed the participants to look at all the cards, one by one, shuffle them and put them facing down on the six green marks on the table. iCub explained that it was going to point each card one by one and they had to take the pointed card, look at it, describe it and then put it back facing down on the table. Then, iCub explained the game rules: "*The trick is this: if the card you take is your secret card, you should describe it in a deceitful and creative way. Otherwise, describe just what you see*". Finally, iCub asked the participants to wear the Tobii Pro Glasses 2 eye tracker, take a deep breath and relax.

iCub randomly pointed to each of the six cards, while listening to participants' description, and acknowledging it with a short greeting sentence (*e.g.*, "ok", "I see", etc.). After the last description, iCub guessed the participants' *secret card* and asked them to put the six cards aside to validate the detection or show to iCub the real *secret card* to reject it. Participants' confirmation is meant to select the correct *secret card* in case iCub fails to detect it. Before the beginning of the *Testing Phase*, the experimenter could manually override the detected *secret card* with the one presented by the participants, in case the robot failed the guess. Finally, iCub asked them to remove the six cards to start a new game.

2.4.2 Testing Phase. As soon as the participants removed the six cards from the table, iCub asked to take the deck again and draw out six new cards. iCub told the participants to look at all the cards, one by one, then shuffle them and place them on the six green marks. Afterward, iCub instructed the participants that it was going to point to all the cards from right to left (with respect to participants' point of view) and instructed them to handle the pointed card as in the first game. However, it added: "*This time you can choose, for each card, whether to describe it in a creative and deceitful way, or to describe just what you see*". While the robot was

explaining the rules, the participants kept wearing the Tobii Pro Glasses 2.

For each card, iCub (i) pointed it, (ii) listened to participants' description, (iii) acknowledged it with a short sentence, (iv) tried to classify the description as truthful or false and, (v) asked for a confirmation. The participants had to show the card they just described to reject iCub's classification or do nothing to validate it.

2.4.3 General Remarks. During the rule explanation of the two phases, iCub instructed the participants to press a button on the keyboard in order to move to the next task (i.e., after shuffling the cards deck, or after memorizing the *secret card*). No time limit was given to shuffle the card, to look at them, to memorize the *secret card* nor to describe them. iCub's pointing has been designed to replicate a human-like gesture: first moving the gaze toward the target, then the arm, fingers, and torso with a biological inspired velocity profile.

After the second game, the experimenter led the participants to the initial room and asked them to fill in a questionnaire meant to evaluate their task load and self-report their performance during the game (see Section 3). Finally, the experimenter deeply debriefed the participants and let them have the chance to ask questions about the experiment before receiving their monetary compensation.

2.3 Participants

39 participants (25 females, 14 males), with an average age of 28 years (SD=8) and a broad educational background took part in the experiment. They signed an informed consent form approved by the ethical committee of the Regione Liguria (Italy) where it was stated that cameras and microphones could record their performance and agreed on the use of their data for scientific purposes. After the experiment, they received a monetary compensation of 10€. Although all participants completed the game, 5 were excluded from further analysis: 2 for technical issues, 2 because they did not follow the rules of the game. The last one was considered an outlier, as she concluded the game in 38 minutes (a duration longer than 3SD plus the average game duration, which lasted 17 minutes). Hence, the final sample includes N=34 participants (22 females, 12 males).

3 Measurements

3.1 Pre-questionnaires

Before the experiment, the participants filled in the following questionnaires: The Big Five personality traits (extroversion, agreeableness, conscientiousness, neuroticism, openness) [49]; the Brief Histrionic Personality Disorder (BHPD) [50]; and the Short Dark Triad (SD3, machiavellianism, narcissism, and psychopathy) [51].

3.2 Post-questionnaires

After the experiment, the participants filled in the NASA-TLX [52] and a set of questions regarding: (i) the experienced fun, (ii) creative effort, (iii) strategies adopted in fabricating a deceitful and creative description during the game, (iv) previous experience about the Dixit Journey card game, (v) previous experience about improvisation and acting, and, (vi) habits on playing deception-related games.

3.3 Gaze Measurements

From the full set of pupillometric features measured by the Tobii Pro Glasses 2 eye tracker, we collected and used only the pupil dilation, in millimeters, for right and left eyes. To avoid any impact on the informality of the social interaction, we avoided the eye tracker calibration phase; indeed, the calibration does not affect the pupil dilation measurement [53]. Pupil dilation data points are synchronized over the YARP robotic platform with the annotation events.

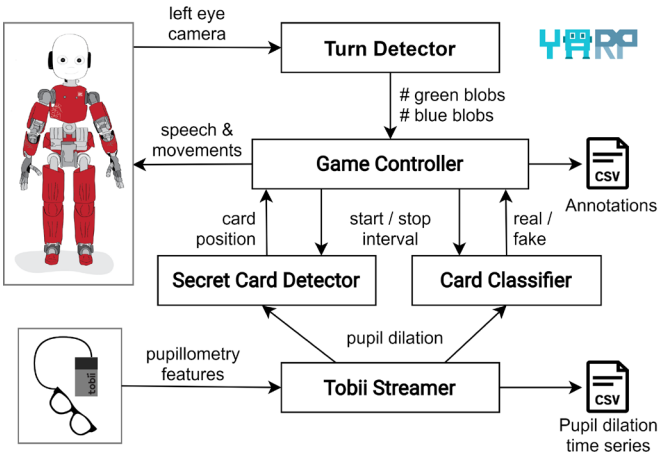

Figure 3: Autonomous end-to-end architecture used in real-time to make iCub able to lead the card game.

4 Robot Architecture

iCub autonomously leads the human-robot interaction thanks to the end-to-end architecture in Figure 3. An initial version of the architecture, designed to handle the *Calibration Phase* only, is described in [46]. With the *Turn Detector* iCub detects the beginning and end of each card description by tracking the number of green (marks) and blue (cards) blobs visible in the scene. This is also used to understand participants' confirmations. The *Tobii Streamer* reads participants' pupillometric features from the Tobii Pro Glasses 2 eye-tracker and streams and logs them in real-time over the YARP robotic platform [54]. The *Game Controller* implements the main game engine: (i) it controls iCub's movements and speech; and (ii) it segments the start and end of each pointing, card description and phases, logging annotation

events. The logged annotation events and pupil data points are synchronized over the YARP robotic platform [54], providing an autonomous annotation for future analysis.

