InterpretableTruthDetection: Interpretable Bimodal Within-Subject Truth and Deception Detection Models

Moritz Maleck[0009-0006-5551-8386], Tom Gross[0000-0001-8353-7388]

Human-Computer Interaction Group, University of Bamberg, 96045 Bamberg, Germany
hci@uni-bamberg.de

Abstract. Identifying truthful answers in web-based questionnaires can drastically increase the validity of the collected data. Approaches based on the cognitive load of the person giving the answers have been successfully applied. Often, they rely on measuring the cognitive load with one modality (e.g., changes in the pupil diameter). In this paper, we present a bimodal approach that combines two modalities (i.e., the pupil diameter, and mouse movements). It automatically generates truth scores and weighting factors of the different modalities and produces human-readable graphs that allow study administrators to understand the background of the produced scores and weights.

Keywords: Deception, Truth Detection System, Lie Detection, Eye Tracking Method, Mouse Movements, Cognitive-Load-Based Deception Detection.

1 Introduction

The detection of truth and deceit is a highly active research area with many different approaches that evolved over the past decades [1]. Such include the observation of behaviour, the analysis of speech, or (non)verbal cues [2]. Most recently, cognitive-load-based approaches are an emerging and promising field [1, 3].

Challenging, but best meeting practical requirements, is the truth and deceit detection within-subject instead of between-subject [2]. This requires relying on within-subject measures [2]. Such can be the measurement of the cognitive load; which can be determined by relying on mouse and eye tracking [3-11]. By using both modalities, one modality can compensate for weaknesses of the other and lead to more reliable results.

Machine learning (ML) is a rapidly growing field, also in in the area of deception detection [12]. An additional trend is the usage of bi- and multimodal approaches; yet many approaches use linguistic features and face the limitation of not being applicable to other languages [12]. A further problem of ML-approaches is the interpretability of the generated results and classifications in general. This also leads to refusal of using ML-approaches in critical fields because of their 'black-box nature' [13, 14]. *Explainable ML* may help here [13, 15], but there is also critique [14]. Instead, *interpretable* models are recommended to be used [14].

In this paper, we provide a comprehensive answer to the above-discussed challenges with the following contributions:

- A within-subject, language-independent bimodal cognitive-load-based deception detection approach for separating truthful from deceptive answers in web-based questionnaires, and an automated generation of per-participant models based on non-invasive eye and mouse tracking.
- An interpretable ML approach providing a logical model by the usage of weightings, using a truth score, and providing an easy to interpret and visualise graph-based feature-wise model. Figure 1 shows a schematic example of a generated model.

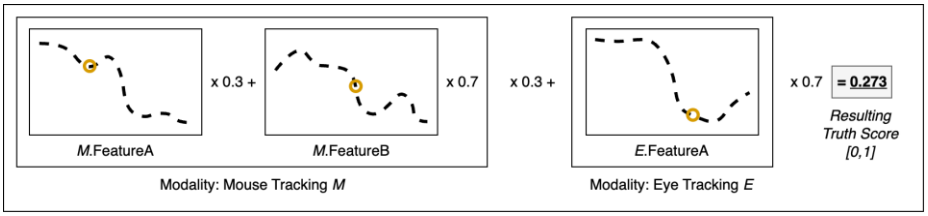

Fig. 1. Schematic example of an interpretable model for truth and deception detection based on two mouse tracking and one eye tracking features. The model consisting of (i) human readable truth-score-graphs for each feature, and (ii) optimal weightings of features and modalities.

We organise this article as follows: First, we give an overview over related work; followed by our concept. Then, we describe the implementation of the concept as the InterpretableTruthDetection system. Finally, we evaluate the system in a systematic user study; and conclude the paper with a discussion and an outlook for future work.

2 Related Work

Cognitive-load based deception detection methods are an emerging field providing great opportunities [2], and particularly for distinguishing truthful from deceptive answers in web-based questionnaires. They complement detecting deception by analysing verbal and nonverbal behaviour, or measuring arousal. Recent studies emphasise the increased accuracy of cognitive-load based truth detection approaches [16]. Cognitive load can be measured by using mouse and eye tracking modalities [17, 18]. Both modalities have been found to be precious additions to the field of truth detection [4-11]. For example, mouse tracking can reveal distinct behavioural types and "insights into the dynamics of the decision process" [19] in order to detect dishonesty. For both eye and mouse modalities detection is possible due to the increased cognitive load for deceptive answers [11, 20].

