

Detecting Lies in the Wild: Creativity and Learning @ the Maker Faire Rome

Dario Pasquali^{1,*}, Francesco Rea² and Alessandra Sciutti¹

¹*COgNiTive Architectures for Collaborative Technologies (CONTACT) - Istituto Italiano di Tecnologia (IIT), Via Enrico Melen 83, Genova (Italy)*

²*Robotics Brains and Cognitive Sciences (RBCS) - Istituto Italiano di Tecnologia (IIT), Via Enrico Melen 83, Genova (Italy)*

Abstract

Creativity is one of the most powerful skills humans can rely on to overcome daily challenges. While most of the research has focused on the positive facet of creativity, like problem-solving and art, a few contributions explored its dark side: lying and deception. Virtual and embodied intelligent agents approaching the real world will soon face humans' deception with poor means to understand and unmask it. In a previous study, we asked participants to describe a set of gaming cards to the humanoid robot iCub, either describing what they saw or producing a creative and deceiving description; The robot autonomously classified players' behavior with a pupillometry-based heuristic method. After collecting an in-laboratory dataset, we trained a Random Forest classifier enabling the humanoid robot iCub to detect lie-related creativity autonomously during informal human-robot interactions. In this manuscript, aiming at real-world applications, we challenged our classifier on a diametrically opposite environment: the Maker Faire Rome 2022. Moreover, we compared its performance with respect to an Adaptive Random Forest twin, able to learn online after each interaction. The performance of the two models and the detection of concept and data drift give relevant insight into how adaptivity would be the key to developing more effective intelligent agents.

Keywords

Lie Detection, Creativity, Human-Robot Interaction, Incremental Learning, In the wild

1. Introduction

Humans' creativity is highly subjective; from problem-solving to art production, humans rely on creativity to survive and develop nowadays' society. Traditionally, creativity has been related to the originality and appropriateness of people's creative products and the ability to generate novel and effective ideas [1]. While literature mainly focuses on the more intuitive, art-related understanding of creativity, recent studies started exploring the creative process embedded in lying and deception [2, 3, 4, 5]. Indeed, creativity could also be used for negative purposes, with different degrees of malice [2]. Focusing on everyday social interaction, the "white lies" - lies not meant to harm others - are the most diffused negative creative attempts. Everybody lies [6], with an average of two lies per day [7] or even higher frequencies - 60% of the participants in


CREAI 2022 - Workshop on Artificial Intelligence and Creativity, November 28 - December 2, 2022, Udine, IT

*Corresponding author.

✉ dario.pasquali@iit.it (D. Pasquali); francesco.rea@iit.it (F. Rea); alessandra.sciutti@iit.it (A. Sciutti)

🆔 0000-0001-8185-8188 (D. Pasquali); 0000-0001-8535-223X (F. Rea); 0000-0002-1056-3398 (A. Sciutti)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

[8] lied at least once in a 10-minutes dialogue. For instance, we lie to present ourselves better than we are [7], to persuade others [9], or to avoid undesired conversations [10].

Despite the high impact of lying on social interactions, humans perform poorly on recognizing liars - the average accuracy is 47% on lie detection and 65% on recognizing true statements [11]. Several technical attempts have been developed to compensate for our poor performance and grasp the lying creative process. State-of-the-art solutions usually monitor the physiological proxies (e.g., skin conductance, respiration rate, heartbeat, blood pressure, or pupil diameter) known to reflect cognitive load and emotional arousal variations. Indeed, it has been proved how the fabrication and maintenance of a creative and coherent lie produce an increased cognitive effort [7, 12] and emotional arousal [13] with respect to truth-telling. Traditional lie detection methods rely on fMRI images [14], skin temperature variation [15], micro-expressions [16], photoplethysmography [17] or acoustic prosody [18]. Finally, the polygraph - the most famous "truth machine", debunked in [12] - merges a multi-modal set of physiological measures [19]. However, most of these methods are invasive, expensive, not autonomous, or dependent on expert figures, making them poorly applicable in everyday social interactions.

The problem becomes even more crucial when considering the novel artificial agents that will soon be part of our society. Artificially intelligent agents, either virtual or embodied in robots, will sooner or later clash with humans' deceptive behaviors, with scarce means to comprehend them or to understand when others are trustworthy [20, 21, 22]. Robots could hardly equip most of the mentioned technical solutions without requiring an expert operator or affecting the (in)formality of the social Human-Robot Interaction (HRI) (e.g., by touching the human partners to assess their skin conductance). Also, intelligent agents are usually trained in closed laboratory environments under strict and controlled context assumptions; hence, they would lack humans' ability to adapt and learn from daily experiences.

1.1. Detecting Lies in Human-Robot Interaction

This study is part of a four-year research project to enable the humanoid robot iCub [23] to detect lies in social human-robot interactions (HRI).