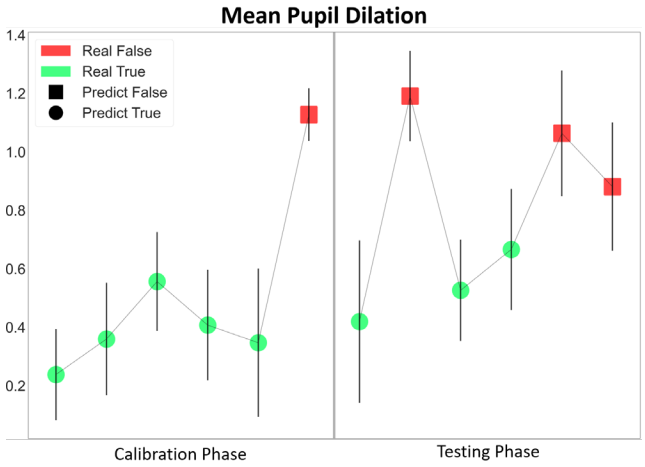

Figure 4: Mean pupil dilation during *Calibration Phase* (left) and *Testing Phase* (right) for participant A. Green circles are truthful card description; red squares are false ones. Bars represent standard errors.

Finally, the *Secret Card Detector* and the *Card Classifier* enable iCub to identify participants' lies during the game. iCub detects (*Calibration Phase*) and classifies (*Testing Phase*) players' lies thanks to a specific Task Evoked Pupillary Response: the fabrication of a credible and consistent deceptive card description triggers an increase in players' cognitive load [55], [56]; this increment reflects on a higher pupil dilation with respect to a truthful card description [19], [21], [57]. iCub aggregates participants' eye pupil dilation data points, computing the mean pupil dilation during each card description and use them to detect players' lies. We focused on right eye's pupil dilation since both Tobii documentation [53] and previous results indicate that pupil dilation is not different between right and left eye [37]. The components implement two heuristic methods:

Calibration Heuristic (Figure 4, left) During the *Calibration Phase*, iCub detects as *secret card* the one related to the highest mean pupil dilation among the six card descriptions. This approach has been described in [46].

Testing Heuristic (Figure 4, right) At the end of the *Calibration Phase*, iCub knows 6 mean pupil dilation data points: 1 related to the *secret card*, and 5 related to truthful cards. With them, it computes two *reference scores*: the *true reference score* is the average of the 5 mean pupil dilations of truthful cards; the *false reference score* is just the *secret card* mean pupil dilation. For each *Testing Phase* card description, the mean pupil dilation was computed and compared to the two *reference scores*. By taking the minimum absolute difference iCub could label the current description as fake or a true.

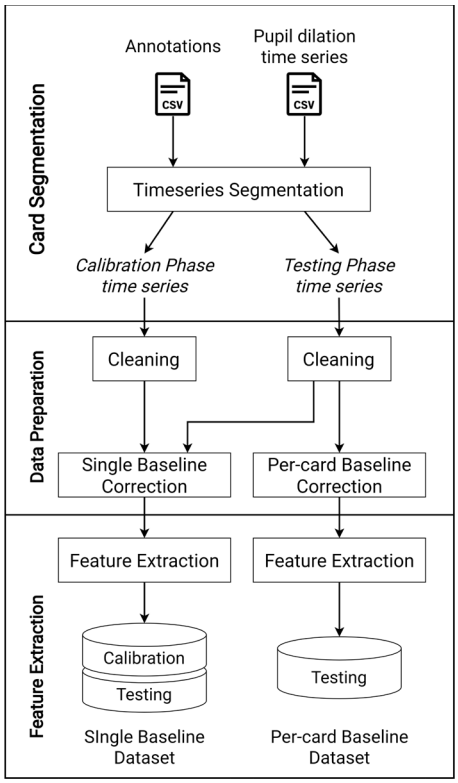

Figure 5: Computational workflow to preprocess the collected data from Tobii Pro Glasses 2 eye-tracker. Two datasets are extracted. The difference depends on the applied baseline correction (single or per-card).

5 Data Preparation

From the pupil dilation data points collected in real-time we built two datasets following the computational workflow in Figure 5.

5.1 Card Segmentation (Figure 5, top)

The card trial annotation is autonomously performed by the *Game Controller* (Figure 4) by rising annotation events on the YARP robotic platform for the beginning and end of each pointing and card description. We segmented the pupil dilation time series into 3 temporal intervals for each card trial: (i) *robot's turn*: iCub's pointing gesture, from the moment iCub starts the pointing gesture till the participant takes the pointed card from the green mark; (ii) *player's turn*: a card description, from the moment the participant takes the card from the green mark, till they put it back on it; (iii) *card trial*: the whole interaction for a single card, from the moment iCub starts the pointing gesture till the participant puts the card back on the green mark.

5.2 Data Preprocessing (Figure 5, center)

We fitted and resampled the time series at 10 Hz to make it consistent with the real-time processing, then applied a median filter to remove the outliers and a rolling window mean filter to smooth the time series and infer any eventual missing data points. We then corrected each time series subtracting a baseline value

for each participant [58]. In this reference system, a positive value represents a dilation, while a negative value represents a contraction with respect to the baseline. We corrected the time series with respect to two different baselines: (i) In the *Single Baseline Correction, the baseline* is computed as the average pupil dilation during the 5 seconds before the first pointing of the *Calibration Phase* and applied to all the cars of both phases; (ii) in the *Per-card Baseline Correction*, a specific baseline is computed for each card as the average pupil dilation during the 5 seconds before each pointing.

5.3 Feature Extraction (Figure 5, bottom)

Finally, we aggregated the time series of each temporal interval, and computed several features. For each *player's turn, robot's turn,* and *card trial* we computed the maximum, minimum, mean and standard deviation of the pupil dilation in millimeters, and the duration in seconds. Moreover, on the whole *card trial* we computed a set of 26 specific time series features using the python module Time Series Feature Extraction Library (TSFEL) [59]. In particular, the TSFEL features are: (i) *Statistical Features*: median, median absolute deviation, mean absolute deviation, kurtosis, skewness and variance; (ii) *Temporal Features*: absolute energy, area under the curve, autocorrelation, centroid, entropy, mean absolute difference, mean difference, median absolute difference, median difference, peak to peak distance, slope, total energy; (iii) *Spectral Features*: fundamental frequency, maximum frequency, median frequency, spectral centroid, spectral entropy, spectral kurtosis, spectral skewness, spectral slope. We considered the features for both eyes as separate data points to augment the datasets. This results in two different datasets:

Single Baseline Dataset. This dataset includes the data points of both phases, replicating the data structure used in real-time. It is meant to explore an incremental learning over multiple interactions with the same individual.

Per-card Baseline Dataset. This dataset, instead, includes only data from the *Testing Phase*; it is meant to train a generic machine learning model, independent from the specific interacting partner.

Shapiro-Wilk and D'Agostino K-squared normality tests showed that some of the features of the datasets were not normally distributed. Therefore, we opted to use non-parametric tests for all the following statistical analyses. Additionally, we decided to focus on data points from participants' right eye (unless otherwise specified), since there is no difference between right and left eye pupillary features [53].

6 Results

In this section we report the in-game and questionnaires results, along with the post-hoc analysis on the collected pupillometric data. In the post-hoc analysis, we mainly focus on the learning from the *Calibration* to the *Testing Phase* and on the second phase per-se; for a deeper analysis of the *Calibration Phase* see [46].