Mouse tracking includes various temporal and spatial features. Spatial mouse features include the *area under the trajectory curve*, *maximum deviation*, *maximum log ratio* and *x-coordinate flips* [3, 21-23]. Eye tracking includes the measurement of the pupil diameter as window into real-time cognitive processes [17, 24].

Combining modalities and features is a promising trend when it comes to measuring cognitive load [25, 26] and distinguishing truthful from deceptive answers [12, 27, 28]. Within-modality combinations such as combining eye fixations and pupil diameter fluctuations show higher effectiveness and reveal more insights [29]. When it comes to

combining eye and mouse tracking, [3] proposed to combine the two modalities to reach higher reliability of truth classifications. They introduced the *AnswerTruthDetector* system, using binary feature-wise thresholds (i.e., deceptive, truthful) for answers.

In all approaches that aim to leverage on a combination of modalities, putting the different modalities in relation to each other is a challenge [30, 31]. Also for truth detection, weighting of features is a common practise, e.g., for combining facial micro-expression features [32] or within text-based web-based deception detection [33]. The AnswerTruthDetector [3] provided a first solution for eye and mouse tracking.

Machine learning (ML) approaches have been suggested to identify weighting factors [9, 10, 27]. General ML concepts include the usage of training and testing data [34]. For small data sets the *k*-fold cross-validation method is recommended [34].

However, ML approaches mostly follow a black-box principle, where the users of the systems do not get feedback on the reasoning behind the weighting [14]. Some ML approaches for truth detection provide interpretable models (e.g., based on facial images and facial cues [35]; or acoustic, visual, temporal and linguistic data [36]). Yet, their application areas are limited [12].

3 Interpretable Truth Detection Concept

In this section we present the automatic generation of interpretable within-subject models for assessing the truthfulness of answers in computer-administered questionnaires based on mouse and eye tracking data. We describe how to train and test the models.

The concept follows two steps: Firstly, we generate *truth score graphs* for each feature. The generation of the graphs uses a *scoring function*. Secondly, optimal weightings are determined using a *rating function*. This generates interesting insights, like comparison between features and modalities within or between participants. Both steps are fully automated and produce interpretable models.

3.1 Weighting of Modalities Features Based on Continuous Truth Scores

Both eye and mouse tracking modalities provide various single features as measurement-based indicators for truth detection. All features are calculated per-question, allowing the comparison of measurements. If a question has a higher value than usual for a single feature, it may indicate a deceptive answer. When only relying on one single feature and/or modality, outlier detection is relatively simple. However, with the combination of features and modalities, analysis is more difficult. We provide a solution to this through an easy-to-understand weighting mechanism with two levels: First level is the weighting of features within some modality, and second level is the weighting of modalities. The adjustable weightings address the challenge that some features or modalities may work better than others do in general, or for a single participant. This logic is not limited to the particular features and modalities, but can be easily extended.

We use a continuous truth score ranging between zero and one (with a value closer to one indicating a truthful answer and vice versa). Such a score avoids a finite statement about the truthfulness of an answer and sensitises for the possibility of false positives or negatives, being especially harmful in the context of truth classifications.

3.2 Truth-Score-Graphs for Features

Instead of using binary classifications on feature-level (as done by [3]; cf. Figure 2 A), we use a continuous score on feature-level. This can be best explained with an example: Imagine, for some participant the number of x-coordinate flips for all questions ranged from zero to six. For most questions, four x-coordinate flips would serve well as threshold for a deceptive classification; this would not be true for one question. Here, the strength of a continuous truth score can be applied by returning a high but not full probability for a deceptive answer for this number of x-coordinate flips. This allows an easy lookup of a calculated feature-value for some question on a pre-generated graph (cf. Figure 2 B). Such a graph-based logic is especially useful in terms of interpretability.

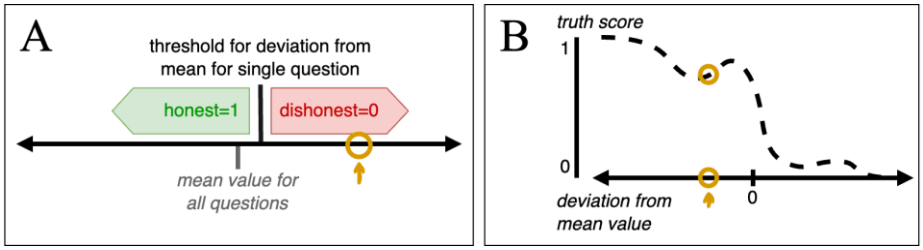

Fig. 2. Binary classification, and new truth-score-graph based logic. (A) A binary truth score {0,1} for classifying an answer as truthful or deceptive (as proposed by [3]). (B) A truth-score-graph with a truth score [0,1]. A is used for the generation of graphs.