We identified pupillometry [24, 25, 26, 27], in particular the Task Evoked Pupillary Responses (TEPRs) [28], as a minimally invasive proxy to detect deceptive behaviors. Such measures have proved to reflect cognitive load increases relative to deception and lie creation with respect to truth-telling [29]. We opted for this metric because it is hardly intentionally controllable and is measurable with unobtrusive devices usually embedded in everyday objects (i.e., glasses) [30, 31]. Furthermore, recent findings prove it would soon be feasible to measure TEPRs from standard RGB cameras [32, 33, 34] (e.g., the ones equipped on robots). Such systems still require users to stay uncomfortably close to the camera; however, they make pupillometry-based lie detection a good candidate for human-robot interaction (HRI). Based on these assumptions, we first proved that the same TEPRs studied in human-human interaction also happen in HRI during a formal interrogatory-like scenario [35, 36]. However, as other attempts in literature [37, 38], our findings were limited to formal interrogatory contexts. Hence, we explored whether the same effects would happen in informal interactions. For this purpose, we generalized the

concept of lying with its core component: *creativity*. To be precise, it is wrong to consider any creative attempt as a lie; however, lying implicitly embeds a creative effort [3]. Hence, modeling humans' creative process could help intelligent systems to better grasp lying and deception during everyday social interaction.

For this purpose, we asked 39 participants to play a game with the humanoid robot iCub. Players described a set of gaming cards from the Dixit Journey card game¹ to the robot either in a *descriptive* - i.e., by narrating what they saw in the card -, or in a *creative* way - i.e., by inventing a fake card description. iCub detected players' *creative* attempts, autonomously and in real-time, only based on their pupil dilation, streamed from a Tobii Pro Glasses 2 eye-tracker. More precisely, the game was composed of two phases. In the first phase, the players described six cards, creating a fake description only for one of them they knew in advance; the iCub achieved an accuracy of 83% on detecting the single *creative* description by selecting the one relative to the highest mean pupil dilation among the six [39]. Moreover, the robot leveraged this brief interaction to learn an internal model of how the specific human partner was *descriptive* and *creative*: it stored the mean pupil dilation for the creative card description (*creative reference score*), and the average of the mean pupil dilation for the other descriptions (*descriptive reference score*). In the second phase, the iCub used this model to classify further card descriptions as *creative* or *descriptive*. Players were asked to describe six new cards, deciding for each one whether to narrate what they saw or fabricate a fake description. The robot computed the absolute distance of the mean pupil dilation for each new card description with respect to the two *reference scores* and assigned the label relative to the closest one. Such a simple heuristic allowed iCub to achieve an accuracy of 73% [20]. Posthoc, we used the second-phase data - 409 datapoints re-balanced through the SMOTE algorithm [40] - to train a Random Forest classifier to detect creativity with an F1-score of 71%. This model is supposed to be more generic and robust than the subjective heuristic employed during the game. Aiming at applying our system in everyday life one important question remains open:

How would our creative detection system perform "in the wild"?

1.2. Online Interactive Adaptation

Moving in-lab trained machine learning systems to a real context is non-trivial. Some potential issues that could arise are relative to the specific solution adopted: in our case, the Tobii eye-tracker is sensitive to environmental illumination; also, players' cognitive load and arousal depend on a multitude of factors (e.g., emotion, stress, presence of distractions, ...), other than the act of being *descriptive* or *creative*. Other issues are known to affect all machine learning models ported in realistic environments: *Data Drift*, the distribution of data in the real world might be different from the ones on which the model is trained; and *Concept Drift*, the relation between the dependent and independent variables (i.e., the pattern learned by the model) can be different [41]. In both cases, the result is a decreasing performance over time. Training the models on a more extensive and representative dataset should help, but it could not be enough or even possible, like in our lie detection scenario.

¹<https://boardgamegeek.com/boardgame/121288/dixit-journey>

A more feasible solution to face data and concept drift is online learning. As students need to adapt what they learned once they start their first job after university, also machine learning models should be able to learn and adapt to the real environment wherein they are employed. This is particularly true for models for social human-robot interaction (HRI). Through experience and interactive feedback from human pals, it would be possible to improve and adapt in-lab machine-learning models to everyday life.



Figure 1: (Left) The Maker Faire game setup with the real-time plot of visitors' pupil dilation on the screen, the Tobii Pro Glasses 2 eye-tracker, and Dixit Journey cards; (Right) and example of the described card.