6.1 In-game Results

The interaction lasted on average 17 minutes (SD=5) from the beginning of iCub explaining the *Calibration* Phase's rules till the final greeting of the *Testing Phase*.

The *Calibration Phase* lasted on average 8 minutes (SD=3), during which, iCub successfully detected the players' *secret card* with an accuracy of 88.2% (against a chance level of 16.6%, N=34). The *Testing Phase* lasted on average 8 minutes (SD=2). The participants were free to choose whether to lie or not, producing on average 2.73 (SD=0.94, 45%) false descriptions among 6 cards. ICub successfully classified each card description as true or false with accuracy = 70.8%, precision = 73.6%, recall = 57% and F1 score = 64.2% (N=34).

Considering the results of the questionnaires, Table I summarizes the results of the Big Five personality traits [49], Brief Histrionic Personality Disorder [50] and Short Dark Triad [51] questionnaires, performed before the experiment. Average scores for the Big Five were Agreeableness: M=0.659, SD=0.113; Conscientiousness: M=0.481, SD=0.072; Neuroticism: M=0.387, SD=0.16; Openness to experiences: M=0.476, SD=0.07 and Extraversion: M=0.486, SD=0.061. Considering the Dark Triad, the scores were Psychopathy: M=0.191, SD=0.113; Machiavellianism: M=0.438, SD=0.129 and Narcissism: M=0.396, SD=0.15. For the Brief Histrionic Personality Disorder, the average score was M=0.481, SD=0.26.

	Participants' psychological profile		
Score %	***Big 5*** *{C, A, N, O, E}*	***Dark Triad*** {M, N, P}	***Histrionic***
0-20%	{0, 0, 2, 0, 0}	{2, 6, 15}	6
20-40%	{5, 1, 17, 4, 3}	{8, 4, 13}	4
40-60%	{22, 6, 6, 24, 23}	{18, 11, 1}	11
60-80%	{2, 21, 3, 1, 1}	{0, 4, 0}	4
80-100%	{0, 1, 1, 0, 0}	{1, 4, 0}	4

Table I: Participants' psychological profile from pre-questionnaires. Big 5 (Conscientiousness, Agreeableness, Neuroticism, Openness to experience, Extraversion); Dark triad (Machiavellianism, Narcissism, Psychopathy); and Histrionic – higher score means higher effect. In brackets the number of participants per each percentage range.

After the experiment, participants filled in the NASA-TLX questionnaire, rating on a 10-points Likert scale their effort on performing the task. On average, they reported a low task load (M=3.717, SD=1.041). Among the components, Mental Effort (M=5.41, SD=1.78), Fatigue (M=5.07, SD=2.14) and Performance (M=5.35, SD=2.32) are slightly higher than Temporal Effort (M=2.59, SD=1.72), Frustration (M=2.72, SD=1.83) and Physical Effort (M=1.21, SD=0.49). This is consistent with the requirements

of the task. Also, we asked participants to self-report, on a 5-points Likert scale the effort put on fabricating creative and deceptive descriptions (Lie Effort: M=4.17, SD=0.71) and the experienced fun (Fun: M=4.59, SD=0.57).

Then, we explored whether pupil dilation features were dependent on participants' personality traits. We considered the *Testing Phase* data from the *Per-card Baseline Dataset*, to minimize the impact of card presentation order on pupil features, normalizing each card for its own baseline. We fit two linear regression models with the personality traits from the pre-questionnaire as independent variables and, as dependent variables the difference between mean pupil dilation for false and true cards or the mean pupil dilation baseline. Results show that only Neuroticism correlates significantly with the mean pupil dilation baseline (t=2.492, p=0.021, Adj. R^2=0.115). We also tested whether pupil features correlated with the average description duration, Fun, Lie Effort, task load or Mental Effort, but we did not find any significant correlation.

6.3 Learning from a Brief Interaction

To investigate in more detail the relationship between pupil dilation and lying observed during the game, we started analyzing the *Single Baseline Dataset* which resembles the data structure used in real-time.

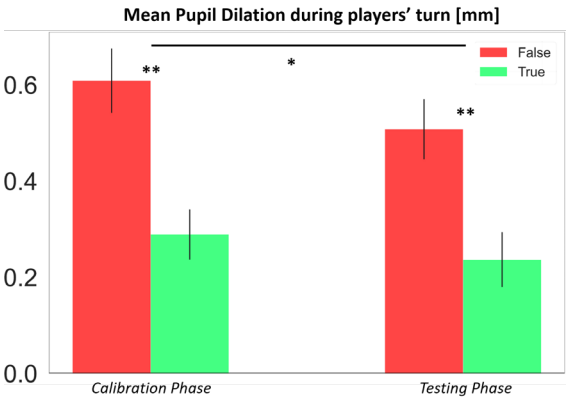

Figure 6: Average of mean pupil dilation during *player's turn* for *Calibration* and *Testing Phases*, with standard errors of the mean. (* = p<0.05, ** = p<0.001).

The *Single Baseline Dataset* presents a multilevel structure (multiple phases for the same participant, nested in card trials, nested in turns) with unbalanced card classes (one *secret card* among six (about 16.6%) in the *Calibration Phase* and on average 45% of false cards in the *Testing Phase*). Since the real-time game was based on participants' mean pupil dilation during the *player's turn*, we decided to focus on such temporal intervals.

We fitted a mixed effects model for the *player turns* with mean pupil dilation as the outcome variable. As fixed effects we entered "card label" (two levels: true, false), "phase" (two levels: calibration, testing) and their interaction into the model. As random effect we had intercept for participants. We set the reference level on the *Testing Phase* and false card label. Results show a highly significant effect of card label (B=-0.223, t=-8.885, p<0.0001) revealing a higher mean pupil dilation for the false card descriptions with respect to the truthful ones. We also found a significant effect of phase (B=0.104, t=2.428, p=0.016), with a significantly lower mean pupil dilation in the T*esting Phase*, and no significance of the interaction between the two factors (B=-0.052, t=-1.023, p=0.307).

As an exploratory analysis, we fit another mixed effects model on the *robot's turn*, with the same abovementioned structure. Results show no effect on the card label (B=-0.035, t=-1.373, p=0.171), but a highly significant effect on the phase (B=0.124, t=3.490, p=0.0005) confirming a lower mean pupil dilation in the *Testing Phase* with respect to the *Calibration* one also for this turn. Finally, we found no effect of the interaction of card label and phase factors (B=-0.014, t=-0.331, p=0.741).

6.3.1 Incremental Testing Heuristic. Even if the *Testing Heuristic* demonstrated a quite good accuracy – humans perform near chance on detecting lies [60] – it has a low recall score (recall = 57%, accuracy = 70.8%, precision = 73.6%, N=34), that is it recognizes only a relatively low proportion of the false statements made by the participants.