The graph-based logic is complex when it comes to the initial formulation (e.g., by the study administrator) of a graph. For this, the study administrator can be supported by ML algorithms for the automatic generation of truth score graphs for each feature. For the generation, we use a brute-forcing mechanism, calculating for each pair of $<x, y>$ (where x is the value tested for and later used for lookup) the truth score y (that is, the result of the scoring function; cf. Section 3.4) for a given x; the concrete values for x are set by applying a configurable sampling rate within a configurable range. This sampling rate can differ for the features and modalities.

We train the model on feature-level using relative x-values, i.e., we use percent deviations from the mean value for a certain feature. This allows independence from absolute measured data and provide a solution to individual within-subject differences.

3.3 Automatic Determination of Weightings

With having graphs and thus, truth scores per question and per feature available, per-modality truth scores and the combined score are calculated. For each feature and modality, weightings are defined. Our automated assessment first generates all possible weighting combinations. Analogue to the sampling rate, a study administrator can configure the weighting steps (e.g., for each feature, use steps of 0.1 to result in the following set: {0, 0.1, 0.2, …, 0.9, 1}); over these sets, the Cartesian product is generated. From the resulting set, all tuples with a sum other than one are removed. The remaining tuples are then applied for a single participant's data, whereby the resulting classifications for each of these iterations are assessed by a *scoring function*, which is returning a value between zero and one (higher values indicating better resulting classifications for a tested tuple than for lower values). The tuple of weightings with the best scoring value is then used for the final proposed configuration for the concrete participant.

3.4 Scoring and Rating Functions for Generating Graphs and Weightings

In the above subsections, the *scoring function* and *rating function* were essential parts. Both algorithms have not yet been used in this domain and are the result of extensive and lengthy adjustments in a trial-and-error approach. The rating function returns a rating score as continuous value between zero and one. The better some distribution of truth scores on the complete truth score scale, the higher the resulting score is. We define the concrete algorithm for the *rating function* as follows:

```
rating_function:
  rating ← 0.5
  R[] ← Ordered list with retrieved results (balanced
        number of truthful/deceptive items) ascending by
        truth score
  c ← rating_score / floor (|R| / 2)
  While |R| > 1:
    f ← first element of R
    l ← last element of R
    If f is a requested deceptive item
    and l is a requested truthful item:
      rating ← rating + c
    Otherwise:
      rating ← rating − c
    R[] ← Remove f and l
  Return rating
```

The *scoring function* consists of three parts: a naive truth score assumption, a correction of the score based on false negatives, and a weighting of the score based on the score from the rating function. The algorithm is as follows:

```
scoring_function:
  dMd ← distance from 0 of the mean truth score for
        requested deceptive items
  dMt ← distance from 1 of the mean truth score for
        requested truthful items
  rS  ← score from rating function for the current
        iteration
  n   ← |all requested items (i.e., all questions)|
  nT  ← |all requested truthful items (i.e., all
         questions where the participant was asked to
         answer truthfully)|

  naive_assumption ← 1 - dMd
  correction_based_on_false_negatives ← dMt * (nT / n)
  weighting_based_on_rating_score ← 2 - rS

  score ← (naive_assumption
           - correction_based_on_false_negatives)
           * weighting_based_on_rating_score
  Return score
```

4 Interpretable Truth Detection Implementation

We integrate and implement the proposed concept as the *Model Generation Service* into the *AnswerTruthDetector* system by [3] (cf. Section 2). The component serves as background service (cf. Figure 3). It is based on JavaScript and node.js (version 21.4.0).

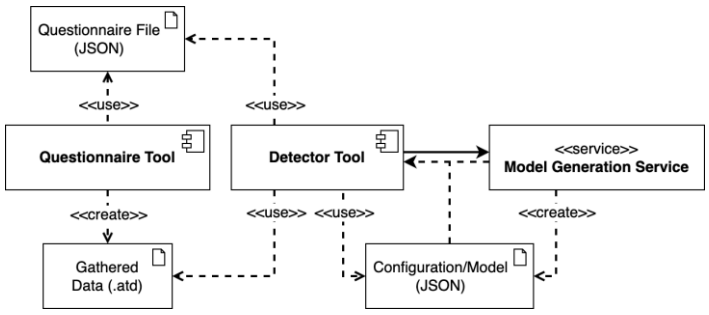

Fig. 3. The system after integration of the Model Generation Service.