For this purpose, we challenged our Random Forest *creativity* classifier and explored the effect of online interactive adaptation on a diametrically opposite context: the 10th Maker Faire Rome 2022² (see Figure 1). Visitors played a simplified version of our interactive card game, adapted as a Human-Computer Interaction demonstration. As in the laboratory, the game tried to classify players' *descriptive* and *creative* card descriptions based on pupil dilation, real-time streamed from the Tobii eye-tracker. Two classifiers processed visitors' pupillometry in parallel:

Random Forest (RF) The model is a simplified version of the one presented in [20]. It is pre-trained on 409 card descriptions (225 truthful and 184 deceptive) collected in our in-lab experiment; for each card, the following features are computed: *duration* of the description; *average*, *minimum*, *maximum*, *skewness*, *absolute energy* and *slope* of the pupil dilation. We randomly selected 75% as training set, re-balanced it with the SMOTE algorithm, and performed a 4-fold grid-search cross-validation to identify the following best parameters: *n_estimators=5*, *split_criterion=entropy*, *max_depth=2*, *max_features=log2*, *bootstrap=True*. The best model achieved an accuracy of 71.8% and an F1-score of 70.7% on the test set. Finally, we trained a comprehensive model on the full dataset, achieving a training accuracy of 70.7% and an F1-score of 68.8%.

²<https://makerfairerome.eu/en/>

Adaptive Random Forest (ARF) As a comparison, we trained an Adaptive Random Forest³ [42] able to both start from a pre-trained knowledge and to adapt online based on visitors' interactive feedback. We used the same dataset and features of the RF classifier; however, we selected the following parameters: $n_estimators=5$, $max_features=log2$, $split_criterion=nba$, and $max_depth=2$ - please notice the entropy-based split criterion is not implemented in the python library we used to implement the ARF. The model, learning from one datapoint after the other of the in-laboratory dataset, achieved an accuracy of 68.5% and an F1-score of 62.8%. Even if the performance is worse than the other model, the Adaptive Random Forest can improve online.

2. Methods

The Maker Faire took place in Rome from Friday the 7th to Sunday the 9th of October 2022. At the expo, we presented the game as one of the possible scientific demonstration visitors could interact with at our stand. The game was meant to collect data in the wild and disseminate our research, teaching visitors about pupillometry, its relation to cognitive load, and its potential application in robotics and other fields. 146 visitors played the game (58 on Friday, 45 on Saturday, and 43 on Sunday).

2.1. Setup

Players played the game while standing. A graphical user interface was presented in front of them on a 15" laptop. Next to the computer lied the Tobii Pro Glasses 2 eye-tracker and the deck of 80 Dixit Journey gaming cards (see Figure 1).

2.2. Procedure

The experimenter explained to the visitors that, through the Tobii Pro Glasses 2 eye-tracker, the game would read their right-eye pupil dilation, stream it in real-time, and store it in an anonymized format. If the visitors agreed to collect such data, the experimenter asked them to wear the eye-tracker and started a new match - we did not perform the eye-tracker calibration as not necessary to measure pupil dilation [43].

The game showed a real-time plot of visitors' right-eye pupil diameter. For dissemination purposes, the experimenter explained how pupils change due to environmental illumination (i.e., asking them to look at a lamp and then watch the pupil decrease) and cognitive effort (i.e., asking them to perform math calculations) - the procedure was also meant to verify the correct measurement of the device. Then, the match started.

For each match, players were asked to describe *at least* three cards: for the first one, they had to describe what they saw (i.e., be *descriptive*); for the second one, they had to be *creative* and to deceive; from the third card onward they could decide whether to be *descriptive* or *creative*. For each card, the game tried to classify it as *descriptive* or *creative*, provided a classification to the visitor, and asked for honest feedback. Please notice that classifications were also produced for the first two cards; however, showing them to the visitors was meaningless since they

³<https://riverml.xyz/dev/api/ensemble/AdaptiveRandomForestClassifier/>

were instructed on how to behave. The controlled procedure was meant to collect a dataset as balanced as possible while letting players challenge the system as they wanted.

Visitors were not limited in the descriptions' duration nor the number of described cards. The experimenter manually clicked a GUI button to start and stop the data collection during the card description and validated or rejected the classification based on visitors' feedback. Once visitors decided to end the match, the experimenter removed the eye-tracker and explained how the data were processed and the classification performed (see next section).

2.3. Data Processing

Visitors' right-eye pupil dilation was streamed from the Tobii eye-tracker at a frequency of 20 Hz. The game continuously accumulates data in a 1-second baseline queue (i.e., 20 datapoints). During the descriptions (i.e., between the start and stop GUI button clicks), the pupil data were instead stored in a separate buffer specific to that card. At the end of the description, both the baseline and the timeseries relative to the card were cleaned with a median outlier filter and smoothing sliding window; the timeseries was baseline-normalized, subtracting the mean pupil dilation during the baseline [27], and the features (see section 1.2) were computed. Both models classified the card description in parallel; however, only the Adaptive Random Forest classification was presented to the visitor. We opted for this solution to disseminate another piece of science: an incremental interactive model able to progressively improve thanks to visitors' interactions. Finally, the data - both raw and cleaned timeseries and baseline, along with the extracted features - and the real label for each description were stored for posthoc analysis.

