Opportunities and Challenges of Automated Verbal Deception Detection

Joint PhD Program in Cognitive and Cultural Systems,
track in
Cognitive, Computational, and Social Neuroscience
XXXVII Cycle

Institutions:
IMT School for Advanced Studies Lucca, Lucca, Italy
Tilburg University, Tilburg, Netherlands

By
Riccardo Loconte
2026

The dissertation of Riccardo Loconte is approved.

PhD Program Coordinator: Prof. Emiliano Ricciardi, IMT School for Advanced Studies Lucca

Supervisors: Prof. Pietro Pietrini, IMT School for Advanced Studies Lucca; Dr. Bennett Kleinberg, Tilburg University

The dissertation of Riccardo Loconte has been reviewed by:

Prof. Emiliano Ricciardi, IMT School for Advanced Studies Lucca, Italy

Dr. Sibilla Di Guida, IMT School for Advanced Studies Lucca, Italy

Prof. Dr. Katrijn Van Deun, Tilburg University, Netherlands

Dr. Emmanuel Keuleers, Tilburg University, Netherlands

Prof. Dr. Kristina Suchotzki, University of Marburg, Germany

Prof. Dr. Shane O'Mara, Trinity College Dublin, Ireland

IMT School for Advanced Studies, Lucca

Tilburg University, Tilburg

2026

To my younger self of ten years ago,
who was lost and afraid of aiming big:
You are doing great!

Contents

Acknowledgements

This thesis is the result of inspiring and enriching thoughts, perspectives, and collaborations that have shaped my PhD journey over the past few years.

A due acknowledgment goes to my advisor, Professor Pietro Pietrini, for placing his trust in me. I am particularly grateful for the chance he gave me to freely explore a research topic that was of special interest to me. Thanks also for the guidance and suggestions throughout these years.

My sincerest gratitude goes to my advisor, Dr. Bennett Kleinberg, for being really life-changing in the way I approached research. In our meetings, he has always been supportive, motivational, and insightful. I always felt how much he valued me and my work. With him, I also learnt how research can be rigorous and fun at the same time. All this was fundamental for my growth as a researcher, and he became my inspiration for the academic I wish to be in the (near) future.

A special thanks also goes to Prof. Giuseppe Sartori. After all, this journey would probably never have happened without his encouragement to apply for this PhD in Lucca and his introduction of me to Bennett for the start of my visiting period.

Furthermore, I would like to thank all my co-authors for the excellent work we did together, which was invaluable to my academic development. Among them, a special thank you goes to Dr. Merylin Monaro for her precious help and advice on my first research project and for providing me with special opportunities to present my findings.

Thanks also to all the members of the PhD committee for the time they spent reviewing this thesis and for the insightful discussions.

However, my PhD path would not have been the same without my people and the new ones I met along this journey. A special thanks goes to my family for their relentless support, to my friends, who have always been there for me, and to my partner, who is my certainty.

Thanks to my colleagues in Tilburg and the whole CPCM Lab for making the lab meetings both insightful and enjoyable.

Ultimately, but because they are the most important, my deepest gratitude goes to the IMT people who made me grow as a researcher, but especially as a person. Thanks to all the amazing colleagues who are now my dearest friends.

On a separate note, I also acknowledge the use of AI-based tools as a support for improving the clarity, readability, and stylistic consistency

of the text. These tools were employed for language refinement and editorial assistance (e.g., rephrasing, grammar, and style). After using these tools, I reviewed and edited the content as needed and take full responsibility for the content of the publication.

Vita

August 29, 1993	Born, Trani, Italy
2016 – 2019	BSc in Psychological Sciences and Techniques Grade: 110 *cum laude*/110 University of Bari Aldo Moro, Bari, Italy
2019 – 2021	MSc in Neuroscience and Neuropsychological Rehabilitation Grade: 110 *cum laude*/110 University of Padova, Padova, Italy
2021 – 2022	Post-graduate internship in Forensic Psychology Department of General Psychology, University of Padova, Padova, Italy
2021 – 2025	PhD Candidate in Cognitive, Computational, and Social Neuroscience (Joint PhD Program) IMT School for Advanced Studies Lucca, Lucca, Italy Tilburg University, Tilburg, Netherlands
2023 – 2025	Visiting PhD Student Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands
November 2025 – ongoing	Postdoctoral Researcher Tilburg School of Social and Behavioral Sciences, Tilburg, Netherlands

Publications

Journal Articles

1. **Loconte, R.**, Monaro, M., Pietrini, P., Verschuere, B., & Kleinberg, B. (2026). Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*, 109019. https://doi.org/10.1016/j.chb.2026.109019
2. **Loconte, R.**, & Kleinberg, B. (2025). Examining embedded lies through computational text analysis. *Scientific Reports*, *15*(1), 26482. https://doi.org/10.1111/lcrp.70001
3. Kleinberg B., **Loconte R.** & Verschuere B. (2025), Effective faking of verbal deception detection with target-aligned adversarial attacks. *Legal and Criminological Psychology*. *00*, 1–24. https://doi.org/10.1111/lcrp.70001
4. **Loconte R.**, Battaglini, C., Maldera, S., Pietrini, P., Sartori, G., Navarin, N., & Monaro, M. (2025). Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Processing. *Journal of Language and Social Psychology*, *0*(0). https://doi.org/10.1177/0261927X251316883
5. **Loconte R.**, Orrù G., Tribastone M., Pietrini P. & Sartori G. (2024). Challenging Large Language Models' "intelligence" with human tools: A Neuropsychological Investigation in Italian language on Prefrontal Functioning. *Heliyon*, *10*(19). https://doi.org/10.1016/j.heliyon.2024.e38911
6. **Loconte R.**, Sesso G., Scarpazza C., & Pietrini P. (2024). A unique case of iatrogenic hebephiliac behavior emerging late in life in a patient with Gordon Holmes Syndrome. *Psychiatry Research Case Report*, *3*(2): 100237. https://doi.org/10.1016/j.psycr.2024.100237
7. **Loconte R.**, Kleinberg B. & Monaro M. (2024). Which methods for the assessment of Large Language Models (LLM)? Comments on the usage of a neuropsychological approach and the development of psychometrics of LLMs. *Giornale Italiano di Psicologia*, *51*(3): 579-584
8. **Loconte R.**, Russo R., Capuozzo P., Pietrini P. & Sartori G. (2024). Verbal lie detection using Large Language Models. *Scientific Reports*, *13*(1): 22849. https://doi.org/10.1038/s41598-023-50214-0
9. Palmisano A., Bossi F., Barlabà C., Febbraio F., **Loconte R.**, Lupo A., Nitsche MA. & Rivolta D. (2021) Anodal tDCS effects over the left dorsolateral prefrontal cortex (L-DLPFC) on the rating of facial expression: evidence for a gender-specific effect. *Heliyon*, *7*(11). https://doi.org/10.1016/j.heliyon.2021.e08267

Presentations

Symposia

Symposium: New strategies for detecting Socially Desirable Responding: assessing the credibility of individuals' reports.
Presenting: Comparing Human and GPT-4 Performance in Scoring Genuine and Deceptive Memory Reports for Credibility Assessments
European Association of Psychology and Law (EAPL), Caparica, Portugal, **July 2024**

Symposium: Novel strategies to assess the credibility of reports. Presenting: Are humans bad evaluators or poor decision-makers? A human vs machine experiment.
Associazione Italiana di Psicologia – Sezione Clinico Dinamica, Univeristy of Florence, Florence, Italy, **September 2023**

Conference presentations

How AI-related accuracy and confidence shape human reliance on AI-based veracity judgments.
International Conference on Decision Making in Medicine and Law: Opportunities and pitfalls of information technologies, **December 2025**

When lies are mostly truthful: automated verbal deception detection for embedded lies.
Lies and Allies Tuesdays - Deception Research Society, **May 2025**

Moving the dial towards embedded lies using computational analysis.
ODISSEI Conference for Social Science in the Netherlands 2024, Utrecht, Netherlands, **December 2024**

Moving the dial towards embedded lies using computational analysis.
The 19th Conference of the International Academy of Investigative Interviewing, Santiago De Compostela, Spain, **June 2024**

Fine-tuning Large Language Models for Verbal Lie Detection.
Associazione Italiana di Psicologia - Sezione Sperimentale, Lucca, Italy, **September 2023**

Large Language Models per la rilevazione della menzogna.
V Convegno Nazionale di Psciologia Giuridica, Milan, Italy, **May 2023**

Abstract (English version)

Verbal deception detection refers to techniques that enable the detection of deceptive intentions or deceptive content in written or transcribed statements. Nowadays, deception detection remains a well-known and unsolved problem with significant implications in various high-stakes contexts, including criminal investigations, financial fraud, and deceptive behaviors in online platforms. This thesis aimed to research the extent to which computational methods from artificial intelligence (AI) can be leveraged for the automated detection of verbal deception. This research question was addressed by investigating opportunities (Chapters 2, 3) and challenges (Chapters 3, 4, and 5) for automated verbal deception detection.

Chapter 1 presents an overview of automated verbal deception detection by systematically reviewing 248 papers and 5,148 machine learning (ML) models. Findings revealed that deception detection research is undergoing a technological shift, transitioning from training statistical models on low-level features (e.g., word frequency, part-of-speech, word statistics) to fine-tuning language models on high-level features (e.g., semantics). However, to advance the reliability and applicability of such models in real-life contexts, important limitations still concern the establishment of a clear ground truth, the investigation of deceptive strategies beyond fabrication, and a sufficient out-of-domain generalization performance.

To further explore the topic of automated verbal deception detection, Chapters 2 and 3 investigate its opportunities for automated coding of statements and prediction of deception. Specifically, Chapter 2 closely compares the performance of naïve judges and expert judges trained on Reality Monitoring with that of theory-led and data-driven machine learning (ML) models in detecting verbal deception. Findings showed that both theory-led (accuracy=69.4%) and data-driven (accuracy=77.3%) ML algorithms significantly outperformed naïve (accuracy=54.7%) and expert judges (accuracy=59.4%), suggesting that such models may represent a valid alternative when psychological manual approaches to deception fall short. Chapter 3 builds upon these findings and investigates whether fine-tuning a large language model for deception detection is effective and robust in cross-domain detection (Chapter 3). Findings revealed that LLMs outperform previous models when trained on a single dataset or a combination of them, reaching an accuracy of up to 79.31%. However, the accuracy rate dropped dramatically to chance level when the model was tested on a novel dataset. These results revealed that there is not a "universal rule" for deception and that previous exposure to deceptive examples is necessary to achieve forms of generalization.

Conversely, Chapters 4 and 5 provide more critical insights into the challenges associated with such automated methods. One limitation of previous studies is that they mostly investigate deception in fabricated statements, overlooking that a more ecological form of deception involves the incorporation of deceptive information into truthful statements. This resulting type of deception is known as embedded lies. By collecting 2,088 truthful and deceptive statements with annotated embedded lies, findings showed that, this time, a fine-tuned language model (Llama-3-8B) could detect embedded lies with 64% accuracy at best. Additional findings on individual differences, linguistic properties, and explainability analysis revealed that embedded lies pose a significant challenge for automated verbal deception detection (and also for deception detection in general), due to their incorporation of truthful information. Finally, with the vision that automated methods may be integrated into real-life settings to aid experts in deception detection, Chapter 5 investigates the extent to which humans would endorse such algorithmic predictions. With only a few available studies on hybrid decision-making, this chapter developed a behavioral experiment to examine how humans rely on or reject AI predictions based on varying degrees of information about the AI model (i.e., accuracy and confidence). Findings showed that the model's accuracy played a role, as humans followed predictions from a highly accurate model more than from a less accurate one. Additionally, confidence had an unexpected effect: human judgments deviated more from highly confident AI predictions, especially if the model predicted deception. Ultimately, human interaction with algorithmic predictions either hindered the machine's performance or was ineffective. These results on human aversion to AI judgments provide practical insights and limitations for future integrations of human oversight in algorithmic decision-making.

Altogether, the findings from these five studies contribute to highlighting not only contexts where computational methods clearly outperform human performance but also their current methodological, conceptual, and practical shortcomings, underscoring the conditions under which model application in real-life settings is constrained.

Abstract (Dutch version)

Verbale leugendetectie verwijst naar technieken die het mogelijk maken om misleidende intenties of misleidende inhoud te detecteren in geschreven of getranscribeerde verklaringen. Tegenwoordig blijft leugendetectie een bekend en onopgelost probleem met belangrijke implicaties in verschillende risicovolle contexten, waaronder strafrechtelijke onderzoeken, financiële fraude en misleidend gedrag op online platforms. Dit proefschrift had tot doel te onderzoeken in hoeverre computationele methoden uit de kunstmatige intelligentie (AI) kunnen worden ingezet voor de geautomatiseerde detectie van verbale misleiding. Deze onderzoeksvraag werd benaderd door zowel de mogelijkheden (Hoofdstukken 2 en 3) als de uitdagingen (Hoofdstukken 3, 4 en 5) van geautomatiseerde verbale leugendetectie te onderzoeken.

Hoofdstuk 1 biedt een overzicht van geautomatiseerde verbale leugendetectie door middel van een systematische review van 248 artikelen en 5.148 machine learning (ML)-modellen. De bevindingen tonen aan dat het onderzoek naar leugendetectie een technologische verschuiving doormaakt: van het trainen van statistische modellen op laag-niveau kenmerken (bijv. woordfrequentie, woordsoorten, woordstatistieken) naar het verfijnen van taalmodellen op hoog-niveau kenmerken (bijv. semantiek). Om de betrouwbaarheid en toepasbaarheid van dergelijke modellen in praktijksituaties te verbeteren, blijven echter belangrijke beperkingen bestaan, zoals het vaststellen van een duidelijke 'ground truth', het onderzoeken van misleidingsstrategieën die verder gaan dan pure verzinsels, en het behalen van voldoende generaliseerbaarheid buiten de trainingscontext.

Om het onderwerp verder te verkennen, onderzoeken Hoofdstukken 2 en 3 de mogelijkheden van geautomatiseerde verbale leugendetectie voor het automatisch coderen van verklaringen en het voorspellen van misleiding. Hoofdstuk 2 vergelijkt specifiek de prestaties van lekenbeoordelaars en experts (getraind in Reality Monitoring) met die van theoriegestuurde en datagedreven machine learning-modellen bij het detecteren van verbale misleiding. De resultaten tonen aan dat zowel theoriegestuurde (nauwkeurigheid = 69,4%) als datagedreven (nauwkeurigheid = 77,3%) ML-algoritmen significant beter presteren dan leken (nauwkeurigheid = 54,7%) en experts (nauwkeurigheid = 59,4%). Dit suggereert dat dergelijke modellen een waardevol alternatief kunnen vormen wanneer psychologische handmatige methoden tekortschieten. Hoofdstuk 3 bouwt hierop voort en onderzoekt of het verfijnen van een groot taalmodel effectief en robuust is voor leugendetectie over verschillende domeinen heen. De bevindingen tonen aan dat LLM's beter presteren dan eerdere modellen wanneer ze worden getraind op één dataset

of een combinatie daarvan, met een nauwkeurigheid tot 79,31%. Wanneer het model echter werd getest op een nieuwe dataset, daalde de nauwkeurigheid drastisch tot kansniveau. Deze resultaten tonen aan dat er geen "universele regel" voor misleiding bestaat en dat eerdere blootstelling aan misleidende voorbeelden noodzakelijk is om generalisatie te bereiken.

Daarentegen bieden Hoofdstukken 4 en 5 een kritischer perspectief op de uitdagingen van dergelijke geautomatiseerde methoden. Een beperking van eerdere studies is dat zij zich voornamelijk richten op verzonnen verklaringen, terwijl een meer realistische vorm van misleiding juist bestaat uit het verweven van misleidende informatie in waarheidsgetrouwe verklaringen. Dit type misleiding staat bekend als 'embedded lies'. Op basis van 2.088 waarheidsgetrouwe en misleidende verklaringen met geannoteerde embedded lies tonen de resultaten aan dat een verfijnd taalmodel (Llama-3-8B) deze embedded lies met maximaal 64% nauwkeurigheid kan detecteren. Verdere analyses naar individuele verschillen, taalkundige eigenschappen en verklaarbaarheid laten zien dat embedded lies een aanzienlijke uitdaging vormen voor geautomatiseerde (en algemene) leugendetectie, juist omdat zij waarheidsgetrouwe informatie bevatten.

Tot slot onderzoekt Hoofdstuk 5, vanuit de gedachte dat geautomatiseerde methoden in de praktijk experts kunnen ondersteunen, in hoeverre mensen dergelijke algoritmische voorspellingen accepteren. Met slechts een beperkt aantal studies over hybride besluitvorming werd een gedragsexperiment opgezet om te analyseren hoe mensen omgaan met AI-voorspellingen, afhankelijk van informatie over de nauwkeurigheid en het vertrouwen van het model. De resultaten tonen aan dat de nauwkeurigheid van het model een rol speelt: mensen volgen voorspellingen van zeer nauwkeurige modellen vaker dan die van minder nauwkeurige modellen. Tegelijkertijd had het vertrouwen (confidence) een onverwacht effect: menselijke oordelen weken juist sterker af van zeer zekere AI-voorspellingen, vooral wanneer het model misleiding voorspelde. Uiteindelijk bleek dat menselijke interactie met algoritmische voorspellingen de prestaties van het model eerder verslechterde of geen effect had. Deze bevindingen over menselijke terughoudendheid ten opzichte van AI-oordelen bieden belangrijke inzichten en beperkingen voor toekomstige integratie van menselijke controle in algoritmische besluitvorming.

Samengevat dragen de bevindingen uit deze vijf studies bij aan het identificeren van contexten waarin computationele methoden duidelijk beter presteren dan mensen, maar ook aan het blootleggen van hun huidige methodologische, conceptuele en praktische tekortkomingen. Daarmee

worden de voorwaarden verduidelijkt waaronder de toepassing van dergelijke modellen in realistische settings beperkt blijft.

List of Figures

List of Tables

Preface

Deception has been defined as the deliberate attempt to convince someone else to accept as true something that the lie-teller knows is false in order to obtain benefits or avoid a loss (Abe, 2009). In some contexts, the adoption of deceptive communication can have significant implications. For example, in legal and forensic scenarios, detecting deception during investigative interviews is paramount as distorted information can mislead investigations and result in wrongful convictions, undermining public trust in the legal system. In the financial services sector, deception can lead to fraud. In online platforms, deceptive practices, such as posting fake reviews, can manipulate public opinion and erode trust.

Given these implications, deception detection has a long-standing research tradition. Among the various approaches developed over the years, detecting deception from verbal cues has been found to be the most promising (Amado et al., 2016; Gancedo et al., 2021; Vrij et al., 2016; Vrij & Granhag, 2012). More specifically, verbal deception detection involves implementing techniques that enable the recognition of deceptive intentions or content from written or transcribed statements. This line of research originated in the psychology field, with authors proposing truth criteria for statement evaluation (e.g., Criteria-Based Content Analysis; Amado et al., 2016), but it still remains a well-known unsolved problem for its application in real-life settings. For example, many approaches have been found to be promising at the group-level, but they cannot be easily translated into single-subject predictions (Cooke & Michie, 2010; McManus et al., 2023; Yarkoni & Westfall, 2017). Additionally, most validation methods lack safeguards against performance overestimation, particularly when studies rely on small sample sizes (Kleinberg et al., 2019). Lastly, many of these approaches rely on human evaluation and coding, which limits the possibility of detecting deception at scale and introduces errors related to inter-rater disagreement (Aroyo & Welty, 2015).

However, these limitations may be overcome by resorting to new approaches that pertain to computer science. In fact, deception detection has also been recently investigated by computer scientists, who have proposed automated methods for detecting verbal deception, leveraging machine learning (ML) and natural language processing (Constancio et al., 2023). In fact, once ML models are trained, predictions can be made at the single-subject level on a large scale, thereby reducing the effort and time required for evaluating thousands of statements.

This possibility opened new avenues for deception detection, shifting to *automated* verbal deception detection. Yet, relying on automated approaches comes with its own limitations, as ML models often fail to generalize to out-of-distribution data and risk introducing new biases (Zhou et al., 2023).

To bridge the gap between research in psychology and computer science, this thesis aims to further explore the topic by replying to the following overarching research question:

> *To what extent can we rely on automated methods for the detection of verbal deception?*

This research question is addressed by investigating both **opportunities** (Chapters 2 and 3) and **challenges** (Chapters 3, 4, and 5) for automated verbal deception detection. The outline of this thesis is provided below.

Outline of this thesis

Chapter 1

In the first Chapter, we provide a broad introduction to psychological research on human abilities in deception detection and existing leading theories of deception, with a focus on how and why verbal cues of deception are the most promising. We also introduce recent advances in computation that have enabled more sophisticated methods for automated verbal deception detection. Finally, we present the state of automated verbal deception detection in a systematic review.

Chapter 2

In Chapter 2, we begin by examining and comparing the performance of naïve judges, expert judges trained on a well-established technique for credibility assessment (i.e., Reality Monitoring; Johnson & Raye, 1981), and theory-led and data-driven statistical learning models in detecting verbal deception. By closely comparing these four conditions, this chapter provides insights into the effectiveness and reliability of each approach, underscoring the advantage of automated methods in overcoming human limitations in manual coding and predictions.

Chapter 3

This chapter builds upon the findings in Chapter 2 and tests the effectiveness and robustness of fine-tuning large language models (LLMs) for automated verbal deception detection. LLMs are known for their ability to learn human language representations and generate text that is often indistinguishable from human writing (Jakesch et al., 2023). For this

work, we developed three scenarios to study the deception detection performance of LLMs within personal opinions, autobiographical memories, and future intentions. This chapter provides further insights into the potential and limitations of LLMs for cross-domain deception detection.

Chapter 4

Moving beyond binary distinctions, this study examines a more nuanced and ecological form of deception, known as embedded lies. Embedded lies consist of the incorporation of deceptive information into otherwise truthful statements (Verigin et al., 2020). To this aim, this chapter introduces a new dataset of 2,088 truthful and deceptive statements with annotated embedded lies. This chapter highlights the challenge of detecting deception when lies are embedded into truthful statements and provides a new resource for advancing research in this area.

Chapter 5

Despite the advantages of relying on fully automated approaches to save time and reduce efforts, human oversight of AI predictions is still necessary in high-stakes contexts, such as in the legal domain, where accusations of deception carry significant consequences. Therefore, this chapter brings humans back into the loop to examine the extent to which humans would rely on AI-based judgments for deception detection. Additionally, it will test whether the human-AI interaction surpasses the performance of an AI model in isolation. This chapter offers practical insights and limitations for future integrations of human oversight of algorithmic decisions.

Chapter 6

This final chapter provides a comprehensive answer to the overarching research questions by discussing the potential and limitations of automated verbal deception detection based on the findings collected throughout this thesis. More specifically, this chapter discusses the promises of automated methods for coding statements and predicting deception, as well as their limitations in generalizing findings across different contexts and types of deception, and the potential human aversion to algorithmic predictions.

To sum up, this thesis contributes to showing applications where automated methods clearly outperform human performance and highlighting the conditions, with a focus on forensic contexts, under which model application in real-life settings is constrained.

References

Abe, N. (2009). The neurobiology of deception: Evidence from neuroimaging and loss-of-function studies. *Current Opinion in Neurology*, *22*(6), 594–600. https://doi.org/10.1097/WCO.0B013E328332C3CF

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/J.IJCHP.2016.01.002

Aroyo, L., & Welty, C. (2015). Truth Is a Lie. *AI Magazine*, *36*(1), 15–24. https://doi.org/10.1609/AIMAG.V36I1.2564

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*(4), 259–274. https://doi.org/10.1007/S10979-009-9176-X

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/EJPALC2021A10

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(11). https://doi.org/10.1073/PNAS.2208839120

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/JOURNAL.PONE.0220228

McManus, R. M., Young, L., & Sweetman, J. (2023). Psychology Is a Property of Persons, Not Averages or Distributions: Confronting the

Group-to-Person Generalizability Problem in Experimental Psychology. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231186615

Verigin, B. L., Meijer, E. H., Vrij, A., & Zauzig, L. (2020). The interaction of truthful and deceptive information. *Psychology, Crime & Law*, *26*(4), 367–383. https://doi.org/10.1080/1068316X.2019.1669596

Vrij, A., Fisher, R. P., Blank, H., Leal, S., & Mann, S. (2016). A cognitive approach to elicit verbal and nonverbal cues to deceit. *Cheating, Corruption, and Concealment: The Roots of Dishonesty*, 284–302. https://doi.org/10.1017/CBO9781316225608.017

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117. https://doi.org/10.1016/J.JARMAC.2012.02.004

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2023). Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3195549

Chapter 1

The state of automated verbal deception detection

Abstract

The limitations of human verbal deception detection abilities have sparked an interest in automated methods enabled by advances in natural language processing and machine learning. Advancing that research is challenging because of its fragmentation into two scientific disciplines (i.e., psychology and computer science) and because of wide variety in research practices. This systematic review maps the research practices of automated verbal deception detection research to highlight important gaps and help future research in the area. We identified 23,773 published and unpublished records that were screened for title-and-abstract using the ASReview Lab software. Of these, a total of 248 records, involving 5,148 models for verbal deception detection, were included for this review. The analysis revealed that deception is predominantly operationalized as fabrication (59.3%) and studied in fake reviews (43.2%). The ground-truth of deception was poorly operationalised, with only the 13.71% providing a clear and fully verifiable ground truth. Most research relies on existing datasets (53.6%), focused on English language (82.3%). For the analytical approach, the adoption of more computationally complex approaches, including embeddings for text-representation and neural networks or transformers for models' architectures, increased over time. However, only 18.05% of models were tested on new data to test the generalizability of performance. Key strengths in automated verbal deception detection are represented by research replicability and shifts toward methodological advances. Key limitations concern the need for more rigorous ground truth standards, broader operationalizations of deception, and rigorous out-of-domain evaluation to advance the reliability and applicability of automated approaches for detecting verbal deception.

Keywords: systematic review, verbal deception detection, machine learning, natural language processing

1. Introduction

Deception has been defined as the "*psychological process by which one individual deliberately attempts to convince another person to accept as true what the liar knows to be false, typically for the liar, or sometimes for others, to gain some type of benefit or to avoid loss*" (Abe, 2009).

Deceptive intentions and communications occur daily (Serota & Levine, 2015) across a wide range of contexts (e.g., in political statements, Bond et al., 2017; airport security, Kleinberg et al., 2017; and criminal proceedings, Steller & Koehnken, 1989Bond et al., 2017). While typical everyday lies are often trivial, such as pretending to be busy to avoid a social event, in other contexts, deception carries significant implications. For instance, in legal and forensic settings, assessing deception is crucial for evaluating the suspect's credibility during investigative interviews. False or distorted information can mislead investigations, delay case resolution, and even result in wrongful convictions, allowing the actual offender to remain free and undermining justice and public trust in the legal system. In financial services, deception can translate into fraud when individuals or organizations manipulate transaction records to gain unfair and unlawful economic advantage. In online platforms, deceptive practices can erode trust, for example, by posting fake reviews that distort consumer perceptions and unfairly promote certain businesses. These important risks and concerns posed by deception have motivated academic researchers to develop effective methods for deception detection.

In this chapter, we first discuss psychological research on human abilities in deception detection and existing leading theories of deception, to then focus on how and why verbal cues of deception are the most promising. Second, we introduce how recent advances in the analysis of textual data, thanks to developments in machine learning (ML) and natural language processing (NLP), have enabled more sophisticated ways to approach verbal deception detection in automated manners. Finally, we detail the state of automated verbal deception detection through an extensive systematic review of journal articles, conference papers, and preprints.

1.1 Psychological deception research

Human ability in deception detection and the rise of the Truth-Default Theory

Decades of research suggest that humans are generally poor at detecting deception. One meta-analysis revealed that in the absence of any prior contextual knowledge, personal familiarity, or specialized training, lay

people's accuracy is only slightly better than what can be expected from chance (average accuracy = 54%; Bond & DePaulo, 2006). This finding was also replicated for presumed lie experts, such as law enforcement professionals and people who work daily with deception (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006; Hartwig & Bond, 2011). One explanation for this poor performance is that humans often rely on cues that are weak in detecting deception (Hartwig & Bond, 2011).

To overcome this limitation, researchers have systematically explored which cues might improve deception accuracy. A seminal meta-analysis examined 158 potential cues for deception and found that behavioral cues (e.g., gaze aversion, fidgeting) were largely ineffective, whereas verbal cues (e.g., response length, level of detail) showed considerable potential (DePaulo et al., 2003). Subsequent meta-analytical work suggested that deception could theoretically be detected with up to 67.86% accuracy when multiple cues would effectively be combined together, and this level of accuracy remained stable across different contexts, such as lab experiments and real-life situations (Hartwig & Bond, 2014).

The debate becomes more complicated as more recent research has challenged the notion that humans are poor at detecting deception. For example, one study showed, in nine experiments, that when humans use a heuristic approach, namely judging the detailedness of statements instead of their veracity, their average judgment accuracy rises to 70% (Verschuere et al., 2023). Moreover, a recent reanalysis of Bond and DePaulo's meta-analysis concluded that the "54% accuracy" claim needs to be contextualized under Truth Default Theory (TDT) and suggested that practical accuracy may even be higher (Levine & Serota, 2025). In fact, according to TDT (Levine, 2014), deception is relatively infrequent and often produced by a few prolific liars. Consequently, people presume honesty in communication by default and become suspicious of deception only when strong triggers disrupt this assumption. Therefore, with the real-world truth–lie base rates being way far from 50–50, and humans being truth-biased (i.e., more accurate detecting truthful statements compared to deceptive ones), this 54% accuracy might actually be an underestimation if translated to real-world scenarios (Levine & Serota, 2025).

However, before the TDT, researchers focused on hunting the most diagnostic cue of deception. This led to the development, over the years, of several theories of deception, each focused on different categories of cues, but the most prominent ones remain the following two: the arousal theory and the cognitive theory of deception.

The arousal theory of deception

The arousal theory posits that the cognitive state of lying can be inferred from the arousal experienced with lying (Vrij et al., 2010). With deception being associated with the stress of being caught, the underlying assumption is that the lie-teller involuntarily displays physiological and emotional signs that are indicative of their mental state of deception. Common ways to detect such changes in arousal include measuring physiological responses with a polygraph, relying on nonverbal indicators (the so-called *body language*), and detecting facial expressions of emotions.

The polygraph measures arousal through physiological responses, such as respiration, pulse, blood pressure, and the electrodermal galvanic response, and is commonly employed in association with questioning techniques. One technique is the control question test (CQT; Kircher & Raskin, 1988) and consists of comparing the level of arousal during questions relevant to the crime (target questions) with questions unrelated to the crime (control questions). The examiner may conclude that the suspect is lying when there is a significant change in arousal during target questions compared to control questions. Another technique is the guilty knowledge test (GKT; Ben-Shakhar & Elaad, 2003) and consists of strategically using pieces of evidence known only to law enforcement and the offender, but not to other innocent people or the media. In these cases, the polygraph measures the arousal associated with multiple-choice questions about that piece of evidence (e.g., is this the gun used in the crime?). If the suspect's arousal increases when the correct answer is shown, rather than the incorrect ones, this indicates that the suspect may have personal knowledge about details of the crime. The big shortcomings of this approach are that i) arousal responses are not universal, with some clinical populations displaying attenuated physiological responses (Stern & Krapohl, 2004); ii) the polygraph is susceptible to simple countermeasures (e.g., counting backwards, scrunching one's toes while answering control questions; Nortje & Tredoux, 2019), making the polygraph ineffective as changes in arousal between target and control questions become negligible; iii) the questioning protocol is problematic as target questions (e.g., "did you commit the murder?") inherently elicit a different emotional response than control questions (e.g., "what's your name?"), regardless of the individual's innocence (Verschuere et al., 2011).