The Model Generation Service is designed as a REST service running on a server. It provides a training and a testing mode. The training mode allows the application of the introduced concept, i.e., generating a per-participant model. The testing mode receives a trained configuration and resulting truth scores and returns the metrics for evaluating the validity of the produced model. The Model Generation Service is used as plugin of the Detector Tool, providing an easy-to-use UI (cf. Figure 4).

With the generation process being initiated, both components autonomously communicate with each other by sending and receiving required data. After the training and testing process has finished, the study administrator gets a first impression of the results, such as interactive boxplot visualisations. All produced data are available for further data analysis and visualisation with other tools (like SPSS, Excel, etc.).

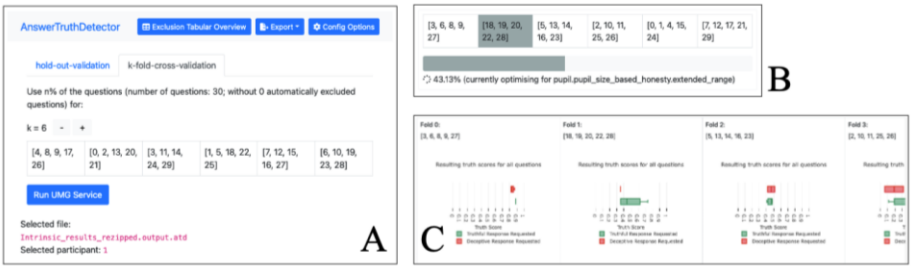

Fig. 4. Screenshots of the integration of the Model Generation Service into the Detector Tool: (A) Selecting hold-out-validation or k-fold cross-validation mode. (B) Visualisation of the current generation progress. (C) Presentation of the results after finished model generation.

By implementing the Model Generation Service on a dedicated server, multiple instances of the Detector Tool can connect to it. This way also, the computing resources lie outside of local machine of the study administrator, resulting in less computational requirements for the computer of the study administrator. For development purposes, the service ran on a local node.js server on an Apple MacBook Pro with a M3 Pro chip and 18 GB RAM. This was sufficient for a single client; yet, for more clients a more advanced set-up is advised.

5 Proof of Concept

5.1 Method

We tested our implementation with 24 participants (18 female, 6 male, 0 diverse) with age from 20 to 29 years (M = 24.42, SD = 2.33). Participants were recruited with mouth-to-mouth sampling and teaching lectures at the university. We applied the following criteria for selecting participants: Only the age spawn between 18 and 29 [37]; no drug consumption (excl. nicotine and caffeine) within the 12 hours before the study execution [38]; no former eye-surgeries with remaining scars on the cornea (due to technical limitations of our eye tracker); no participants with glasses (due to possible accuracy and precision losses by the eye tracker); and right-handed participants only [39]. For future studies, a reduction of these criteria should be considered; due to the limited sample size of this study, the above criteria were applied to reduce additional noise.

For the conducted study half of the participants had previous experience with the used Questionnaire Tool (random allocation to the groups). The study included a prebriefing and informed consent, followed by a calibration of the eye tracker and an introduction to the used tool. Then, the participants completed an intrinsic questionnaire

with the provided tool. For the main evaluation question—to evaluate how well we can distinguish truthful from deceptive answers—we chose a within-subject study design based on the answers within the intrinsic questionnaire (please note: we do not evaluate the participants, we only evaluate our proposed concept and implementation).

The study was conducted in a lab with identical circumstances for all participants (esp. lighting and noise). The setup included a standard PC with *Windows 10*, using a conventional mouse for answering the questions (and recording the mouse paths). Furthermore, a *Tobii Pro Spectrum* eye tracker was installed with 1,200 Hz.

The intrinsic questionnaire task included answering simple questions without the need of preparation. We used 30 general knowledge questions from standardised questionnaires of related work—the *Sheffield lie test* ([37]). Example questions are 'Is water wet?' or 'Is grass blue?'. We asked the participants to answer half of the questions deceptively and the other half truthfully; applying a random distribution for each participant. Further, the order of the questions was randomised for each participant [37]. The sequence of steps for some question within the questionnaire was identical for all questions (cf. Figure 5).