3. Results

The 146 visitors described 532 cards - an average of 3.75 (SD=0.90) cards each. They were free to decide whether to be *creative* or *descriptive* from the third card onward; hence, the card dataset is slightly unbalanced with 274 *creative* and 258 *descriptive* attempts.

Starting from the stored features, we applied a standardized outlier detection procedure: we discarded (i) all the descriptions associated with average pupil dilation, slope and duration far more than three times the subjective mean for each feature; and (ii) all the descriptions shorter than 1 second. Also, we excluded 4 visitors since they decided to leave after performing only one or two descriptions. The resulting dataset comprises 515 descriptions (269 *creative* and 246 *descriptive*).

3.1. In-Game Performance

The Random Forest (RF) classifier achieved an accuracy of 59.1%, while the Adaptive Random Forest (ARF) only achieved a 54.4% of accuracy. However, a Chi-square Pearson's test showed that the latter performance was not statistically lower ($z=1.51$, $p=0.065$). On the F1-score instead, the RF classifier achieved a performance of 55.2%, significantly lower than the 66.3% achieved by the ARF ($z=-2.70$, $p=0.003$). With respect to their in-lab performances, both models worsened: the Chi-square test showed significantly lower accuracy for both RF ($z=3.66$, $p<0.001$) and

ARF ($z=4.40$, $p<0.001$); however, while the Random Forest F1-score was significantly lower in-the-wild ($z=2.86$, $p=0.002$), the difference was higher - but not significantly - for the Adaptive counterpart ($z=-0.79$, $p=0.22$).

To better understand such performance differences, we explored the presence of data and concept drift on the fair with respect to the lab dataset. We composed a comprehensive dataset, including data from both environments. A Shapiro-Wilk normality test showed that the data were not normally distributed, so we opted for a non-parametric analysis.

3.1.1. Data Drift Analysis

In our context, a *data drift* would represent a different distribution of the behavioral and pupillometry features between the two environments. We fitted a set of mixed effect models, one for each feature (*duration* of the description; *average*, *minimum*, *maximum*, *skewness*, *absolute energy* and *slope* of the pupil dilation), as dependent variable; we entered a fixed effect "environment" (two levels: *lab*, *faire*; with reference on the *lab*), and participants' ID as random effect. Mixed effect models can represent the differences in distribution given by an independent fixed factor (i.e., the environments), while taking into account and grouping datapoints characterized by the same subjective random difference (i.e., the set of cards described by the same participant). Furthermore, they allow using the entire dataset, even if the distributions are unbalanced (e.g., the different number of described cards and participants in the two environments)

Participants took slightly more time to describe the cards at the *faire* ($B=4.54$, $t=2.55$, $p=0.012$); we speculate this could be due to the more informal context and the lack of the turn-taking mechanic of the original game. Regarding the pupillometry features, participants' average ($B=-0.151$, $t=-3.68$, $p<0.001$), minimum ($B=-0.317$, $t=-6.75$, $p<0.001$) and absolute energy ($B=-81.9$, $t=-9.24$, $p<0.001$) of the pupil dilation were lower at the *faire*. Finally, the maximum ($B=-0.018$, $t=-0.358$, $p=0.721$), skewness ($B=0.014$, $t=0.315$, $p=0.753$) and slope ($B=-2.11$, $t=-1.51$, $p=0.134$) of the pupil dilation were not statistically different. We speculate this difference could have been caused by the higher illumination of the fair, inducing, and averagely lower pupil dilation.

3.1.2. Concept Drift Analysis

Then, we looked for *concept drifts*, i.e., differences in the relationship between *descriptive* and *creative* attempts in the two environments. We fitted another set of mixed effects models on the same features; we entered two fixed effects "environment" (two levels: *lab*, *faire*; with reference on the *lab*) and "behavior" (two levels: *descriptive*, *creative*; with reference on *descriptive*), along with a random effect on participants' ID.

Interestingly, only the maximum - *lab*:($B=0.194$, $t=6.32$, $p<0.001$); *faire*:($B=0.92$, $t=3.61$, $p<0.001$) - and the slope - *lab*:($B=12.0$, $t=10.9$, $p<0.001$); *faire*:($B=3.6$, $t=3.42$, $p<0.001$) - of the pupil dilation preserved the same pattern among the two environments: they were higher when being *creative*, even if the effect was stronger in the *lab*. On the duration, the difference between *creative* and *descriptive*, not significant in the *lab* ($B=-0.134$, $t=0.121$, $p=0.904$), was highly significant at the *faire* ($B=5.542$, $t=5.7$, $p<0.001$), with longer descriptions for *creative* attempts. On the contrary, the average - *lab*:($B=0.243$, $t=9.2$, $p<0.001$); *faire*:($B=0.01$, $t=0.521$, $p=0.603$) -, absolute energy - *lab*:($B=34.84$, $t=4.99$, $p<0.001$); *faire*:($B=0.03$, $t=0.01$, $p=0.996$) -, and skewness - *lab*:($B=-0.289$,