Figure 7 provides two examples of correct (left graph) and wrong (right graph) classifications. The two panels show the mean pupil dilations of participant A (left graph) and participant B (right graph) as processed by the *Testing Heuristic*. In each graph, the two data points on the left represent the two *reference scores*: the red square is the mean pupil dilation for the *secret card*, while the green circle is the average of the mean pupil dilations for the truthful cards. On the right there are the mean pupil dilation data points for each card of the *Testing Phase*. For participant A, pupil dilations for false and true descriptions remain consistent with the average values measured during the previous phase and the classification is always successful. Conversely, all the *Testing Phase* mean pupil dilations of participant B (right graph) fall in the range of the *true reference score*. Hence all the false card descriptions have been misclassified as false positives (red circles).

The observed errors are determined by two assumption on which the heuristic is based: (i) the difference in pupil dilation between false and true sentences remains almost the same between the two phases; and (ii) participants' pupil dilation remains almost stable between the two phases. The first assumption is confirmed by the non-significant difference in the interaction of "phase" and "card labels" in both turns. However, the statistical analysis showed that participants' pupil dilation is on average lower during the *Testing Phase*.

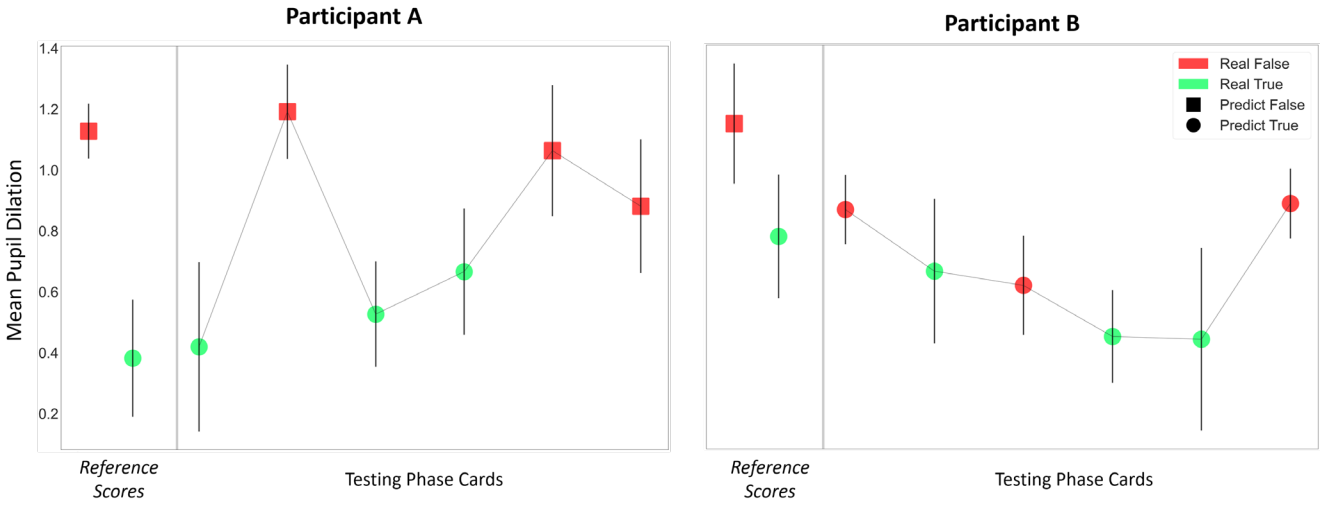

Figure 7: (Left) Mean pupil dilation data points as seen by the *Testing Heuristic* for participant A (left) and B (right). Color represents the real class (green = true, red = false); shape represents the predicted class (circle = true, square = false); bars represent standard deviation.

To compensate for this effect and increase the robustness of the heuristic, we explored the possibility to incrementally adapt the *reference scores* for truthful and false card description. After each card classification, the new card value is aggregated with the *reference scores*. This way iCub incrementally learns how the human partner lies and tells the truth, improving the classification performances trial by trial. We simulated the *Testing Heuristic* based on mean pupil dilation during the *player's turn*, as in the real-time game, but including the incremental learning. For each *Testing Phase* card trial, both the *reference scores* are updated computing the mean between each score and the novel mean pupil dilation data point. The heuristic performance increases to accuracy = 76.7%, precision = 76.1%, recall = 73.7% and F1 score = 75.6%.

Then, we simulated the *Testing Heuristic* performing a grid search on several parameters: (i) all the possible combinations of the available features (limited to a maximum of 3 features considered at the same time, see section 5.3 for the full list); (ii) methods to compute the *true reference score* (mean, median, minimum); (iii) methods to update the *reference scores* (mean, difference, quadratic error); (iv) whether to update both scores or just the one of the correct class; (v) whether to update the *reference scores* only if the card trial is misclassified. Since we assume that for a lie detection system it is preferable to detect a greater amount of true negative (i.e., spot a larger amount of lies) even at the expenses of having a few false positives, we prioritized the recall score. The best heuristic has an accuracy = 78.7%, precision = 76%, recall = 80% and F1 score = 77.9%. It is based on both the mean and minimum pupil dilation during *player's turn*, which are compared by a 2D Euclidean distance with the *reference scores*; the *true reference score* is computed as the minimum among mean pupil dilations for the truthful cards descriptions during the *Calibration Phase*; both the *reference scores* are updated in any case, averaging each score with the new value.

6.3.2 Random Forest classifier. Even if the new heuristic method performs better than the one exploited in real-time, it is still not generic and robust enough to describe the variability of participants' pupil dilation between the two phases. Indeed, the *Testing Heuristic* is meant to adapt to each specific individual. We supposed that, by relaxing this constraint, it would be possible to compensate for the variability between the two phases. We trained a machine learning model able to learn from the *Calibration Phase* on the whole participants sample, and to exploit the gained knowledge on the *Testing Phase*. The classification problem is a binary classification defined by a couple [X, Y] where: X (42 x 1) is the vector of input behavioral features and Y ∈ [0: *true*; 1: *false*] is the vector of desired outputs. We defined a within-participant split, considering the *Calibration Phase* data as training set and the *Testing Phase* data as validation and test (with a splitting ratio of 50:50). *Calibration Phase* data points have two main issues: (i) they are unbalanced (1 *secret card* among 6 cards); and (ii) the set is relatively small (6 cards, for 34 participants, 2 eyes for participants, for a total of 408 data points). We considered features from both eyes to augment the dataset. Due to these limitations we selected a Random Forests algorithm [61]. This kind of model should not overfit when increasing the number of trees, even with relatively small datasets. Also, we tackled the unbalancing problem by oversampling the *Calibration Phase* data points with the synthetic minority oversampling technique (SMOTE) [62]. We did not oversample the *Testing Phase* data points validating and testing on realistic data. Even if not strictly required by the Random Forest algorithm, we applied a min max normalization [63] to all the features within the data points of each participant in both phases. The idea is that a value that is relevant for a participant could be not relevant for another. We

performed a grid search validation, with fixed validation set, searching the best hyper-parameters and feature set for the random forest classifier. Due to the unbalanced dataset, we rely on the F1 score, precision, recall and AUROC score. The best random forest classifier trained on the full features set achieved an F1 score of 56.5%, a precision of 57.1%, a recall of55.9% and AUCROC score of 59.6%.