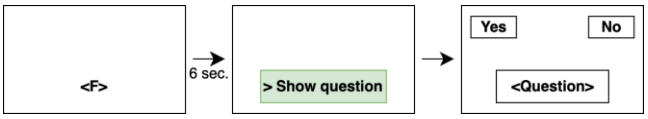

Fig. 5. Scheme of the sequence for each question. First, a fixation and instruction letter (lie/truth) is shown to the participant. After six seconds, the letter is replaced by a button to show the next question. By clicking this button, the question and its answer options appear.

After having conducted the intrinsic questionnaire task with all participants, we generated optimal configurations for each participant (i.e., per-participant training), using the recorded mouse and eye tracking data. We used the *k*-fold cross-validation with k=7, i.e. for each participant's questions were split into seven folds and then our implementation of our InterpretableTruthDetection concept was applied within seven iterations. The resulting truth scores for each requested question from the testing folds were stored, as well as the mean and median truth scores for requested truthful and requested deceptive items. Finally, mean values over all testing-folds were stored for each participant.

5.2 Results and Discussion

We formulate following hypotheses based on the continuous truth score characteristics:

- H_0: The truth scores for requested truthful and deceptive items do not differ.
- H_a: The truth scores for requested truthful items are higher than for deceptive items.

From in total 24 participants * 30 questions per participant = 720 questions, 64 questions were excluded automatically because of invalid mouse movements (i.e., the mouse was not moved fast enough after the question presentation—if selecting an answer by moving the mouse after the completion of the decision process, the traced path would not reveal the mental processes), with remaining N = 656 questions. For these

we reach significantly higher truth scores ($p = 0.00102 < .01$; one-tailed) for requested truthful items ($M = 0.4962$, $SD = 0.2396$) than for requested deceptive items ($M = 0.4388$, $SD = 0.2341$). Thus, H_0 can be rejected.

Comparing the mean values for each participant (i.e., the mean *M*- and *Mdn*-values over all testing-folds for the requested deceptive and requested truthful items) further supports our previous observation. For $83.\overline{3}\%$ of the participants, the mean *M* truth score for requested truthful items was higher than for requested deceptive items; and for $79.1\overline{6}\%$ of the participants, the mean *Mdn* truth score for requested truthful items was higher than for the requested deceptive items. Our approach performs well compared to different approaches; in a recent literature review from 81 selected approaches, performance range was from 51% to 100%—with an often stated limitation of the applicability to the English language only [12]. The authors recommended future research to focus on language-independent approaches, which is the case for our approach.

Summarised, our proposed concept and implementation does work as intended—we reach higher truth scores for truthful items than for deceptive items. Yet, we can still observe a large number of false positives and false negatives (cf. Figure 6). This emphasises once again to see the truth score only as a cautious assumption. Furthermore, it stresses out the need for a more detailed investigation on the differences between the individual participants (including investigating why the concept worked better for some participants than for others).

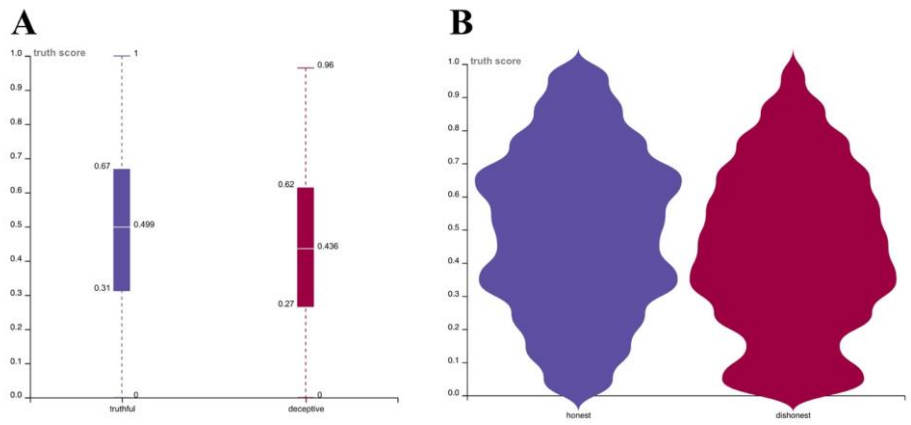

Fig. 6. Visualised truth scores for requested truthful and requested deceptive items. (A) Box plots and (B) violin plots for all requested truthful and requested deceptive items.

In the future, the analysis and comparison of the per-participant models and graphs will be highly interesting. This can be the comparison of the weightings (i.e., compare how well some modality or feature works for some group of participants compared to the other groups), or the comparison of the generated truth graphs for single features. As promising first insights serve the interpolated truth graphs of the different features. Even though the model training was done per-participant, we can observe very similar generated graphs, which strengthens the first assumption, that universal as well as group-based models can be derived from the per-participant models.