Nonverbal indicators have also been extensively studied, as the stress associated with lying is believed to be manifested in behavioral and emotional displays. Most laypeople and practitioners, indeed, believe that such nonverbal cues, including gaze aversion, fidgeting, and increased movement, are clear indicators of deception (Aamodt & Custer, 2006; Bogaard et al., 2016; C. F. Bond & DePaulo, 2006; Hartwig & Bond, 2011).

Interestingly, due to their self-reported lie-detection ability and experience, practitioners tend to be even more confident in these beliefs than laypeople (Bogaard et al., 2016). However, it is now a robust finding that these beliefs are wrong. One meta-analytical study (Sporer & Schwandt, 2007) of 41 studies concluded that, of eleven behavioral cues, most are unreliable indicators of deception, and the few significant ones (e.g., nodding, leg, and hand movements) were, conversely, indicators of truthfulness. The study also concluded that the effect sizes of the relationships between nonverbal cues and deception are generally small and heterogeneous, suggesting that context is an important moderator and these correlations are far weaker than most people assume (Sporer & Schwandt, 2007).

Switching the focus from the body to the face, a line of research focused on detecting deception from facial micro-expressions. The term "*micro-expressions*" was first introduced by Paul Ekman (Ekman, 2003) and refers to brief displays of facial expressions (i.e., lasting less than 40 ms) that can be detected from the analysis of video footage and reveal concealed emotions. The underlying theory is that humans evolved with six basic emotions that are uniquely associated with facial expressions beyond voluntary control. Therefore, concealing emotions can only result in shallow masking, with the true underlying emotion leaving traces in the form of micro-expressions. Over time, training programs to infer deception from the detection of such micro-expressions have been developed (Matsumoto & Hwang, 2011), but independent research suggests there is little supporting evidence in their favour (Lahay et al., 2025; Vrij & Granhag, 2012). The problem with this approach is that micro-expressions do not measure deception itself but rather potential discrepancies between expressed and experienced emotions (Vrij et al., 2010).

To conclude, despite the spark of interest in the arousal theory of deception, the big shortcoming of this theory is that arousal represents a direct indicator of an emotional response or stress, but only indirectly suggests the presence of deception (Sternglanz et al., 2019). Put differently, truth-tellers might also display arousal during investigations because of anxiety or fear, for example, simply because they are suspected of murder despite being innocent, and regardless of deception. In addition, the conflicting evidence found for the arousal theory led researchers to seek alternative explanations. Increased efforts to unravel the cognitive mechanisms behind deception shifted the deception detection community towards a cognitive theory of deception.

The cognitive theory of deception

The cognitive theory posits that the increased cognitive load resulting from lying leads to the leakage of cues of deception (Vrij et al., 2008, 2015, 2016). Behavioral evidence of this cognitive load was shown in a meta-analysis of 114 studies examining reaction times in computerized tasks for deception detection (Suchotzki et al., 2017). Their findings indicated that "lying takes time", as lie-tellers needed longer reaction times than truth-tellers to perform the task, explaining this delay as the behavioral leakage of the cognitive cost of lying. Other research has provided additional evidence of the cognitive demands of deception through behavioral cues, such as pupil dilation (Dionisio et al., 2001), blink rate (Monaro et al., 2020), and mouse and keyboard dynamics (Monaro et al., 2017, 2018). Neuroscientific research provided support to this theory, too. Deception has been found to consistently engage multiple brain regions associated with executive control, particularly within the prefrontal cortex (Christ et al., 2009). Among executive functions, working memory appears to play a central role, with activation of the dorsolateral prefrontal cortex and posterior parietal cortex reflecting the need to maintain and manipulate information while lying (Christ et al., 2009). Moreover, socially interactive deception was found to recruit additional areas involved in theory of mind and moral reasoning, such as the anterior cingulate cortex and posterior superior temporal gyrus, highlighting that interpersonal lies require conflict monitoring and perspective-taking, further increasing cognitive complexity (Lisofsky et al., 2014).

The cognitive theory has also been translated into strategic approaches to elicit information (Vrij et al., 2016), shifting the focus from behavioral to verbal cues. This approach, known as the information-gathering approach, relies on designing interview strategies that maximize the amount of elicited information from truth-tellers and diagnostic cues of deception in terms of errors and contradictions from lie-tellers. Among these, the most promising ones are the Strategic Use of Evidence (SUE) and the Increase in Cognitive Load (CL) approach.

The SUE can be used in situations where investigators possess incriminating evidence, and the suspect is unaware of it. It consists of withholding knowledge of incriminating evidence from suspects until after they provide their own account, and it is based on the idea that guilty individuals tend to avoid mentioning incriminating details, while innocent ones tend to share freely. By delaying the disclosure of evidence, investigators can significantly enhance deception detection through the identification of omissions and inconsistencies between the suspect's story and external evidence (Hartwig et al., 2014).

The CL approach places greater emphasis on the increased cognitive demand that characterizes lie-tellers to develop interviewing strategies (Vrij et al., 2015). While a truth-teller simply has to recollect and describe existing memories, a lie-teller must keep in mind both the original and alternative versions of the event, thereby increasing the use of working memory (Christ et al., 2009). When implementing a CL approach to deception detection, the interviewer might i) intentionally increase the suspect's cognitive load (e.g., by asking them to tell the story backward or to perform a second task meanwhile), ii) encourage them to add more details, or iii) ask unexpected questions (Vrij et al., 2016). A meta-analysis indicated that the CL approach was more accurate than the standard approach in detecting deception, with an overall accuracy of 71% (Vrij et al., 2015).

Drawing from research on memory detection, deception detection has also been investigated from the content-analysis perspective. Of these, the most investigated approaches are derived from Criteria-Based Content Analysis (CBCA; Steller & Koehnken, 1989) and Reality Monitoring (RM; Johnson & Raye, 1981). Both approaches are based on previous evidence showing that truthful statements about past personal experiences include more contextual and perceptual details than fabricated stories, which rely more on cognitive processes, such as imagination. Based on this evidence, a list of criteria was developed to evaluate statements. A meta-analytical study testing the effectiveness of CBCA and RM in detecting deception reported an accuracy rate of 70% with no significant difference between the two techniques (Oberlader et al., 2021), underscoring their potential for the evaluation of statements in forensic contexts. Building on these principles, the Verifiability Approach (VA) has been later developed as a coding scheme on top of RM, extending the idea that truthful accounts should contain more details that can be subsequently verified, whereas lie-tellers refrain from providing such details (Nahari et al., 2012). Two independent meta-analyses confirmed that truth-tellers provide a higher absolute and relative number of verifiable details, while unverifiable details were not significantly more present in deceptive statements (Palena et al., 2021; Verschuere et al., 2021).

In conclusion, the cognitive theory represents a more valid alternative to the arousal theory and has paved the way for further research into verbal cues of deception, stemming from either strategic interviewing or content analysis. In general, verbal cues have been shown to be the most effective indicators of deceit (DePaulo et al., 2003; Hartwig & Bond, 2014; Vrij et al., 2016). Their promise, however, lies not only in their diagnostic accuracy but also in their practical accessibility, as verbal data are easy to collect and often readily available across various contexts (e.g., from forensic settings where law enforcement and judges are accustomed to

collecting and analyzing statements to online scenarios where fake reviews are presented almost exclusively in their textual format). This promise underscores the need for future research in the area, especially given the recent advances in computational approaches that are able to work with linguistic data, opening new avenues for an automated detection of verbal deception.

1.2 Computational approaches to verbal deception detection

Aware of the limitations of human judgment, recent research has focused on automated methods for deception detection. Advances in computing and machine learning (ML) have enabled approaches that aim to increase the scalability and reliability of traditional manual techniques for detecting verbal deception. Among the most prominent computational approaches, those leveraging textual data, combining Natural Language Processing (NLP) with ML models, are particularly promising.

On the one hand, NLP methods enable the representation of texts in a numerical vector form at various levels of granularity. For instance, dictionary-based approaches, such as the Linguistic Inquiry Word Count (LIWC; Boyd et al., 2022; Tausczik et al., 2010), count the frequency of words in a statement that match words in validated dictionaries, reflecting various psychological and linguistic processes. Machine-learning approaches rely on pre-trained models to automatically tag words as grammatical categories (e.g., part-of-speech tagging; POS) or label specific information (e.g., named-entities recognition; NER) to broader categories (e.g., 27.11.2025 into DATES, Trafalgar Square into LOCATIONS, Apple into ORGANIZATIONS; see Kleinberg, Mozes, et al., 2017,for an application to deception detection). Other approaches rely on term frequency, for example, bag-of-words models, but lack the context and semantics of texts. Finally, embeddings generate a vectorial numerical representation of texts that incorporates semantic and contextual information, allowing similar terms to have closer vector representations in a multidimensional space (Joulin et al., 2017; Mikolov et al., 2013; Pennington et al., 2014). In general, this numerical representation of text can be grounded in theories and automate verbal credibility assessment techniques, or can be extracted following a completely data-driven approach.

On the other hand, ML leverages features to identify patterns and make predictions. ML is an umbrella term that, depending on taxonomies, includes not only statistical learning models (e.g., logistic regression, support vector machine, k-means) and ensemble models (e.g., random forest), but also deep learning models, such as artificial neural networks, transformer-based models, and the most recent large language models.

To recap, applications in verbal deception detection draw from the interplay between NLP and ML. NLP methods enable the extraction of numerical features from textual data, which will be used to train ML models in distinguishing between truthful and deceptive statements in a supervised classification task. A previous meta-analytical work on the topic showed the potential of early computational approaches, primarily based on LIWC, for automated detection of verbal deception (Hauch et al., 2015). By integrating operational definitions for 79 cues across 44 studies, the meta-analysis revealed systematic differences between liars and truth-tellers. For example, liars exhibited a greater cognitive load, distanced themselves from the events, and used fewer sensory-perceptual words. Although the small effect sizes and the impact of some moderators (e.g., even type, emotional valence, and motivation), these findings provide empirical support for the notion that linguistic and diagnostic cues of deception can be detected by computer programs.

Expanding on this, a subsequent systematic review revealed a growing body of research applying ML to deception detection in the last decade, involving diverse datasets, algorithms, and multimodal approaches (Constancio et al., 2023). However, this review focused more on how ML can be employed to predict deception from any cues, including non-verbal, verbal, and paraverbal cues, and did not specifically investigate how NLP was employed for automated detection of verbal deception. This leaves open the need for a systematic review dedicated to automated verbal deception detection, which would consolidate existing approaches, highlight methodological gaps, and guide future research in this area.

1.3 Aim and research questions

In this thesis, automated verbal deception detection is defined as the application of computational techniques (e.g., stemming from NLP) to extract verbal features that will be used to predict human deception in statements through ML models.

Research on automated verbal deception detection is currently fragmented across various disciplines, but mostly within psychology and computer science. These two disciplines differ in terms of goals, methods, and traditions. For example, psychological research has mostly focused on explaining phenomena by relying on experiments that allow hypothesis testing and the development of new theories. On the contrary, research in computer science aims at predicting outcomes by developing ML models trained on previously collected data and with a greater focus on achieving the highest performance. Furthermore, re-

search practice in these areas involves several degrees of freedom, including choices related to dataset characteristics, extraction of linguistic features, and analytical approaches (e.g., algorithmic modeling, training procedure, and evaluation metric). These variations complicate efforts to synthesize findings and establish best practices in the domain, underscoring the need for a systematic mapping of current research practices in automated verbal deception detection.

Furthermore, given the recent advances in NLP and its related applications to verbal deception detection, updates to the field are necessary. Recent developments have moved beyond the static dictionary-based approaches analyzed in Hauch et al. (2015). For example, the development of machine-learning approaches, such as NER, for a richer representation of text is particularly relevant for deception detection and has been previously explored by Kleinberg, Mozes, et al. (2017). Transformer-based architectures, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), moved the field toward more dynamic and context-aware representations of text (see Fornaciari et al., 2021, for applications in deception detection). Advances in pre-trained language models also facilitate transfer learning across domains (Chung et al., 2022), enabling the fine-tuning of models for deception detection purposes (Loconte et al., 2023). These innovations justify revisiting earlier findings and assessing how modern NLP methods are applied to investigate the detection of verbal deception.

The present chapter, therefore, builds on and expands previous works (Constancio et al., 2023; Hauch et al., 2015) by providing a structured overview of the state of automated verbal deception detection by addressing the following research questions:

i. How was deception operationalized?
ii. Which datasets have been used to investigate deception, and what are their characteristics?
iii. What types of linguistic features have been employed for deception detection?
iv. Which analytical approaches have been employed?

2. Methods

2.1 Protocol and Preregistration

This review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) protocol (Page et al., 2021) and the whole protocol was preregistered following the Generalized Systematic Review Registration in Open Science Framework

(https://osf.io/wxme6/). Figure 1 shows the adapted and condensed version of the PRISMA flow diagram (Page et al., 2021), that summarizes the records selection process (the version of the PRISMA flow diagram with full details is available at https://osf.io/rmuq8/files/jt5c4) .

2.2 Eligibility criteria

Identified records were selected according to predefined inclusion and exclusion criteria[1]. First, eligible records were empirical works, published in English in peer-reviewed journal, conference proceeding, book chapters, and preprints. Opinion papers, systematic reviews and meta-analysis were excluded.

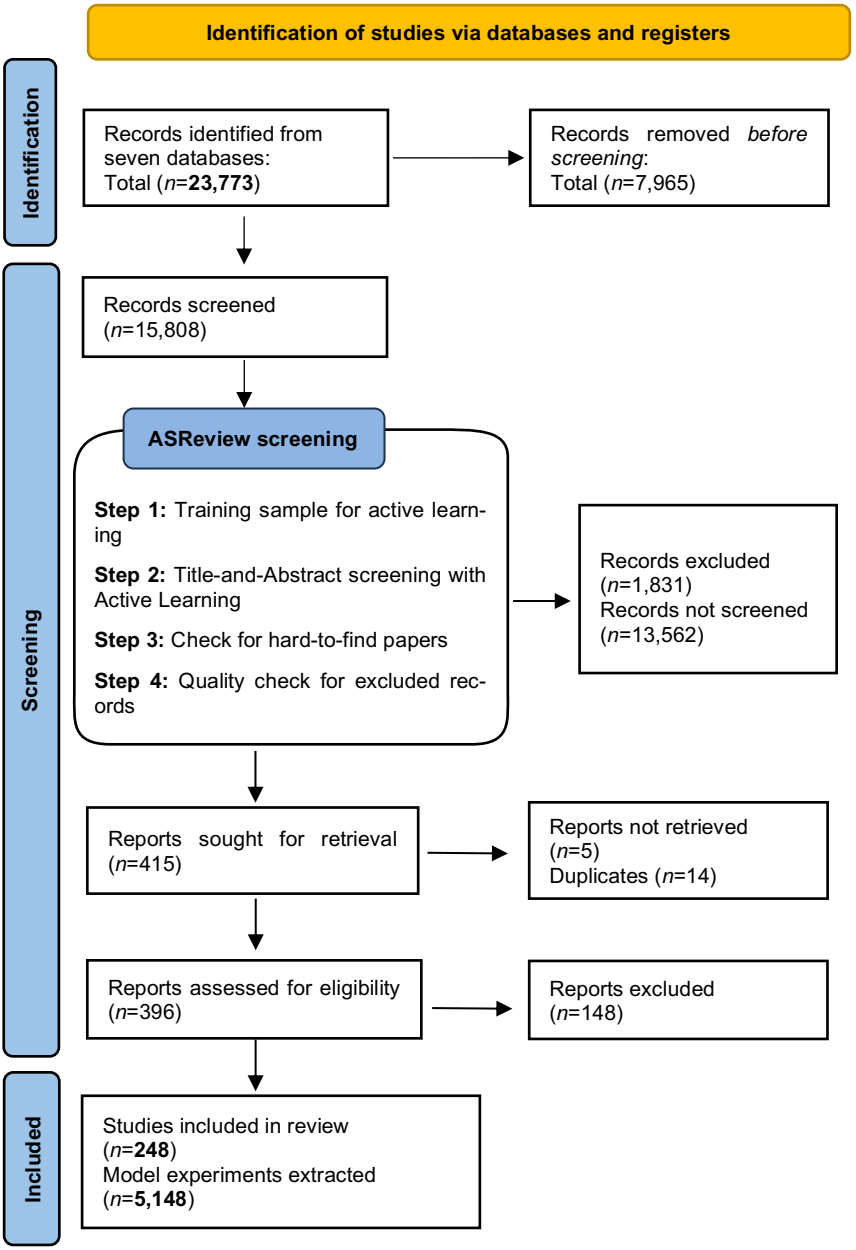
Second, records eligible for this review had to address the topic of automated detection of human deceptive statements (either typed or transcribed) from verbal cues. Only records that applied automated techniques for both the extraction of features or verbal cues and for the prediction judgment were included. Specifically, all types of verbal cues were considered eligible, as long as they were automatically extracted. In contrast, research targeting non-verbal indicators—such as vocal, facial, body, or physiological signals—was omitted. Records on multimodal deception detection were included if they also reported the performance of models trained exclusively on textual features.

Furthermore, eligible records had to utilise predictive models and report classification metrics in their analysis. Research solely reporting explanatory findings (e.g., using ANOVA for explaining main and interaction effects of factors), relying on unsupervised methods (e.g., clustering), or focusing on the accuracy of human judgment - in the absence of a computational model - was excluded.

Third, all forms of human deception, such as fabrication, omissions, embedded lies, were eligible. However, following our definition of deception, research related to i) fake news or misinformation, ii) deception by LLMs, and iii) the detection of human-written vs AI-generated content was excluded.

[1] Records addressing the topic of automated verbal deception detection but with very low quality of reporting were excluded. Low quality of reporting consisted in either big inconsistencies about the methods and findings within the main text or between what written in the sections and what reported in tables/figures or performance metrics reported only in figures and hardly derivable. Among 396 eligible records, only 23 were excluded for this reason.

FIGURE 1. Adapted and condensed version of the PRISMA flow diagram.



2.3 Search strategy

Seven databases were searched for relevant research. Literature in computer science and computational linguistics was searched in (1) the *ACL Anthology*, (2) the *ACM Digital Library*, (3) *IEEEXplore*. Literature in psychology was searched in (4) *Web of Science*, (5) *Scopus*, and (6) *PsychINFO*. Preprints were searched on (7) *ArXiv*. The search strategy included terms associated with the following semantic areas: (1) automated, (2) verbal, (3) deception, and (4) detection. The employed full search string is reported in Supplementary Materials (SM). A total of 23,773 records were identified. After removing 7,520 duplicates, 153 books, 147 dissertations, 13 retracted papers, and 132 records marked ineligible by Zotero, 15,808 records remained for title-and-abstract screening. We confirmed the quality of our search strategy by double-checking the inclusion of key records from two past relevant reviews and meta-analyses on deception detection (Constancio et al., 2023; Hauch et al., 2015). Key records were defined and preregistered before starting the search (see SM2 for the reference list) and used later to also assess the quality of the title-and-abstract screening phase.

2.4 Data selection with ASReview LAB

ASReview LAB v.1 (https://asreview.nl) was used for the title-and-abstract screening. ASReview LAB is a free, open-source tool for screening and labelling a large collection of records for systematic reviews and meta-analysis (van de Schoot et al., 2021). The core of the tool relies on active learning, wherein the researcher – as human-in-the-loop - labels records as relevant or irrelevant for the review and is in interaction with a ML model, which updates a ranking algorithm to continuously re-rank the literature records from the most to the least relevant. Each time the researcher labels a new record, the rank order is updated. The active learning cycle is repeated until a stopping rule is reached (i.e., a predefined criterion that makes the screeners sufficiently confident that all relevant records were seen).

For this review, the title-and-abstract screening through ASReview was conducted by two screeners, who independently screened records, following the SAFE procedure (detailed below, Boetje & van de Schoot, 2024). The whole screening eventually resulted in 415 records deemed eligible for retrieval.

Phase 1: Screening a random set for training data

The screening with AS Review requires adding prior knowledge to a model (i.e., a small set of records labelled as relevant or irrelevant). The

best procedure recommended in the SAFE guidelines (Boetje & van de Schoot, 2024) consisted of labeling a random and stratified 1% of total records (n=158 out of 15,808) for the training data. Stratification was done for years of publication, database (e.g., ACL, ArXiv, Psychinfo) and type of document (e.g, preprint, full paper, conference proceeding). Two researchers independently screened the records from the same training set and solved disagreements by discussion. Of the 158 records, eight were marked as relevant and 150 as irrelevant.

Phase 2: Screening titles-and-abstracts with active learning

Two screeners uploaded the same training data labelled in Step 1 as prior knowledge to train the same active learning model but started screening records independently. For this stage, a naïve-bayes model was trained on a TF-IDF document representation (for details see SM3). Both screeners stopped the screening after all of the following criteria were met (Boetje & van de Schoot, 2024):

1. All key records were marked as *relevant*;
2. At least 10% of the total records was screened (n=1,500);
3. At least twice the estimate of relevant records (ERR)[2] was screened (n=1,600);
4. No relevant records were identified in the last consecutive 50 records.

After the independent screening ended, the records that were seen by only one of the two screeners were evaluated by the other screener and all disagreements were resolved by discussion. A total of 1,693 titles-and-abstracts were screened, of which 381 (22.50%) were labelled as relevant.

Phase 3: Locating hard-to-find records

As a safeguard against missing records in the screening procedure described above, a final round of automated screening was performed by switching to a more advanced text classification model (see Teijema et al., 2023). With the partly labelled dataset obtained from Step 2 as prior knowledge, we used a XGBoost model trained on a TF-IDF[3] text representation to initiate a new active-learning cycle. The stopping rule was

---

[2] The ERR is a crude estimate of the number of relevant records in the total dataset computed by dividing the number of relevant records found in the training set by the number of records in the training set, multiplied by the total number of records in the whole dataset. In this case, the estimate of relevant records was equal to ERR = (8/158) * 15,808 = 800.

[3] We preregistered this phase slightly differently as we aimed to employ a logistic regression trained on a sBERT document representation to rely on embeddings instead of word-

set to 100 consecutive irrelevant records. After the independent screening ended, records seen by only one of the two screeners were evaluated by the other screener and all disagreements were resolved by discussion. In this phase, an additional 375 title-and-abstracts were screened, with six records identified as relevant.

Phase 4: Quality check for excluded records

Lastly, to assess whether records were incorrectly excluded (e.g., due to screening fatigue), a second screening was conducted on excluded records (see Boetje & van de Schoot, 2024). This quality check was run by using as prior knowledge the ten highest- and lowest-ranked records from Phase 3 with a simple model (here: naïve bayes + TF-IDF). In other words, the two screeners double-checked for the initially excluded records but rank-ordered on relevance scores until the stopping rule of 50 consecutive excluded records was reached. After the independent screening ended, records seen by only one of the two screeners were evaluated by the other screener and disagreements were solved by discussion. Of 1,831 initially excluded records, 220 title-and-abstracts were doubled-checked, and 26 records were marked back as *relevant*.

2.5 Full-text screening

After the titles-and-abstracts screening, 415 eligible records were assessed for full-text screening. Of these, five could not be retrieved and 14 were duplicates, leaving 396 records for full-text eligibility assessment. After full-text review, 148 records were excluded for reasons such as: language other than English (*n*=1), focus on human deception detection (*n*=3), use of synthetic/generated data (*n*=7), non-empirical work (*n*=9), manual extraction of verbal cues (*n*=5), low quality of reporting (*n*=23), focus on fake news/rumor detection (*n*=25), no use of supervised ML models (*n*=26), and lack of focus on verbal deception detection (*n*=49). Ultimately, 248 records were included in the review: 121 (48.79%) conference papers, 116 (46.77%) peer-reviewed journal articles, and 11 (4.44%) preprints. Full list of included records is available at https://osf.io/rmuq8/files/r6kyp.

frequency. However, given that the switch to a transformer-based model was too computationally intense for our resources, we decided to keep the TF-IDF text representation and switch to a more computationally comple model for training (here: XGBoost instead of logistic regression). The stopping rule was initially set to 50 irrelevant consecutive records but then increased to 100 for a more extensive screening.

2.6 Coding scheme

Data extraction was conducted on the full-text versions of the final set of 248 records, using the coding scheme presented in Table 1. We piloted the initial coding scheme and trained three reviewers by independently coding 42 records (16.93%), which allowed us to refine the coding scheme for its final form and confirm the agreement of included records among coders (percentage agreement = 0.78 - 0.79). Based on a satisfactory agreement and the large number of records, we assigned a subset of records to three independent reviewers[4]. All reviewers had the possibility to discuss with each other how to handle difficult cases.

TABLE 1. List of overarching categories, labels, descriptions, and levels of coded variables for the included records.

Item category	Item label	Description	Levels
General information	Author(s)	Author's last name or first author's last name with the abbreviation *et al.* when appropriate	For example, *Ott, Ott et al.*
	Years of publication	Year of publication (YYYY)	For example, *2024*
	Title	Title of the publication	For example, *Examining embedded lies through computational text analysis*
	Publication type	Type of publication ranging from journal article, conference paper, and preprints	*Journal article* and *Conference paper*
Deception operationalization	Type of deception	Label of the type of deception investigated	*Concealment, Embedded lies, Exaggeration or Minimization, Fabrication, Falsification, Social desirability, Mixed,* and *Unclear*
	Topic of investigation	Label of the topic employed to study deception	For example, *fake reviews, personal opinions, past experiences*
	Ground truth	Label of the level of clarity of the ground truth for the data used	*Clear and verifiable, Clear but not verifiable, Directly inferred, Indirectly inferred, Mixed,* and *None*
Dataset	Data reuse	Label of whether the record collected new data or analyzed	*Yes, No,* and *Unclear*

[4] Initially, we planned to have two independent coders coding full texts of all included studies, followed by verification through inter-rater agreement and discussion. However, given the large number of included records, we decided to split the task between three coders, after reaching substantial agreement on an independent training set of 42 records.

Item category	Item label	Description	Levels
		already available dataset	
	Language	Language of the dataset used	For example, *English, Italian, Dutch*
	Data source	Source of the data collection	*Online*, and *Offline*
	Research design	Research design for data collection	*Experimental, Quasi-experimental, Naturalistic, Mixed,* and *Unclear*
	Size	No. of statements/ utterances used for training the model or available in the dataset	For example, *1,200*
Linguistic features	Features category	Label of the category that groups a type of linguistic features used to train the model	*Linguistic and statistical features, Term frequencies, Embeddings, Topic and latent semantic features,* and *Hybrid approaches*
	Connection to theories	Label of whether the features extracted are connected to theories	*Yes, No, Unclear*
Analytical approach	Model category	Label of the category of ML model employed for the experiment	*Statistical models, Ensemble models, Neural networks, Transformer-based models,* and *Large language models*
	Evaluation	Label of the procedure adopted to evaluate the quality of the training	*Hold-out test-set, Within cross-validation,* and *Different dataset*
	Metrics	Metric used to evaluate the quality of the training	*Accuracy, F-1 score, Precision,* and *Recall*

Note. All levels of coded variables are reported in the Levels column (in *italic*), except for those variables that contained too many levels (e.g., Language), for which only examples are provided.

Using a spreadsheet (available at: https://osf.io/rmuq8/files/8pxv2), full texts were coded at the level of the individual ML model, meaning that we created a separate entry for each model trained to perform automated verbal deception detection, with multiple entries within the same record. This approach allowed us to capture model-specific features while maintaining the ability to aggregate results at the record-level when appropriate. Following an inductive (bottom-up) approach, variables, such as type of deception and topic of investigation, dataset reuse, language, and size, features and model categories, were first coded following what was reported in the full-text and then eventually mapped

into overarching categories. In contrast, a deductive (top-down) approach, where full texts were coded using pre-defined codes, was employed for coding the variables of data source, ground truth, and evaluation procedure.

Data source and research design

A hierarchical coding scheme was developed to systematically capture the source of datasets. At the first level, data sources were classified as either online or offline and further divided into research design within three levels: experimental, quasi-experimental, and naturalistic design (see definitions in SM 4). Records that combined multiple data origins were coded as mixed sources.

Ground truth

The term ground truth (GT) originates from remote sensing science and geography and refers to when humans verify digitalized information (e.g., pictures) about a landmark taken from the air (Hoffer, 1972). By extension, this term has become very popular in deception detection research, and it has been intended as what we know to be true as proved by empirical evidence (e.g., direct observation and measurement) and not by inference (Kühne et al., 2024). To code GT in the included records, we developed a deductive coding scheme that coded GT within six levels (Table 2):

TABLE 2. Explanations and examples of the six levels used to code ground truth.

Level	Explanation	Examples
Clear and verifiable	The GT is fully transparent and verifiable, allowing confirmation of whether each statement is truthful or deceptive	Object descriptions, videos, mock crimes
Clear but not verifiable	The GT is clearly defined through experimental manipulation, but the experimenter has not the full control to verify individual statements. This means GT is mostly based on participants compliance to instructions.	Assigning participants to truthful vs. deceptive condition to write personal opinions or recount a past holiday, without the possibility of double-checking their statements.
Directly inferred	The GT is inferred from plausible but indirect indicators, making its validity error prone and open to interpretation	Court trial transcripts where deception is inferred from the final verdict, assuming the judge is always correct.

Level	Explanation	Examples
Indirectly inferred	The GT is based on weak inference methods, reducing confidence in its reliability	Relying on human impressions of deception or on classifications stemming from a filtering algorithm, without objective validation and clear specifications
Mixed	The record combines multiple datasets with varying levels of GT clarity resulting in heterogeneous GT reliability	A record working on multiple datasets, merging experimentally manipulated data with real-world court trial transcripts
None	No documentation is available regarding how GT was established	Lack of methodological details in the record or use of datasets not documented in an academic outlet.

Abbreviations: GT = ground truth

Evaluation

Three overarching categories were predefined to code the evaluation approach of trained ML models:

1. *Hold-out test set*, including models tested on a random subset of the dataset that was not used during training.
2. *Within cross-validation,* including models evaluated through cross-validation, in which the dataset is randomly split into k partitions (folds). The model is iteratively trained on $k - 1$ folds and tested on the remaining one, with the overall performance computed as the average of the evaluation metrics across all k iterations. This procedure is typically more robust than the hold-out test set approach, as it reduces the impact of random variation in test data composition.
3. *Tested on different data*, including models trained on one dataset and tested on a separate, independent dataset.

3. Results

The final dataset for this review consisted of 248 records, of which 48.79% of records (n=121) were conference papers, 46.77% (n=116) consisted of peer-reviewed journal articles, and 4.44% (n=11) were preprints. The number of included records per publication year spanned from 1982 to 2025, with an increased number of publications in the last 15 years (see Figure S1 in SM5).

In the following sections, findings related to deception operationalization and dataset characteristics are reported at the record level, while findings related to linguistic features and the analytical approach (e.g., model categories, evaluation procedure, and performance metrics) are reported at the model level.