6.3.3 Lying as an anomaly: one-class support vector machine. Given the low performance of the random forest classifier we changed approach and we considered the lie detection task as an anomaly detection problem. In this frame, the model knows just the values associated to true descriptions and learns to consider as a lie what is not truthful. We trained a one-class support vector machine (SVM) anomaly detector on the *Calibration Phase* data points, validating and testing it on the *Testing Phase* data points. We considered as training set the truthful card description of the *Calibration Phase* and we carefully balanced *Testing Phase* data points, preserving the ratio between true and false card descriptions in the validation and test sets. We performed a grid search validation, with fixed validation set, searching the best hyper-parameters and feature set for the one-class SVM model. Due to the nature of the anomaly detection problem, we evaluate it based on precision, recall and F1 score. The best one-class SVM model achieved a F1 score of 67.7%, a precision of 60% and a recall of 77.8%. It is based on features from both the *player's turn* (minimum, maximum and mean pupil dilation); and the whole *card trial* (minimum, maximum, mean, and median pupil dilation; total energy, absolute energy, and autocorrelation).

6.4 Detecting lies from novel human partners

After having analyzed how previous knowledge gained during an interaction, can be used to improve lie detection in a subsequent task, we explored the possibility of building a pupil-dilation based lie detector able to classify false card descriptions from novel human partners. This could be the first step toward a minimally invasive and ecological lie detector able to classify a generic sentence as true or false, without any previous interaction with the specific partner. In this sense, it is important to consider the card descriptions as independent as possible from the specific participant and the description order. Hence, we focused on the *Per-card Baseline Dataset* which includes only *Testing Phase* data points. In the *Per-card Baseline Dataset*, the baseline is computed as the average of the pupil dilation, for each eye separately, during the 5 seconds before each card trial. This baseline is subtracted to the pupil dilation time series of the relative card description (see Section 5.2). We considered only the data from the *Testing Phase* since the nature of the task – "*This time, you can choose, for each card, if describe it in a deceitful and creative way, or describe what you see*"makes each card description more similar to a generic and standalone lie.

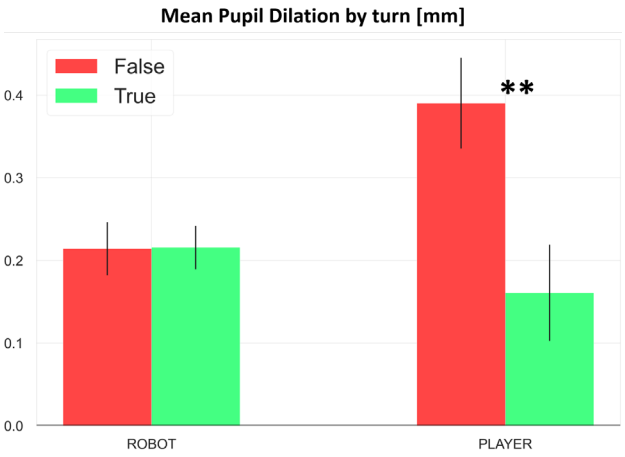

Figure 8: Average of mean pupil dilation during *robot's* and *player's turns* in the *Testing Phase*, with standard deviation. (= p<0.001).**

First, we analyzed whether the use of a *Per-card* baseline determined substantial differences with respect to the descriptive and statistical analysis conducted with the *single* baseline. We fitted a mixed effects model with mean pupil dilation as the outcome variable. We considered "card label" (two levels: true, false) and "turn" (two levels: robot, player) and their interaction as fixed factors, and we had as random effect the intercept for participants. We set the reference level on the *player's turn* and false card label. Results show a highly significant effect on the card label (B=-0.234, t=-6.58, p<0.001), the turn (B=-0.321, t=-6.39, p<0.001) and their interaction (B=0.255, t=5.205, p<0.001). This pattern of results (Figure 8) is similar to that observed for the *Testing phase* in the analysis with the "same" baseline (cf. Figure 6).

We also analyzed whether the other features differed significantly between the true and false card descriptions. We computed the average of each feature for true and false cards and performed Wilcoxon signed-rank tests. Results show that also the minimum pupil dilation (Z=570.0, p<0.001) and the maximum pupil dilation (Z=530.0, p<0.001) during *player's turn* were significantly different. Regarding the whole *card trial*, the mean pupil dilation (Z=555.0, p<0.001), the median pupil dilation (Z=561.0, p<0.001), the minimum pupil dilation (Z=542.0, p<0.001), the maximum pupil dilation (Z=500.0, p<0.001) and the slope (Z=550.0, p<0.001) were significantly different. Also, the total energy (Z=477.0, p=0.001), the absolute energy (Z=457.0, p=0.003), the autocorrelation (Z=458.0, p=0.003), and the area under the curve (Z=442.0, p=0.007) on the whole *card trial* were significantly different. Finally, we found no significance on *robot's turn* features.

6.4.1 Random Forest classifier. To design a lie detector that could classify a card description as true or false with not prior knowledge of the participants, we started from the statistical findings: we selected a subset of the 42 available features, excluding the one related to the *robot's turn*. The classification problem is a binary classification defined by a couple [X, Y] where: X (37 x 1) is the vector of input behavioral features and Y ∈ [0:

true; 1: *false*] is the vector of desired outputs. Considering data points from both participants' eyes, we split *Testing Phase* data between-participants. We considered 25 randomly selected participants (75%) as training and validation set and the remaining as test set. We did not apply any within-participant normalization of the features. We ran a 4-fold grid search cross validation looking for the best values of the hyper-parameters for the classifier. Even if *Testing Phase* data points are more balanced (47% of false card description, against 16.6% during the *Calibration Phase*), we still embedded the SMOTE algorithm [62] in the cross validation. This way it is possible to oversample just the training set, avoiding any synthetized sample in the validation set. The best model achieves a precision, recall and F1 score of 71.1% and AUCROC score of 73.3%.