6 Conclusions and Future Work

We introduced a concept and its implementation for the fully automatic generation of within-subject models for distinguishing truthful from deceptive answers in web-based questionnaires. This approach is independent from the spoken language and allows a barrier-free application due to its cognitive-load based nature, relying on non-invasive eye and mouse tracking modalities. More precisely, we proposed an interpretable ML approach—that means, not only providing some classifications that were produced within some black-box system, but instead providing interpretable classifications. The model is made up of two pillars—feature-wise truth-score-graphs and dynamic weightings of features and modalities (i.e., eye and mouse tracking). Further, the model uses a continuous truth score that gives weight to the challenging problem of false positive and false negative classifications in the critical field of truth classifications. Therefore, it only provides a cautious assumption that assists a study administrator with analysing within-subject's recorded data and nudges to not only rely on the returned score.

ML has found a broad application when it comes to the separation of truthful from deceptive answers. Many approaches are of a black-box nature and entail several problems. Such can be blind faith in the resulting classification of a ML system, not understanding the actual logic of the implementation and possible weaknesses (e.g., biases); or contrary, not using such approaches at all due to ethical and other concerns within this highly sensitive and critical field. On the other hand, there already exist interpretable ML approaches for truth detection, trying to address the above problems. Yet, the majority of existing approaches use linguistic features and thus are working exclusively for a single language—which we address by using cognitive-load based features relying on eye and mouse tracking data. We further provide a within-subject solution, which is best matching practical requirements. Yet, most existing research focuses on between-subject truth detection due to the less complexity. To the best of our knowledge, we are the first to provide an interpretable language-independent within-subject approach that relies on eye and mouse tracking modalities.

We tested our implementation of our concept with 24 participants. The evaluation results are highly promising and confirm that our concept works as intended. We trained our model per-participant. We pay attention to the problem of few training-data by using the k-fold cross-validation method. We reached significantly higher truth scores for requested truthful items than for requested deceptive items (higher truth scores indicate a higher probability of a truthful answer and vice versa). Yet, we also faced a large number of false positives and false negatives, to be addressed with future optimisations.

A highly promising insight of the evaluation results is, that the generated feature-wise truth-score-graphs (one of the two pillars of the overall model) are very similar for all participants. This is a confident sign, that a definition of a universal model for all participants might be possible in the future. Further, we plan to investigate the performance of the approach across the different participants (including personality characteristics, and language and cultural differences), which will then be followed by providing various pre-sets for the identified groups of participants to further assist study administrators and to reduce the number of false positives and false negatives within a

certain group. Another helpful addition could be a tailored visualisation concept of the generated models to more extensively support the study administrator with interpreting.

Acknowledgments. We thank the members of the Cooperative Media Lab at the University of Bamberg. We also thank the anonymous reviewers for insightful comments.