3.1 Deception operationalization

Types of deception

The type of deception investigated varied across the included records. The most frequently examined form was fabrication (e.g., production of a made-up statement), reported in 59.68% of records (n=148). For 35.48% (n=88) of records, mixed forms of deception were investigated as a result of either tasks where participants were free to choose and combine any forms of deception or combinations of different datasets for the same experiment. The remaining 4.44% of records involved less frequently studied types of deception, such as embedded lies (n=5), falsification (n=3), concealment or omission (n=2), and social desirability (n=1). Only one record was unclear in the type of deception investigated. This distribution indicates a strong research focus on fabrication, with other forms of deception receiving limited attention. Noteworthy, coding the type of deception proved challenging, as this information was rarely stated explicitly.

TABLE 3. Number of records investigating different forms of deception by source and research design.

	Experimental		**Quasi-experimental**		**Naturalistic**	
	Offline	Online	Offline	Online	Offline	Online
Concealment	2	-	-	-	-	-
Embedded lies	2	2	-	-	1	-
Fabrication	16	16	2	54	7	37
Falsification	2	-	-	-	-	1
Social Desirability	1	-	-	-	-	-
Mixed	18	5	-	1	37	11
Unclear	-	-	-	-	-	1

Note. Records (n=32) that relied on multiple sources of data collection were excluded from this table as the type of deception could not be derived.

When examining the interplay between the type of deception, data source, and research design, Table 3 reveals that offline experimental designs addressed the widest range of deception types, including concealment, embedded lies, falsification, and social desirability, although fabrication remained the most frequently investigated. In contrast, quasi-experimental designs focused almost exclusively on fabrication, while naturalistic designs predominantly examined fabrication and mixed forms of deception. Overall, experimental approaches appeared to be the most versatile research design for covering different types of deception.

Topic of deception

Concerning the topic of deception, substantial variability was found among the included records. Most research focused exclusively on online fake reviews (n=107, 43.15%). In addition, 17.74% of records focused on past events (n=44), including either real-life data such as trial hearings and experimental settings where participants were asked to recollect memorable events or take part in a mock crime on campus. Most of these records aimed to investigate deception in episodic and autobiographical memories in order to derive insights relevant to the legal and forensic domains. Other commonly studied topics included personal opinions on sensitive topics (e.g., abortion, death penalty, and immigration policy; n=18, 7.26%) and lies about identity and personal information (7.26%, n=18). In contrasts, topics rarely addressed included financial or scientific fraud (3.23%, n=8), email or sms spam (2.02%, n=5), and cheating on a test (n=5, 2.02%), personal feelings (n=4, 1.61%), future intentions and promises (n=3, 1.21%), object description (n=3, 1.21%), and open-domain statements (n=3, 1.21%). A total of 12% of works (n=30) combined multiple datasets, resulting in multiple and mixed topics of deception. This distribution highlights a strong emphasis on online deception in fake reviews and past experiences.

Ground truth

The quality of GT was coded across records within six levels. Only 13.71% (n=34) of records provided a *clear and verifiable* GT (e.g., object description, video recollection), while 41.13% (n=102) of records offered a *clear but not verifiable* GT. A typical experimental manipulation for this GT level consists of assigning participants to two different conditions (i.e., truthful vs deceptive), but the veracity of statements relies on participants' compliance to follow the instructions (e.g., personal opinions, past holidays), without the possibility to fully double-check what is reported by participants. Another 14.11% (n=35) of records were classified as *directly inferred* GT, given that GT was inferred from plausible but indirect indicators (e.g., trial hearings of false testimony). Moreover,

11.69% (*n*=29) of records were labeled as *indirectly inferred* GT because they relied on uncommon or questionable inference methods, such as on human impression of deception for classifying job posts (whose accuracy is known to be at chance level) or on a non-transparent filtering algorithm of the *Yelp* website for fake reviews. Finally, 15.32% (*n*=38) of records were categorized as mixed, incorporating datasets with varying levels of GT reliability. For only 4% of records (*n*=10), no information on GT was available. These findings reveal significant variability in the rigor of GT manipulation and documentation across deception detection research.

3.2 Datasets used in automated verbal deception detection

Automated verbal deception detection was investigated using different data collection approaches. Among the included records, the majority (*n*=133, 53.63%) reused existing datasets, while a minority (*n*=89, 35.89%) collected new datasets. For 4.43% of records (*n*=11), it was unclear whether they reused available datasets or collected a new one. Additionally, 6.05% of records (*n*=15) used a mixed approach, combining newly collected data, previously available datasets, or data of unclear origin.

A total of 51 datasets were reused and reanalyzed for new research (see Table S2 for full details). The most frequently reused dataset was the Deceptive Opinion Spam dataset (Ott et al., 2011, 2013)[5], which accounted for 46.61% of the reused datasets and was used in 25% (*n*=62) of all records. This dataset contains 800 genuine reviews scraped from TripAdvisor and 800 crowdsourced fake reviews of the 20 most popular hotels in Chicago. The second most reused dataset was the Real-life trial dataset (Pérez-Rosas et al., 2015), accounting for 16.54% of reused datasets and being used in 8.87% (*n*=22) of records. It consists of 121 transcripts from videos of public court trial hearings. The third most reused dataset entails customer-generated truthful reviews, crowdsourced-generated deceptive reviews, and employee-generated deceptive reviews (Li et al., 2014). The dataset is composed of 3,032 reviews about hotels, restaurants, and doctors, and accounted for 11.27% of re-used datasets, but was used in only 6.05% (*n*=15) of records.

[5] Despite this being two different datasets published in two different conference papers and years, most research incorrectly referred to them as one single dataset, sometimes citing only one of the two references. Therefore, we decided to combine the two datasets together for easiness of computation.

Language distribution

Most records (82.26%, *n*=204) worked with data in English. Other languages were underrepresented, with only 2.42% (*n*=6) of records in Chinese and 2.42% (*n*=6) of records in Italian. Additional languages included Arabic (*n*=3, 1.2%), Polish (*n*=3, 1.2%), Spanish (*n*=3, 1.2%), Thai (*n*=2, 0.8%), Mandarin (*n*=2, 0.8%), Russian (*n*=2, 0.8%), and single instances of Bengali (*n*=1, 0.4%), Dutch (*n*=1, 0.4%), German (*n*=1, 0.4%), Korean (*n*=1, 0.4%), Indonesian (*n*=1, 0.4%), Roman Urdu (*n*=1, 0.4%), and Serbian (*n*=1, 0.4%). Finally, six records (2.42%) utilized mixed-language datasets, and in four records (1.6%), the language of the datasets used was unspecified and unclear. This distribution underscores the strong dominance of English-language resources in the field.

Data source and research design

Data collection methods were distributed across offline and online settings. Most records (*n*=128, 51.61%) relied on online sources, while 35.48% of records (*n*=88) gathered data offline. Only 12.90% of records (*n*=32) utilized mixed sources, integrating both online and offline data. Of those collecting data online, 42.97% of records (*n*=55) relied on online quasi-experiments, lacking full control on conditions randomization, 39.06% of records (*n*=50) used naturalistic designs by scraping data from online platforms (e.g., Twitter, Amazon, Yelp), and only 17.97% of records (*n*=23) conducted online controlled experiments (e.g., via Prolific or Amazon Mechanical Turk). In contrast, of those collecting data offline, 51.13% of records (*n*=45) gathered data from real-life contexts using naturalistic designs (e.g., trial hearings, TV game shows), 46.59% of records (*n*=41) relied on laboratory-controlled experiments, and only 2.27% of records (*n*=2) employed quasi-experimental designs.

Size

Dataset sizes varied substantially across sources and research designs (Table 4; see Table S1 in SM6 for descriptive statistics reported at the model-level). Overall, offline sources yielded smaller datasets compared to online sources, reflecting how online platforms enable large-scale studies thanks to greater data availability and accessibility, whereas achieving similar volumes offline requires more substantial effort. Among online sources, naturalistic designs that scraped data from platforms (e.g., Facebook, Amazon, Yelp) produced larger datasets (*Median*=9,216; range: 103–2,542,553). For offline sources, larger datasets came from quasi-experimental studies (*Median*=142; range: 18–17,880), although these accounted for only 0.81% of all records and 2.27% of offline sources, followed by offline experiments (*Median*=300; range: 20–

9,104). This wide range highlights the heterogeneity of resources used in deception detection research, spanning small-scale controlled experiments to large-scale online data, which likely reflect differences in research goals, feasibility, and data accessibility.

TABLE 4. Descriptive statistics of no. of statements (M, SD) by source of data collection and research design.

Source	Design	*N* (%)	*M*	*SD*	*Median*	Range
Offline	Experiment	35 (14.11%)	1,430.54	2182.78	300	20 - 9,104
	Quasi-experiment	2 (0.81%)	1,328.00	667.51	1,328	856 - 1,800
	Naturalistic	43 (17.34%)	1,271.68	3,650.42	142	18 - 17,880
Online	Experiment	22 (8.87%)	1,928.14	2,211.78	1,014.50	38 - 7,168
	Quasi-experiment	54 (21.77%)	2,692.62	5,115.70	1,600	640 – 31,146.68
	Naturalistic	49 (19.76%)	94,018.12	374,226.33	9,216	103 – 2,542,553
Mixed	Mixed	32 (12.90%)	54,791.86	287,698.32	928.10	121 – 1,630,263.25

Note. No. of statements was averaged per record before computing the descriptive statistics by source and design.
Abbreviations. *N* = number of records that reported the number of statements in full text.

3.3 Linguistic features

Linguistic feature categories were coded at the model level (n=5,148) by first following a bottom-up inductive approach and then recoding those codes into overarching categories. We found the following five feature categories:

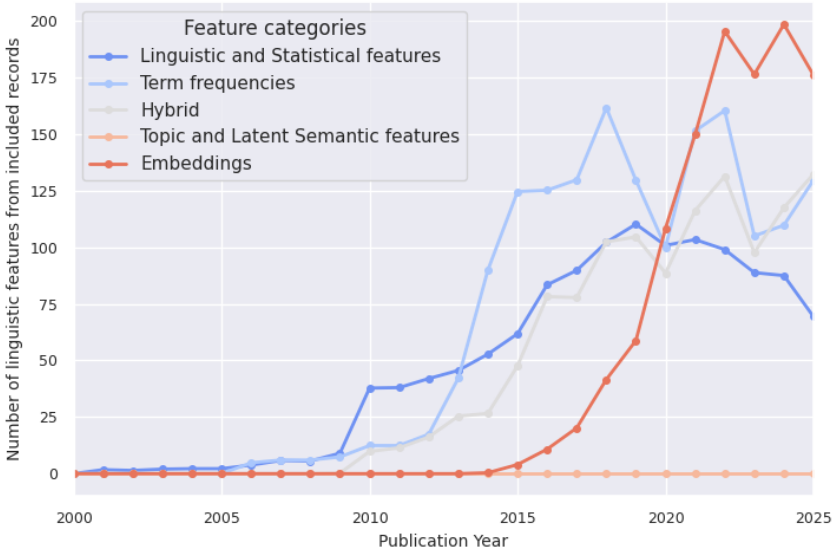
1. *Linguistic and statistical features:* features derived from dictionary-based methods, machine-learning analysis, or statistical properties of text. Examples include Linguistic Inquiry and Word Count (LIWC) features, Part-of-Speech (POS) tags, Named Entity Recognition (NER), syntactic dependencies, readability scores, and lexical diversity measures;
2. *Term frequencies:* text-representation based on word frequency, ignoring word order. It includes simple counts, and binary indicators (i.e., bag-of-words), weighted schemes (e.g., term frequency-inverted document frequency; TF-IDF), and often extended with *n*-grams for limited context (e.g., bigrams and trigrams);

3. *Embeddings*: dense vector representations capturing semantic aspects of textual data. It includes static embeddings (e.g., Word2Vec, Mikolov et al., 2013; GloVe, Pennington et al., 2014; FastText, Joulin et al., 2017), contextual embeddings (e.g., BERT, GPT, ELMo), and sentence/document embeddings (e.g., Sentence-BERT, Doc2Vec);
4. *Topic and latent semantic features*: hidden themes or semantic structures in text using dimensionality reduction or probabilistic models such as topic modeling or Latent Dirichlet Allocation (LDA).
5. *Hybrid Approaches:* combining multiple strategies (e.g., embeddings + linguistic features, or TF-IDF + topic modeling).

Linguistic and statistical features (e.g., LIWC features, POS-tags, and text statistics) accounted for 23.7% (*n*=1,222) of models. Among these, 46.2% (*n*=564) of models explicitly connected feature selection to theoretical frameworks or previous research, while 35.7% (*n*=436) of models were purely data-driven. The remaining 18.2% (*n*=222) of models were unclear regarding their theoretical grounding. However, the most frequent linguistic features relied on *term frequencies*, such as n-grams, bag-of-words, and TF-IDF, representing 30.7% of all models (*n*=1,580). Embedding-based features (e.g., word2vec, BERT embeddings, Doc2vec) were used in 21.8% of models (*n*=1,121). In contrast, *topic and latent semantic features* appeared in only 1.3% of models (*n*=69), and 0.4% of models (*n*=21) were classified as unclear due to insufficient methodological detail. Finally, *hybrid approaches*, which combine multiple strategies such as linguistic features with term frequency, comprised 22.0% (*n*=1,135) of models.

Figure 2 illustrates temporal trends in the use of different linguistic feature categories in deception detection research from 2000 to 2025. Term frequencies and linguistic and statistical features dominated early work, showing growth since 2010, but remaining steady after 2018. Topic and latent semantic features remained less prevalent overall. In contrast, embeddings exhibit a sharp and sustained increase starting around 2018, surpassing all other feature types by 2021 and maintaining the highest frequency through 2025. Hybrid approaches emerged gradually after 2010, reaching moderate levels compared to term frequencies and embeddings. These patterns indicate a methodological shift from traditional lexical and statistical features toward representation learning approaches, reflecting the technological evolution of NLP techniques for text representation.

FIGURE 2. Distribution, at the model-level, of feature categories over time in verbal deception detection research.



Note. One model from 1982, included in the systematic review, was excluded from the plot for a better visualization. Displayed values were computed using a 5-year rolling average to minimize volatility in the visualization.

3.4 Analytical approach

Model categories

Related to the analytical approach of the included records, the model categories and evaluation approaches were coded at the model level (n=5,148). Model categories were first coded following a bottom-up deductive approach and then recoded into overarching categories. The following five model categories were found:

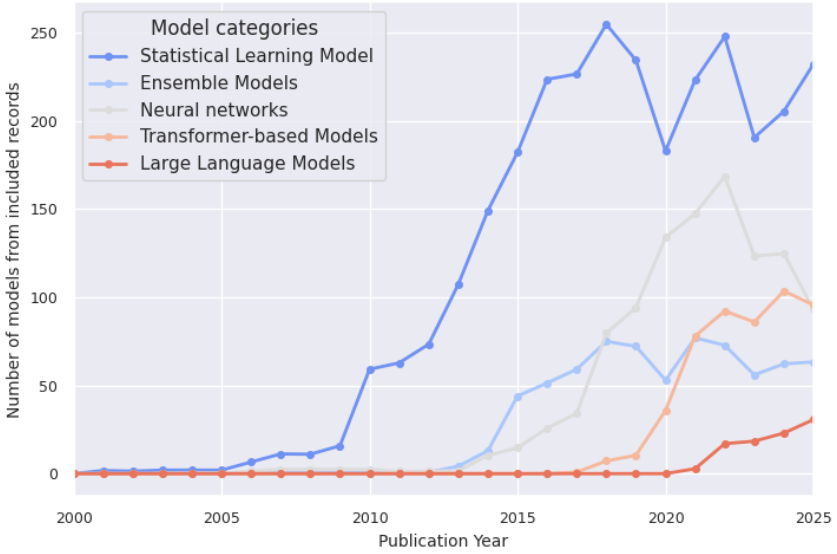
1. *Statistical learning models,* including models that require explicit text representations (e.g., linguistic and statistical features, bag-of-words, and embedding representations) and are based on statistical principles, such as logistic regression, naïve bayes, and support vector machine (SVM).

2. *Ensemble models*, including models that combine predictions from multiple models to enhance overall accuracy and robustness. Common ensemble techniques include bagging, boosting, and voting. Some algorithms, such as Random Forest, XGBoost, and LightGBM, inherently implement these ensemble strategies.
3. *Neural networks,* including models that learn representations through interconnected layers and neurons through shallow (e.g., feedforward networks) and deep architectures (e.g., convolutional neural networks, CNN; recurrent neural networks, RNN; long-term short memory, LSTM)
4. *Transformer-based language models,* including deep learning models pretrained on large corpora capturing long-range dependencies and contextual meaning thanks to self-attention mechanisms (e.g., Bidirectional Encoder Representation from Transformers, BERT; Robustly Optimized BERT Pretraining Approach, RoBERTa). These models can be fine-tuned on smaller, task-specific datasets for classification without requiring retraining from scratch.
5. *Large language models,* including large-scale, transformer-based models pretrained on extensive corpora and capable of performing few-shot or zero-shot classification through prompt engineering (e.g., GPT-4.5, FLAN-T5, Llama-2).

The analysis revealed that statistical learning models, such as logistic regression, SVM, and naïve bayes, were the most frequent (n=2,842, 55.22%) of all models. Models using neural network architectures, such as CNN, RNN, and LSTM, represented the second largest group at 20.52% (n=1,056). Ensemble Models, which combine predictions from different models, constituted 13.19% (n=673) of models. Transformer-based models, such as BERT models and their variants (e.g., RoBERTa, distilBERT), accounted for only 9.17% (n=472), while LLMs were the least used models, comprising only 1.90% (n=98).

In terms of models' adoption over time, Figure 3 shows clear trends driven by the technological evolution in NLP and the availability of large-scale datasets. Statistical learning models dominated from 2010 onward, peaking around 2018, and remained, overall, the most frequently used models. Ensemble models gained traction after 2015, reaching their peak near 2020 before experiencing a slight decline. Neural networks emerged around 2015 and have continued to grow steadily, marking a shift toward deep learning approaches. Transformer-based models emerged around 2019, followed shortly by LLMs, which appear only in the most recent years.

FIGURE 3. Distribution, at the model-level, of model categories over time in verbal deception detection research.



Note. One model from 1982, included in the systematic review, was excluded from the plot for a better visualization. Displayed values were computed using a 5-year rolling average to minimize volatility in the visualization.

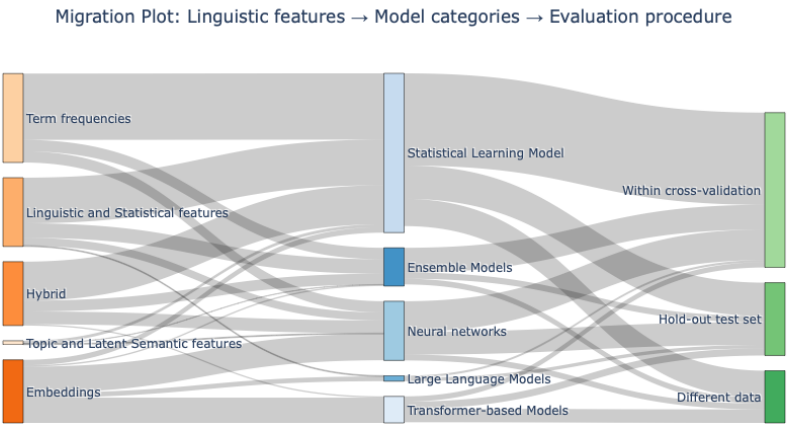
Evaluation procedure and performance metrics

Concerning the evaluation procedures, 53.44% of models (n=2,751) were evaluated using within cross-validation approaches, indicating a strong preference for methods that provide more robust and reliable estimates of model performance. Another 25.14% (n=1,294) employed a hold-out test set for evaluation, and 18.05% of models (n=929) were tested on different data, including both new data from the same domain and data from different domains. Only 0.52% of models (n=27) employed an in-sample evaluation procedure and, for the remaining 2.85% of models (n=147), the evaluation approach was not specified in the record. Evaluation performance was reported in terms of accuracy (n=2,246, 43.62% of models) or in terms of F1 score (n=1,255, 24.38% of models). Only 27.39% (n=1,410) of models reported both metrics, while the remaining 4.6% (n=237) used other metrics, such as the area under the curve (AUC), precision, or recall

Methodological interplay

We finally investigated the interplay between the three key methodological choices in automated verbal deception detection: i) linguistic features categories (e.g., term frequencies, linguistic and statistical features, embeddings), ii) model categories (e.g., statistical learning models, ensemble models, neural networks), and iii) evaluation procedures (i.e., cross-validation, hold-out testing, testing on different datasets). Figure 4 illustrates the heterogeneity of feature–model–evaluation pathways and the predominant use of statistical learning models trained on term frequencies and linguistic and statistical features, typically evaluated within cross-validation frameworks. In contrast, embedding representations were mostly used to train neural networks, transformer-based models, and LLMs.

FIGURE 4. Sankey diagram illustrating the interplay between three methodological choices: linguistic features, model category, and evaluation procedure.



4. Discussion

Given the recent rapid advances in computational methods for analyzing textual data, this systematic review updates the state of automated verbal deception detection by focusing on deception conceptualization and current research practices. With this aim, we reported the combined results of 248 records, including journal articles, conference papers, and preprints, with a total of 5,148 included models. To the best of our knowledge, this is the largest systematic review available on the topic. While a previous meta-analysis of 44 studies investigated automated

verbal deception detection primarily through dictionary-based approaches (Hauch et al., 2015), this work covers a broader array of methods that have been developed and employed over the last decade. Furthermore, we build upon a previous systematic review on the use of ML for deception detection (Constancio et al., 2023) by providing a more extensive and focused knowledge on automated verbal deception detection. Specifically, we include a larger number of records from seven databases, covering both published journal articles and grey literature, and focus exclusively on the application of computational methods for identifying deception through verbal cues, rather than any other potential cues. We discuss below the key strengths of research practices and knowledge gaps in the area, underscoring the need for further research on less-investigated forms of deception and on safeguards to improve the generalizability of findings.

4.1 Deception investigation

Internal validity

In deception research, the level of ground truth (GT) is crucial, as it defines what we know to be true, as proven by empirical evidence, and in contrast to what can be simply inferred (Kühn et al., 2024). Establishing a reliable GT is fundamental for internal validity because it ensures that observed effects can be attributed to actual deception rather than measurement error or ambiguous labelling (Vrij et al., 2010). Without a robust GT, the causal relationship between linguistic cues and deceptive behavior becomes questionable (Vrij et al., 2010), introducing confounds, such as different data origins or errors in data labelling, that undermine the reliability of findings. High internal validity depends on minimizing such threats, as it directly impacts the strength of causal inference and the reproducibility of results (Cook et al., 2002).

Our findings show that only 13.71% (n=34) of records provided a *clear and verifiable* GT. These studies employed tasks that, for example, relied on object description (e.g., Soldner et al., 2019), video recollection (e.g., Bond & Lee, 2005), and mock crimes (Matsumoto & Hwang, 2015), allowing the researcher to fully verify participants' statements. In contrast, most research (41.13%, n=102) studied deception under a clear but not fully verifiable ground truth. Examples of this research include studies that experimentally manipulate deception by assigning participants to two different conditions (i.e., truthful vs. deceptive), but lack control over what participants report, thus relying mostly on participants' compliance with the instructions. For example, when participants are asked to fabricate a past holiday in Mexico to resemble those who have actually

visited the country, researchers typically have limited control over how the fabricated statement is generated and to what extent participants truly make up details or draw on similar previous truthful experiences to fabricate (see Loconte & Kleinberg, 2025 for findings on sources of embedded lies). Conversely, previous research has already shown that even truthful statements may incorporate deceptive details, albeit to a lesser extent than in deceptive statements (Markowitz, 2024), thereby undermining the solid assumption that manipulating instructions is sufficient to determine a valid GT.

Even more concerning is that a minority of research worked with a level of GT that was either inferred from direct or indirect indicators or whose operationalization was not documented. For example, for the Real-life trial dataset (Pérez-Rosas et al., 2015), one of the most widely used datasets for investigating deception in ecological settings, video clips of defendants accused of guilt were labeled as deceptive, and clips of witnesses in the same trial were labeled as truthful. This GT labelling, however, overlooks the possibility that witnesses, law enforcement, and even judges may be error-prone and commit mistakes in trials, thereby possibly introducing bias and confounders in the data. Other works relied on even more uncommon or questionable inference methods. For example, one study asked a so-called "expert" practitioner in lie detection to disentangle truthful from deceptive police reports (Quijano-Sánchez et al., 2018), despite the well-established knowledge that experts, as well as laypeople, perform close to chance level in deception detection tasks (Aamodt & Custer, 2006; C. F. Bond & DePaulo, 2006; Hartwig & Bond, 2011). Another study (Belbachir & Alkan, 2022) classified reviews as reliable (i.e., truthful) when they were labeled as *helpful* by at least two users, and as unreliable (i.e., deceptive) otherwise, again assuming that humans are good lie detectors.

These findings highlight a significant lack of rigor in the manipulation and documentation of GT across deception detection research, with important implications for the internal validity of these studies. Without reliable GT, conclusions about linguistic markers of deception risk being speculative rather than evidence-based, as we potentially introduce confounders and weaken causal inferences. Therefore, strengthening the operationalization of GT becomes a fundamental requirement for future research in this area to advance the reliability and theoretical integration of findings.

External validity

External validity, the extent to which research findings generalize beyond the context of the study, is a critical yet often overlooked dimension in deception detection studies (Konečni & Ebbesen, 1979).

Deception was mostly examined in terms of fabrication (n=148, 59.68% of records), which is defined as the production of completely made-up statements and represents the most extreme form of deception. The second most investigated type of deception (n=88, 35.48%) was the mixed form. Mixed forms of deception emerged either because participants were free to use any deceptive strategies, such as combining embedded lies with omissions, mixing exaggeration with minimization, or fabricating statements driven by social desirability, or because multiple datasets were merged, thereby testing models on several forms of deception simultaneously. Research focusing exclusively on other forms of deception (i.e., embedded lies, falsification, concealment, and social desirability) accounted for only 5.25% (n=11) of records.

While focusing on fabrication may facilitate research in magnifying differences between truthful and deceptive statements, real-life deception has been shown to occur mostly in more nuanced forms (Loconte & Kleinberg, 2025; Markowitz, 2024; Verigin et al., 2019), underscoring the need for research that has greater external validity and that covers a broader array of deception forms. To address this need, our findings suggest that experimental research designs are the most versatile, enabling researchers to investigate various deceptive strategies within controlled yet adaptable settings.

Furthermore, we found a strong emphasis in research studying deception in online fake reviews and in past experiences, reflecting two distinct disciplinary focuses. Research on fake reviews originates from computer science and aims to examine deception to develop scalable and automated solutions that mitigate deceptive content in online environments. Conversely, most research on past experiences comes from psychology and aims to investigate deception in episodic and autobiographical memories, using experimental tasks that involve mock crimes or real-life data from trial hearings, to derive insights relevant to the legal and forensic domain.

Finally, most studies investigated deception detection using English-language datasets (n=204, 82.26%). Our findings confirm what was previously found in other studies (Constancio et al., 2023; Hauch et al., 2015): automated verbal deception detection is an English-centric research area. This limits the applicability of findings in other languages and even cul-

tures (Papantoniou et al., 2022), underscoring the need for more multilingual, culturally diverse datasets and novel approaches that move beyond English.

Taken together, these findings reveal that deception research often prioritizes methodological convenience over external generalizability, with a greater preference for fabrication, deception in fake reviews or past events, and English-language contexts. This narrow scope constrains external validity, reducing the applicability of models to real-world settings where deception is typically subtle (Loconte & Kleinberg, 2025; Markowitz, 2024), context-dependent (Blair et al., 2010), and culturally mediated (Papantoniou et al., 2022). To advance the field, future research must engage more in experimental and naturalist designs that capture the full spectrum of deceptive strategies (e.g., embedded lies, omissions, denials), integrate cross-cultural perspectives, and focus on different contexts of deception (e.g., malicious intentions, self-presentations, personal opinions and feelings).

Ecological validity

Ecological validity extends the concept of external validity by concerning the extent to which findings can be generalized to real-world settings (Schmuckler, 2001). In other words, it measures how well the conditions, tasks, and behaviors observed in research reflect what happens in everyday life.

Data sources and research designs are relevant for the ecological validity of findings. We found that data collection methods were evenly distributed across both offline and online settings, as well as various research designs, including controlled experiments, quasi-experiments, and real-life data collection. This diversity is beneficial for enhancing both external and ecological validity and ensuring that findings are not limited to artificial laboratory conditions. By incorporating multiple contexts and designs, studies can better capture the complexity of real-world deception, which strengthens the robustness and applicability of models.

Furthermore, dataset sizes varied substantially across sources and research designs, spanning from small-scale controlled experiments to large-scale online scraped data, with offline sources generally yielding smaller datasets compared to online sources. This reflects how online platforms enable large-scale studies due to greater data availability and accessibility, whereas achieving similar volumes offline requires more substantial effort. This heterogeneity in scale and data origin may also reflect differences among disciplines. In fact, psychology and computer science often differ in research goals, methodology, and data accessibil-

ity, with psychology focusing more on controlled lab experiments to examine the effect of manipulated variables on the outcome, and computer science research aiming at scraping real-life data at a large scale to build robust and efficient models.

In terms of ecological validity, investigating deception in naturalistic settings, such as scraping fake reviews on online platforms (Fornaciari & Poesio, 2014), may differ substantially from lab-based psychological experiments where deception is investigated in mock crimes (Matsumoto & Hwang, 2015). Outside the lab, lie-tellers self-select and engage deception as a problem-solving activity and are not instructed by researchers (Levine, 2018); on the other hand, receivers of deception generally face a truth-lie base rate that differs from the 50–50 distribution, typical of experimental settings, and decisions about honesty are not immediate or prompted by the experimenter (Levine, 2018). However, despite the advantage of capturing different aspects of deception, field studies often face challenges in establishing a reliable ground truth (Vrij et al., 2010). Deception labels are, in fact, frequently inferred rather than verified, compromising internal validity (Vrij et al., 2010). In contrast, laboratory experiments typically allow for greater control over conditions and verification of truthfulness, thereby strengthening causal inference (Cook et al., 2002; Vrij et al., 2010).

These differences underscore the importance of balancing ecological validity with internal validity (i.e., clear definitions and operationalizations of ground truth) when interpreting findings and assessing their generalizability.

4.2 Research practice

Research replicability

Automated verbal deception detection has been primarily investigated through the reanalysis of existing datasets (n=133, 53.63%). In contrast, a minority of studies collected new datasets (n=89, 35.89%) or employed a mixed approach (n=15, 6.05%), combining newly collected data, previously available datasets, or data of unclear origin. Among the studies that reused datasets, we identified a total of 51 unique and open-source datasets (see Table S1 for full details). This reliance on existing resources is often seen as beneficial for addressing the replicability crisis in psychology (Anvari & Lakens, 2018), as it facilitates the verification and comparison of findings across studies. We believe this represents a valuable resource for future research to conduct mega analytical studies involving a diverse array of deceptive settings, topics, and strategies. Conversely, the advantage of creating new datasets lies in guaranteeing data

enrichment, thereby facilitating the development of more robust and generalizable models and helping broaden the external validity of findings. Together, the practice of reanalyzing existing datasets and collecting new ones represents a good balance between enhancing the replicability of findings and generating new knowledge to move forward with research in automated verbal deception detection.