7 Discussion

In this study we endowed iCub with the capability to detect lies in the context of a natural game-like interaction, using pupil responses to detect cognitive load associated to lying. Games are known to provide ecological assessments, preserving the relationship between the interacting partners [6], [64]–[66]. Also in the context of HRI, games have been successfully exploited to perform diverse types of measurements, even related to cognitive load assessment [40], [67]–[69]. In the current work, the game is a perfect scenario to demonstrate that our lie detection method based on a heuristic function is quick, interactive and does not depend on invasive measures. The game results also provide evidence of the feasibility of our approach, with an overall accuracy of 70.8% (F1 score of 64.2%) during the *Testing Phase*, when basing the lie detection on mean pupil dilation alone. We also show that such accuracy can increase up to 78.7% (F1 score of 77.9%) by enabling an iterative adaptation to each individual partner and by leveraging on a combination of different pupil-related features. The effect on which the lie detection heuristic was based, i.e., the difference in pupil dilation during false or true card descriptions, was relatively robust and did not depend on participants' personality traits, nor on the characteristics of the game (e.g., the experienced fun or the description duration),

Moreover, we explored the possibility to extend the lie detection (i) over multiple interactions with the same individual and (ii) with novel partners. First, we trained a random forest classifier splitting within-subject over the two phases. However, the model did not perform better than the heuristic (F1 score = 56.5%). We assume that this depends on the unimodality of the features, the small number of data points and the strong reliance on synthesized data on the training set. We expect that a machine learning model trained on more real data would be more robust and generic with respect to a real-world human-robot interaction. We try to overcome these issues by tackling the problem as an anomaly detection: we trained a one-class SVM anomaly detector on the truthful examples of the *Calibration Phase* and tested on the whole *Testing Phase* (F1 score = 67.7%). Needing only truthful examples makes the models independent from collecting lying examples. This could facilitate the learning, considering, for instance, a humanoid robot that wants to improve the lie detection model online in a supervised way. Finally, thinking about a generic lie detection system, we trained a random forest classifier (F1 score = 71.7%) between-subject to classify false card descriptions from novel individuals. The main difference between the heuristic methods and the machine learning models is in that, the heuristics' knowledge is limited to a single individual. Hence, even if the machine models' performances are worse than the heuristic methods' ones, the formers should be more robust against unexpected behaviors from the participants. Additionally, they offer features that ease their portability on a real-world human-robot interaction i.e., the need of truthful examples only for the one-class SVM model or the ability to classify lies without any previous interaction for the last random forest classifier.

The proposed models are light and independent from any network connections; this makes them suitable to be implemented with extreme simplicity in the context of HRI and avoiding untreatable computation demand. The other advantage of the presented contribution is that the robot can autonomously address all the stages of the interaction keeping the human partner engaged and assessing deceptive behavior in real-time. At the current development stage, the only potential intervention is required if the robot fails to detect the *secret card* at the end of the *Calibration Phase*. However, also this intervention could be made autonomously by the robot: after iCub's detection, the participants have to show the correct *secret card* in order to reject it; iCub could detect the correct card position, thanks to the HSV (Hue, Saturation, Value) color threshold of cards and marks, and hence self-learn the correct *false reference score*.

The current implementation relies on the players' pupil dilation measured with a head mounted eye-tracker, such as the Tobii Pro Glasses 2. This device tends to be dependent on the environmental light condition and could impact the naturalness of the human-robot interaction. We tried to limit the latter factor by removing the calibration step (not needed to measure participants' pupil dilation). However, skipping the calibration, we could not use the other features from the eye-tracker (e.g., gaze orientation). The ideal solution would be to measure a full set of pupillometric features from the RGB cameras embedded on the robotic platform. Recent findings suggest that this approach could be feasible [43]–[45], [70]; hence, we look forward to removing this limitation, making the system completely non-intrusive.

The analysis of pupil dilation revealed that 38% of the participants (N=13 out of 34) presented a lower pupil dilation, during the second phase with respect to the first one. We speculate that this is associated to a reduction in cognitive load and that this effect depends on several factors that contribute to making the *Testing Phase* less stressful. First, in this phase participants do not need to remember the *secret card* and can freely choose how to play the game. As a result, there is no need to prepare in advance the deceptive and creative card description. Moreover, participants are also more used to play the game, even if there are small

differences, and they are more aware of their role and iCub's behavior and capabilities. Additionally, iCub provides a feedback after each card description, eliminating the need to wait for the phase completion to know if the lie had been discovered or not. All these factors could have contributed to decrease of participants' cognitive load. However overall, the interaction has been judged as entertaining and not too cognitively demanding in the questionnaires, suggesting that also the *Calibration phase* was not too challenging for the participants.

We designed the human-robot interaction to validate our lie detection method in an informal interactive scenario. Since the game is based on 84 different cards, with complex and diverse drawings, we speculate that the results we obtain cannot be explained by artifacts on pupil dilation based on the nature of the cards (e.g., different colors, or emotions in the cards' pictures). Hence, we think our approach is modular and generic enough to be ported to different application fields. For instance, in an elderly caregiving scenario, the cards could be pills bottles a patient has to take; the robot could ask the patient if he took the medication, detecting a lie from the patient. Also, the modularity of the end-to-end architecture makes it easy to replace iCub with other robotic platforms, developing a consistent way to present the items based on the application context.

By detecting lies a humanoid robot could evaluate whether the interacting partner is trustworthy or not. Furthermore, the robot could adapt its social behavior over multiple interactions based on this evaluation. However, the system is not perfectly accurate; hence, how the robot should perform its judgment and adaptation should be carefully managed to minimize the impact on the partners' trust toward the humanoid. For instance, in the abovementioned elderly caregiving case, if a caregiver robot detected patient's lies several times it might need to report the patient's behavior to the doctor along with its confidence about the performed measure, rather than accusing the patient to be a liar. In the future, it would be necessary to explore the impact of a misclassification of both truthful and false sentences, on the interacting partners, along with the effect on their trust toward the robot.

Besides the practical applications of detecting lies to assess trustworthiness, the proposed setup, interaction, and methods are based on measuring the task-evoked cognitive load related to creativity. The evaluation is performed in real-time, providing entertainment [46]. This is novel with respect to the long, strictly controlled, and tedious cognitive-load measurement tasks from the literature [37], [38], [41]. For instance, the system could be used to assess creative thinking abilities, before and after a creativity training session [71]. Also, one could use the system to monitor patients' cognitive load during a training task in order to provide a correct support [72], adapt task difficulty [69], evaluate their progress [73] or schedule proper resting sessions [74].

8 Conclusion

In the current manuscript we proposed novel methods to enable robots to detect whether the human partner is lying in a quick and entertaining interaction. We have shown that the detection works and that it is possible to improve it if the model can adapt to each partner during the interaction (F1 score=78%). The approach, however, could still succeed in a first encounter with a new participant (F1 score=71%). The ability to autonomously detect lies could be relevant for robots as a basis to build a model of its human partners' trustworthiness. The naturalness of the approach proposed here would allow to do so without impacting on the sociality of the human-robot interaction. Mutual trust is important to ensure healthy and stable social interactions and this should hold also for HRI. Hence, we believe that novel methods to understand human deceptive behavior will be more and more important in pursuing effective human-robot cooperation.

KEYWORDS

Lie detection, machine learning, adaptation, human-robot interaction, entertainment.

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.

CONFLICTS OF INTERESTS

Not appliable

AVAILABILITY OF DATA AND MATERIAL

The collected dataset will be made available on a public repository (i.e., Zenodo) upon manuscript publication.