$t=-9.73$, $p<0.001$); *faire*:($B=0.02$, $t=0.69$, $p=0.493$) - of the pupil dilation were higher when being *creative* in the *lab*, but not at the *faire*. Finally, the minimum pupil dilation showed a significant effect in both environments, but with opposite valence: it was higher when being *creative* in the *lab* ($B=0.262$, $t=8.86$, $p<0.001$), and the other way round at the *faire* ($B=-0.06$, $t=-2.47$, $p=0.014$).

3.2. Temporal Evolution and Learning

Given such different distributions and patterns, it is not surprising that the two models performed worst at the *faire*. The crucial factor is whether the classifiers can learn and adapt to the novel environment. To better understand the learning process of the Adaptive Random Forest, we observed the temporal evolution of the accuracy and F1-score curves.

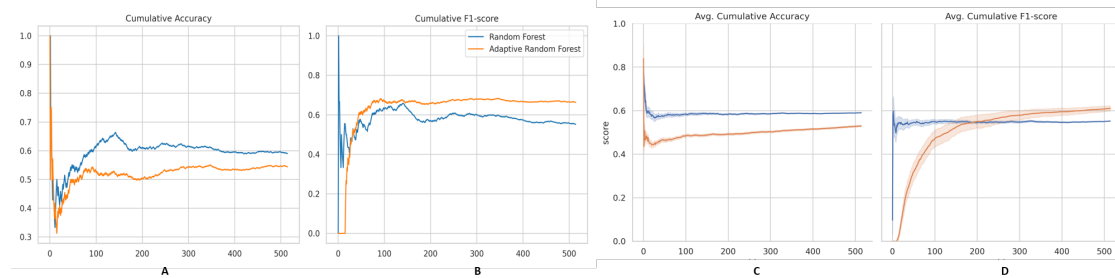


Figure 2: (A and B) Cumulative Accuracy (A) and F1-score (B) at the Maker Faire Rome 2022 for the Random Forest (blue) and Adaptive Random Forest (ARF) models. (C and D) Simulated Accuracy and F1-score over 30 randomized permutations of the visitors

Figure 2 (A-B) shows the cumulative accuracy, and F1-score of the two models as visitors played the game. Please remember that the accuracy metric balances the performances on detecting both *creative* and *descriptive* attempts; the F1-score instead focuses on how good the models are on detecting the *creative* class only. Both metrics of the Random Forest classifier converge to a plateau; for the Adaptive Random Forest instead, the two metrics are increasing, particularly visible for the F1-score. To verify whether the adaptability of the latter model produces the progressive increase it is first necessary to mitigate the order effect of how visitors interacted with the game. Indeed the order by which examples are provided could impact the learning rate and the convergence time of an online model [44].

Hence, using the *faire* dataset only, we simulated 30 random permutations of the visitors interacting with the game (i.e., keeping the order of the described card unaffected). For each permutation, we generated and evaluated two new models. Figure 2 (C-D) shows the simulation’s cumulative accuracy and F1-score. As it is possible to see, the ARF model’s final accuracy ($M=52.9\%$, $SD=0.02\%$) and F1-score ($M=61\%$, $SD=0.07\%$) are poorly affected by the order - the final RF metrics are unaffected, as expected. Also, the increasing and converging patterns of the RF and ARF curves are preserved for accuracy and F1-score. To better quantify this increasing trend, we analyzed the average slope of the two metric curves, for both models, among the 30 simulated permutations. Given the normal distribution of the slope averages, we opted for a parametric test. A paired t-test showed that the slopes of both accuracy ($t=3.91$, $p<0.001$) and F1-score ($t=13.22$, $p<0.001$) of the ARF were higher than the RF ones - please notice the same

statistically significant difference was present also for the precision ($t=8.25$, $p<0.001$), recall ($t=15.51$, $p<0.001$) and ROCAUC ($t=3.8$, $p<0.001$). Hence, with more interacting players, the ARF model's learning and adaptation ability would make it surpass and overcome the other static classifier.

4. Discussion & Conclusion

In this manuscript, we tested the performance of our trained lying (i.e., creativity) classifier in the chaotic environment of the Maker Faire Rome 2022. Grasping humans' creative process is a highly complex problem in a closed and controlled laboratory, even more when approaching realistic environments. As proved by humans, the key to "surviving" in the wild is learning and adaptation. Hence, we compared the performance of our static model with respect to an adaptive counterpart trained on the same data and features.