Furthermore, to strengthen reproducibility and transparency, future work should also adopt standardized practices that make models and methodologies more accessible and verifiable. For example, future research should resort to model cards (Mitchell et al., 2019), which consist of concise yet detailed reports of models, to disclose details about datasets, training procedures, evaluation methods, and contexts of use. In addition, adapted protocols for pre-registering predictive modeling should be used to enhance transparency in ML practices (Hofman et al., 2023). One example of such pre-registration flow for predictive modelling include: i) pre-registering the problem statement and research questions, ii) training models, iii) choosing the one that yields the best performance (i.e., validation), iv) pre-registering details of such finalized model and of the evaluation approach, and finally v) test this model on a test set of new collected data. By adopting practices such as model cards and pre-registration protocols, reproducibility and transparency in automated verbal deception detection can be enhanced, fostering greater trust, accountability, and scientific progress in the domain.

The case of deceptive hotel reviews

The Deceptive Opinion Spam datasets (Ott et al., 2011, 2013) were the most frequently reused datasets for examining automated verbal deception detection, accounting for 46.61% of the reused datasets and 25% (n=62) of all records. This dataset comprises 1600 genuine and fake reviews of the 20 most popular hotels in Chicago.

Despite its popularity, this dataset has big limitations that limit the generalizability of findings. One limitation entails the source of data collection. Genuine reviews were scraped from Tripadvisor, while fake reviews were crowdsourced by asking participants to write a fake review that matched a genuine one. This artificial setup introduces a confound, as fake reviews may differ from genuine ones, not because of their veracity, but due to a different data collection procedure. The second limitation is that individuals who wrote these fake reviews lack the experience to lie about it in a credible way. In fact, it is likely that crowdsourced participants, who are paid a few dollars for completing surveys or experimental tasks, represent a different population from those who can

afford an expensive hotel in Chicago, thus introducing a second confound in the dataset. Previous research experimentally manipulated data origin (i.e., if the dataset comes from one or multiple sources) and product ownership (i.e., if reviewers had already some experience with the product) to test their effect on performance (Soldner et al., 2022). Findings showed that these variables constitute important confounds, and when combined, they can inflate the deception detection performance by 24.89–46.23% (Soldner et al., 2022). The fact that 25% of all reported findings stemming from datasets with these inherent limitations raises concerns about the generalizability of these findings and possible overestimations. This suggests that we may have 62 records on automated verbal deception detection, with overestimated findings up to 46.23%. We therefore suggest that when designing new datasets or reanalyzing existing ones, it is paramount to control for the quality of the datasets and the presence of potential confounds to avoid performance overestimation that may compromise the validity of results.

Linguistic features and model interplay

Research in automated verbal deception detection extracted linguistic features that fall within five main categories that primarily differ in their ability to capture information and nuances from text: (1) *Linguistic and statistical features*, (2) *Term frequencies*, (3) *Embeddings*, (4) *Topic and latent semantic features*, and (5) *Hybrid Approaches*. Once features were extracted, studies relied on five clusters of model categories that vary in their computational complexity: (1) *Statistical learning*, (2) *Ensemble models*, (3) *Neural networks*, (4) *Transformer-based language models*, and (5) *Large language models*.

Regarding the linguistic features, all categories were employed in approximately equal measure, with the exception of the *topic and latent semantic features*, which were used to a lesser extent. However, over time, and specifically after 2015, we observed a methodological shift from *traditional lexical and statistical features* toward representation learning approaches, such as *embeddings*. This transition reflects the evolution of NLP techniques for improved text representations and marks a substantial deviation from earlier practices, as highlighted in Hauch et al. (2015), where dictionary-based approaches (e.g., LIWC) dominated verbal credibility research. The transition to embeddings reflects not only technological progress but also a growing emphasis on capturing nuanced semantic and contextual information that static word lists cannot provide.

Regarding model categories, statistical learning models remain the most frequently used models (n=2,842, 55.22%). However, recent years have also seen the emergence of deep learning, transformer architectures, and

the most recent LLMs. This trend reflects the gradual adoption of increasingly complex computational models over time, driven by technological evolution in NLP and the availability of large-scale datasets. This progression mirrors the broader trend reported in previous research (Constancio et al., 2023) and further extends such previous findings by highlighting the recent emergence of transformers and LLM, which were neither yet available nor prominent in previous years.

One advantage of relying on such pre-trained language models (e.g., BERT, Devlin et al., 2018; FLAN-T5, Chung et al., 2022; Llama models, Grattafiori et al., 2024) lies in the possibility of leveraging the pre-existing language representation of these models and fine-tuning it for deception detection purposes. Previous research that already adopted this approach showed how it reduces the need for a labor-intensive process of manually engineering features or determining which features to extract and select (Fornaciari et al., 2021; Loconte et al., 2023; Loconte & Kleinberg, 2025), which we argue is a practice that often introduces variability and fragmentation within research methodologies. By fine-tuning large language models, researchers can directly work with raw text inputs, streamlining the workflow and enhancing consistency across studies.

However, a notable limitation of this approach is its lack of interpretability compared to simpler, feature-based models. While *term frequency*, *embeddings*, and *topic or latent semantic* approaches are primarily data-driven, *linguistic and statistical features* offer the advantage of being theoretically grounded. For instance, prior studies have employed dictionary-based methods to automate traditional manual approaches, such as the RM (Kleinberg, van der Vegt, et al., 2019; Schutte et al., 2021), linked statistical properties of texts to cognitive load (Sarzynska-Wawer et al., 2023), or leveraged ML-based techniques like NER as proxies for extracting verifiable details (Kleinberg et al., 2017, 2018). All these examples highlight how a features-based text representation, despite its simplicity, may be more effective in terms of connecting machine predictions of deception to specific and theoretically based verbal cues. Translational research in high-stakes contexts (e.g., court proceedings, airport security, fraud prevention) emphasizes that, whenever possible, simpler and more explainable models should be favored over highly accurate but opaque alternatives (Oswald et al., 2018). Additionally, a previous study demonstrated that fine-tuned language models are not necessarily more robust than simpler bag-of-words models against adversarial attacks that leverage text paraphrasing (Kleinberg et al., 2025). In such scenarios, understanding the rationale behind algorithmic decisions is as crucial as the decision itself (Oswald et al., 2018); therefore, employing models grounded in psychological theories of deception provides a significant advantage for enhancing interpretability and explainability.

"Being accurate about accuracy"[6]

Regarding the evaluation procedures, our findings reveal a marked difference compared to practices employed in psycholegal research. A previous investigation highlighted that 81% of studies in psycholegal research evaluated the effectiveness of manual approaches using in-sample estimation, leading to inflated accuracy estimations (Kleinberg et al., 2019). The solution proposed by the authors was to rely on cross-validation when testing on an independent sample is not feasible (Kleinberg et al., 2019). Interestingly, we observe that for automated verbal deception detection, in-sample validation was only used in 0.52% of models (n=27). In contrast, cross-validation was already common practice (n=2,751, 53.44% of models), followed by the use of a hold-out test set (n=1,294, 25.14% of models), and with a minority of models being tested even on an independent sample (n=929, 18.05% of models). This widespread adoption of out-of-sample validation, on one hand, reduces the risk of overestimating performance and increases the reliability of reported results. On the other hand, it increases the reliability of results in automated verbal deception detection research compared to those found in psycholegal research on manual approaches.

Despite this positive note, the poor reliance on independent samples for evaluation (e.g., testing on new data or new domains), raises concerns about the external validity of these models. Without testing on unseen or out-of-domain data, it is difficult to assess whether models can generalize beyond the conditions under which they were trained. Previous studies have already showed how testing apparent accurate models on new domains brings their performance back to chance level (Velutharambath & Klinger, 2023), and this happens also when testing models across different cultural dimensions of individualism and collectivism (Papantoniou et al., 2022), with some researchers even wondering whether deception can actually be detected at all (Velutharambath et al., 2025). Consequently, future research should consider testing their models using both cross-validation and independent samples to validate the robustness of their findings and ensure a fair performance estimation.

[6] This expression comes from Kleinberg et al. (2019), and has been borrowed to refer to the importance of using correct validation procedures to obtain fair estimates of performance. The results in this paragraph cite and take into consideration the results of Kleinberg et al. (2019).

4.3 Limitations and future outlooks

This systematic review is subject to three limitations that should be noted. First, as with every systematic review, the choices regarding eligibility criteria and search strategy inevitably shape the scope and comprehensiveness of our findings. For example, we excluded all research related to fake reviews and misinformation because, although they constitute a form of deception in the media and news domain (Galeotti & Meini, 2022), they represent a separate research line that would have overly influenced our review. Moreover, excluding records that discussed fake news was sometimes challenging, as some works explicitly conceptualized fake news as deception (e.g., studies using datasets of fake news in political statements; Wang, 2017), and this may create disagreements among researchers who aim to replicate our findings. Second, some full texts, especially those found in conference proceedings, showed substandard reporting quality. This may have introduced errors in the coding phase, particularly for variables that were rarely explicitly detailed, such as the type of deception or the ground truth, and that we often had to infer. Finally, while using ASReview Lab to assist human reviewers enabled an extensive coverage of the available literature, both manual and semi-automated approaches introduce measurement error, with disagreements and edge cases invariably remaining. However, the hunt for the very last paper is a myth, as previous research has shown that, regardless of the methods employed, some relevant papers will inevitably be missed (Schoot et al., 2025).

An important avenue for future research involves conducting meta-analyses to estimate the average accuracy of automated verbal deception detection methods under varying conditions, such as research design, data sources, or model types. Such analysis would enable a more nuanced understanding of performance variability and identify methodological factors that systematically influence outcomes. Moreover, given the substantial amount of research that has shared data online, future studies should leverage these resources to systematically re-analyze existing datasets to i) test multiple models for robustness and ii) evaluate verbal cues that consistently contribute to deception detection across contexts. Finally, beyond meta-analysis and re-analysis of findings, there is also a pressing need for a mega-analysis that aggregates all available datasets on deception to create a unified, large-scale resource. Some studies have already attempted to address this issue. For example, one study examined deception in relation to cultural dimensions of individualism and collectivism across multiple datasets (Papantoniou et al., 2022), while another work focused on developing a unified corpus of deceptive texts for cross-corpus deception detection (Velutharambath & Klinger, 2023).

Such efforts would facilitate robust benchmarking and support the development of more generalizable models.

Conclusion

This Chapter mapped deception operationalization and research practices from 248 records and 5,148 models, being, to the best of our knowledge, the most extensive systematic review on automated verbal deception detection. Key strengths include the balanced approach of reanalyzing existing datasets, reducing the replicability crisis, and collecting new data, which enrich the available resources for training new models. Research diversity in design, data sources, text representations, and model categories further enables opportunities for future research to test generalizability across contexts and approach ecological validity. Notably, the adoption of more advanced computational techniques, such as text embeddings and neural architectures like transformers, has increased over time, suggesting methodological progress. However, important limitations remain. Deception is predominantly operationalized as fabrication and studied in fake reviews and English datasets, limiting linguistic and contextual variety. Moreover, ground truth was poorly operationalized, with only 13.71% of records providing a clear and fully verifiable ground truth, undermining internal research validity. Finally, despite methodological improvements compared to research practices in validating manual approaches, only 18.05% of models were tested on new data to test the generalizability of performance out-of-sample. These gaps underscore the need for more rigorous ground truth standards, broader operationalizations of deception, and rigorous out-of-domain evaluation to advance the reliability and applicability of automated approaches for detecting verbal deception.

In the next Chapter, we begin by highlighting the advantages of automated approaches for detecting verbal deception as a means to overcome the limitations of manual coding of statements and human prediction of verbal deception. To this aim, the next Chapter will explore and compare, in four experiments, the performance of naïve judges, expert judges trained on Reality Monitoring (Johnson & Raye, 1981), and theory-led and data-driven statistical learning models in detecting verbal deception.

References

Aamodt, M., & Custer, H. (2006). Who can best catch a liar? *Forensic Examiner*, *15(1)*, 6.

Abe, N. (2009). The neurobiology of deception: Evidence from neuroimaging and loss-of-function studies. *Current Opinion in Neurology*, *22*(6), 594–600. https://doi.org/10.1097/WCO.0B013E328332C3CF

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Ben-Shakhar, G., & Elaad, E. (2003). The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. *The Journal of Applied Psychology*, *88*(1), 131–151. https://doi.org/10.1037/0021-9010.88.1.131

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in Context Improves Deception Detection Accuracy. *Human Communication Research*, *36*(3), 423–442. https://doi.org/10.1111/J.1468-2958.2010.01382.X

Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, *13*(1), 1–10. https://doi.org/10.1186/S13643-024-02502-7

Bogaard, G., Meijer, E. H., Vrij, A., & Merckelbach, H. (2016). Strong, but Wrong: Lay People's and Police Officers' Beliefs about Verbal and Nonverbal Cues to Deception. *PLOS ONE*, *11*(6), e0156615. https://doi.org/10.1371/JOURNAL.PONE.0156615

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personal. Soc. Psychol. Rev.*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, G. D., Holman, R. D., Eggert, J. A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., & Rustige, R. (2017). 'Lyin' Ted', 'crooked hillary', and 'Deceptive Donald': Language of lies in the 2016 US presidential debates. *Appl. Cognit. Psychol.*, *31*(6), 668–677. https://doi.org/10.1002/acp.3376

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*(3), 313–329. https://doi.org/10.1002/ACP.1087

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-analyses. *Cerebral Cortex*, *19*(7), 1557–1566. https://doi.org/10.1093/CERCOR/BHN189

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). *Scaling Instruction-Finetuned Language Models*. http://arxiv.org/abs/2210.11416

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (Vol. 1195). Boston, MA: Houghton Mifflin.

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychol. Bull.*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Dionisio, D. P., Granholm, E., Hillix, W. A., & Perrine, W. F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, *38*(2), 205–211. https://doi.org/10.1111/1469-8986.3820205

Ekman, P. (2003). Darwin, Deception, and Facial Expression. *Annals of the New York Academy of Sciences*, *1000*(1), 205–221. https://doi.org/10.1196/ANNALS.1280.010

Fornaciari, T., Bianchi, F., Poesio, M., & Hovy, D. (2021). BERTective: Language Models and Contextual Information for Deception Detection. *EACL 2021 - 16th Conference of the European Chapter of the*

Association for Computational Linguistics, Proceedings of the Conference, 2699–2708. https://doi.org/10.18653/V1/2021.EACL-MAIN.232

Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, 279–287. https://doi.org/10.3115/V1/E14-1030

Galeotti, A. E., & Meini, C. (2022). Scientific Misinformation and Fake News: A Blurred Boundary. *Social Epistemology*, *36*(6), 703–718. https://doi.org/10.1080/02691728.2022.2070788

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783v3

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/A0023589

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/ACP.3052

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic Use of Evidence During Investigative Interviews: The State of the Science. *Credibility Assessment: Scientific Research and Applications*, 1–36. https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Hoffer, R. M. (1972). The importance of ground truth data in remote sensing. In *LARS Print* (p. 120371).

Hofman, J. M., Chatzimparmpas, A., Sharma, A., Watts, D. J., & Hullman, J. (2023). *Pre-registration for Predictive Modeling*. https://arxiv.org/abs/2311.18807v1

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:*

Volume 2, Short Papers (Vol. 2, pp. 427–431). https://aclanthology.org/E17-2068/

Kircher, J. C., & Raskin, D. C. (1988). Human Versus Computerized Evaluations of Polygraph Data in a Laboratory Setting. *Journal of Applied Psychology*, *73*(2), 291–302. https://doi.org/10.1037/0021-9010.73.2.291

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/JOURNAL.PONE.0220228

Kleinberg, B., Loconte, R., & Verschuere, B. (2025). Effective faking of verbal deception detection with target-aligned adversarial attacks. *Legal and Criminological Psychology*, *00*, 1–24. https://doi.org/10.1111/LCRP.70001

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2017). Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., Nahari, G., Arntz, A., & Verschuere, B. (2017). An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/COLLABRA.80

Konečni, V. J., & Ebbesen, E. B. (1979). External validity of research in legal psychology. *Law and Human Behavior*, *3*(1–2), 39–70. https://doi.org/10.1007/BF01039148

Kühne, S., Aachen, R., & Paul, G. B. (2024). Gut Feelings and Algorithms: Searching for Harmful Intentions in Airport Security Processes. *Engaging Science, Technology, and Society*, *10*(3), 120–146–120–146. https://doi.org/10.17351/ESTS2023.2337

Lahay, R., Leach, A. M., Cutler, B. L., Woolridge, L. R., & Elliott, E. (2025). (MIS)measuring cognitive load and arousal in deception: A multitrait–multimethod analysis. *Legal and Criminological Psychology*, *30*(1), 127–142. https://doi.org/10.1111/LCRP.12299

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R. (2018). Ecological Validity and Deception Detection Research Design. *Communication Methods and Measures*, *12*(1), 45–54. https://doi.org/10.1080/19312458.2017.1411471

Levine, T. R., & Serota, K. B. (2025). A Fresh View of the Veracity Effect in Deception Research: Bond and DePaulo Re-examined. *Communication Research*. https://doi.org/10.1177/00936502251316927

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, *1*, 1566–1576. https://doi.org/10.3115/V1/P14-1147

Lisofsky, N., Kazzer, P., Heekeren, H. R., & Prehn, K. (2014). Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. *Neuropsychologia*, *61*(1), 113–122. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2014.06.001

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., & Allen, P. G. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692v1

Loconte, R., & Kleinberg, B. (2025). Examining embedded lies through computational text analysis. *Scientific Reports 2025 15:1*, *15*(1), 26482-. https://doi.org/10.1038/s41598-025-11327-w

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. *Scientific Reports 2023 13:1*, *13*(1), 1–19. https://doi.org/10.1038/s41598-023-50214-0

Markowitz, D. M. (2024). Deconstructing deception: Frequency, communicator characteristics, and linguistic features of embeddedness. *Applied Cognitive Psychology*, *38*(3), e4215. https://doi.org/10.1002/ACP.4215

Matsumoto, D., & Hwang, H. C. (2015). Differences in Word Usage by Truth Tellers and Liars in Written Statements and an Investigative Interview After a Mock Crime. *Journal of Investigative Psychology and Offender Profiling*, *12*(2), 199–216. https://doi.org/10.1002/JIP.1423

Matsumoto, D., & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, *35*(2), 181–191. https://doi.org/10.1007/S11031-011-9212-2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness,*

Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Monaro, M., Capuozzo, P., Ragucci, F., Maffei, A., Curci, A., Scarpazza, C., Angrilli, A., & Sartori, G. (2020). Using blink rate to detect deception: A study to validate an automatic blink detector and a new dataset of videos from liars and truth-tellers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12183 LNCS*, 494–509. https://doi.org/10.1007/978-3-030-49065-2_35

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Sci Rep*, *8*(1), 1976. https://doi.org/10.1038/s41598-018-20462-6

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, *12*(5). https://doi.org/10.1371/JOURNAL.PONE.0177851

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal Criminol. Psychol.*, *19*(2), 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2021). Validity of content-based techniques for credibility assessment—How telling is an extended meta-analysis taking research bias into account? *Applied Cognitive Psychology*, *35*(2), 393–410. https://doi.org/10.1002/ACP.3776

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223–250. https://doi.org/10.1080/13600834.2018.1458455

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative Deceptive Opinion Spam. *Association for Computational Linguistics*, 497–501. https://aclanthology.org/N13-1053/

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 309–319.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., …

Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/BMJ.N71

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2021). The Verifiability Approach: A Meta-Analysis. *Journal of Applied Research in Memory and Cognition*, *10*(1), 155–166. https://doi.org/10.1016/J.JARMAC.2020.09.001

Papantoniou, K., Papadakos, P., Patkos, T., Flouris, G., Androutsopoulos, I., & Plexousakis, D. (2022). Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering*, *28*(5), 545–606. https://doi.org/10.1017/S1351324921000152

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 59–66. https://doi.org/10.1145/2818346.2820758

Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J., & Camacho-Collados, M. (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowledge-Based Systems*, *149*, 155–168. https://doi.org/10.1016/J.KNOSYS.2018.03.010

Schmuckler, M. A. (2001). What Is Ecological Validity? A Dimensional Analysis. *Infancy*, *2*(4), 419–436. https://doi.org/10.1207/S15327078IN0204_02

Schoot, R. van de, Coimbra, B. M., Evenhuis, T., Lombaers, P., Weijdema, F., Bruin, L. de, Neeleman, R., Grandfield, E., Sijbrandij, M., Teijema, J. J., Jalsovec, E., Bron, M. P., Winter, S., Bruin, J. de, & Zuiden, M. van. (2025). The hunt for the last relevant paper: blending the best of humans and AI. *European Journal of Psychotraumatology*, *16*(1), 2546214. https://doi.org/10.1080/20008066.2025.2546214

Serota, K. B., & Levine, T. R. (2015). A Few Prolific Liars. *Journal of Language and Social Psychology*, *34*(2), 138–157. https://doi.org/10.1177/0261927X14528804

Soldner, F., Kleinberg, B., & Johnson, S. D. (2022). Confounds and overestimations in fake review detection: Experimentally controlling

for product-ownership and data-origin. *PLOS ONE*, *17*(12), e0277869. https://doi.org/10.1371/JOURNAL.PONE.0277869

Soldner, F., Pérez-Rosas, V., & Mihalcea, R. (2019). Box of Lies: Multimodal Deception Detection in Dialogues. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 1768–1777. https://doi.org/10.18653/V1/N19-1175

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law*, *13*(1), 1–34. https://doi.org/10.1037/1076-8971.13.1.1

Steller, M., & Koehnken, G. (1989). Criteria-Based Content Analysis. *The Suggestibility of Children's Recollections*. https://doi.org/10.1037/T27704-000

Stern, B. A., & Krapohl, D. J. (2004). The Efficacy of Detecting Deception in Psychopaths Using a Polygraph 1. *Stern & Krapohl Polygraph*, *4*, 33.

Sternglanz, R. W., Morris, W. L., Morrow, M., & Braverman, J. (2019). A review of meta-analyses about deception detection. *The Palgrave Handbook of Deceptive Communication*, 303–326. https://doi.org/10.1007/978-3-319-96334-1_16

Suchotzki, K., Verschuere, B., Bockstaele, B. Van, Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/BUL0000087

Tausczik, Y., and, J. P.-J. of language, & 2010, undefined. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journals.Sagepub.Com*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Teijema, J. J., Hofstee, L., Brouwer, M., de Bruin, J., Ferdinands, G., de Boer, J., Vizan, P., van den Brand, S., Bockting, C., van de Schoot, R., & Bagheri, A. (2023). Active learning-based systematic reviewing using switching classification models: the case of the onset, maintenance, and relapse of depressive disorders. *Frontiers in Research Metrics and Analytics*, *8*. https://doi.org/10.3389/FRMA.2023.1178181

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature*

Machine Intelligence 2021 3:2, *3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

Velutharambath, A., & Klinger, R. (2023). UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 39–51. https://doi.org/10.18653/V1/2023.WASSA-1.5

Verigin, B. L., Meijer, E. H., Bogaard, G., & Vrij, A. (2019). Lie prevalence, lie characteristics and strategies of self-reported good liars. *PLoS ONE*, *14*(12). https://doi.org/10.1371/JOURNAL.PONE.0225566

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). Memory Detection: Theory and Application of the Concealed Information Test. *Memory Detection: Theory and Application of the Concealed Information Test*, 1–319. https://doi.org/10.1017/CBO9780511975196

Verschuere, B., Bogaard, G., & Meijer, E. (2021). Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis. *Applied Cognitive Psychology*, *35*(2), 374–384. https://doi.org/10.1002/ACP.3775

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 2023 7:5*, *7*(5), 718–728. https://doi.org/10.1038/s41562-023-01556-2

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/JIP.82

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal Criminol. Psychol.*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Fisher, R. P., Blank, H., Leal, S., & Mann, S. (2016). A cognitive approach to elicit verbal and nonverbal cues to deceit. *Cheating, Corruption, and Concealment: The Roots of Dishonesty*, 284–302. https://doi.org/10.1017/CBO9781316225608.017

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117. https://doi.org/10.1016/J.JARMAC.2012.02.004

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public*

Interest, Supplement, 11(3), 89–121. https://doi.org/10.1177/1529100610390861

Supplementary Materials

1. Search string

The employed full search string is reported below. In bold the main keyword, followed by the other semantically-related terms.

((**automat*** OR comput* OR "machine learning" OR "deep learning" OR "natural language processing" OR AI OR "artificial intelligence" OR "language models" OR LLMs)

AND

(**verbal** OR text* OR narrative* OR written OR statement* OR content)

AND

(**decept*** OR lie* OR lying OR deceit* OR dishonest* OR credibility OR veracity OR truth* OR believability)

AND

(**detect*** OR identif*))

NOT
("fake news" OR disinformation OR misinformation).

2. List of key records

Key records from Hauch et al (2015)

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*, 313-329. doi:10.1002/acp.1087

Evans, A. D., Brunet, M. K., Talwar, V., Bala, N., Lindsay, R. C. L., & Lee, K. (2012). The effects of repetition on children's true and false reports. *Psychiatry, Psychology and Law*, *19*, 517-529. doi:10.1080/13218719.2011.615808

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*, 585-594. doi:10.1016/j.dss.2010.08.009

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, *29*, 665-675. doi:10.1177/0146167203029005010

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics* (pp. 309-319). Portland, Oregon: Omnipress Incompany.

Key records from Costancio et al. (2022)

Papantoniou, K., Papadakos, P., Patkos, T., Flouris, G., Androutsopoulos, I., & Plexousakis, D. (2022). Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering*, *28*(5), 545-606.

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta psychologica*, *213*, 103250.

Barsever, D., Singh, S., & Neftci, E. (2020, July). Building a better lie detector with BERT: The difference between truth and lies. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

Hu, S. (2019, July). Detecting concealed information in text and speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 402-412).

Kopev, D., Ali, A., Koychev, I., & Nakov, P. (2019, December). Detecting deception in political debates using acoustic and textual features. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 652-659). IEEE.

Mbaziira, A. V., & Murphy, D. R. (2018, March). An empirical study on detecting deception and cybercrime using artificial neural networks. In *Proceedings of the 2nd International Conference on Compute and Data Analysis* (pp. 42-46).

Kleinberg, B., Van Der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied cognitive psychology*, *32*(3), 354-366.

Hosomi, N., Sakti, S., Yoshino, K., & Nakamura, S. (2018, November). Deception detection and analysis in spoken dialogues based on FastText. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 139-142). IEEE.

Levitan, S. I., Maredia, A., & Hirschberg, J. (2018, June). Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1941-1950).

Pak, J., & Zhou, L. (2015, May). A comparison of features for automatic deception detection in synchronous computer-mediated communication. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 141-143). IEEE.

Pérez-Rosas, V., & Mihalcea, R. (2015, September). Experiments in open domain deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1120-1125).

Briscoe, E. J., Appling, D. S., & Hayes, H. (2014, January). Cues to deception in social media communications. In *2014 47th Hawaii international conference on system sciences* (pp. 1435-1443). IEEE.

Mihalcea, R., Pérez-Rosas, V., & Burzo, M. (2013, December). Automatic detection of deceit in verbal communication. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 131-134).

Rubin, V. L., & Conroy, N. (2012). Discerning truth from deception: Human judgments and automation efforts. *First Monday*, *17*(5).

Fornaciari, T., & Poesio, M. (2012, April). On the use of homogenous sets of subjects in deceptive language analysis. In *Proceedings of the Workshop on Computational Approaches to Deception Detection* (pp. 39-47).

Almela, A., Valencia-García, R., & Cantos, P. (2012, April). Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the workshop on computational approaches to deception detection* (pp. 15-22).

Feng, S., Banerjee, R., & Choi, Y. (2012, July). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 171-175).

Rubin, V. L., & Conroy, N. J. (2011). Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1-4.

3. Active learning model's details

ASReview allows the researcher to select between multiple combinations of features extractors and classifiers.

For feature extraction, the simplest approach relies on Term Frequency-Inverse Document Frequency (TF-IDF). Term frequency (TF) measures how frequently a term occurs in the abstract. Inverse document frequency (IDF) measures how often the term appears across all the abstracts in my records. By multiplying the TF and IDF, we obtain the TF-IDF of a specific word and it represents the importance of that word in the abstract relative to the whole dataset. By repeating this process with every word, the TF-IDF converts the textual version of the abstract into its numerical format. The inherent limitations of this approach are that the TF-IDF does not incorporate information related to words order, contexts, and semantics of words.

In contrast, the Sentence-Bidirectional Encoder Representations from Transformers (S-BERT) relies on transformer architectures to generate a numerical vector representation of sentences, known as embeddings, that also captures semantics. One limitation for this approach is that computing embeddings can be time-consuming and computationally expensive.

After processing the abstracts into their numerical format, classifiers can be used to sort data into categories (here: relevant or irrelevant for the systematic review). Different models were available in asreview, such as logistic regression, naïve bayes, and XGBoost. Logistic regression estimates the probability that a given input point belongs to a certain class. Naïve Bayes Classifier is a classifier based on Bayes' theorem and assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. XGBoost works by building an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones using gradient boosting and regularization to improve accuracy and prevent overfitting. Both logistic regression and naïve bayes are very fast in computation, while XGBoost are more expensive. Further information are available at https://asreview.nl/blog/asreview-model-selection-guide/

4. Definitions of sources and research designs

Online sources: Data gathered through internet-based sources, including web platforms, social media, online surveys, or data scraping from digital environments.

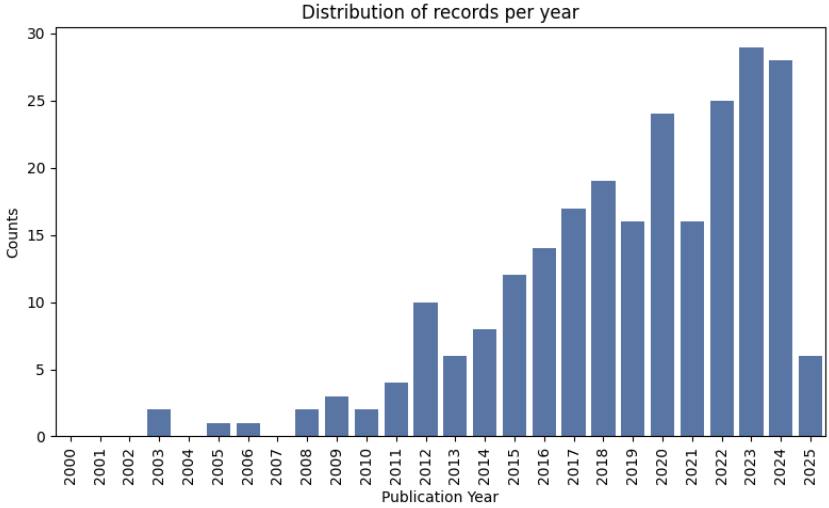
Offline sources: In contrast to online data, offline data collected are collected in vivo, either in controlled or natural settings, without using the internet. Examples include laboratory experiments, field studies, or face-to-face interviews.