CODE AVAILABILITY

The code of both the robot architecture and the post-hoc analysis will be made available on a public repository (i.e., GitHub) upon manuscript publication.

REFERENCES

[1] S. A. Mccornack and M. R. Parks, "Deception Detection and Relationship Development: The Other Side of Trust," *Ann. Int. Commun. Assoc.*, vol. 9, no. 1, pp. 377–389, Jan. 1986, doi: 10.1080/23808985.1986.11678616.

[2] G. Lucas, S. Lieblich, and J. Gratch, "Trust Me : Multimodal Signals of Trustworthiness," in *ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 5–12, doi: https://doi.org/10.1145/2993148.2993178.

[3] M. Rueben, A. M. Aroyo, C. Lutz, J. Schmolz, P. Cleynenbreugel, A. Corti, S. Agrawal, and W. Smart, "Themes and Research Directions in Privacy-Sensitive Robotics," *EEE Work. onAdvanced Robot. its Soc. Impacts*, 2018.

[4] 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," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.

[5] A. Freedy, D. Ph, G. Weltman, D. Ph, and U. S. A. Rdecom-sttc, "Measurement of Trust in Human-Robot Collaboration," in *International Symposium on Collaborative Technologies and Systems*, 2007, pp. 106–114, doi: 10.1109/CTS.2007.4621745.

[6] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti, "Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble?," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3701–3708, 2018, doi: 10.1109/LRA.2018.2856272.

[7] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum. Comput. Stud.*, vol. 58, pp. 697–718, 2003, doi: 10.1016/S1071-5819(03)00038-7.

[8] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. C. Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," in *PIE - The International Society for Optical Engineering*, 2014, no. February 2015, p. 90840E, doi: 10.1117/12.2050622.

[9] N. Wang, D. V Pynadath, S. G. Hill, N. Wang, and D. V Pynadath, "Building Trust in a Human-Robot Team with Automatically Generated Explanations Building Trust in a Human-Robot Team with Automatically Generated Explanations," *Proc. Interservice/Industry Training, Simul. Educ. Conf.*, no. 15315, pp. 1–12, 2015.

[10] S. Agrawal and H. Yanco, "Feedback Methods in HRI: Studying their effect on Real-Time Trust and Operator Workload," *HRI'12 - Proc. 7th Annu. ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 73–80, 2012, doi: 10.1145/2157689.2157702.

[11] B. M. Muir, "Trust between humans and machines , and the design of decision aids," *Int. J. Man. Mach. Stud.*, vol. 27, no. 5–6, pp. 527–539, 1987, doi: https://doi.org/10.1016/S0020-7373(87)80013-5.

[12] A. M. . Aroyo, D. . Pasquali, A. . Koting, F. . Rea, G. . Sandini, and A. Sciutti, "Perceived differences between on-line and real robotic failures," 2020.

[13] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," *Proc. seventh Annu. ACM/IEEE Int. Conf. Human-Robot Interact. - HRI '12*, p. 73, 2012, doi: 10.1145/2157689.2157702.

[14] K. E. Schaefer, "The Perception Measurement of Human-Robot Trust," 2013.

[15] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *Int. J. Soc. Robot.*, vol. 8, no. 2, pp. 193–209, 2016, doi: 10.1007/s12369-015-0333-8.

[16] R. E. Yagoda and D. J. Gillan, "You Want Me to Trust a ROBOT? The Development of a Human–Robot Interaction Trust Scale," *Int. J. Soc. Robot.*, no. April 2014, 2012, doi: 10.1007/s12369-012-0144-0.

[17] S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi, "Would a robot trust you? Developmental robotics model of trust and theory of mind," *CEUR Workshop Proc.*, vol. 2418, p. 74, 2019, doi: https://doi.org/10.1098/rstb.2018.0032.

[18] M. Patacchiola and A. Cangelosi, "A Developmental Cognitive Architecture for Trust and Theory of Mind in Humanoid Robots," *IEEE Trans. Cybern.*, pp. 1–13, 2020, doi: 10.1109/TCYB.2020.3002892.

[19] 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.

[20] 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.

[21] S. M. Kassin, "On the Psychology of Confessions: Does Innocence Put Innocents at Risk?," *Am. Psychol.*, vol. 60, no. 3, pp. 215–228, Apr. 2005, doi: 10.1037/0003-066X.60.3.215.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] J. G. May, R. S. Kennedy, M. C. Williams, W. P. Dunlap, and J. R. Brannan, "Eye movement indices of mental workload," *Acta Psychol. (Amst).*, vol. 75, no. 1, pp. 75–89, 1990, doi: 10.1016/0001-6918(90)90067-P.

[29] M. Nakayama and Y. Shimizu, "Frequency analysis of task evoked pupillary response and eye-movement," in *Proceedings of the Eye tracking research & applications symposium on Eye tracking research & applications - ETRA '2004*, 2004, pp. 71–76, doi: 10.1145/968363.968381.

[30] K. F. Van Orden, W. Limbert, S. Makeig, and T. P. Jung, "Eye activity correlates of workload during a visuospatial memory task.," *Hum. Factors*, vol. 43, no. 1, pp. 111–21, 2001, doi: 10.1518/001872001775992570.

[31] J. A. Stern, L. C. Walrath, and R. Goldstein, "The endogenous eyeblink.," *Psychophysiology*, vol. 21, no. 1, pp. 22–33, Jan. 1984, Accessed: Jul. 07, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6701241.

[32] B. C. Goldwater, "Psychological significance of pupillary movements," *Psychol. Bull.*, vol. 77, no. 5, pp. 340–355, May 1972, doi: 10.1037/h0032456.

[33] Andreassi, *Psychophysiology: Human Behavior and Physiological Response.* Psychology Press, 2010.

[34] S. Mathôt, "Pupillometry: Psychology, Physiology, and Function," *J. Cogn.*, vol. 1, no. 1, Feb. 2018, doi: 10.5334/joc.18.

[35] J. Beatty and B. Lucero-Wagoner, "The pupillary system.," *Handb. Psychophysiol. 2*, 2000.

[36] 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.

[37] J. Gonzalez-Billandon, A. M. Aroyo, A. Tonelli, D. Pasquali, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a Robot Catch You Lying? A Machine Learning System to Detect Lies During Interactions," *Front. Robot. AI*, vol. 6, Jul. 2019, doi: 10.3389/frobt.2019.00064.

[38] A. M. Aroyo, J. Gonzalez-Billandon, A. Tonelli, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a Humanoid Robot Spot a Liar?," *IEEE-RAS 18th Int. Conf. Humanoid Robot.*, 2018.

[39] A. Szulewski, N. Roth, and D. Howes, "The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise," *Acad. Med.*, vol. 90, no. 7, pp. 981–987, Jul. 2015, doi: 10.1097/ACM.0000000000000677.