As expected, the Maker Faire dataset was both data and concept drifted with respect to the in-laboratory collected data. As a result, even if the two models discriminated *descriptive* and *creative* card descriptions better than chance, their performance decreased with respect to the testing ones. However, by analyzing the average slope of the accuracy and F1-score metrics, we showed how the Adaptive Random Forest was learning and adapting to the novel environment. It already surpassed the static counterpart on the F1-score metrics (i.e., the goodness on recognizing the *creative* attempts only); moreover, the positive slopes of the curves suggest it would also improve the accuracy and other metrics.

In the manuscript, we omitted the case where the static Random Forest is retrained with the novel examples (e.g., at the end of each day). Even if such a solution is reasonable from a pure computer science point of view, it could not be the best option thinking about an intelligent agent (e.g., a humanoid robot) embedding the lie detection system in an actual application. Indeed, the speed by which a model adapts and improves is crucial in human-robot and human-computer interaction. Aiming to maximize the quality of each interactive session, a daily-retrained Random Forest would not compete with a per-interaction adaptive model.

Both models are still limited in the timing they provide classifications: before classifying players' behavior, they have to observe the full card description. It could be more effective - and interesting - to recognize creativity and deception from sub-segments of the interactions, i.e., by incrementally gaining confidence until one of the classes is grounded. Focusing on the adaptive interactive classifier, the model still depends on players' explicit feedback (i.e., validating or rejecting the classification). In a realistic environment, it would not always be possible to access an explicit ground truth, especially in the lie detection field. For this purpose, we speculate that subjective adaptation and reliance on past experiences could effectively recognize implicit signals to validate the classifications.

Generally speaking, our system is still limited by relying on pupillometry only. Even if literature proves pupillometry is affected by lying [29] and creativity [13], pupil fluctuations are not one-to-one bound to such behaviors. They rather reflect variations in cognitive effort and emotional arousal. To be reliable in evaluating humans' creativity, they must be considered with respect to a specific context or task (i.e., Task Evoked Pupillary Responses [28]). Hence, intelligent systems aiming to grasp humans' reactions based on pupillometry,

or other physiological effects, must be able to model the context in which the interaction takes place, reporting humans' reactions to it. The second general limitation is the usage of gaming cards as a creative medium. We opted for Dixit gaming cards because they are designed to stimulate creativity and divergent thinking. However, aiming for realistic interactions, the medium should be generalized (e.g., using generic pictures of photographs) or even removed.

To improve our system, we are pursuing this research in four parallel directions: (i) classifying humans' behavior as nearly real-time by processing sequential chunks of the card descriptions; (ii) developing subjective models able to learn how a specific human partner lies and reacts when a classification is provided, using such knowledge to both improve the classification and recognize implicit feedback; (iii) including multiple modalities in the classification (i.e., posture and verbal prosody) to understand better the context in which the interaction happens and to improve the robustness of the model; and (iv) looking for innovative solutions[32, 33, 34] to measure pupillometry from standard RGB cameras. Besides the applications in the lie detection and creativity understanding fields, our research is based on evaluating humans' cognitive effort when facing diverging and stressful tasks. We speculate that our findings would help in diverse fields like security, teaching, and caregiving by endowing intelligent virtual and robotic agents to understand humans' behavior and better support us.

Acknowledgments

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. GA No 804388, wHiSPER.

References

- [1] M. A. Runco, G. J. Jaeger, The standard definition of creativity, *Creativity Research Journal* 24 (2012) 92–96. doi:10.1080/10400419.2012.650092.
- [2] A. J. Cropley, The dark side of creativity: What is it?, 2010. doi:10.1017/CBO9780511761225.001.
- [3] J. J. Walczyk, M. A. Runco, S. M. Tripp, C. E. Smith, The creativity of lying: Divergent thinking and ideational correlates of the resolution of social dilemmas, *Creativity Research Journal* 20 (2008) 328–342. doi:10.1080/10400410802355152.
- [4] N. Hao, M. Tang, J. Yang, Q. Wang, M. A. Runco, A new tool to measure malevolent creativity: The malevolent creativity behavior scale, *Frontiers in Psychology* 7 (2016) 1–7. doi:10.3389/fpsyg.2016.00682.
- [5] M. L. Beaussart, C. J. Andrews, J. C. Kaufman, Creative liars: The relationship between creativity and integrity, *Thinking Skills and Creativity* 9 (2013) 129–134. URL: <http://dx.doi.org/10.1016/j.tsc.2012.10.003>. doi:10.1016/j.tsc.2012.10.003.
- [6] B. M. DePaulo, S. E. Kirkendol, D. A. Kashy, M. M. Wyer, J. A. Epstein, Lying in everyday life, *Journal of Personality and Social Psychology* 70 (1996) 979–995. doi:10.1037/0022-3514.70.5.979.