For either data collected online or offline, we defined the research designs as follows:

- **Experiments**: Studies in which researchers actively manipulate one or more variables under controlled conditions to establish cause-and-effect relationships.
- **Quasi-experiments**: Research designs that examine causal relationships without full randomization or control, often using naturally occurring groups or conditions or by mixing observational data with experimentally manipulated data.
- **Naturalistic**: Information collected from real-life settings without experimental manipulation, reflecting authentic behaviors, events, or interactions as they occur in everyday contexts. Usually, data points are assigned to different conditions (here: truthful vs deceptive) based on predefined inferred criteria. For online naturalistic data, we mean data that are naturally provided in online platforms (e.g., on Facebook) in contrast to online experiments where the researcher collects new data via crowdsourcing (e.g., Prolific, Amazon Turk).
- **Mixed**: Records that combined multiple data origins and designs.

5. Distribution of included records per publication year

FIGURE S1. Distribution of published records over time (2000-2025).



Note. One model from 1982, included in the systematic review, was excluded from the plot for a better visualization.

6. Number of statements by source and research design

TABLE S1. Descriptive statistics of no. of statements (M, SD) by source of data collection and research design at the model-level.

Source	Design	*N* (%)	*M*	*SD*	*Median*	Range
Offline	Experiment	724 (14.06%)	584.19	1,237.39	217	20 - 9,104
	Quasi-experiment	14 (0.27%)	1,665.14	342.80	1,800	856 - 1,800
	Naturalistic data	504 (9.79%)	1,470.10	3,804.44	202	18 - 17,880
Online	Experiment	1,044 (20.28%)	840.49	1,513.01	346	38 - 14,343
	Quasi-experiment	1,918 (37.26%)	3,763.31	28,628.84	1,600	200 - 501,472
	Naturalistic data	829 (16.10%)	82,342.34	435,110.78	5,854	78 - 3,868,306
Mixed	Mixed	108 (2.10%)	647.58	998.75	220	200 - 4018

Note. Seven models with unclear data source are not reported in this table.
Abbreviations. *N* = number of models from included records.

7. Reused datasets

Table S2. Reference list and details of reused datasets that were reported in an academic outlet and fit within the first three levels of ground truth.

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Bluff the Listener (BLUFF; Skalicky et al., 2020)	753	0.5	Typed	English	Radio game broadcast	Fabrication	Offline naturalistic	Clear and verifiable
Boulder Lies and Truth corpus (Salvetti et al., 2016)	1,492	0.43	Typed	English	Reviews	Fabrication and embedded lies	Offline naturalistic	Clear but not verifiable
Box of Lies corpus (Soldner et al., 2019)	1,049	0.22	Transcribed	English	Object description	Fabrication	Offline naturalistic	Clear and verifiable
CliPS stylometry investigation coprus (Verhoeven & Daelemans, 2014)	1298	1	Typed	Dutch	Reviews	Fabrication	Offline naturalistic	Clear but not verifiable
Columbia X Cultural Deception (CXD)	139	1	Transcribed	English	Interview on biographical questions	mixed	Offline experiment	Clear but not verifiable

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Corpus (Levitan et al., 2015)								
Columbia-SRI-Colorado (CSC) Deceptive Speech dataset (Hirschberg et al., 2005)	32	0.5	Transcribed	English	Interview on task performance	Exaggeration, minimization	Offline experiment	Clear and verifiable
Cross-cultural deception (Pérez-Rosas & Mihalcea, 2014)	1550	1	Typed	American English, Indian English, Spanish	Personal opinions	Fabrication	Online experiment	Clear but not verifiable
Daily Deceptive Dialogues corpus of Mandarin (Huang et al. 2019)	2764	-	Transcribed	Mandarin	Dyadic interactions, asking questions about past experiences	mixed	Offline experiment	Clear but not verifiable
Deception Corpus (Almela et al., 2012)	600	1	Typed	Spanish	Personal opinions	fabrication	Offline experiment	Clear but not verifiable
Deception in Reviews (DeRev; Fornaciari & Poesio, 2014)	236	1	Typed	English	Reviews	fabrication	Online naturalistic	Directly inferred

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Deception in Reviews (DeRev; Fornaciari & Poesio, 2020)	1552	1	Typed	English	Reviews	fabrication	Online experiment	Clear but not verifiable
Deception on Facebook (Songram et al., 2016)	2387	1	Typed	Thai	Job posts	fabrication	Online naturalistic	Directly inferred
Deception Speech Database (DSD; Schuller et al., 2016)	1555	2.6	Transcribed	English	Mock crime	fabrication	Offline experiment	Clear and verifiable
Deceptive Interview Corpus (Burgoon, et al., 1999)	732	1	Transcribed	English	Self-presentation	mixed	Offline experiment	Clear but not verifiable
Deceptive Opinion Spam (Ott et al., 2011)	800	1	Typed	English	Reviews	fabrication	Online quasi-experiment	Clear but not verifiable
DeCop (Capuozzo et al., 2020)	2500	1	Typed	English	Personal opinions	fabrication	Online experiment	Clear but not verifiable
DeCop (Capuozzo et al., 2020)	2500	1	Typed	Italian	Personal opinions	fabrication	Online experiment	Clear but not verifiable

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
DECOUR - DEception in COUR (Fornaciari and Poesio; 2012)	3015	1.27	Transcribed	Italian	Trial hearings	mixed	Offline naturalistic	Directly inferred
Deeb et al. (2020)	243	1	Transcribed	English	Past memorable experience	fabrication	Offline experiment	Clear but not verifiable
Deeb et al., (2022a)	112	1	Transcribed	English	Lab mission	fabrication	Offline experiment	Clear but not verifiable
Deeb et al., (2022b)	211	1	Transcribed	English	Lab mission	fabrication	Offline experiment	Clear but not verifiable
Deeb, Vriij, Leal, & Burkhardt (2021)	243	1	Transcribed	English	Past memorable experience	fabrication	Offline experiment	Clear but not verifiable
Deeb, Vrij, Leal & Mann (2021)	175	1	Transcribed	English	Video footage	fabrication	Offline experiment	Clear but not verifiable
DeFaBel (Velutharambath, Wuhrl & Klinger, 2024)	1031	0.61	Typed	German	Personal opinions	fabrication	Online experiment	Clear but not verifiable

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Desert Survival Problem (Zhou et al., 2004)	30	0.88	Typed	English	Personal opinions	fabrication	Offline experiment	Clear but not verifiable
Diederik Stapel dataset (Markowitz & Hancock, 2014)	49	1.04	Typed	English	Scientific fraud	fabrication	Offline naturalistic	Clear and verifiable
Essays Corpus (Mihalcea & Strapparava, 2009)	600	1	Typed	English	Personal opinions	fabrication	Online experiment	Clear but not verifiable
Hippocorpus (Sap et al., 2022)	6854	1	Typed	English	Past memorable experience	fabrication	Online experiment	Clear but not verifiable
Intentions dataset (Kleinberg & Verschuere 2021)	1640	0.91	Typed	English	Future intentions	fabrication	Online experiment	Clear but not verifiable
Li et al., (2014)	3032	0.46	Typed	English	Reviews	fabrication	Online quasi-experiment	Clear but not verifiable
Mafia Game Dataset (Ibraheem et al., 2022)	2162	1.81	Typed	English	Mafia game	mixed	Online experiment	Clear and verifiable
Mafiascum dataset (de Ruiter & Kachergis, 2019)	10,000	-	Typed	English	Mafia game	mixed	Online naturalistic	Clear and verifiable

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Miami University Deception Detection Dataset (MU3D; Lloyd et al., 2019)	320	1	Transcribed	English	Personal opinions	fabrication	Offline experiment	Clear but not verifiable
Monaro et al., (2022)	62	1	Transcribed	Italian	Past trip	fabrication	Offline experiment	Clear but not verifiable
Negative Deceptive Opinion Spam (Ott et al., 2013)	800	1	Typed	English	Reviews	fabrication	Online quasi-experiment	Clear but not verifiable
Open Domain Dataset (Pérez-Rosas & Mihalcea, 2015)	7168	1	Typed	English	Open domain	fabrication	Online experiments	Clear but not verifiable
Real-life trial dataset (Pérez-Rosas et al., 2015)	121	0.98	Transcribed	English	Trial hearings	fabrication	Offline naturalistic	Directly inferred
Russian Deception Bank (Litvinova et al., 2017)	226	1	Typed	Russian	Past experiences	fabrication	Lab experiment	Clear but not verifiable
Samsung Dataset (Chen & Chen 2015)	251,984	80.87	Typed	Chinese	Reviews	fabrication	Online data	Directly inferred

Dataset	No of statements or utterances	truthful / deceptive proportion	Format	Language	Topic	Type of deception	Source and research design	Ground truth
Stony Brook University (SBU) deception dataset (Banerjee et al., 2020)	2600	1	Typed	English	Personal Opinions	fabrication	Online experiment	Clear but not verifiable
Vrij et al., (2020)	201	1	Transcribed	English	Past trip	fabrication	Offline experiment	Clear but not verifiable
Vrij, Leal, Deeb, Castro Campos et al., (2022) – experiment 1 and 2	430	1	Transcribed	English	Past trip	fabrication	Offline experiment	Clear but not verifiable

Chapter 2

Detecting deceptive narratives through Natural Language Processing: comparing naïve and expert judges vs. theory-led and data-driven models

This chapter is based on: Loconte, R., Battaglini, C., Maldera, S., Pietrini, P., Sartori, G., Navarin, N., & Monaro, M. (2025). Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Processing. *Journal of Language and Social Psychology*, *44*(3-4), 523-552. https://doi.org/10.1177/0261927X251316883

Abstract

Detecting deception in interpersonal communication is a pivotal issue in social psychology, with significant implications for court and criminal proceedings. In this study, four experiments were designed to compare the performance of natural language processing (NLP) techniques and human judges in detecting deception from linguistic cues in a dataset of 62 transcriptions of videotaped interviews (32 genuine and 30 deceptive). The results showed that machine-learning (ML) algorithms significantly outperform naïve (accuracy=54.7%) and expert judges (accuracy=59.4%) when trained on features from the reality monitoring (RM) and cognitive load (CL) frameworks (accuracy=69.4%) or on features automatically extracted through NLP techniques (accuracy=77.3%), but not when trained on the RM criteria alone. This evidence suggests that NLP algorithms, due to their ability to handle complex patterns of linguistic data, might be helpful for better disentangling truthful from deceptive narratives, outperforming traditional theoretical models.

Keywords: deception, Reality Monitoring, Natural Language Processing, lie detection, deception linguistic cues

1. Introduction

Detecting deception in interpersonal communication is a pivotal issue in social psychology, with significant implications for criminal investigations, court proceedings, and criminal trials. For example, assessing the credibility of a suspect during interrogation is relevant, as false or distorted information may lead the investigation in the wrong direction. Moreover, in some cases, such as in the Italian court, allegations of sexual harassment are often based on the victims' declarations, which are treated as evidence, and most often without clear and independent evidence of the offence. In such situations, a police officer or a judge is tasked with determining the veracity of the information provided and of the alleged accusation (Oberlader et al., 2021). However, humans exhibit inherent biases when it comes to detecting lies. In the absence of prior knowledge of the context, human intuitive judgment in deception detection has been shown to be only slightly above chance level (Bond & DePaulo, 2006; Ekman & O'Sullivan, 1991). Even experts in the field, such as police officers, tend to commit false-negative and false-positive errors (Elaad, 2009; Vrij et al., 2008). For naïve judges, the truth bias (i.e., the human inclination to presume others as honest) has been proposed as one possible explanation for this poor performance (Levine, 2014; Street & Masip, 2015). For expert judges, instead, the debate remains open on whether the problem stems from the identification of the cue (i.e., during the evaluation process) or the difficulty of combining several cues (due to limited cognitive resources) to then make a straightforward decision (Verschuere et al., 2023).

Recently, researchers have increasingly relied on automated approaches to deception detection based on machine learning (ML) techniques, i.e., computational methods that enable computer algorithms to identify patterns in data and make predictions based on these patterns (see Constancio et al., 2023 for a review). One of the most widely exploited techniques is natural language processing (NLP), a field of artificial intelligence (AI) that focuses on enabling machines to interpret, analyze, and respond to human language. NLP has been applied mainly in detecting deception online, such as for identifying fake reviews (Ott et al., 2011) or misinformation (Pérez-Rosas et al., 2018). Typically, NLP-based approaches are heavily data-driven. This means that they rely on extracting features directly from textual data, such as word frequencies, syntactic patterns, and word embeddings (i.e., numerical representations of word co-occurrences), without necessarily incorporating insights from psychological theories that have been used to identify cues of deception in language.

In this context, NLP techniques can be leveraged to transform textual data into numerical features based on theoretical frameworks, which will

then be used to feed ML models trained to identify subtle verbal indicators in datasets where truthful and deceptive examples are already labeled (i.e., supervised learning). The advantage of this approach is that, after the training is complete, a well-trained ML model can be used to predict whether new statements are likely to be deceptive or truthful, based on learned patterns, and evaluate the efficacy of a specific psychological theory of deception.

1.1 Investigating the veracity of verbal content using reality monitoring

Deception may imply reporting fabricated details or intentionally omitting relevant information conveyed in such a way as to seem truthful to the interlocutor (Newman et al., 2003). The Undeutsch hypothesis suggests that deceitful information is qualitatively different in form and content from truthful information (Vrij et al., 2000). Nevertheless, there is no verbal cue that is inherently associated with deception (Vrij, 2008).

Among the several verbal lie detection tools, reality monitoring (RM; Johnson & Raye, 1981) was developed on the basis of evidence from cognitive psychology and currently stands out in the literature for its theoretical robustness, being the most commonly employed approach by researchers. This approach is grounded in the notion that memories of actual experiences exhibit stronger connections to external stimuli than memories of imagined events. Accordingly, memories originating from perceptual experiences should feature contextual, sensory, and affective details, whereas internally generated memories, stemming from thoughts or imagination, should be marked by references to cognitive processes. Eight criteria were outlined to distinguish between these two types of memories, with the presence of cognitive operations being the only lie-criterion (Johnson & Raye, 1981; Sporer, 1997, 2004). Research indicates that when scores are combined from these eight criteria, the average accuracy RM rate is comparable to that of other content-based techniques (e.g., the criteria-based content analysis), ranging between 61% and 83%, with an average of 69% (Vrij, 2008). Among the individual criteria, contextual factors, such as temporal and spatial criteria, appear to hold the most significant diagnostic value (Masip et al., 2005).

Systematic reviews (Masip et al., 2005; Vrij, 2005, 2008) and meta-analyses (Amado et al., 2015, 2016; Hauch et al., 2017; Oberlader et al., 2016) have demonstrated that RM exhibits satisfactory inter-rater reliability and empirical validity across diverse study designs and populations, possibly because of its straightforward application (Sporer, 1997; Strömwall et al., 2004; Vrij et al., 2000). Indeed, RM is not time-consuming, involves less subjective decision-making (Oberlader et al., 2016), and

holds precise criteria that are easy to operationalize. Despite these findings, caution is advised when using RM due to the lack of an objective decision rule, namely a numerical cutoff for scores that allows the user to classify a narrative as honest or fake (Amado et al., 2015, 2016).

1.2 Investigating the veracity of verbal content by imposing cognitive load

In interviews, lie-tellers are known to consume more cognitive resources because they need to fabricate responses that are congruent with other fabricated information while maintaining their credibility during the examination (Vrij et al., 2008). This cognitive load (CL) is often reflected in several indices (e.g., behavioural, physiological, verbal, and neural, among others) that can be leveraged to distinguish truthful from deceptive statements (Vrij et al., 2008).

The "imposing cognitive load approach" (Vrij et al., 2015) exploits this vulnerability in lie-tellers, utilizing interviewing strategies that further deplete their cognitive resources while maintaining a manageable demand for truth-tellers (Walczyk et al., 2013). These strategies may involve asking the examinee to perform a second task during the interview or to continuously switch between two tasks, imposing time restrictions on answering questions, recalling events in reverse order, or asking the examinee unexpected questions (Vrij et al., 2009; Walczyk et al., 2013).

Among these, the strategy of asking unexpected questions has proven effective, achieving accuracy rates ranging from 80% to 95% (e.g., Lancaster et al., 2013; Monaro et al., 2017, 2018). It involves the examiner initially asking anticipated questions, i.e., questions that the lie-tellers expect and prepare in advance, and then switching to questions that cannot be foreseen and for which the responses were not prepared. For example, in the context of fake identities, lie-tellers may prepare answers about the name, surname, and date of birth of the stolen identity, but it is unlikely they would prepare the answer for their zodiac sign (Monaro et al., 2017). Responses to unexpected questions in lie-tellers are associated with slower reaction times and a higher number of inconsistencies compared to truth-tellers (Melis et al., 2024; Monaro et al., 2017). Lie detection approaches that impose cognitive load yield higher accuracy rates compared to standard approaches in detecting deception (Vrij et al., 2015).

1.3 Investigating the veracity of verbal content using natural language processing

The advent of NLP for analyzing human language has provided new opportunities for the automatic detection of deception (Fitzpatrick et al., 2015). These techniques allow to extract features at different level of granularity: i) the *n*-gram model breaks the text into linear sequences of tokens and reveal frequent patterns; ii) part-of-speech (POS) tagging assigns grammatical categories (nouns, verbs, and adjectives) to words, thus informing on shallow syntactic text structure; iii) word/sentence length and number are extracted to evaluate text complexity; iv) the Linguistic Inquiry and Word Count (LIWC; Boyd et al., 2022; Pennebaker et al., 2015) categorizes words into psychological, social, and emotional dimensions (e.g., positive/negative affect, social words, etc) providing the semantic content of the text; v) named-entity recognition (NER), automatically identifies and categorizes proper nouns (such as names of people, places, and organizations) within texts.

Moreover, these computational techniques can be applied in conjunction with various methodologies. Data-driven approaches are based on statistical procedures to perform feature extraction and selection. Theory-led approaches investigate samples of features derived from psychological theoretical models of deception. Lastly, hybrid models rely on theory to select variables that are restricted to those found to be statistically significant.

The RM and CL frameworks have been proven suitable for investigation using this computational perspective. Recent studies highlighted how manual coding is not necessarily superior to automated coding of RM (Deeb et al., 2024; Schutte et al., 2021). In addition, a meta-analysis by Hauch et al. (2015) demonstrated the effectiveness of RM in detecting verbal deception when using LIWC features. The same meta-analysis also found evidence in favor of the CL theory, showing that lie-tellers produce shorter, less elaborate, and less complex statements (Hauch et al., 2015). These characteristics can be automatically captured in statements through linguistic features such as the number of words, number of sentences, average sentence length, type-token ratio, word length, and use of exclusive words (e.g., but, except, without).

Furthermore, a recent study found that verbal cues of CL effectively distinguished truthful from deceptive statements in a mixed dataset that included different contexts of deception (i.e., personal opinions vs. autobiographical memories vs. future intentions), suggesting the potential for these features to serve as more generalizable cues compared to others (Loconte et al., 2023).

TABLE 1. Studies employing ML techniques on verbal content for lie-detection tasks, as reported in Constancio et al. (2023).

Authors	Dataset	Features	Algorithm	Accuracy
Rubin and Conroy (2011)	Stories written by volunteers	LIWC categories, Lexical measures	Support Vector Machine (SVM)	0.65
Fentg et al. (2012)	Reviews of 35 Italian restaurants	Bigrams, POS tags, Syntax complexity, Unigrams	SVM	0.91
Fornaciari et al. (2012)	DeCour corpus	LIWC categories, Lexical measures, N-grams, POS tags	SVM	0.69
Rubin and Conroy (2012)	Stories written by volunteers	LIWC categories, Lexical measures	Decision Tree	0.65
Perez-Rosas et al. (2013)	Video recordings from volunteers	Unigrams	SVM	0.74
Briscoe et al. (2014)	Statements provided by volunteers in a mock chat room	Emoticons, Informality, Sentiment, Syntax complexity	Gradient Boosting	0.91
Pak and Zhou (2015)	Communications during sessions of the online Mafia game	LIWC categories, Syntax complexity, Unigrams	Decision Tree	0.98
Kleinberg et al. (2018)	Interviews on weekend plans collected from volunteers	LIWC categories, Named entities, Psychological processes	SVM	0.77
Mbaziira et. al (2018)	Combination of four publicly available datasets	Syntax complexity	Neural Network	0.80
Barsever et al. (2020)	Ott Deceptive Opinion Spam Corpus	BERT embeddings	Neural Network	0.94
Kleinberg et al. (2021)	True and deceptive statements collected by a web application from volunteers	LIWC categories, POS tags	Random Forest	0.69

Note. The type of dataset, the features extracted, the algorithm employed, and the highest accuracy reached in the test set are reported as they appear in the original study, rounded to two decimal places.

Finally, previous works employed NER as a proxy for the automated scoring of details and verifiable details, accurately classifying truthful and deceptive hotel reviews and future intentions (Kleinberg, Mozes, et al., 2018; Kleinberg, van der Toolen, et al., 2018).

Parallel to these theory-led approaches, the detection of verbal deception has also been investigated using a data-driven approach. For example, Mihalcea & Strapparava (2009) and Ott et al. (2011) detected deceptive opinions by training a naïve Bayes and a support vector machine model on n-grams and a combination of n-grams and LIWC features, respectively. Pérez-Rosas & Mihalcea (2015) investigated an open-domain dataset using n-grams, shallow and deep syntactic features (using POS tagging), semantic features from LIWC, and readability and syntactic complexity metrics.

A recent review of studies from Constâncio et al. (2023) on ML and NLP techniques for lie detection showed that automatic verbal analysis significantly outperformed chance and human-level performance across various datasets (see Table 1). Furthermore, a short review of the literature on the use of LIWC software for lie detection (Van Der Zee et al., 2022) showed that studies using a data-driven approach reached an accuracy rate ranging from 65% to 74%, whereas studies using a theory-led approach reached an accuracy rate ranging from 51% to 69%. Studies that employed a hybrid approach, training ML models only on statistically significant theoretical variables, reached an accuracy rate ranging from 50% to 69%. These findings underscore the importance of selecting the most appropriate approach, as it can significantly impact the performance of ML models (Van Der Zee et al., 2022).

1.4 The current study

This study examines deception detection through the theoretical lenses of RM and CL in the context of transcripts of interviews with unexpected questions. Through four experiments, this study compared the performances of naïve judges, expert judges trained on RM, and both theory-led and data-driven ML models in detecting deception, offering critical insights into the relative effectiveness and reliability of each approach.

Specifically, in **Experiment 1,** we compared the performance of laypeople (i.e., naïve judges) to that of individuals with expertise in the forensic field (i.e., expert judges). Expert judges were trained in the application of RM criteria for lie detection. The main hypothesis (Hyp. 1a) posits that expert judges achieve higher accuracy than naïve judges because studies have demonstrated the effectiveness of RM criteria in distinguishing

truth from deception in verbal content (Amado et al., 2016; Gancedo et al., 2021; Vrij, 2008; Vrij et al., 2022).

However, we saw expert judges performing poorly in Experiment 1. Therefore, in **Experiment 2,** we aimed to investigate the reasons behind this poor performance, attempting to disentangle whether it was associated with limitations in the effectiveness of RM when applied to specific datasets or whether the problem lay in experts' limitations in their evaluation and decision-making skills. We employed a computational approach for this purpose. Specifically, four ML models were trained on two sets of RM ratings: those assigned to each transcription by expert judges in Experiment 1 and those provided by leveraging NLP techniques. Three alternative hypotheses were defined for this study:

- Hyp. 2a) Expert judges performed poorly in Experiment 1 because they were poor evaluators. Specifically, expert judges may not be able to assess RM criteria effectively. If this hypothesis is true, a poor performance is expected from ML models trained on expert ratings, similar to that obtained by forensic experts in Experiment 1. Moreover, they are expected to show lower accuracy than those trained on ratings derived through NLP techniques.
- Hyp. 2b) Expert judges performed poorly in Experiment 1 because RM criteria were poorly informative for this type of dataset. If this hypothesis is true, both ML models trained on expert and NLP-based ratings of RM are expected to underperform, with an accuracy similar to that obtained by forensic experts in Experiment 1.
- Hyp. 2c) The RM criteria were informative, and expert judges were effective evaluators but may have struggled in effectively combining all the information to derive a final decision. If this hypothesis is true, ML models trained on expert ratings of RM are expected to outperform the accuracy obtained by expert judges in Experiment 1 (Monaro et al., 2020).

Finally, Experiments 3 and 4 concerned the performance of ML models. Specifically, we examined the effectiveness of theory-led vs. data-driven approaches in feature extraction.

Experiment 3 employed linguistic features from two theoretical frameworks: RM and CL. This procedure was adopted for two reasons: i) a previous meta-analysis on deception detection indicated that the CL approach yields higher accuracy rates than standard approaches (Vrij et al., 2015); ii) the dataset under analysis was specifically designed to establishing a minimum of 10 judges per statement, replicating the original recruiting and evaluation procedure as in Monaro et al., (2022). The first

hypothesis for this experiment posits that adding linguistic features associated with the CL framework potentially increases ML models' accuracy in detecting verbal deception compared to models trained solely on features from RM (Hyp. 3a). The second hypothesis postulates that various interview phases (free recall vs. unexpected questions vs. full text) influence ML models' performance. Specifically, ML models trained on features extracted from unexpected questions (or full text) will yield higher accuracy than those trained on features extracted solely from free speech, based on the assumption that CL features are more prevalent in responses to unexpected questions (Hyp. 3b).

Experiment 4 was designed to investigate whether a data-driven approach could outperform the previous theory-based methods. To this end, NLP techniques were employed to extract a broad set of linguistic features, along with a data-driven feature selection strategy, to identify a subset of highly informative features (Ghosh, 2022). The main hypothesis of this study posits that a data-driven approach may outperform theory-led approaches, particularly in scenarios where theory-based methods have already shown limited effectiveness. Specifically, a data-driven method is hypothesized to achieve superior performance by directly extrapolating rules from data, rather than relying on general theories (Hyp. 4a).

Experiments 1 to 4 altogether allow us to test a final hypothesis for this study, which posits that theory-led and data-driven ML approaches are expected to outperform naïve and expert human judges in identifying deception (Hyp. 4b). This hypothesis stems from ML models' computational ability to integrate and analyze complex datasets more comprehensively than humans.

2. Experiment 1: Naïve vs Expert judges

2.1 Materials and methods

Participants

The sample size was determined through an a priori power analysis using G*Power (Faul et al., 2007). For the sample of naïve judges, the results indicated that a sample size of 42 is sufficiently large to achieve a statistical power ($1\text{-}\beta$)=0.8 in a one-tailed Wilcoxon signed-rank test (one-sample case), given a significance level α=0.05 and a medium effect size (d=0.4; Bond & DePaulo, 2006). Since we had access to a larger sample size, we collected a significantly higher number of participants, ensuring

a minimum of 10 judges per statement, thereby replicating the original recruiting and evaluation procedure as described in Monaro et al. (2022). Therefore, the sample of naïve judges consisted of 121 Italian-speaking participants (75 females). Age ranged from 18 to 62 years (M=33.92, SD=12.40), with years of education ranging from 8 to 21 (M=15.84, SD=2.40). Participants were recruited as volunteers through a snowball sampling procedure. One participant did not complete the task and was excluded from the analysis.

For the sample of expert judges, the results of the power analysis indicated that a sample size of 23 is sufficiently large to achieve a statistical power (1-β)=0.8 in a one-tailed Wilcoxon signed-rank test (one-sample case), given a significance level α=0.05 and a medium effect size (d=0.55; Gancedo et al., 2021). Since having only 23 participants would result in fewer than two expert judgments per stimulus, we decided to expand the sample size to ensure at least three judges for each statement. Therefore, the sample of expert judges resulted in 36 Italian-speaking participants (27 females). Experts were recruited among psychology students attending the Master's course in Forensic Psychology and among the authors' professional network. Participation in the study was on a voluntary basis. Age ranged from 22 to 55 years (M=30.17, SD=7.77), with years of education ranging from 13 to 21 (M=18.86, SD=2.18). Nineteen experts were psychology students, and 17 were practitioners (i.e., psychologists, psychotherapists, psychiatrists, and lawyers). Experts were also asked to specify their level of expertise in forensic psychology via a multiple-choice question: 11.1% had only completed a course in forensic psychology, 27.8% had undergone additional training in the field, 47.2% held a master's degree in forensic psychology, and the remaining 13.9% were currently employed in the field.

All participants provided their informed consent before starting the experiment. The Board of the School of Psychology at the University of Padova approved the experimental procedure.

Dataset

The dataset consisted of 62 videotaped interviews with Italian participants (43 females, aged 20-29, who voluntarily participated in the study) in a low-stakes scenario (recalling a holiday experience). The dataset was collected by Monaro et al. (2020) in a previous study and was analyzed to detect deception through blink rate (Monaro et al., 2020) and facial expressions (Monaro et al., 2022).

The dataset comprised 32 participants allocated to the truthful condition, who were instructed to describe a holiday experience that had occurred within the preceding 12 to 18 months. Thirty participants were assigned

to the deceptive condition and were required to describe an entirely fabricated holiday. Each videotaped interview comprised three distinct phases:

1. **Baseline**, in which the interviewee provided their autobiographical information
2. **Free speech**, in which the interviewee freely recalled their holiday experience for approximately two minutes
3. **Unexpected Questions**, in which the interviewer asked unexpected questions about the holiday experience to increase the interviewee's CL (e.g., "*Did a particular event occur during the holiday that made it necessary to revise the initial plans?*").

The average length of the videos was 9.56 minutes. A more detailed description of the dataset is reported in Monaro et al. (2020) and in the Supplementary Materials.

Narrative transcription procedure

Interviews were manually transcribed verbatim. Then, a linguistic expert checked and modified raw transcriptions, following the guidelines in CLIPS (Savy, 2006). Adaptations tailored for the present study were made to ensure readability for the naïve readers. Regionalisms were substituted with the standard Italian alternative. Hesitations, false starts, and repetitions were transcribed as accurately as possible. False starts were reported with the symbol +, following Savy (2006): "abbiamo spos+, cioè abbiamo trovato" ("we have mov+, I mean we have found"). Pauses were signaled using punctuation. Sentence boundaries, signaled with a full stop and commas, were inserted using the standard Italian rules for punctuation. Hesitations and laughter were reported using standardized formulas (see Table 1S in Supplementary Materials).

Reality monitoring scoring

Following Sporer (1997, 2004), the RM framework consisted of eight criteria, as outlined in Table 2. Previous research employed three primary units of measurement to evaluate a statement according to the RM criteria: rating scales, categorical measures (presence vs. absence), and frequency/density counts (see Gancedo et al., 2021, for a meta-analytical review on the use of RM in the forensic context). For this experiment, RM criteria were evaluated using a 7-point rating scale (1=none, 7=very much). Previous findings showed that frequency counts may offer better performance and reliability than rating scales (Nahari, 2016). However, rating scales present other advantages: they require less training for hu-

man raters, are quick to apply, and provide a clear minimum and maximum score, helping the raters understand whether the obtained score falls within a higher or lower range. This is particularly helpful, given that the RM approach does not rely on a standardized cutoff to ultimately determine whether an account is truthful or deceptive (Amado et al., 2015, 2016). Moreover, using rating scales (or categorical measures) ensures consistency across all RM criteria. Indeed, when we use relative or absolute frequencies, only five out of eight criteria can be scored by frequency (i.e., perceptual information, spatial information, temporal information, affective information, and cognitive operations). Conversely, the remaining three criteria (i.e., vividness, reconstructability, and realism) are commonly evaluated on a rating or a categorical scale (see Table 2). The formula for calculating the overall RM score is reported in formula (1) (Sporer, 2004). The total RM score ranged from 0 to 48, with higher scores indicating higher narrative genuineness.