References

1. R. Weylin Sternglanz, Wendy L. Morris, Marley Morrow, Joshua Braverman. 2019. A review of meta-analyses about deception detection. *The Palgrave handbook of deceptive communication* (2019), 303-326. https://doi.org/10.1007/978-3-319-96334-1_16
2. Aldert Vrij. 2019. Deception and truth detection when analyzing nonverbal and verbal cues. *Applied Cognitive Psychology*, 33, 2 (2019), 160-167. https://doi.org/10.1002/acp.3457
3. Moritz Maleck, Tom Gross. 2023. AnswerTruthDetector: A Combined Cognitive Load Approach for Separating Truthful from Deceptive Answers in Computer-Administered Questionnaires. *i-com - Journal of Interactive Media*, 22, 3 (2023), 241-251. https://doi.org/10.1515/icom-2023-0023
4. Daphne P. Dionisio, Eric Granholm, William A. Hillix, William F. Perrine. 2001. Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, 38, 2 (2001), 205-211. https://doi.org/10.1111/1469-8986.3820205
5. R. E. Lubow, Ofer Fein. 1996. Pupillary size in response to a visual guilty knowledge test: New technique for the detection of deception. *J. Exp. Psychol.-Appl.*, 2, 2 (1996), 164-177. https://doi.org/10.1037/1076-898x.2.2.164
6. Ira Heilveil. 1976. Deception and pupil size. *Journal of Clinical Psychology*, 32, 3 (1976), 675-676. https://doi.org/10.1002/1097-4679
7. Xinyue Fang, Yiteng Sun, Xinyi Zheng, Xinrong Wang, Xuemei Deng, Mei Wang. 2021. Assessing Deception in Questionnaire Surveys With Eye-Tracking. *Frontiers in Psychology*, 12 (2021). https://doi.org/10.3389/fpsyg.2021.774961
8. Andrea K. Webb, Douglas J. Hacker, Dahvyn Osher, Anne E. Cook, Dan J. Woltz, Sean Kristjansson, John C. Kircher. 2009. Eye Movements and Pupil Size Reveal Deception in Computer Administered Questionnaires. *Proceedings of the FAC 2009: Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (San Diego, CA, USA). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02812-0_64
9. Cristina Mazza, Merylin Monaro, Franco Burla, Marco Colasanti, Graziella Orrù, Stefano Ferracuti, Paolo Roma. 2020. Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific Reports*, 10, 1 (2020). https://doi.org/10.1038/s41598-020-61636-5
10. Merylin Monaro, Francesca Ileana Fugazza, Luciano Gamberini, Giuseppe Sartori. 2017. How Human-Mouse Interaction can Accurately Detect Faked Responses About Identity. Springer International Publishing. https://doi.org/10.1007/978-3-319-57753-1_10Merylin Monaro, Francesca Ileana Fugazza, Luciano Gamberini, Giuseppe Sartori. 2017. How Human-Mouse Interaction can Accurately Detect Faked Responses About Identity. Springer International Publishing. https://doi.org/10.1007/978-3-319-57753-1_10
11. Jeffrey Jenkins, Jeffrey Proudfoot, Joseph Valacich, G. Grimes, Jr Jay F. Nunamaker. 2019. Sleight of Hand: Identifying Concealed Information by Monitoring Mouse-Cursor Movements. *Journal of the Association for Information Systems*, 20, 1 (2019-01-31 2019). https://doi.org/10.17705/1jais.00527
12. Alex Sebastião Constâncio, Denise Fukumi Tsunoda, Helena de Fátima Nunes Silva, Jocelaine Martins da Silveira, Deborah Ribeiro Carvalho. 2023. Deception detection with

machine learning: A systematic review and statistical analysis. *Plos one*, 18, 2 (2023), e0281323. https://doi.org/10.1371/journal.pone.0281323

13. Khishigsuren Davagdorj, Jang-Whan Bae, Van-Huy Pham, Nipon Theera-Umpon, Keun Ho Ryu. 2021. Explainable artificial intelligence based framework for non-communicable diseases prediction. *IEEE Access*, 9 (2021), 123672-123688. https://doi.org/10.1109/ACCESS.2021.3110336
14. Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1, 5 (2019), 206-215. https://doi.org/10.1038/s42256-019-0048-x
15. Vaishak Belle, Ioannis Papantonis. 2021. Principles and practice of explainable machine learning. *Frontiers in big Data* (2021), 39. https://doi.org/10.3389/fdata.2021.688969
16. Adrianna Wielgopolan, Kamil K Imbir. 2023. Cognitive load and deception detection performance. *Cognitive Science*, 47, 7 (2023), e13321. https://doi.org/10.1111/cogs.13321
17. Maria K. Eckstein, Belén Guerra-Carrillo, Alison T. Miller Singley, Silvia A. Bunge. 2017. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, 25 (2017), 69-91. https://doi.org/10.1016/j.dcn.2016.11.001
18. Hansol Rheem, Vipin Verma, D. Vaughn Becker. 2018. Use of Mouse-tracking Method to Measure Cognitive Load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62, 1 (2018), 1982-1986. https://doi.org/10.1177/1541931218621449
19. Carina I Hausladen, Olexandr Nikolaychuk. 2024. Color me honest! Time pressure and (dis) honest behavior. *Frontiers in Behavioral Economics*, 2 (2024), 1337312. https://doi.org/10.3389/frbhe.2023.1337312
20. Joseph Tao-Yi Wang, Michael Spezio, Colin F. Camerer. 2010. Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games. *American Economic Review*, 100, 3 (2010), 984-1007. https://doi.org/10.1257/aer.100.3.984
21. Jonathan B. Freeman, Nalini Ambady. 2010. MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42, 1 (2010), 226-241. https://doi.org/10.3758/BRM.42.1.226
22. Rick Dale, Jennifer Roche, Kristy Snyder, Ryan McCall. 2008. Exploring Action Dynamics as an Index of Paired-Associate Learning. *PLoS ONE*, 3, 3 (2008), e1728. https://doi.org/10.1371/journal.pone.0001728
23. Mora Maldonado, Ewan Dunbar, Emmanuel Chemla. 2019. Mouse tracking as a window into decision making. *Behavior Research Methods*, 51, 3 (2019), 1085-1101. https://doi.org/10.3758/s13428-018-01194-x
24. Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. *Proceedings of the CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose California, USA). https://doi.org/10.1145/2858036.2858117
25. Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M Asif Khawaja, Ronnie Taib, Bo Yin, Yang Wang. 2013. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2, 4 (2013), 1-36. https://doi.org/10.1145/2395123.2395127
26. Pieter Vanneste, Annelies Raes, Jessica Morton, Klaas Bombeke, Bram B Van Acker, Charlotte Larmuseau, Fien Depaepe, Wim Van den Noortgate. 2021. Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work*, 23 (2021), 567-585. https://doi.org/10.1007/s10111-020-00641-0

27. Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, Erik Cambria. 2023. A Deep Learning Approach for Multimodal Deception Detection. *Proceedings of the CICLing 2018* (Mexico City, Mexico). Springer, Cham. https://doi.org/10.1007/978-3-031-23793-5_8
28. Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, Mihai Burzo. 2014. Deception detection using a multimodal approach. *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey). Association for Computing Machinery. https://doi.org/10.1145/2663204.2663229
29. Valentin Foucher, Anke Huckauf. 2024. Unveiling Deceptive Intentions: Insights from Eye Movements and Pupil Size. *Proceedings of the ACM on Human-Computer Interaction*, 8, ETRA (2024), 1-17. https://doi.org/10.1145/3655612
30. Haibin Liu, Shengyu Fang, JI Jianhua. 2020. An improved weighted fusion algorithm of multi-sensor. *Proceedings of the 2019 2nd International Conference on Computer Information Science and Artificial Intelligence (CISAI 2019)* (Xi'an, China). IOP Publishing. https://doi.org/10.1088/1742-6596/1453/1/012009
31. Shaik Shehanaz, Ebenezer Daniel, Sitaramanjaneya Reddy Guntur, Sivaji Satrasupalli. 2021. Optimum weighted multimodal medical image fusion using particle swarm optimization. *Optik*, 231 (2021). https://doi.org/10.1016/j.ijleo.2021.166413
32. Mengting Chen, Heather T Ma, Jie Li, Huanhuan Wang. 2016. Emotion recognition using fixed length micro-expressions sequence and weighting method. *Proceedings of the 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (Angkor Wat, Cambodia). IEEE. https://doi.org/10.1109/RCAR.2016.7784067
33. Lina Zhou, Yongmei Shi, Dongsong Zhang. 2008. A statistical language modeling approach to online deception detection. *IEEE Transactions on Knowledge and Data Engineering*, 20, 8 (2008), 1077-1081. https://doi.org/10.1109/TKDE.2007.190624
34. Sebastian Raschka. 2020. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv:1811.12808 [cs.LG]* (2020). https://doi.org/10.48550/arXiv.1811.12808
35. Borum Nam, Joo Young Kim, Beomjun Bark, Yeongmyeong Kim, Jiyoon Kim, Soon Won So, Hyung Youn Choi, In Young Kim. 2023. FacialCueNet: unmasking deception-an interpretable model for criminal interrogation using facial expressions. *Applied Intelligence*, 53, 22 (2023), 27413-27427. https://doi.org/10.1007/s10489-023-04968-9
36. Hamid Karimi. 2018. Interpretable multimodal deception detection in videos. *Proceedings of the 20th ACM international conference on multimodal interaction* (Boulder CO, USA). https://doi.org/10.1145/3242969.3264967
37. Evelyne Debey, Maarten De Schryver, Gordon D. Logan, Kristina Suchotzki, Bruno Verschuere. 2015. From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta Psychologica*, 160 (2015), 58-68. https://doi.org/10.1016/j.actpsy.2015.06.007
38. Sunpreet S. Arora, Mayank Vatsa, Richa Singh, Anil Jain. 2012. Iris recognition under alcohol influence: A preliminary study. *Proceedings of the 2012 5th IAPR International Conference on biometrics (ICB)* (New Delhi, India). IEEE. https://doi.org/10.1109/ICB.2012.6199829
39. Eric Hehman, Ryan M. Stolier, Jonathan B. Freeman. 2015. Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, 18, 3 (2015), 384-401. https://doi.org/10.1177/1368430214538325