- [7] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, H. Cooper, Cues to deception, 2003. URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.129.1.74>. doi:10.1037/0033-2909.129.1.74.
- [8] O. FeldmanHall, P. Glimcher, A. L. Baker, E. A. Phelps, Emotion and decision-making under uncertainty: Physiological arousal predicts increased gambling during ambiguity but not risk, *Journal of Experimental Psychology: General* 145 (2016) 1255–1262. doi:10.1037/xge0000205.
- [9] C. Hadnagy, Social engineering: The art of human hacking, *The Art of Human Hacking* 3 (2010) 408. doi:10.1504/ijipsi.2018.10013213.
- [10] C. Tosone, Living everyday lies: The experience of self, *Clinical Social Work Journal* 34 (2006) 335–348. doi:10.1007/s10615-005-0035-z.
- [11] C. F. Bond, B. M. DePaulo, Accuracy of deception judgments, *Personality and Social Psychology Review* 10 (2006) 214–234. doi:10.1207/s15327957pspr1003_2.
- [12] C. R. Honts, D. C. Raskin, J. C. Kircher, Mental and physical countermeasures reduce the accuracy of polygraph tests, *Journal of Applied Psychology* 79 (1994) 252–259. URL: <http://www.ncbi.nlm.nih.gov/pubmed/8206815>. doi:10.1037/0021-9010.79.2.252.
- [13] M. M. Bradley, L. Miccoli, M. A. Escrig, P. J. Lang, The pupil as a measure of emotional arousal and autonomic activation, *Psychophysiology* 45 (2008) 602–607. doi:10.1111/j.1469-8986.2008.00654.x.
- [14] M. Gamer, Detecting of deception and concealed information using neuroimaging techniques, 2011, pp. 90–113. URL: https://www.cambridge.org/core/product/identifier/CBO9780511975196A018/type/book_part. doi:10.1017/CBO9780511975196.006.
- [15] B. A. Rajoub, R. Zwigelaar, Thermal facial analysis for deception detection, *IEEE Transactions on Information Forensics and Security* 9 (2014) 1015–1023. URL: <http://ieeexplore.ieee.org/document/6797879/>. doi:10.1109/TIFS.2014.2317309.
- [16] C. Y. Ma, M. H. Chen, Z. Kira, G. AlRegib, Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition, *Signal Processing: Image Communication* 71 (2019) 76–87. URL: <http://arxiv.org/abs/1703.10667>. doi:10.1016/j.image.2018.09.003.
- [17] V. Karpova, P. Popenova, N. Glebko, V. Lyashenko, O. Perepelkina, "was it you who stole 500 rubles?" - the multimodal deception detection, 2020, pp. 112–119. doi:10.1145/3395035.3425638.
- [18] X. L. Chen, S. I. Levitan, M. Levine, M. Mandic, J. Hirschberg, Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies, *Transactions of the Association for Computational Linguistics* 8 (2020) 199–214. doi:10.1162/tac1_a_00311.
- [19] A. Gaggioli, Beyond the truth machine: Emerging technologies for lie detection, *Cyberpsychology, Behavior, and Social Networking* 21 (2018) 144–144. URL: <http://www.liebertpub.com/doi/10.1089/cyber.2018.29102.csi>. doi:10.1089/cyber.2018.29102.csi.
- [20] D. Pasquali, J. Gonzalez-Billandon, A. M. Aroyo, G. Sandini, A. Sciutti, F. Rea, Detecting lies is a child (robot)'s play: Gaze-based lie detection in hri, *International Journal of Social Robotics* (2021) 1–16. URL: <https://link.springer.com/article/10.1007/s12369-021-00822-5>. doi:10.1007/s12369-021-00822-5.
- [21] S. Vinanzi, M. Patacchiola, A. Chella, A. Cangelosi, Would a robot trust you? developmental