(1) **RM score =** *Vividness* + *Realism* + *Reconstructability* + *Perceptual information* + *Sensory information* + *Temporal information* + *Affective information* - *Cognitive Operation*

Experimental procedure

Twelve questionnaires were created on the Qualtrics platform. The transcriptions of the 62 videoclips were randomly distributed among the 12 questionnaires, as in Monaro et al. (2022), to balance the number of truthful and deceptive transcriptions for each questionnaire. Therefore, each questionnaire consisted of 5 transcriptions, with the exception of only two questionnaires with 6 transcriptions. All participants provided informed consent and demographic information before starting the questionnaire. Moreover, expert judges were assessed for their level of expertise in the forensic psychology field. After providing the instructions, the experimenter randomly gave each participant one of the 12 questionnaires.

Specifically, naïve judges were instructed to read each transcript carefully and rate its credibility on a 10-point scale (1=totally fabricated, 10=totally genuine). Forensic experts were first trained on the use of RM criteria for lie detection through a video lesson, which explained the theoretical background of the RM framework, the operationalization of the eight criteria, and how to compute the final RM score; moreover, it provided two practical examples of application of the RM criteria to transcripts. After the training session, the experimenter remained available to answer any questions the experts had related to the RM procedure. Then, they were randomly assigned one of the 12 questionnaires. For

each transcription, they were instructed to read the text carefully, evaluate it based on the eight RM criteria, and calculate the overall RM score using formula (1). Based on the final RM score, they were then asked to rate the transcripts' credibility on a 10-point scale (1=totally fabricated, 10=totally genuine). Other measures were also collected for naïve and expert judges, but were not included in the analysis (see Supplementary Materials for a detailed description).

TABLE 2. List of Reality Monitoring criteria adapted from Sporer (2004)

Reality Monitoring criteria	Automatic score	Human scoring
Vividness	-	rating scales, categorical measures (presence vs. absence), absolute / relative frequency
Realism	-	
Reconstructability	-	
Temporal information	LIWC "Tempo" + NER "DATE + NER "Time" + NER "Event"	
Spatial information	LIWC "Spazio" + NER "GPE + NER "LOC" + NER "FAC"	
Perceptual information	LIWC "Proc_Sen"	
Affective information	LIWC "Affett"	
Cognitive operations	LIWC "Mec_Cog"	

Note. LIWC and NER features selected for Experiment 2 to automatically compute RM criteria are provided in the second column. The general human scoring procedure for each criterion is provided in the third column.
Abbreviations: LIWC = Linguistic Inquiry and Word Count; NER = Named-Entity Recognition

2.2 Results

Accuracy was first computed at the subject level (i.e., the number of correct classifications divided by the total number of transcripts) and then averaged across naïve judges and forensic experts. The random baseline was established using the zero rule (see the Supplementary Materials for further details). Nonparametric analyses were conducted after verifying that the data distribution did not violate normality assumptions. Data were preprocessed in Python using the Google Colab interface, and statistical analyses (i.e., Wilcoxon signed-rank and Mann-Whitney U tests) were conducted in RStudio.

The results showed that naïve judges and forensic experts achieved an average accuracy slightly above chance level ($accuracy_{NJ}$ = 54.1% ± 20.1, $accuracy_{FE}$ = 59.4% ± 19.9) in identifying lie-tellers and truth-tellers. However, a Wilcoxon signed-rank test revealed that the naïve judges' performance was not significantly higher than chance level (V=3831.00, p=0.30, r_{rb}=0.05, 95% CI [- 0.15, 0.26]). A Wilcoxon signed-rank test also showed that forensic experts' performance was not significantly higher than chance level (V=431.00, p=0.06, r_{rb}=0.30 [- 0.07, 0.59]). Contrary to expectations, a Mann-Whitney U test also showed that the forensic experts' average accuracy was not significantly higher than that of naïve judges (U=2417.5, p=0.134, r_{rb}=0.12 [- 0.10, 0.32]), suggesting there is not enough evidence in favor of Hypothesis 1a. In the Supplementary Materials, we report the accuracy achieved by naïve judges and forensic experts in each experimental condition (truth-tellers vs. lie-tellers).

3. Experiment 2: machine-learning models trained on expert vs computerized reality monitoring scores

3.1 Methods and Materials

Text preprocessing

Before we extracted linguistic features, two raters manually preprocessed each transcription by removing semantic repetitions ("*we left at noon… ~~yes at noon~~*") and false starts ("*~~What I remember uhm~~… we went to London.*"). Using a two-way mixed-effects model for single measures (Shrout & Fleiss, 1979) with the JASP software (JASP Team., 2024), the intraclass correlation coefficient ($ICC_{3,1}$) was found to be ICC=0.99 [0.98, 0.99] for repetitions and ICC=0.98 [0.96, 0.99] for false starts, indicating excellent reliability among raters.

NER and feature extraction of the CL were computed after this preprocessing step. For the LIWC scoring, an additional preprocessing step was conducted before the computation. It consisted of automatic lowercase and tokenization of text using SpaCy and manual adjustment of bigrams (Italian: "non so," bigram: "nonso," English: "I don't know") and trigrams (Italian: "al di fuori," trigram: "aldifuori," English: "outside"). Word stemming was already included in the LIWC dictionary.

Feature extraction for reality monitoring

In this experiment, the RM criteria were automatically computed for each transcription using the LIWC software in combination with the NER technique.

LIWC is the gold standard software for analyzing word usage to identify psychosocial processes (Tausczik et al., 2010). It calculates the percentage of words in a text corresponding to more than 80 categories related to linguistic and psychosocial dimensions, as defined in validated dictionaries (a detailed description of the LIWC-15 software's functioning and categories is reported in Pennebaker et al., 2015). Using the Italian dictionary (software version LIWC-15), semantic features related to time, space, affect, sensory processes, and cognition were computed to reflect five of the eight RM criteria. The RM criteria of vividness, reconstructability, and realism are subjective scores that do not fit any LIWC categories (see Table 2) and are therefore often excluded from computation.

NER is an NLP technique that identifies and extracts information (named entities) from texts and classifies them into predefined categories, such as people, locations, organizations, and times. Named entities were automatically extracted using SpaCy, a Python library for NLP, on preprocessed text using a transformer-based model for the Italian language, available in Huggingface (*ita_nerIta_trf*, https://huggingface.co/bullmount/it_nerIta_trf). Table 2S in the Supplementary Materials lists all entities available in the *ita_nerIta_trf* model with their descriptions and examples. Figure 1S in the Supplementary Materials depicts a common way to represent text with annotated named entities.

Named entities related to "DATE," "TIME," and "EVENT" and "GPE", "LOC," and "FAC" were added to the LIWC features related to time (Tempo) and space (Spazio), respectively. This procedure was adopted because the LIWC software was mainly focused on words with a meaning associated with time and space (e.g., adverbs such as *"then"* and *"now"* and verbs such as "*to go*") without taking into account specific information about space and time (e.g., toponymies, such as "*Ibiza,*" and terms indicating time, such as "*Monday*" and "*11am*") that were detected using the NER technique. Table 2 provides a summary of how the RM features using LIWC and NER scores were computed.

Machine-Learning Models and Training

Logistic regression, support vector machine (SVM), decision tree, and random forest were employed for the computational analysis, conceptualizing the lie detection task as a binary classification problem. The in-

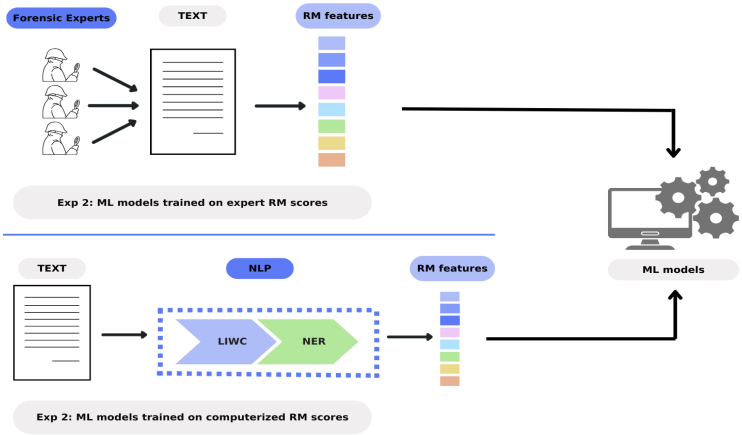
clusion of four ML models was intended to ensure that the obtained results were not dependent on the specific model assumptions and were stable across classifiers. The performance of the aforementioned models was evaluated and discussed in terms of accuracy. A random baseline was set using the zero rule. In the Supplementary Materials, we provide a brief description of each model and include additional metrics to add interpretations, such as the Area Under the Curve (AUC), precision, recall, and F1 score.

A nested cross-validation (NCV) framework was employed to evaluate the performance of ML models. NCV is a robust method for model evaluation and hyperparameter tuning in ML, especially in scenarios in which unbiased estimation of model performance is required (Müller & Guido, 2016). This method incorporated two layers: an inner loop for hyperparameter optimization and an outer loop for evaluating model performance. The inner loop, dedicated to hyperparameter optimization, utilized Grid Search for a systematic exploration of hyperparameter space to identify the optimal hyperparameter combination for each model. This process was repeated across the 10 folds of the inner cross-validation, ensuring that the hyperparameter selection was based solely on the training subset and not influenced by the test data. The best hyperparameter combination identified in the inner loop was then used to train the model on the entire training set of the outer loop. The outer loop then assessed the model's generalizability using a 10-fold cross-validation. To enhance the robustness and reliability of the performance estimates, the NCV process was repeated with three random seeds, thereby mitigating the effects of random variations in data partitioning and providing a robust, unbiased estimation of the model's performance. Following this rigorous procedure, the assessment of the model was safeguarded against overfitting and accurately reflected the model's capability to classify truthful and deceptive statements in our dataset.

Procedure

Figure 1 depicts the steps adopted to conduct the computational analyses in this experiment. Specifically, ML models were trained on two sets of ratings. For the first set, the scores provided by the three experts in Experiment 1 for the eight RM criteria were averaged for each story. This resulted in a vector of eight scores per story, which was then used to train the ML models with an NCV procedure. For the second set, the RM features were extracted using NLP techniques following the procedure described in section 3.1.2. This resulted in a vector of five scores per story, which was employed to train the ML models with an NCV procedure.

FIGURE 1. Procedure employed in Experiment 2 to obtain two sets of features to train ML models.



Note. The first set of features was obtained by averaging the ratings provided by three forensic experts for each RM criterion (upper part of the figure) for each text. A second set of features was obtained by leveraging NLP techniques (i.e., LIWC and NER) to extract linguistic features that mimic RM criteria on each text.
Abbreviations: ML= Machine Learning; RM = Reality Monitoring; NLP = Natural Language Processing; LIWC = Linguistic Inquiry and Word Count; NER = Named-Entity Recognition.

3.2 Results

Table 3 provides the results of Experiment 2. Table 3S (in the Supplementary Materials) reports the average performance and standard deviation in terms of accuracy, AUC, precision, recall, and F1 score obtained from the four ML models when they are trained on expert ratings of RM and computerized RM applied to full text.

When we used expert ratings of RM, the SVM and random forest models produced the highest average accuracy, 57.9% (±17.4) and 56.1% (±20.0), respectively. However, these performances were only slightly above chance level. Similarly, when we used computerized RM scores, the decision tree model exhibited the highest average accuracy, 57.1% (±15.8), but this performance was only slightly above the chance level. A Kruskal-Wallis test showed that the average accuracy of forensic experts from Experiment 1 was not significantly different from that of the best ML

models trained with expert and computerized RM scores in Experiment 2 ($H(2)=0.01$, $p=0.997$, $\eta^2(H)=0$ [0, 0.05]). These findings support the second hypothesis (Hyp. 2b), namely that expert judges performed poorly in the lie detection task because the RM criteria were poorly informative for this type of dataset.

TABLE 3. The performance of ML models is reported in terms of average accuracy using a 10-fold nested cross-validation.

		Experiment 2		Experiment 3	Experiment 4
Dataset	ML models	RM - Expert Scores	RM - Computerized Scores	RM + CL	Data-driven approach
Free speech	Logistic regression	-	-	68.3 (±20.6)	68.9 (±21.1)
	SVM	-	-	64.1 (±19.7)	**69.0 (±18.3)**
	Decision Tree	-	-	60.3 (±18.5)	59.6 (±18.8)
	Random Forest	-	-	**69.4 (±16.5)**	68.2 (±17.9)
Unexpected questions	Logistic regression	-	-	56.9 (±19.1)	66.2 (±19.5)
	SVM	-	-	53.3 (±19.2)	64.7 (±17.3)
	Decision Tree	-	-	**60.8 (±14.6)**	57.5 (±18.6)
	Random Forest	-	-	60.6 (±18.5)	**67.4 (±18.8)**
Full text	Logistic regression	53.4 (±19.8)	40.2 (±18.0)	**67.2 (±19.1)**	73.3 (±18.6)
	SVM	**57.9 (±17.4)**	48.8 (±22.9)	62.3 (±17.6)	**77.3 (±17.2)**
	Decision Tree	49.9 (±19.7)	**57.1 (±15.8)**	57.2 (±16.6)	53.5 (±21.7)
	Random Forest	56.1 (±20.0)	52.8 (±20.4)	64.9 (±20.3)	75.1 (±17.5)

Note. Standard deviations are reported in brackets. The best accuracy achieved in each experiment for each part of the dataset analyzed is in bold.
Abbreviations: ML = Machine Learning; RM = Reality Monitoring; CL = Cognitive Load; SVM = Support Vector Machine.

4. Experiment 3: Theory-led approach combining RM and CL

4.1 Methods and Materials

Feature Extraction for Cognitive Load

Previous research employed statistics regarding the text's length, readability, and complexity to extract linguistic features associated with CL in deception studies (Hauch et al., 2015; Pérez-Rosas & Mihalcea, 2015; Sarzynska-Wawer et al., 2023; Solà-Sales et al., 2023; Zhou et al., 2004). Statistics associated with CL were automatically computed on preprocessed text using the Python library TEXTSTAT and are reported in Table 4.

TABLE 4. List of the linguistic features associated with the cognitive load framework and their operational definition.

Features associated with cognitive load	Operational definition
num_sentences	Total number of sentences
word_count	Total number of words
num_unique_words	Total number of unique words
type-token ratio	Total number of unique words divided by the total number of words
num_syllables	Total number of syllables
avg_num_syllables_per_word	Average number of syllables per word
num_content_words	Total number of words that express lexical meaning
num_unique_content_words	Total number of unique content words
content-word diversity	Total number of unique content words divided by the total number of content words
fk_grade	The Flesch-Kincaid Grade Level Index expressing the grade level required to understand the text
fk_read	The Flesch Reading-Ease Level Index expressing the texts's readability

Procedure

Figure 2 depicts the procedure adopted for the computational analyses in Experiments 3 and 4. Specifically, the dataset was first split into three sections:

i. Free Speech, which contained the transcription of the free recall of the holiday;
ii. Unexpected Questions, which contained the responses to unexpected questions;
iii. Full Text, intended as the combination of text from the Free Speech and Unexpected Questions sections.

Then, linguistic features associated with RM and CL were automatically extracted following the procedure defined in sections 3.1.2 and 4.1.1. Subsequently, the same ML models and NCV procedure employed in Experiment 2 were applied to each section of the dataset.

4.2 Results

Table 3 reports the results from this experiment (see also Table 4S in the Supplementary Materials). Considering the four ML models trained on linguistic features extracted from the Full Text dataset, using the RM and CL framework, we observed a general increase in the obtained predictive performance. In fact, after combining features from two theoretical frameworks, we reached an accuracy of 69.4% (±16.5), with an improvement of up to 9.3% over models trained solely on expert or computerized RM scores. The results show that the inclusion of linguistic features associated with the CL framework resulted in enhanced accuracy of ML models in detecting verbal deception compared to models trained solely on features from RM, confirming our hypothesis (Hyp. 3a).

Additionally, features from RM and CL were specifically extracted from statements in the Free Speech, Unexpected Questions, and Full Text sections to investigate their potential informative and predictive role. Findings comparing the performance obtained from ML models trained on each section showed that linguistic features from the Free Speech section significantly contributed to an increase in overall accuracy across the four models. Contrary to our expectations (Hyp. 3b), linguistic features from the Unexpected Questions section yielded lower accuracy rates, similar to those achieved by models trained exclusively on expert and computerized RM scores. Interestingly, when we leveraged the Full Text section for feature extraction, there was a slight decline in performance, with the highest average accuracy recorded at 67.2% (±19.1).

5. Experiment 4: Data-driven approach using NLP features

5.1 Methods and Materials

Feature Extraction and Selection

This experiment involved the extraction of a comprehensive set of 128 linguistic features, using a combination of NLP techniques on preprocessed texts. The Python library TEXTSTAT was employed to compute 11 basic textual features related to the text's length, readability, and complexity, as in Experiment 3. The LIWC software was employed to extract 85 psychological, linguistic, and affective dimensions from texts. The Python SpaCy library was employed to extract 15 named entities (with the NER technique) and 17 grammatical and syntactical parts of speech (with the POS-tagging technique) in the text.

Using the scikit-learn library in Python, the original set of 128 features was narrowed down to a more manageable and informative set of 20 features that best captured the nuances of the textual data, following these steps:

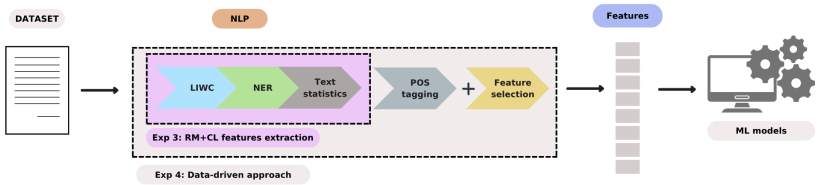
1. Removing POS features unrelated with the purpose of the task, such as the number of spaces (using spacebar; SPACE), punctuation usage (PUNCT), the number of symbols (SYM), and other noncanonical pos tags (X); those features were more related to the transcription process than to telling a truthful or a deceptive story and could represent a confounder if included in the analysis.
2. Removing duplicates, such as the numerical features in LIWC and POS tagging that were already detected with the NER technique and the LIWC "Non_flu" category, which was a duplicate of the POS tag "INTJ."
3. Removing LIWC linguistic features overlapping with the grammatical and syntactic features extracted with the POS-tagging technique, given that the latter is more efficient and complete in extracting these features than the LIWC software.
4. Feature-engineering a new variable named "fillers" by summing filler words and non-fluencies, typical of hesitation and oral speech patterns, extracted with the LIWC software and the POS tagging. Specifically, the LIWC "riempiti" category was added to the POS tag "INTJ."
5. Removal of features that showed more than 60% of zero values across the dataset.

6. Selection of the best 20 features after testing for mutual information, which measures the linear and nonlinear dependency between random variables, ensuring that each selected feature contributed significantly to the predictive models (Ross, 2014). This selection process was performed using the function *sklearn.feature_selection.SelectKBest* (Pedregosa et al., 2011).

Procedure

Figure 2 depicts the procedure adopted for the computational analyses in Experiments 3 and 4. As in Experiment 3, the dataset was first divided into three sections (i.e., Free Speech vs. Unexpected Questions vs. Full Text). Then, following the feature-extraction and selection process described in the previous section, four ML models were trained using an NCV procedure on the best 20 linguistic features from each section of the dataset. The ML models and the NCV procedure are described in Section 3, subsection ML Models and Training.

FIGURE 2. Procedures employed in Experiments 3 and 4 to create a set of linguistic features to feed ML models.



Note. In Experiment 3, ML models were trained on linguistic features that mimic RM and CL using NLP techniques. In Experiment 4, a wider range of linguistic features was extracted from texts using NLP techniques; then, an automatic feature selection step was applied to obtain a final set of features used to train ML models.
Abbreviations: NLP = Natural Language Processing; LIWC= Linguistic Inquiry and Word Count; NER = Named Entity Recognition; POS = Part-of-Speech; RM = Reality Monitoring; CL = Cognitive Load.

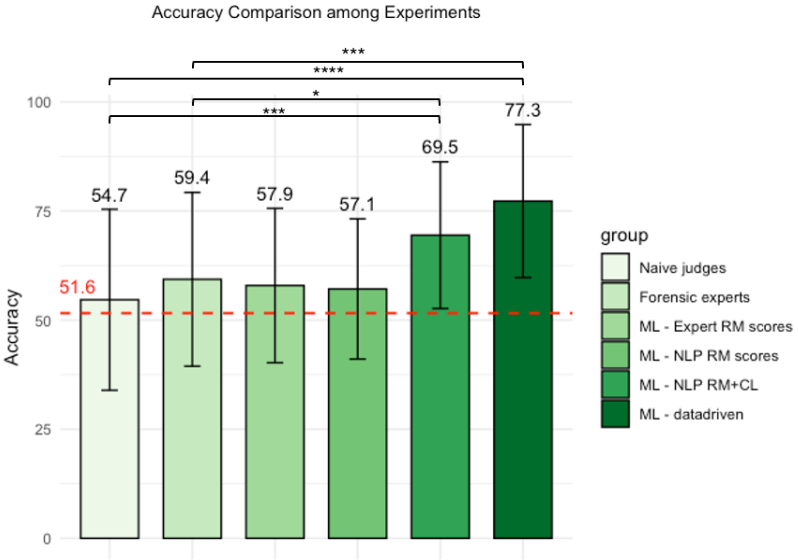
5.2 Results

Data-driven approach

The data-driven approach demonstrated an overall improvement in performance compared to Experiment 3 (Table 3; see also Table 5S in Supplementary Materials), providing evidence in support of the superior performance of data-driven approaches compared to theory-led ap-

proaches (Hyp. 4a). Specifically, the Full Text section showed a significant leap in accuracy, particularly with the SVM, which reached the highest accuracy (77.3% ±17.2). The Unexpected Questions section, despite having the lower performance, showed a noticeable improvement in accuracy, with the random forest model performing the best, at 67.4% (±18.8). In the Free Speech section, the SVM achieved the highest accuracy, 69.0% (±18.3), but it did not surpass the performance of the best model trained with RM and CL in the same section.

FIGURE 3. Bar plot of the average accuracy (and standard deviation) obtained from the four experiments.



Note. Error bars represent standard deviations. The red dashed line represents the chance level (51.6%) as defined using the zero rule. Significant comparisons are reported only for human vs. machine approaches. Naive judges' and forensic experts' accuracy are derived from Experiment 1; ML-Expert RM scores: best accuracy achieved in Experiment 2 using ML models trained on RM scores provided by forensic experts in Experiment 1. ML-NLP RM scores: best accuracy achieved in Experiment 2 using ML models trained on RM scores computed with NLP techniques; ML-NLP RM+CL: best accuracy achieved in Experiment 3 using ML; ML-data-driven: best accuracy achieved in Experiment 4 using ML. *Abbreviations:* ML = machine learning; RM = Reality Monitoring; CL = Cognitive Load
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Comparing accuracy among experiments

Figure 3 presents the accuracy achieved by human judges and the best

ML models in each experiment. A Kruskal-Wallis test revealed a statistically significant difference in the average accuracy scores across experiments ($H(5)=39.21$, $p< 0.01$, $\eta^2(H)=0.13$ [0.07, 0.24]). Dunn's post hoc tests with false-discovery rate correction were applied to assess differences between pairs of conditions. Notably, the average accuracy of the best ML model trained on RM and CL features was significantly higher than the average accuracy achieved by naïve judges ($z=3.92$, $p<0.01$) and forensic experts ($z=2.39$, $p=0.02$) in Experiment 1. Similarly, the average accuracy of the best ML model trained on features extracted using a data-driven approach was found to be significantly higher than that of naïve judges ($z=5.47$, $p<0.001$) and forensic experts ($z=3.67$, $p<0.01$) in Experiment 1. Table 6S of the Supplementary Materials presents the remaining post hoc comparisons. These findings support our last hypothesis, proving that theory-led and data-driven approaches leveraging ML and NLP techniques are more effective in detecting verbal deception than human judges (Hyp. 4b).

6. General Discussion

This series of studies contributes to deception detection research by examining the theoretical frameworks of Reality Monitoring (RM) and Cognitive Load (CL) through computational methods, advancing our understanding of how these frameworks function in the analysis of deceptive language. In addition, through four experiments, we assessed the effectiveness of human (naïve vs. experts) and ML-based (theory-led vs. data-driven) approaches in deception detection when applied to a dataset of interviews with unexpected questions.

In the first experiment, we tested the RM framework by comparing the performance of naive and expert judges, with the latter having been trained in RM. Neither naïve judges nor forensic experts surpassed the chance level ($accuracy_{NJ} = 54.1\% \pm 20.1$, $accuracy_{FE} = 59.4\% \pm 19.9$). Additionally, the average accuracy of forensic experts was not significantly higher than that of naïve judges. Although this result was expected for naïve judges and aligns with previous studies (Bond & DePaulo, 2006; Curci et al., 2019; DePaulo et al., 2003; Pérez-Rosas & Mihalcea, 2015), it was contrary to expectations for forensic experts (Hyp. 1a). In fact, previous research proved the effectiveness of RM criteria in verbal-deception detection, reaching approximately 70% accuracy (Amado et al., 2016; Gancedo et al., 2021; Vrij, 2008; Vrij et al., 2022). Additionally, a meta-analysis of studies has shown that the average accuracy that can be obtained by following different cues is 67% (Hartwig & Bond, 2014).

To address the reasons behind this poor performance of experts in Ex-

periment 1, we introduced a second experiment that leverages computational techniques and ML models. ML models were trained on two sets of ratings, those given by expert judges using the RM criteria (i.e., expert ratings) in Experiment 1 and those obtained by computerized methods using NLP techniques for RM (i.e., computerized ratings), to determine whether experts' poor performance was due to an inaccurate assessment of RM criteria, a lack of informativeness of those criteria for this dataset, or a decision-making problem in combining all the information. Our findings showed that the average accuracy of forensic experts was not significantly higher than those of the best ML models trained on experts ratings (accuracy = 57.9% ± 17.4) or computerized ratings of RM (accuracy = 57.1% ± 15.8), supporting the hypothesis that RM criteria might be poorly informative for the dataset under analysis, regardless of whether they were evaluated by forensic experts or derived from computational methods (Hyp. 2b).

The results of Experiments 1 and 2, collectively, challenged the presumed robustness of RM in deception detection and raises questions about its sensitivity across different datasets and contexts. Furthermore, these results challenge the efficacy of computational approaches built on theoretical frameworks that ultimately exhibit limited effectiveness, as demonstrated in this case with the RM. Considering these premises, we tested in a third experiment whether the combination of multiple theoretical frameworks, specifically the RM and CL frameworks, could enhance ML models' accuracy in detecting verbal deception (Hyp. 3a).

The results from Experiment 3 demonstrated that the combination of features from two theoretical frameworks resulted in an accuracy of 69.4% (±16.5), with an improvement of 9.3% compared to models trained solely on expert or computerized RM scores. Furthermore, the average accuracy of the ML model trained on RM and CL features was significantly higher than that of naïve and expert judges in Experiment 1. This finding may be attributed to the fact that the dataset under analysis was specifically designed to increase CL in lie-tellers by posing unexpected questions (Monaro et al., 2020, 2022) and to the inherently higher accuracy of the CL approach (Vrij et al., 2015). Alternatively, the simple inclusion of a higher number of relevant features might have led to this higher accuracy. Because CL features were more prevalent in responses to unexpected questions, we hypothesized that ML models trained on features extracted from Unexpected Questions would yield higher accuracy than those trained on features extracted from Free Speech (Hyp. 3b). Contrary to this hypothesis, the results indicated that linguistic features derived from the Free Speech dataset significantly contributed to an increase in overall accuracy across all models. Linguistic features derived from the Unexpected Questions dataset yielded lower accuracy, similar to that

achieved by models trained exclusively on expert and computerized RM scores. Notably, when we employed the Full Text dataset for feature extraction, performance slightly declined, with the highest average accuracy recorded at 67.2% (± 19.1). One possible interpretation of this result is that the extraction of linguistic markers useful for detecting deceptive narratives is more effective with longer texts, as seen in the Free Speech and Full Text datasets.

Finally, when NLP techniques are employed, various methodologies are available for extracting features from textual data, including theory-led, data-driven, and hybrid approaches (Van Der Zee et al., 2022). Although hybrid models are generally preferred over data-driven models in forensic contexts, given that data-driven models may be effective at prediction but ineffective at explanation, studies have shown a lower effectiveness of hybrid approaches compared to data-driven models (Van Der Zee et al., 2022). Therefore, in the fourth experiment, we investigated the performance of a data-driven approach in this lie detection task and compared the results with those of previous experiments. Specifically, comparisons from Experiments 3 and 4 allow us to examine the effectiveness of theory-led vs. data-driven approaches. NLP techniques were employed to extract a broad set of linguistic features, and a data-driven feature selection strategy (Ghosh, 2022) was used to identify a subset of highly informative features.

The results from Experiment 3 showed that training ML models on combined linguistic features from two deception frameworks (i.e., RM and CL) yielded higher but moderate accuracy ($best\_accuracy_{freespeech} = 69.4\% \pm 16.5$). However, in Experiment 4, there was a significant leap, particularly with the SVM, which reached the highest accuracy ($best\_accuracy_{fulltext} = 77.3\% \pm 17.2$). This performance was also significantly better than that of naïve judges and forensic experts (Experiment 1). These findings underscore the efficacy of a data-driven approach in discerning patterns in comprehensive textual data, compared to theory-led approaches that combine linguistic features derived from the RM alone or in combination with the CL framework. We confirmed our hypothesis that a data-driven approach may be particularly relevant in contexts in which theory-based methods have demonstrated limited effectiveness (Hyp. 4a). Indeed, while previous studies have shown that RM typically achieves around 70% accuracy in distinguishing truth from deception (Vrij, 2008), it yielded lower accuracy in our study (from around 57% to 59%). In contrast, our data-driven NLP approach surpassed the expected 70% accuracy, suggesting that it could serve as a reliable alternative in cases where traditional and theory-based methods, like RM, fall short.

Most importantly, the results from Experiments 1 to 4 suggest that training ML models on features extracted using NLP techniques may represent a more advantageous approach in detecting deception from narratives, overcoming the modest accuracy achieved by naïve and expert judges because of their ability to handle complex patterns of language data (Hyp. 4b). Moreover, they could help identify which linguistic features are more informative to derive a final decision.

6.1 Limitations and future perspectives

Although this study's results highlight significant advancements in the field of deception detection, comparing human judgments to computational predictions, several limitations must be acknowledged to contextualize the results and guide future research properly.

First, the reliance on a relatively small dataset significantly constrains the generalizability of the findings. The dataset, comprising only 62 narratives and solely in Italian, limits our ability to confidently extend these results to broader and more heterogeneous contexts. Future studies replicating our experiments using larger and more varied datasets in different languages would enhance the robustness of our findings and potentially reveal cultural and linguistic nuances in deception detection. Additionally, the dataset under analysis was designed to collect outright false statements. However, a more frequent and ecological form of deception is constituted by embedded lies (Caso et al., 2023; Verigin et al., 2019), where people interweave truth and lies together. As a consequence, the detection rates found in our series of studies may be even lower when considering this type of deception.