[40] M. Ahmad, B. Jasmin, L. Katrin, and F. Eyssel, "Trust and Cognitive Load During Human-Robot Interaction," 2019.

[41] J. Klingner, "Measuring Cognitive Load During Visual Task by Combining Pupillometry and Eye Tracking," *Ph.D. Diss.*, no. May, 2010, doi: http://purl.stanford.edu/mv271zd7591.

[42] G. Hossain and M. Yeasin, "Understanding effects of cognitive load from pupillary responses using hilbert analytic phase," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 381–386, 2014, doi: 10.1109/CVPRW.2014.62.

[43] C. Wangwiwattana, X. Ding, and E. C. Larson, "PupilNet, Measuring Task Evoked Pupillary Response using Commodity RGB Tablet Cameras," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–26, Jan. 2018, doi: 10.1145/3161164.

[44] S. Rafiqi, E. Fernandez, C. Wangwiwattana, S. Nair, J. Kim, and E. C. Larson, "PupilWare: Towards pervasive cognitive load measurement using commodity devices," *8th ACM Int. Conf. PErvasive Technol. Relat. to Assist. Environ. PETRA 2015 - Proc.*, no. August, 2015, doi: 10.1145/2769493.2769506.

[45] S. Eivazi, T. Santini, A. Keshavarzi, T. Kübler, and A. Mazzei, "Improving real-time CNN-based pupil detection through domain-specific data augmentation," *Eye Track. Res. Appl. Symp.*, 2019, doi: 10.1145/3314111.3319914.

[46] 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, doi: https://doi.org/10.1145/3434073.3444682.

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

[48] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems - PerMIS '08*, 2008, p. 50, doi: 10.1145/1774674.1774683.

[49] G. B. (Universita di M.-B. Flebus, "Versione italiana dei big five markers di goldberg," 2006.

[50] C. J. Ferguson and C. Negy, "Development of a brief screening questionnaire for histrionic personality symptoms," *Pers. Individ. Dif.*, vol. 66, pp. 124–127, 2014, doi: 10.1016/j.paid.2014.02.029.

[51] D. N. Jones and D. L. Paulhus, "Introducing the Short Dark Triad (SD3): A Brief Measure of Dark Personality Traits," *Assessment*, vol. 21, no. 1, pp. 28–41, 2014, doi: 10.1177/1073191113514105.

[52] F. Bracco and C. Chiorri, "Versione Italiana del NASA-TLX."

[53] Tobii Pro, "Quick Tech Webinar - Secrets of the Pupil." https://www.youtube.com/watch?v=I3T9Ak2F2bc&feature=emb_title.

[54] P. Fitzpatrick, G. Metta, and L. Natale, "Towards long-lived robot genes," *Rob. Auton. Syst.*, vol. 56, no. 1, pp. 29–45, Jan. 2008, doi: 10.1016/j.robot.2007.09.014.

[55] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory.* Psychology of learning and motivation, volume 55. Elsevier., 2011.

[56] J. Leppink, "Cognitive load theory: Practical implications and an important challenge," *J. Taibah Univ. Med. Sci.*, vol. 12, no. 5, pp. 385–

391, 2017, doi: 10.1016/j.jtumed.2017.05.003.

[57] A. K. Webb, C. R. Honts, J. C. Kircher, P. Bernhardt, and A. E. Cook, "Effectiveness of pupil diameter in a probable-lie comparison question test for deception," *Leg. Criminol. Psychol.*, vol. 14, no. 2, pp. 279–292, Sep. 2009, doi: 10.1348/135532508X398602.

[58] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel, "Safe and sensible preprocessing and baseline correction of pupil-size data," *Behav. Res. Methods*, vol. 50, no. 1, pp. 94–106, 2018, doi: 10.3758/s13428-017-1007-2.

[59] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "TSFEL: Time Series Feature Extraction Library," *SoftwareX*, vol. 11, Jan. 2020, doi: 10.1016/j.softx.2020.100456.

[60] 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.

[61] L. Breiman, "Random forests," *Mach. Learn.*, 2001, doi: 10.1023/A:1010933404324.

[62] Nitesh V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, 2006, doi: 10.1613/jair.953.

[63] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," *Iarjset*, no. April, pp. 20–22, 2015, doi: 10.17148/iarjset.2015.2305.

[64] H. S. Ahn, I. K. Sa, D. W. Lee, and D. Choi, "A playmate robot system for playing the rock-paper-scissors game with humans," *Artif. Life Robot.*, vol. 16, no. 2, pp. 142–146, 2011, doi: 10.1007/s10015-011-0895-y.

[65] I. Gori, S. R. Fanello, G. Metta, and F. Odone, "All gestures you can: A memory game against a humanoid robot," *IEEE-RAS Int. Conf. Humanoid Robot.*, pp. 330–336, 2012, doi: 10.1109/HUMANOIDS.2012.6651540.

[66] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing Models of Disengagement in Individual and Group Interactions," *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2015-March, no. March, pp. 99–105, 2015, doi: 10.1145/2696454.2696466.

[67] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "The design and development of a Lie Detection System using facial micro-expressions," in *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, Dec. 2012, pp. 33–38, doi: 10.1109/ICTEA.2012.6462897.

[68] K. Kobayashi and S. Yamada, "Human-Robot interaction design for low cognitive load in cooperative work," *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, no. April, pp. 569–574, 2004, doi: 10.1109/ROMAN.2004.1374823.

[69] S. M. Al Mahi, M. Atkins, and C. Crick, "Learning to assess the cognitive capacity of human partners," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 63–64, 2017, doi: 10.1145/3029798.3038430.

[70] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE : Real-Time Eye Gaze Estimation in Natural Environments," in *European Conference on Computer Vision*, 2018, pp. 339–357, doi: 10.1007%2F978-3-030-01249-6_21.

[71] J. L. Redifer, C. L. Bae, and M. Debusk-lane, "Implicit Theories , Working Memory , and Cognitive Load : Impacts on Creative Thinking," 2019, doi: 10.1177/2158244019835919.

[72] G. Belgiovine, F. Rea, J. Zenzeri, and A. Sciutti, "A Humanoid Social Agent Embodying Physical Assistance Enhances Motor Training Experience," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, no. ii, doi: 10.1109/RO-MAN47096.2020.9223335.

[73] A. Koenig, D. Novak, X. Omlin, M. Pulfer, E. Perreault, L. Zimmerli, M. Mihelj, and R. Riener, "Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training.," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 4, pp. 453–64, Aug. 2011, doi: 10.1109/TNSRE.2011.2160460.

[74] A. Westbrook and T. S. Braver, "Cognitive effort: A neuroeconomic approach," *Cognitive, Affective and Behavioral Neuroscience*, vol. 15, no. 2. Springer New York LLC, pp. 395–415, Jun. 22, 2015, doi: 10.3758/s13415-015-0334-y.