- robotics model of trust and theory of mind, *CEUR Workshop Proceedings* 2418 (2019) 74. doi:<https://doi.org/10.1098/rstb.2018.0032>.
- [22] M. Patacchiola, A. Cangelosi, A developmental cognitive architecture for trust and theory of mind in humanoid robots, *IEEE Transactions on Cybernetics* (2020) 1–13. doi:10.1109/TCYB.2020.3002892.
- [23] G. Metta, G. Sandini, D. Vernon, L. Natale, F. Nori, The icub humanoid robot: An open platform for research in embodied cognition, *ACM Press*, 2008, pp. 50–56. URL: <http://portal.acm.org/citation.cfm?doid=1774674.1774683>. doi:10.1145/1774674.1774683.
- [24] J. G. May, R. S. Kennedy, M. C. Williams, W. P. Dunlap, J. R. Brannan, Eye movement indices of mental workload, *Acta Psychologica* 75 (1990) 75–89. doi:10.1016/0001-6918(90)90067-P.
- [25] M. Nakayama, Y. Shimizu, Frequency analysis of task evoked pupillary response and eye-movement, *ACM Press*, 2004, pp. 71–76. URL: <http://portal.acm.org/citation.cfm?doid=968363.968381>. doi:10.1145/968363.968381.
- [26] B. C. Goldwater, Psychological significance of pupillary movements, *Psychological Bulletin* 77 (1972) 340–355. doi:10.1037/h0032456.
- [27] S. Mathôt, J. Fabius, E. V. Heusden, S. V. der Stigchel, Safe and sensible preprocessing and baseline correction of pupil-size data, *Behavior Research Methods* 50 (2018) 94–106. doi:10.3758/s13428-017-1007-2.
- [28] J. Beatty, B. Lucero-Wagoner, The pupillary system, *Handbook of psychophysiology* 2 (2000). URL: <https://psycnet.apa.org/record/2000-03927-005>.
- [29] D. P. Dionisio, E. Granholm, W. A. Hillix, W. F. Perrine, Differentiation of deception using pupillary responses as an index of cognitive processing, *Psychophysiology* 38 (2001) 205–211. URL: <http://www.ncbi.nlm.nih.gov/pubmed/11347866>. doi:10.1017/S0048577201990717.
- [30] A. Szulewski, N. Roth, 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, *Academic Medicine* 90 (2015) 981–987. doi:10.1097/ACM.0000000000000677.
- [31] M. I. Ahmad, J. Bernotat, K. Lohan, F. Eyssel, Trust and cognitive load during human-robot interaction, 2019. URL: <https://arxiv.org/abs/1909.05160v1>.
- [32] C. Wangwiwattana, X. Ding, E. C. Larson, Pupilnet, measuring task evoked pupillary response using commodity rgb tablet cameras, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2018) 1–26. URL: <http://dl.acm.org/citation.cfm?doid=3178157.3161164>. doi:10.1145/3161164.
- [33] S. Rafiqi, C. Wangwiwattana, J. Kim, E. Fernandez, S. Nair, E. C. Larson, Pupilware, *ACM*, 2015, pp. 1–8. URL: <https://dl.acm.org/doi/10.1145/2769493.2769506>. doi:10.1145/2769493.2769506.
- [34] S. Eivazi, T. Santini, A. Keshavarzi, T. Kübler, A. Mazzei, Improving real-time cnn-based pupil detection through domain-specific data augmentation, *Eye Tracking Research and Applications Symposium (ETRA)* (2019). doi:10.1145/3314111.3319914.
- [35] J. Gonzalez-Billandon, A. M. Aroyo, A. Tonelli, D. Pasquali, A. Sciutti, M. Gori, G. Sandini, F. Rea, Can a robot catch you lying? a machine learning system to detect lies during interactions, *Frontiers in Robotics and AI* 6 (2019). doi:10.3389/frobt.2019.00064.

- [36] A. Aroyo, J. Gonzalez-Billandon, A. Tonelli, A. Sciutti, M. Gori, G. Sandini, F. Rea, Can a humanoid robot spot a liar?, *IEEE*, 2018, pp. 1045–1052. URL: <https://ieeexplore.ieee.org/document/8624992/>. doi:10.1109/HUMANOIDS.2018.8624992.
- [37] D. O. Iacob, A. Tapus, First attempts in deception detection in hri by using thermal and rgb-d cameras, *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication (2018)* 652–658. doi:10.1109/ROMAN.2018.8525573.
- [38] D. O. Iacob, A. Tapus, Detecting deception in hri using minimally-invasive and noninvasive techniques, *2019 28th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2019 (2019)* 1–7. doi:10.1109/RO-MAN46459.2019.8956384.
- [39] D. Pasquali, A. M. Aroyo, J. Gonzalez-billandon, F. Rea, G. Sandini, A. Sciutti, Your eyes never lie: A robot magician can tell if you are lying, 2020. doi:<https://doi.org/10.1145/3371382.3378253>.
- [40] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research* 16 (2002) 321–357. doi:10.1613/jair.953.
- [41] J. Montiel, M. Halford, S. M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. M. Gomes, J. Read, T. Abdessalem, A. Bifet, River: machine learning for streaming data in python (2020). URL: <http://arxiv.org/abs/2012.04740>.
- [42] H. M. Gomes, A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, T. Abdessalem, Adaptive random forests for evolving data stream classification, *Machine Learning* 106 (2017) 1469–1495. doi:10.1007/s10994-017-5642-8.
- [43] T. Pro, Quick tech webinar - secrets of the pupil, ????. URL: https://www.youtube.com/watch?v=I3T9Ak2F2bc&feature=emb_title.
- [44] A. Cornu6jols, Getting order independence in incremental learning, ????