Secondly, forensic experts assessed the RM criteria on a 7-point scale to judge the narratives' veracity. However, other approaches are available in the literature. For example, one approach involves evaluating the absence and presence of each criterion on a 3-point scale (*0=absent, 1=partially present, 2=totally present*), and another counts the frequency of details for at least five of the eight criteria. The study's results provide insights limited to the qualitative assessment of RM criteria on a 7-point scale and may not be generalizable to methodologies that utilize the frequency of details. Future research could employ a different approach for RM assessment, for instance, by taking into account the frequency of details.

Third, by focusing exclusively on specific deception cues, such as those provided by the RM criteria, people may overlook other potentially informative cues. For instance, the verifiability of details plays a crucial role in deception detection, suggesting that truth-tellers provide a higher

proportion of verifiable details (Nahari et al., 2012; Palena et al., 2021; Verschuere et al., 2021). Accordingly, a recent study found that asking judges to assess narratives for their verifiability or detailedness, rather than veracity, yields higher accuracy, up to 70% (Verschuere et al., 2023). In our study, forensic experts may have underperformed relative to ML models because they employed an ineffective heuristic, as also evidenced by the results from Experiment 2, which demonstrated the RM's limited informativeness in determining the veracity of our dataset. Our research can therefore be extended by asking forensic experts to use different deception cues, such as assessing the verifiability of details or employing criteria-based content analysis (Steller & Koehnken, 1989).

Lastly, we found that a data-driven approach yielded the highest accuracy when testing the models using nested cross-validation. However, it is essential to recognize the limitations of these results, especially considering the potential impact of error rates in forensic contexts. Although our model achieved a significant improvement, the approximate 30% error rate remains a concern, particularly given the serious implications of misclassification in legal settings where credibility assessments can influence case outcomes. These findings underscore the need for ongoing research and refinement of NLP and machine learning methods to enhance reliability in high-stakes applications. Indeed, there is substantial room to explore more sophisticated ML approaches in future studies. For example, techniques such as word embeddings offer a promising avenue for future research. Word embeddings offer a means to capture semantic relationships between words by representing them in a high-dimensional space (Loconte et al., 2023), thereby uncovering subtle linguistic patterns associated with deceptive speech that traditional models often overlook. Moreover, using neural-network architectures, such as long short-term memory networks and transformers, would allow future research to process sequential data more effectively, potentially achieving higher accuracy in models trained on textual data. Finally, fine-tuning large language models has also been proven to be effective in detecting deception in raw texts (Loconte et al., 2023).

However, a significant limitation of data-driven approaches is their lack of explainability, which is particularly relevant in forensic settings, in which understanding the rationale behind an algorithm's decision is as crucial as the decision itself. Although data-driven methods can efficiently identify patterns, make predictions, and sometimes explain which specific linguistic features contributed to those predictions, these outputs often are not easy to interpret. This opacity makes it challenging to align these findings with general theories of memory and deception, which is necessary for forensic credibility. Ensuring that computational techniques not only predict but also explain their predictions in terms

that relate to established psychological theories will be essential for their acceptance and ethical application in legal contexts.

Overall, these future perspectives suggest a trajectory toward more integrated and sophisticated systems that leverage a combination of theoretical insights and cutting-edge ML techniques. By broadening the theoretical frameworks and enhancing the computational methods used in deception detection, researchers can provide more accurate, reliable, and explainable tools for forensic and other critical applications. This progression promises not only to advance the understanding of deception but also to enhance practical lie detection capabilities in real-world settings.

Conclusion

To conclude, the experimental results from four experiments provided theoretical and practical considerations for advancing research on verbal deception detection.

From a theoretical perspective, the exploration of multiple theoretical frameworks, such as RM and CL, through computational methods, has demonstrated the potential to enhance accuracy in identifying deceptive narratives and calls for the fusion of more diverse theoretical perspectives to offer more robust tools for deception detection, especially when one framework alone falls short.
From a practical perspective, the integration of computational methods in deception detection has significant implications in forensic contexts, particularly when credibility assessment is required in criminal proceedings. Potentially, computational methods may aid forensic experts when they perform only slightly above chance level, even after being trained on well-established frameworks.

However, the ethical implications of deploying computational methods in such sensitive settings are significant. Ethical considerations must include discussions on the transparency of the algorithms used (von Eschenbach, 2021; Zerilli et al., 2019), the potential for overreliance on automated systems without adequate human oversight, and the need for ongoing evaluation of the efficacy and fairness of these systems, especially when they influence judicial outcomes. Although our results in Experiment 3 were modest compared to those in Experiment 4, we emphasize the importance of employing hybrid approaches that combine data-driven and theory-led methodologies. Such approaches would provide a more explainable model, which is crucial in forensic contexts in which the reasoning behind decisions must be transparent and justifiable.

The results of this series of studies suggest the need for future studies that aim to integrate advanced computational techniques for deception detection, as well as to provide transparent algorithms for translating results in high-stakes scenarios, like forensic contexts.

In the next Chapter, we build upon the findings from this Chapter and test whether resorting to embedding representations of statements, stemming from fine-tuned large language models, can improve performance in deception detection.

Data availability statement

Data and scripts used to run the experiments are available at https://osf.io/usz26/

References

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *European Journal of Psychology Applied to Legal Context*, *7*(1), 3–12. https://doi.org/10.1016/J.EJPAL.2014.11.002

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/J.IJCHP.2016.01.002

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/S15327957PSPR1003_2

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app

Caso, L., Cavagnis, L., Vrij, A., & Palena, N. (2023). Cues to deception: can complications, common knowledge details, and self-handicapping strategies discriminate between truths, embedded lies and outright lies in an Italian-speaking sample? *Frontiers in Psychology*, *14*. https://doi.org/10.3389/FPSYG.2023.1128194

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

Curci, A., Lanciano, T., Battista, F., Guaragno, S., & Ribatti, R. M. (2019). Accuracy, confidence, and experiential criteria for lie detection through a videotaped interview. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/FPSYT.2018.00748

Deeb, H., Vrij, A., Palena, N., Hypšová, P., Dib, G., Leal, S., & Mann, S. (2024). Honesty repeats itself: comparing manual and automated coding on the veracity cues total details and redundancy. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716424000298

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*(9), 913–920. https://doi.org/10.1037/0003-066X.46.9.913

Elaad, E. (2009). Lie-detection biases among male police interrogators, prisoners, and laypersons. *Psychological Reports*, *105*(3), 1047–1056. https://doi.org/10.2466/PR0.105.F.1047-1056

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). *Automatic Detection of Verbal Deception*. https://doi.org/10.1007/978-3-031-02158-9

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/EJPALC2021A10

Ghosh, C. (2022). Data Analysis with Machine Learning for Psychologists: Crash Course to Learn Python 3 and Machine Learning in 10 hours. *Data Analysis with Machine Learning for Psychologists: Crash Course to Learn Python 3 and Machine Learning in 10 Hours*, 1–161. https://doi.org/10.1007/978-3-031-14634-3

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/ACP.3052

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of Criteria-Based Content Analysis. *Psychological Assessment*, *29*(6), 819–834. https://doi.org/10.1037/PAS0000426

JASP Team. (2024). *JASP (Version 0.18.3) [Computer software]*. https://jasp-stats.org

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018). Using Named Entities for Computer-Automated Verbal Deception Detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Appl. Cognit. Psychol.*, *32*(3), 354–366. https://doi.org/10.1002/acp.3407

Lancaster, G. L. J., Vrij, A., Hope, L., & Waller, B. (2013). Sorting the Liars from the Truth Tellers: The Benefits of Asking Unanticipated Questions on Lie Detection. *Applied Cognitive Psychology*, *27*(1), 107–114. https://doi.org/10.1002/ACP.2879

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. *Scientific Reports 2023 13:1*, *13*(1), 1–19. https://doi.org/10.1038/s41598-023-50214-0

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law*, *11*(1), 99–122. https://doi.org/10.1080/10683160410001726356

Melis, G., Ursino, M., Scarpazza, C., Zangrossi, A., & Sartori, G. (2024). Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions. *Scientific Reports*, *14*(1). https://doi.org/10.1038/S41598-024-63156-Y

Mihalcea, R., & Strapparava, C. (2009). The lie detector. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, 309. https://doi.org/10.3115/1667583.1667679

Monaro, M., Capuozzo, P., Ragucci, F., Maffei, A., Curci, A., Scarpazza, C., Angrilli, A., & Sartori, G. (2020). Using blink rate to detect deception: A study to validate an automatic blink detector and a new dataset of videos from liars and truth-tellers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12183 LNCS*, 494–509. https://doi.org/10.1007/978-3-030-49065-2_35

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics.

Scientific Reports, *8*(1). https://doi.org/10.1038/S41598-018-20462-6

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, *12*(5). https://doi.org/10.1371/JOURNAL.PONE.0177851

Monaro, M., Maldera, S., Scarpazza, C., Sartori, G., & Navarin, N. (2022). Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. *Computers in Human Behavior*, *127*. https://doi.org/10.1016/J.CHB.2021.107063

Müller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python: A guide for data scientists* (1st ed.). O'Reilly Media, Inc. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

Nahari, G. (2016). When the long road is the shortcut: a comparison between two coding methods for content-based lie-detection tools. *Psychology, Crime and Law*, *22*(10), 1000–1014. https://doi.org/10.1080/1068316X.2016.1207770

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*(2), 227–239. https://doi.org/10.1111/J.2044-8333.2012.02069.X

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior*, *40*(4), 440–457. https://doi.org/10.1037/LHB0000193

Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2021). Validity of content-based techniques for credibility assessment—How telling is an extended meta-analysis taking research bias into account? *Applied Cognitive Psychology*, *35*(2), 393–410. https://doi.org/10.1002/ACP.3776

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *ACL-HLT 2011* -

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, *1*, 309–319.

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2021). The Verifiability Approach: A Meta-Analysis. *Journal of Applied Research in Memory and Cognition*, *10*(1), 155–166. https://doi.org/10.1016/J.JARMAC.2020.09.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*.

Pennebaker, J. W., Botth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic Inquiry and Word Count: LIWC2015. In *Austin, TX: Pennebaker Conglomerates*.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 3391–3401.

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1120–1125. https://doi.org/10.18653/V1/D15-1133

Sarzynska-Wawer, J., Pawlak, A., Szymanowska, J., Hanusz, K., & Wawer, A. (2023). Truth or lie: Exploring the language of deception. *PLOS ONE*, *18*(2 February), e0281179. https://doi.org/10.1371/journal.pone.0281179

Savy, R. (2006). Specifiche per la trascrizione ortografica annotata dei testi raccolti. In *Italiano parlato. Analisi di un dialogo* (pp. 1–37). Liguori.

Schutte, M., Bogaard, G., Mac Giolla, E., Warmelink, L., Kleinberg, B., & Verschuere, B. (2021). *Man versus Machine: Comparing manual with LIWC coding of perceptual and contextual details for verbal lie detection*. https://doi.org/10.31234/OSF.IO/CTH58

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Solà-Sales, S., Alzetta, C., Moret-Tatay, C., & Dell'Orletta, F. (2023). Analysing Deception in Witness Memory through Linguistic Styles in Spontaneous Language. *Brain Sciences*, *13*(2). https://doi.org/10.3390/BRAINSCI13020317

Sporer, S. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Appl. Cognit. Psychol.*, *11*(5), 373–397.

Sporer, S. L. (2004). Reality monitoring and detection of deception. *The Detection of Deception in Forensic Contexts*, 64–102. https://doi.org/10.1017/CBO9780511490071.004

Steller, M., & Koehnken, G. (1989). Criteria-Based Content Analysis. *The Suggestibility of Children's Recollections*. https://doi.org/10.1037/T27704-000

Street, C. N. H., & Masip, J. (2015). The source of the truth bias: Heuristic processing? *Scandinavian Journal of Psychology*, *56*(3), 254–263. https://doi.org/10.1111/SJOP.12204

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. *Applied Cognitive Psychology*, *18*(6), 653–668. https://doi.org/10.1002/ACP.1021

Tausczik, Y., and, J. P.-J. of language, & 2010, undefined. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journals.Sagepub.Com*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Van Der Zee, S., Poppe, R., Havrileck, A., & Baillon, A. (2022). A Personal Model of Trumpery: Linguistic Deception Detection in a Real-World High-Stakes Setting. *Psychological Science*, *33*(1), 3–17. https://doi.org/10.1177/09567976211015941

Verigin, B. L., Meijer, E. H., Bogaard, G., & Vrij, A. (2019). Lie prevalence, lie characteristics and strategies of self-reported good liars. *PLoS ONE*, *14*(12). https://doi.org/10.1371/JOURNAL.PONE.0225566

Verschuere, B., Bogaard, G., & Meijer, E. (2021). Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis. *Applied Cognitive Psychology*, *35*(2), 374–384. https://doi.org/10.1002/ACP.3775

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 2023 7:5*, *7*(5), 718–728. https://doi.org/10.1038/s41562-023-01556-2

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/S13347-021-00477-0

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law*, *11*(1), 3–41. https://doi.org/10.1037/1076-8971.11.1.3

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior*, *24*(4), 239–263. https://doi.org/10.1023/A:1006610329284

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/JIP.82

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/LCRP.12088

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal Lie Detection: Its Past, Present and Future. *Brain Sciences*, *12*(12). https://doi.org/10.3390/BRAINSCI12121644

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior*, *33*(2), 159–166. https://doi.org/10.1007/S10979-008-9143-Y

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing Lie Detection by Inducing Cognitive Load on Liars: A Review of Relevant Theories and Techniques Guided by Lessons from Polygraph-Based Approaches. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/FPSYG.2013.00014

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology*, *32*(4), 661–683. https://doi.org/10.1007/S13347-018-0330-6

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6F

Supplementary Materials

Materials and methods

Dataset

The dataset employed for this study was previously collected by Monaro et al. (2020). The same dataset was also analyzed in previous studies to detect deception through blink rate (Monaro et al., 2020) and facial expressions (Monaro et al., 2022), comparing the performance of human judges and ML models.

The dataset consisted of 62 videotaped interviews of Italian participants (43 females, age range 20-29, who voluntarily participated in the study) talking about a previous holiday experience. A total of 32 participants were allocated to the truthful condition, wherein they were instructed to describe a real holiday experience that occurred within the preceding 12 to 18 months. Conversely, the remaining 30 participants were designated to the deceptive condition, where they were required to describe an entirely fabricated holiday. Notably, previous studies already used past holidays as a target event for lying (Sartori et al., 2008; Curci et al., 2019) because the recollection of a holiday involves the same cognitive processes as telling an alibi during a criminal investigation. Therefore, in this context, recalling a past holiday can be considered the analogue of telling a false alibi in a low-stakes scenario.

Participants in the truthful condition were asked to recall some information about their holiday by filling out a form before the interview. This procedure was adopted to avoid truth-tellers producing biased memories because of the time elapsed between the original holiday experience and the interview. Additionally, they were instructed to exclude any details that they could not recall accurately and were allowed to use supporting materials such as photographs and videos to aid in memory recall and ensure the veracity of their recollections. Similarly, to minimize the likelihood of participants in the deceptive condition incorporating factual details into their fabricated narratives, they were supplied with a pre-filled form containing specific, predetermined information about the fictitious holiday they were required to simulate as their own experience. Each videotaped interview was composed of three distinct phases:

1. **Baseline**, in which the interviewee provided their autobiographical information.

2. **Free speech**, in which the interviewee freely recalled their holiday experience for approximately two minutes.
3. **Unexpected Questions**, in which the interviewer asked unexpected questions about the holiday experience to increase the interviewee's cognitive load (e.g., "Did a particular event occur during the holiday that made it necessary to revise the initial plans?").

The average length of the videos was 9.56 minutes.

Each videotaped interview was manually transcribed following the procedure described in the Manuscript (Methods section in Experiment 1). In Table S1, the standardization of hesitations and laughter utterances is reported.

TABLE 1S. Standardized transcription of hesitations and laughter utterances.

Original utterance	Standardized transcription
mmm	uhm
eee	eh
aaa	ah
ooo	oh
uuu	uh
emm	ehm
ahahah	[risata]

Procedure - Experiment 1

After providing the credibility rating, naïve judges were also asked: (i) to provide their confidence level, rated on a 10-point scale (1=totally unconfident, 10= totally confident); ii) to indicate which elements supported their decisions (open-ended question); (iii) to point out which interview phase (Free speech vs. Unexpected questions vs. both vs. I didn't pay attention to it) was the most useful for making the decision; (iv) to indicate on which of the eight RM criteria they based their judgment. Notably, naïve judges did not know that the criteria listed were referred to a psychological framework associated with deception.

Forensic experts, after providing the credibility rating, were asked to rate on a 10-point scale how difficult was the task (1=totally easy, 10= totally difficult), how much the total RM score helped in driving the credibility

judgment (1=totally unhelpful, 10=totally helpful), and how much the credibility judgment driven by the total RM score agreed with their intuitive judgment (1=totally disagree, 10= totally agree). These measures were not included in the analyses for this study.

Machine-learning models

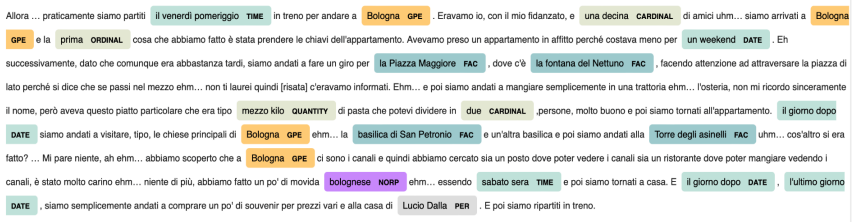
Four machine learning models (Alzubi et al., 2018) were included in this study for Experiments 2, 3, and 4:

- **Logistic Regression,** which estimates the probability that a given input belongs to a particular category using a logistic function;
- **Support Vector Machine** (SVM), which is a classifier that identifies the hyperplane that best separates different classes in the feature space;
- **Decision Tree,** which is a non-parametric method that models decisions and their possible consequences as a tree-like structure, where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label;
- **Random Forest,** which is an ensemble learning technique that consists of developing multiple decision trees for the training phase and considering, as a final output, the most frequently occurring class from those outputted by the individual trees.

Named-Entity Recognition

In Figure 1S, a common way of visualizing automated named-entity extractions in Python using the SpaCy library is depicted. The list of the possible named entities detected is in Table 2S.

FIGURE 1S. Narrative of an autobiographical event annotated with named entities using Named-Entities Recognition in SpaCy.



Note. Examples of entities automatically detected in this narrative refer to named people (e.g., Lucio Dalla), locations (e.g., Piazza Maggiore, Bologna), dates (e.g., weekend), and times (e.g., venerdì pomeriggio).

TABLE 2S. Labels, brief descriptions, and a few examples of the extracted named entities with the Python library SpaCy.

Entity	Description	Example
DATE	Absolute or relative dates or periods	"December 25, 2022", "10th August 1998", "Yesterday"
TIME	Times smaller than a day	"2:30 PM", "9 o'clock", "morning"
GPE	Countries, cities, states	"United States", "Paris", "Tokyo"
LOC	Non-GPE locations, mountain ranges, bodies of water	"Central Park", "Mount Everest", "Amazon Rainforest"
PER	People, including fictional	"Steve Jobs", "Emma Johnson", "Harry Potter"
ORDINAL	"first", "second", etc.	"First", "Third", "Tenth"
ORG	Companies, agencies, institutions, etc.	"Google", "Apple Inc.", "United Nations"
QUANTITY	Measurements, as of weight or distance	"10 kilograms", "5 liters", "100 meters"
WORK_OF_ART	Titles of books, songs, etc.	"Mona Lisa", "Hamlet", "Gone with the Wind"
PRODUCT	Objects, vehicles, foods, etc. (not services)	"iPhone", "Coca-Cola", "Nike shoes"
CARDINAL	Numerals that do not fall under another type	"Five", "Twenty", "One hundred"
NORP	Nationalities or religious or political groups	"American", "Muslim", "Republican"
MONEY	Monetary values, including unit	"$10", "€50", "¥1000"
LANGUAGE	Any named language	"English", "Spanish", "French"
FAC	Buildings, airports, highways, bridges, etc.	"Eiffel Tower", "White House", "Golden Gate Bridge"
EVENT	Named hurricanes, battles, wars, sports events, etc.	"Olympic Games", "Wedding ceremony", "Concert"
PERCENT	Percentage, including "%"	"50%", "10.5%", "75.2%"
LAW	Named documents made into laws.	"Constitution", "Copyright Act", "Traffic laws"

Results

Random Baseline

Given that the dataset was unbalanced for the number of transcripts per condition (truthful=32 vs. deceptive=30), the random baseline was computed using the Zero Rule (ZeroR), which considers the probability prediction of the majority class and provides an unbiased performance estimation (Kubat & Matwin, 1997). Using the formula (1) reported below, the chance level was computed as the ratio between the number of transcripts in the majority class (truthful=32) and the total number of transcripts (N_{total}=62), and was set to 0.516.

$$1. \quad ZeroR = \frac{Nmajoriry\_class}{Ntotal} = 0.516$$

Experiment 1

When analyzing the performance for transcripts in the two experimental conditions (truth-tellers vs. liars), both naïve judges and forensic experts reached a better performance in detecting truth-tellers than liars (naïve judges: $accuracy_{truth}$= 63.54 ± 29.57, $accuracy_{liars}$=44.72 ± 29.39; forensic experts: $accuracy_{truth}$= 68.51 ± 24.89, $accuracy_{liars}$=51.39 ± 34.36). This was in line with what the literature considers as *truth-bias*, namely, the tendency to presume others as honest and to be more accurate in evaluating honesty (Levine et al., 1999; Levine, 2014; Street & Masip, 2015).

Experiment 2

In Table 3S, the performance of the ML models from Experiment 2 in terms of Accuracy and Area Under the Curve (AUC), Precision, Recall, and F1 score is reported.

TABLE 3S. Performance of ML model trained on Expert and Computerized RM scores (Experiment 2)

ML models	Trained on	Accuracy	AUC	Precision	Recall	F1 score
Logistic regression	Expert RM	53.4 (±19.8)	56.3 (±21.5)	61.7 (±24.5)	53.4 (±19.8)	52.0 (±21.3)
	Computerized RM	40.2 (±18.0)	46.6 (±17.7)	59.7 (±26.3)	40.2 (±18.0)	34.1 (±20.5)
SVM	Expert RM	**57.9 (±17.4)**	**59.1 (±20.3)**	**64.3 (±20.1)**	**57.9 (±17.4)**	**57.8 (±17.6)**
	Computerized RM	48.8 (±22.9)	52.3 (±23.1)	58.8 (±26.1)	48.8 (±22.9)	48.5 (±23.2)
Decision Tree	Expert RM	49.9 (±19.7)	53.4 (±20.9)	61.1 (±25.5)	49.9 (±19.7)	46.7 (±20.6)
	Computerized RM	**57.1 (±15.8)**	**59.7 (±16.6)**	**67.8 (±17.9)**	**57.1 (±15.8)**	**55.5 (±18.7)**
Random Forest	Expert RM	**56.1 (±20.0)**	**58.9 (±18.8)**	**68.5 (±19.1)**	**56.1 (± 20.0)**	**54.0 (±21.6)**
	Computerized RM	52.8 (±20.4)	54.3 (±19.7)	62.5 (±22.4)	52.8 (±20.4)	51.3 (±22.5)

Note. Performance is evaluated in terms of Accuracy, AUC (= Area Under the Curve), Precision, Recall, and F1 Score. In **bold** are reported the models with the best accuracy.

Experiment 3

In Table 4S, the performance of the ML models from Experiment 3 in terms of Accuracy and Area Under the Curve (AUC), Precision, Recall, and F1 score is reported.

TABLE 4S. Performance of ML model trained on Computerized RM and CL scores (Experiment 3)

Dataset	ML models	Accuracy	AUC	Precision	Recall	F1 score
Free speech	Logistic regression	68.3 (±20.6)	70.4 (±20.7)	78.0 (±18.6)	68.3 (±20.6)	68.3 (±20.6)
	SVM	64.1 (±19.7)	63.6 (±21.8)	72.2 (±19.7)	64.1 (±19.7)	64.3 (±19.1)
	Decision Tree	60.3 (±18.5)	62.0 (±18.0)	72.9 (±17.6)	60.3 (±18.5)	59.0 (±18.6)
	Random Forest	**69.4 (±16.5)**	**71.0 (±15.0)**	**78.3 (±15.3)**	**69.4 (±16.5)**	**68.9 (±17.2)**
Unexpected questions	Logistic regression	56.9 (±19.1)	55.6 (±23.2)	63.1 (±20.3)	56.9 (±19.1)	57.6 (±18.5)
	SVM	53.3 (±19.2)	54.6 (±22.5)	64.0 (±22.0)	53.3 (±19.2)	53.6 (±18.4)
	Decision Tree	**60.8 (±14.6)**	**64.1 (±16.6)**	**70.2 (±17.5)**	**60.8 (±14.6)**	**60.1 (±16.1)**
	Random Forest	60.6 (±18.5)	63.3 (±20.9)	70.2 (±20.3)	60.6 (±18.5)	61.4 (±18.0)
Full text	Logistic regression	**67.2 (±19.1)**	**67.1 (±21.9)**	**74.6 (±18.3)**	**67.2 (±19.1)**	**67.4 (±18.4)**
	SVM	62.3 (±17.6)	61.1 (±21.0)	70.9 (±19.5)	62.3 (±17.6)	61.3 (±17.5)
	Decision Tree	57.2 (±16.6)	59.6 (±19.9)	67.4 (±19.6)	57.2 (±16.6)	57.4 (±16.5)
	Random Forest	64.9 (±20.3)	67.4 (±20.5)	75.6 (±18.8)	64.9 (±20.3)	65.0 (±20.0)

Note. Performance is evaluated in terms of Accuracy, AUC (= Area Under the Curve), Precision, Recall, and F1 Score. In bold are reported the models with the best accuracy.

Experiment 4

In Table 5S, the performance of the ML models from Experiment 4 in terms of Accuracy and Area Under the Curve (AUC), Precision, Recall, and F1 score is reported.

TABLE 5S. Performance of ML model trained on features extracted and selected through a data-driven approach (Experiment 4).

Dataset	ML models	Accuracy	AUC	Precision	Recall	F1 score
Free speech	Logistic regression	68.9 (±21.1)	69.3 (±22.2)	77.9 (±18.9)	68.9 (±21.1)	68.8 (±20.6)
	SVM	69.0 (±18.3)	69.9 (±19.7)	78.0 (±16.5)	69.0 (±18.3)	68.6 (±18.5)
	Decision Tree	59.6 (±18.8)	57.9 (±21.4)	66.4 (±20.1)	59.6 (±18.8)	59.4 (±18.5)
	Random Forest	**68.2 (±17.9)**	**68.3 (±19.2)**	**76.6 (±16.9)**	**68.2 (±17.9)**	**68.0 (±17.5)**
Unexpected questions	Logistic regression	60.9 (±18.7)	61.9 (±20.2)	67.8 (±22.2)	60.9 (±18.7)	60.2 (±19.6)
	SVM	**69.4 (±20.5)**	**69.3 (±23.5)**	**75.5 (±21.4)**	**69.4 (±20.5)**	**69.3 (±20.6)**
	Decision Tree	52.1 (±20.8)	51.8 (±23.5)	59.1 (±26.0)	52.1 (±20.8)	50.4 (±21.7)
	Random Forest	64.9 (±19.2)	64.4 (±21.8)	71.0 (±19.5)	64.9 (±19.2)	65.2 (±18.7)
Full text	Logistic regression	73.3 (±18.6)	73.7 (±21.0)	82.2 (±14.4)	73.3 (±18.6)	72.9 (±19.4)
	SVM	**77.3 (±17.2)**	**78.6 (±19.2)**	**83.8 (±16.4)**	**77.3 (±17.2)**	**77.3 (±17.5)**
	Decision Tree	53.5 (±21.7)	53.9 (±23.3)	61.0 (±23.9)	53.5 (±21.7)	54.0 (±21.4)
	Random Forest	75.1 (±17.5)	76.5 (±18.8)	82.6 (±15.9)	75.1 (±17.5)	75.3 (±17.5)

Note. Performance is evaluated in terms of Accuracy, AUC (= Area Under the Curve), Precision, Recall, and F1 Score. In bold are reported the models with the best accuracy.

Comparing accuracy among Experiments

TABLE 6S. Dunn's post-hoc tests comparing pairwise conditions among the four experiments.

group1	group2	n1	n2	*t*	*p*
Naive judges	Forensic experts	120	36	1.10	4.08e-01
Naive judges	ML - Expert RM scores	120	30	0.93	4.76e-01
Naive judges	ML - NLP RM scores	120	30	0.88	4.76e-01
Naive judges	ML - NLP RM+CL	120	30	3.92	6.67e-04***
Naive judges	ML - datadriven	120	30	5.47	6.90e-07***
Forensic experts	ML - Expert RM scores	36	30	-0.08	9.68e-01
Forensic experts	ML - NLP RM scores	36	30	-0.12	9.68e-01
Forensic experts	ML - NLP RM+CL	36	30	2.39	3.38e-02*
Forensic experts	ML - datadriven	36	30	3.67	9.99e-04***
ML - Expert RM scores	ML - NLP RM scores	30	30	-0.04	9.68e-01
ML - Expert RM scores	ML - NLP RM+CL	30	30	2.36	3.38e-02*
ML - Expert RM scores	ML - datadriven	30	30	3.59	9.99e-04***
ML - NLP RM scores	ML - NLP RM+CL	30	30	2.41	3.38e-02*
ML - NLP RM scores	ML - datadriven	30	30	3.63	9.99e-04***
ML - NLP RM+CL	ML - datadriven	30	30	1.22	3.69e-01

Note. The reported *p*-values are adjusted using False Discovery Rate (FDR) correction. *Naive judges* and *Forensic experts'* accuracy is derived from Experiment 1; *ML-Expert RM scores*: best accuracy achieved in Experiment 2 using ML models trained on RM scores provided by Forensic Experts in Experiment 1. *ML-NLP RM scores*: best accuracy achieved in Experiment 2 using ML models trained on RM scores computed with NLP techniques; *ML-NLP RM+CL*: best accuracy achieved in Experiment 3 using ML; *ML-data-driven*: best accuracy achieved in Experiment 4 using ML.

Abbreviations: ML = Machine Learning; RM = Reality Monitoring; CL = Cognitive Load

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Chapter 3

Fine-tuning large language models for verbal deception detection

This chapter is based on: Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using large language models. *Scientific reports*, *13*(1), 22849. https://doi.org/10.1038/s41598-023-50214-0

Abstract

Human accuracy in detecting deception with intuitive judgments has been proven not to go above the chance level. Therefore, several automated verbal lie detection techniques employing machine learning and Transformer models have been developed to reach higher levels of accuracy. This study is the first to explore the performance of a Large Language Model, FLAN-T5 (small and base sizes), in a lie-detection classification task in three English-language datasets encompassing personal opinions, autobiographical memories, and future intentions. After performing stylometric analysis to describe linguistic differences in the three datasets, we tested the small- and base-sized FLAN-T5 in three Scenarios using 10-fold cross-validation: one with train and test set coming from the same single dataset, one with train set coming from two datasets and the test set coming from the third remaining dataset, one with train and test set coming from all the three datasets. We reached state-of-the-art results in Scenarios 1 and 3, outperforming previous benchmarks. The results also revealed that model performance depended on model size, with larger models exhibiting higher performance. Furthermore, stylometric analysis was performed to carry out explainability analysis, finding that linguistic features associated with the Cognitive Load framework may influence the model's predictions.

Keywords: verbal deception detection, large language models, fine-tuning, natural language processing

1. Introduction

Lie detection involves the process of determining the veracity of a given communication. When producing deceptive narratives, lie-tellers employ verbal strategies to create false beliefs in the interacting partners and are thus involved in a specific and temporary psychological and emotional state (Walczyk et al., 2014). For this reason, the Undeutsch hypothesis suggests that deceptive narratives differ in form and content from truthful narratives (Amado et al., 2015). This topic has always been under constant investigation and development in the field of cognitive psychology, given its significant and promising applications in the forensic and legal setting (Vrij et al., 2022). Its potential pivotal role is in determining the honesty of witnesses and potential suspects during investigations and legal proceedings, impacting both the investigative information-gathering process and the final decision-making level (Vrij & Fisher, 2016).

Decades of research have focused on identifying verbal cues for deception and developing effective methods to differentiate between truthful and deceptive narratives, with such verbal cues being, at best, subtle and typically resulting in both naive and expert individuals performing just above chance levels (Bond & DePaulo, 2006; DePaulo et al., 2003). A potential explanation coming from social psychology for this unsatisfactory human performance is the intrinsic human inclination to the truth bias (Levine, 2014; Levine et al., 1999), i.e., the cognitive heuristic of presumption of honesty, which makes people assume that an interaction partner is truthful unless they have reasons to believe otherwise (Levine, 2014; Street & Masip, 2015). However, it is worth mentioning that a more recent study challenged this solid result, finding that instructing participants to rely only on the best available cue, such as the detailedness of the story, enabled them to consistently discriminate lies from the truth with accuracy ranging from 59-79% (Verschuere et al., 2023). This finding moves the debate on 1) the proper number of cues that judges should combine before providing their veracity judgment, with the suggestion that the use-the-best heuristic approach is the most straightforward and accurate, and thus on 2) the diagnosticity level of this cue.

More recently, the issue of verbal lie detection has also been tackled by employing computational techniques, such as stylometry. Stylometry refers to a set of methodologies and tools from computational linguistics and artificial intelligence that allow for conducting quantitative analysis of linguistic features within written texts to uncover distinctive patterns that can infer and characterize authorship or other stylistic attributes (H. Chen, 2011; Chen et al., 2011; Daelemans, 2013). Albeit with some limita-

tions, stylometry has been proven to be effective in the context of lie detection (Hauch et al., 2015; Tomas et al., 2022). The main advantage is the possibility of coding and extracting verbal cues independently from human judgment, hence reducing the problem of inter-coder agreement, as researchers using the same technique for the same data will extract the same indices (Tomas et al., 2022).

Alongside this trend, several recent studies have explored computational analysis of language in different domains, such as fake news (Conroy et al., 2015; Pérez-Rosas et al., 2018), transcriptions of court cases (Fornaciari & Poesio, 2012, 2013; Pérez-Rosas et al., 2015), evaluations of deceptive product reviews (Fornaciari & Poesio, 2014; Kleinberg, Mozes, et al., 2018; Ott et al., 2011), investigations into cyber-crimes (Mbaziira & Jones, 2017), analysis of autobiographical information (Levitan et al., 2018), and assessments of deceptive intentions regarding future events (Kleinberg et al., 2017). Taken together, most of those studies focused on the usage of machine learning and deep learning algorithms combined with Natural Language Processing (NLP) techniques to detect deception from verbal cues automatically (see Constâncio et al., 2023, for a systematic review of the computerized techniques employed in lie-detection studies).

More recently, a great step in advance has been made in the field of AI and NLP with the advent of Large Language Models (LLMs). LLMs are Transformer-based language models with hundreds of millions of parameters trained on a large collection of corpora (i.e., pre-training phase, Zhao et al., 2023). Thanks to this pre-training phase, LLMs have proven to capture the intricate patterns and structures of language and develop a robust understanding of syntax, semantics, and pragmatics, being able to generate coherent text resembling human natural language. In addition, once pre-trained, these models can be fine-tuned on specific tasks using smaller task-specific datasets. Fine-tuning refers to the process of continuing the training of a pre-trained model on a new dataset, allowing it to adapt its previously learned knowledge to the nuances and specificities of the new data, thereby achieving state-of-the-art results (Zhao et al., 2023). Common tasks for LLMs fine-tuning include NLP tasks, such as language translation, text classification (e.g., sentiment analysis), question-answering, text summarization, and code generation. Therefore, LLMs excel at a wide range of NLP tasks, as opposed to models uniquely trained for one specific task (Zhao et al., 2023). However, to the best of our knowledge, despite the extreme flexibility of LLMs, the procedure of fine-tuning an LLM on small corpora for a lie-detection task has remained unexplored.

1.1 Related works in the Psychology field

Among previous psychological frameworks aimed at identifying reliable cues of verbal deception, the Distancing framework, the Cognitive Load (CL) theory, the Reality Monitoring (RM) framework, and the Verifiability Approach (VA) have been extensively studied, gaining empirical support for their efficacy not only from primary research but also from meta-analytic studies.

The Distancing framework (Newman et al., 2003; Vrij et al., 2022) of deception states that liars tend to distance themselves from their narratives as a mechanism to handle the negative emotions experienced while lying by using fewer self-references (e.g., "I," "me") and employing more other-references (e.g., "he," "they").

The CL framework states that liars consume more cognitive resources while fabricating their fake responses, checking their congruency with other fabricated information, and maintaining credibility and consistency in front of the examiner (Monaro et al., 2018), resulting in shorter, less elaborate, and less complex statements. A meta-analysis found that approaches based on CL theories produce higher accuracy rates in detecting deception than standard approaches (Vrij et al., 2015).

The RM framework bases its assumptions on the memory characteristics literature, hypothesizing that truthful recollections are based on experienced events, while deceptive recollections are based on imagined events (Johnson & Raye, 1981). Therefore, RM derives its predictions about truthful narratives from sensory, spatial, and temporal information and from emotions and feelings experienced during the event. On the contrary, predictions about deception are drawn from the number of cognitive operations (e.g., thoughts and reasonings; Masip et al., 2005; Sporer, 1997, 2004). The total RM scores appear to be diagnostic (d=0.55) in the detection accuracy of truthfulness (Amado et al., 2016; Gancedo et al., 2021; see also Vrij et al., 2022, for an extensive review of verbal lie-detection methods). More recently, the RM framework was investigated through concreteness in language (Kleinberg et al., 2019). In this study, one underlying and partially supported assumption was the truthful concreteness hypothesis, which suggests that truthful statements usually consist of concrete, specific, and contextually relevant details. In contrast, deceptive or false statements often include more abstract and less specific information, being more associated with the RM criterion of cognitive operations.

The VA in verbal lie detection suggests that truthful statements are more likely to be verifiable than false or deceptive statements, as liars avoid mentioning details that could be verified with independent evidence to

conceal their deception (Nahari et al., 2012; Vrij & Nahari, 2019). Verifiable details may be represented by activities involving or witnessed by identified individuals, documented through video or photographic evidence, or leaving digital or physical traces (e.g., phone calls or receipts; Nahari et al., 2012; Vrij & Nahari, 2019).

Notably, these frameworks offer detectable linguistic cues that can be readily identified using NLP techniques and have been extensively studied in this sense. Hauch et al. (2015) conducted a meta-analysis of studies on computer-based lie detection, with most of the included studies relying on the Linguistic Inquiry and Word Count software (LIWC; Boyd et al., 2022; Pennebaker et al., 2015). LIWC is the gold standard tool for studying lexical diversity and text semantic content. Given a text, LIWC calculates the percentage of total words corresponding to more than 100 categories in the dictionary related to different psychosocial dimensions, which have been validated by human evaluators using rigorous procedures. Among Houch's meta-analysis findings, LIWC metrics reflecting Distancing, CL, and RM frameworks of deception found support from the results and can detect verbal deception through computerized techniques. Usually, for distancing metrics, researchers compute the number of self and other-references by summing the frequency of first-person pronouns in contrast with second and third-person pronouns (Newman et al., 2003; Vrij et al., 2022). When employing CL theory in texts, researchers usually employ and analyze statistics about the number of words and sentences, the readability, and the complexity of texts (Chen, 2011; Chen et al., 2011; Hauch et al., 2015). RM is often investigated with LIWC (Bond et al., 2017; Bond & Lee, 2005; Kleinberg et al., 2017). Schutte et al. (2021) provided evidence that human coding of perceptual and contextual details in discriminating lies from truths is not conclusively superior, thereby highlighting the potential advantages of automated techniques. Additionally, recent studies extracted verifiable details by using named-entity recognition (NER), proving to be an effective automated procedure for the detection of deception in hotel reviews (Kleinberg, Mozes, et al., 2018) as well as in participants' intentions on their weekend plans (Kleinberg, van der Toolen, et al., 2018).

The promising results in applying NLP techniques for psychological research suggest the possibility of combining metrics from different psychological frameworks in a new theory-based stylometric analysis, offering the possibility to investigate verbal lie detection from multiple perspectives in one shot.

1.2 Related works in the AI field

Previous works from the AI field have applied machine learning and deep learning models in a binary classification task for data-driven verbal deception detection. One study developed a database of future intentions to investigate whether combining machine and human judgments may improve accuracy in predicting deception (Kleinberg & Verschuere, 2021). While finding that human judgment impairs automated deception detection accuracy, the authors implemented two machine learning models (i.e., vanilla random forest) trained respectively on LIWC and Part-of-Speech features (e.g., frequency of names, adjectives, adverbs, verbs) reaching an accuracy of 69% (95% CI: 63% - 74%) and 64 (95% CI: 58%, 69%), respectively. On the same dataset, another work evaluated six deep-learning models, including combinations of BERT (and RoBERTa), MultiHead Attention, co-attentions, and Transformers models, achieving up to 70.61% (± 2.58%) using a BERT with co-attention model (Ilias et al., 2022). The authors also provided an explainaibility analysis to understand how the models reached their decisions using a combination of LIME (a tool used to explain deep learning predictions in more straightforward and understandable terms by showing which specific words of the text influenced the outcome) and LIWC.

Capuozzo et al. (2020) developed a new cross-domain and cross-language dataset of opinions, asking English-speaking and Italian-speaking participants to provide truthful or deceptive opinions on five different topics. After encoding the texts with FastText word-embedding, they trained Transformers models in multiple scenarios using 10-fold cross-validation, with averaged accuracy ranging from 63% (± 8.7%) in the "within-topic" scenario to a high of 90.1% (± 0.16%) in the "author-based" scenario.

In contrast, Sap et al. (2022) developed a new dataset of narratives generated from memories and imagination and used an LLM (GPT-3) to compute a new metric called "sequentiality". Sequentiality is a metric of narrative flow that compares the probability of a sentence with and without its preceding story context. While providing insights into the cognitive processes of storytelling with an innovative computational approach, the authors did not employ a fine-tuning procedure for an LLM to classify different narratives.

The findings in the AI domain indicate that as the model's complexity increases, there is a heightened accuracy in predicting deception from texts. However, this increase in accuracy often comes at the expense of explainability for these predictions. LLMs are currently among the most cutting-edge models capable of handling vast amounts and complexities

of linguistic data, and the lack of literature on fine-tuning LLMs for lie-detection tasks provides worthwhile reasons to investigate this area.

1.3 Aims and hypotheses of the study

The main objectives and hypotheses of this study are outlined as follows:

> Hypothesis 1a): Fine-tuning an LLM can effectively classify the veracity of short narratives from raw texts, 1b) outperforming classical machine learning and deep learning approaches in verbal lie detection.
>
> Hypothesis 2): Fine-tuning an LLM on deceptive narratives enables the model to also detect new types of deception;
>
> Hypothesis 3): Fine-tuning an LLM on a multiple-context dataset enables the model to obtain successful predictions on a multi-context test set;
>
> Hypothesis 4): Model performance depends on model size, with larger models showing higher accuracy;
>
> Hypothesis 5a): The linguistic style distinguishing truthful from deceptive statements varies across different contexts, 5b) and can be a significant feature for model prediction.

To test Hypothesis 1a, we fine-tuned an open-source LLM, FLAN-T5, using three datasets: personal opinions (the Deceptive Opinions dataset, Capuozzo et al., 2020), autobiographical experiences (the Hippocorpus dataset, Sap et al., 2022) and future intentions (the Intention dataset, Kleinberg & Verschuere, 2021). Given the extreme flexibility of LLMs, this approach is hypothesized to detect deception from raw texts above the chance level. To test the advantage of our approach compared to classical machine and deep learning models (Hypothesis 1b), we decided to compare the results with two benchmarks, further described in the Methods and Materials section.

With regard to Hypotheses 2 and 3, according to empirical evidence, classical machine learning models tend to experience a decline in performance when trained and tested on the aforementioned scenarios (Hernández-Castañeda et al., 2016; Mihalcea & Strapparava, 2009; Pérez-Rosas & Mihalcea, 2014). In contrast, LLMs have acquired a comprehensive understanding of language patterns during the pre-training phase. We posit that a fine-tuned LLM is capable of generalizing its learning across various contexts. Related to Hypothesis 4, we believe this generalization ability is further enhanced in larger models, as their size is associated with a more sophisticated representation of language.

Finally, to test Hypothesis 5, we introduced a new theory-based stylometric approach, named **D**e**CL**a**R**ati**VE** stylometry, to extract linguistic features related to the psychological frameworks of Distancing (Newman et al., 2003), Cognitive Load (Vrij et al., 2015), Reality Monitoring (Johnson & Raye, 1981), and Verifiability Approach (Nahari et al., 2012; Vrij & Nahari, 2019), providing a pragmatic set of criteria to extract features from utterances. We will apply **D**e**CL**a**R**ati**VE** stylometry to compare truthful and deceptive statements in the three aforementioned datasets in order to explore potential differences in terms of linguistic style. Our hypothesis suggests that the linguistic style distinguishing truthful from deceptive statements may vary across the three datasets, as these types of statements originate from distinct contexts. We also applied the **D**e**CL**a**R**ati**VE** stylometry technique to provide explainability analysis of the top-performing model.

2. Methods and Materials

2.1 Datasets

Three datasets were employed for this study: the Deceptive Opinions dataset (Capuozzo et al., 2020), from now on Opinion Dataset, the Hippocorpus dataset (Sap et al., 2022), from now on Memory Dataset, and the Intention dataset (Kleinberg & Verschuere, 2021). For each dataset, participants were required to provide genuine or fabricated statements in three different domains: personal opinions on five different topics (Opinion dataset), autobiographical experiences (Memory dataset), and future intentions (Intention Dataset). Notably, the specific topic within each domain was counterbalanced among liars and truth-tellers. A more detailed description of each dataset is available in the SM as well as in the method section of each original article.

Table 1 displays an example of truthful and deceptive statements about opinions, memories, and intentions. Table 2 reports descriptive statistics for each dataset, both overall and when grouped by truthful and deceptive sets of statements. These statistics include the minimum, maximum, average, and standard deviation of word counts. Word counts were computed after text tokenization using spaCy, a Python library for text processing. Additionally, Table 2 provides Jaccard similarity index values between truthful and deceptive vocabulary sets. Jaccard's index was derived by calculating the intersection (common words) and union (total words) of these two sets (Ilias et al., 2022; Ríssola et al., 2020). The resulting index ranges from 0, indicating a completely different vocabulary between the two sets, to 1, indicating a completely identical vocabulary

between the two sets. We reported the Jaccard similarity index to provide a measure of similarity or overlap between the word choices of truthful and deceptive statements within the respective datasets. In SM, we offer a detailed methodology for calculating the Jaccard similarity index.

TABLE 1. Truthful and deceptive example statements about opinions, memories, and intentions. In brackets, the topic assigned to the participant in the deceptive condition to fabricate the narrative.

	TRUTHFUL	DECEPTIVE
OPINION (Abortion)	While I am morally torn on the issue, I believe that ultimately it is a woman's body and she should be able to do with it as she pleases. I belive people should not dehumanize the fetus tough, to make themselves feel better. The decision about laws regarding this issue should be left up to the states to decide. To combat this problem, birth control should be easily accessible.	Abortion is the termination of a life and should not be allowed. If a fetus has made it to the point of being able to survive "on its own" outside its mother's body, what right do we have to cut its life short. If the mother's life is in danger, she already chose that she was willing to sacrifice her life to have a child when she consented to procreating.
MEMORY (My boyfriend and I went to a concert together and had a great time. We met some of my friends there and really enjoyed ourselves watching the sunset.)	The day started perfectly, with a great drive up to Denver for the show. Me and my boyfriend didn't hit any traffic on the way to Red Rocks, and the weather was beautiful. We met up with my friends at the show, near the top of the theater, and laid down a blanket. The opener came on, and we danced our butts off to the banjoes and mandolins that were playing on-stage. We were so happy to be there. That's when the sunset started. It was so beautiful. The sky was a pastel pink and was beautiful to watch. That's when Phil Lesh came on, and I just about died. It was the happiest moment of my life, seeing him after almost a decade of not seeing him. I was so happy to be there, with my friends and my love. There was nothing that could top that night. We drove home to a sky full of stars and stopped at an overlook to look up at them. I love this	Concerts are my most favorite thing, and my boyfriend knew it. That's why, for our anniversary, he got me tickets to see my favorite artist. Not only that, but the tickets were for an outdoor show, which I love much more than being in a crowded stadium. Since he knew I was such a big fan of music, he got tickets for himself, and even a couple of my friends. He is so incredibly nice and considerate to me and what I like to do. I will always remember this event and I will always cherish him. On the day of the concert, I got ready, and he picked me up and we went out to a restaurant beforehand. He is so incredibly romantic. He knew exactly where to take me without asking. We ate, laughed, and had a wonderful dinner date before the big event. We arrived at the concert and the music was so incredibly beautiful. I loved

	TRUTHFUL	DECEPTIVE
	place I live. And I love live music. I was so happy.	every minute of it. My friends, boyfriend, and I all sat down next to each other. As the music was slowly dying down, I found us all getting lost just staring at the stars. It was such an incredibly unforgettable and beautiful night.
INTENTION **INTENTION** (Going swimming with my daughter)	We go to a Waterbabies class every week, where my 16-month-old is learning to swim. We do lots of activities in the water, such as learning to blow bubbles, using floats to aid swimming, splashing and learning how to save themselves should they ever fall in. I find this activity important as I enjoy spending time with my daughter and swimming is an important life skill.	I will be taking my 8-year-old daughter swimming this Saturday. We'll be going early in the morning, as it's generally a lot quieter at that time, and my daughter is always up early watching cartoons anyway (5 am!). I'm trying to teach her how to swim in the deep end before she starts her new school in September as they have swimming lessons there twice a week.

TABLE 2. Summary statistics of the number of words and Jaccard similarity for each dataset and subset of truthful and deceptive statements.

Dataset (total number)	Min-Max number of words	Average number of words (*SD*)	Jaccard Similarity Index (qualitative interpretation)
All Opinions (2500)	6 - 338	59.05 (30.66)	0.35
Truthful Opinions (1250)	7 - 338	66.74 (31.95)	(low similarity)
Deceptive Opinions (1250)	6 - 232	51.36 (27.24)	
All Intentions (1640)	15 - 251	50.44 (30.11)	0.34
Truthful Intentions (783)	15 - 206	47.04 (28.36)	(low similarity)
Deceptive Intentions (857)	15 - 251	53.55 (31.31)	
All Memories (5506)	22 - 625	255.24 (92.36)	0.34
Truthful Memories (2770)	22 - 625	269.78 (94.14)	(low similarity)
Deceptive Memories (2736)	22 - 609	240.51 (88.12)	

Note. Jaccard Similarity Index (with qualitative interpretation in brackets) refers to the similarity between truthful and deceptive vocabulary sets for each dataset.
Abbreviation: *SD* = standard deviation.

2.2 FLAN-T5

We adopted FLAN-T5, an LLM developed by Google researchers and freely available through HuggingFace Python's library Transformers (https://huggingface.co/docs/transformers/model_doc/flan-t5). HugginFace is a company that provides free access to state-of-the-art LLMs through Python API. Among the available LLMs, we chose FLAN-T5 because of its valuable trade-off between computational load and goodness of the learned representation. FLAN-T5 is the improved version of MT-5, a text-to-text general model capable of solving many NLP tasks (e.g., sentiment analysis, question answering, and machine translation), which has been improved by pre-training (Chung et al., 2022). The peculiarity of this model is that every task they were trained on is transformed into a text-to-text task. For example, while performing sentiment analysis, the output prediction is the string used in the training set to label the positive or negative sentiment of each phrase rather than a binary integer output (e.g., 0=positive; 1=negative). Hence, their power stands in both the generalized representation of natural language learned during the pre-training phase and the possibility of easily adapting the model to a downstream task with little fine-tuning, without adjusting its architecture.

2.3 DeCLaRatiVE stylometric analysis

This study employed stylometric analysis to achieve two primary objectives. First, we aimed to describe the linguistic features that distinguished the three datasets before initializing the fine-tuning process. Second, we conducted explainability analysis to gain insights into the role of linguistic style that differentiated truthful and deceptive statements in the model's classification process. For this purpose, a new framework that we referred to as **DeCLaRatiVE** stylometry was adopted, which involved the extraction of 26 linguistic features in conjunction with the psychological frameworks of **D**istancing (Newman et al., 2003), **C**ognitive **L**oad (Vrij et al., 2015), **R**eality monitoring (Johnson & Raye, 1981), and **V**erifiability **A**pproach (Nahari et al., 2012; Vrij & Nahari, 2019). A full list of the 26 linguistic features with a short description is shown in Table 3. This comprehensive approach enabled the analysis of verbal cues of deception from a multidimensional perspective.

Features associated with the CL framework consisted of statistics about the length, readability, and complexity of the text (Hauch et al., 2015; Sarzynska-Wawer et al., 2023; Solà-Sales et al., 2023; Zhou et al., 2004) and were extracted using the Python library TEXTSTAT.

TABLE 3. List and short description of the 26 linguistic features pertaining to the **D**e-**CLa**ra**ti**V**E** Stylometry technique.

Label	Description
num_sentences	Total number of sentences
num_words	Total number of words
num_syllables	Total number of syllables
avg_syllabes_per_word	Average number of syllables per word
fk_grade	Index of the grade level required to understand the text
fk_read	Index of the readability of the text
Analytic	LIWC summary statistic analyzing the style of the text in term of analytical thinking (0 - 100)
Authentic	LIWC summary statistic analyzing the style of the text in term of authenticity (0 - 100)
Tone	Standardized difference (0-100) of 'tone_pos' - 'tone_neg'
tone_pos	Percentage of words related to a positive sentiment (LIWC dictionary)
tone_neg	Percentage of words related to a negative sentiment (LIWC dictionary)
Cognition	Percentage of words related to semantic domains of cognitive processes (LIWC dictionary)
memory	Percentage of words related to semantic domains of memory/forgetting (LIWC dictionary)
focuspast	Percentage of verbs and adverbs related to the past (LIWC dictionary)
focuspresent	Percentage of verbs and adverbs related to the present (LIWC dictionary)
focusfuture	Percentage of verbs and adverbs related to the future (LIWC dictionary)
Self-reference	Sum of LIWC categories 'i' + 'we'
Other-reference	Sum of LIWC categories 'shehe' + 'they' + 'you'
Perceptual details	Sum of LIWC categories 'attention' + 'visual' + 'auditory'+ 'feeling'
Contextual Embedding	Sum of LIWC categories 'space' + 'motion' + 'time'
Reality Monitoring	Sum of Perceptual details + Contextual Embedding + Affect - Cognition
Concreteness score	Mean of concreteness score of words
People	Unique named-entities related to people: e.g., 'Mary', 'Paul', 'Adam'
Temporal details	Unique named-entities related to time: e.g., 'Monday', '2:30 PM', 'Christmas'
Spatial details	Unique named-entities related to space: e.g., 'airport', 'Tokyo', 'Central park'
Quantity details	Unique named-entities related to quantities: e.g., '20%', '5 $', 'first', 'ten', '100 meters'

Features related to the Distancing and RM framework were computed using LIWC (Boyd et al., 2022; Pennebaker et al., 2015), the gold standard software for analyzing word usage. Using the English dictionary, we scored each text along with all the categories present in LIWC-22. LIWC scoring was computed on tokenized text using the English dictionary. The selection of the LIWC categories related to the Distancing and RM framework was guided by previous research on computerized verbal lie-detection (Ilias et al., 2022; Kleinberg & Verschuere, 2021; Newman et al., 2003; Ríssola et al., 2020; Sap et al., 2022) and a previous meta-analysis (Hauch et al., 2015). RM was also investigated through the linguistic concreteness of words (Kleinberg et al., 2019). To determine the average level of concreteness for each statement, we utilized the concreteness annotation dataset developed by Brysbaert et al. (2014). For the calculation of concreteness scores, a preprocessing pipeline was applied to textual data using the Python library SpaCy: text was converted to lowercase and tokenized; then stop words were removed, and the remaining content words were lemmatized. These content words were then cross-referenced with the annotated concreteness dataset to assign the respective concreteness value when a match was found. The concreteness score for each statement was then computed as the average of the concreteness scores for all the content words in that statement. For what concerns verifiable details, they were estimated by the frequency of named entities. Named entities were extracted with the NER technique using Python's library SpaCy through the Transformer algorithm for the English language (en_core_web_trf, https://spacy.io/models/en#en_core_web_trf). Further details on how the 26 linguistic features were computed are provided in the Supplementary Materials (SM).

2.4 Experimental set-up

In this section, we describe the methodology that we applied in this work. As a first step, we wanted to perform a descriptive linguistic analysis of our datasets, trying to provide a response to Hypothesis 5a), i.e., whether the linguistic style distinguishing truthful from deceptive statements varies across different contexts. To achieve this result, we employed the **DeCLaRatiVE** stylometric analysis. As a second step, we proceeded to test the capacity of the FLAN-T5 model to be fine-tuned on a Lie Detection task. To do so, we provided three scenarios to verify the following hypothesis:

> Hypothesis 1a): Fine-tuning an LLM can effectively classify the veracity of short narratives from raw texts, 1b) outperforming classical machine learning and deep learning approaches in verbal lie detection.

Hypothesis 2): Fine-tuning an LLM on deceptive narratives enables the model to also detect new types of deception;

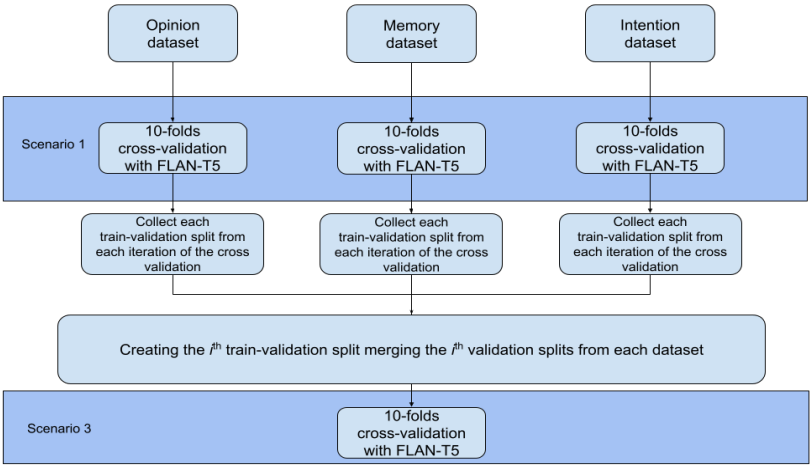
Hypothesis 3): Fine-tuning an LLM on a multiple-context dataset enables the model to obtain successful predictions on a multi-context test set;

Hypothesis 4): Model performance depends on model size, with larger models showing higher accuracy;

We expected hypotheses 1a,1b, 3, and 4 to be verified, while we did not have any a priori expectation for the second hypothesis. The scenarios are described below:

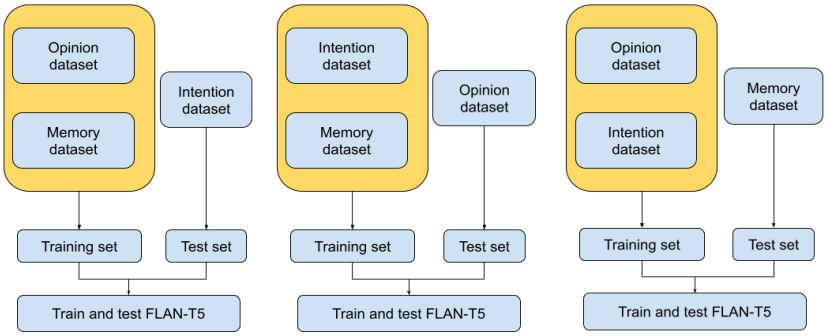
Scenario 1: The model was fine-tuned and tested on a single dataset. This procedure was repeated for each dataset with a different copy of the same model each time (i.e., the same parameters before the fine-tuning process) (Fig. 1). This Scenario assesses the model's capacity to learn how to detect lies related to the same context and responds to Hypothesis 1a;

FIGURE 1. Visual illustration of Scenarios 1 and 3.



Scenario 2: The model was fine-tuned on two out of the three datasets and tested on the remaining unseen dataset. As for the previous Scenario, this procedure was iterated three times, employing separate instances of the same model, each time with a distinct combination of pairings (Fig. 2). This Scenario assesses how the model performs on samples from a new context to which it has never been exposed during the training phase and provides a response for Hypothesis 2;

FIGURE 2. Visual illustration of Scenario 2.



Scenario 3: We first aggregated the three train and test sets from Scenario 1. Then, we fine-tuned the model on the aggregated datasets and tested the model on the aggregated test sets (Fig. 1). This Scenario assesses the capacity of the model to learn and generalize from samples of truthful and deceptive narratives from multiple contexts and provides a response for Hypothesis 3.

In Scenarios 1 and 3, each experiment underwent a 10-fold cross-validation. N-fold cross-validation is a statistical method used to estimate the performance of a model by dividing the dataset into n partitions (n=10 for this study). For each partition i, we created a training set composed of the remaining n-1 partitions using the i partition as a test set (i.e., 90% of the data belongs to the training set, and 10% of the remaining data belongs to the test set). For each iteration, performance metrics are computed on the test set, stored, and then averaged. This procedure ensures an unbiased performance estimation and allows a fair comparison between different models. For our study, we employed identical train-test splits within scenarios 1 and 3 and for both model sizes to guarantee a fair performance comparison. The average test accuracy from each fold and its corresponding standard deviation are presented as performance metrics. Conversely, in Scenario 2, each pairing combination underwent fine-tuning using the entire two paired datasets as a training set, while the model's performance was assessed using the complete unseen dataset as a test set.

Notably, the Opinion dataset was developed to have each participant's truthful and deceptive statements for a total of five opinions. Therefore, we treated each opinion as a separate sample. In order to avoid the

model exhibiting inflated performance on the test set as a result of learning the participants' linguistic style, we adopted the following precautionary measure. Specifically, we ensured an exclusive division of participants between the training and test sets, such that any individual whose opinions were assigned to the training set did not have their opinions assigned to the test set, and vice versa.

Together, Scenarios 2 and 3 provide evidence about the generalized capabilities of the fine-tuned FLAN-T5 model in a lie-detection task when tested on unseen data and on a multi-domain dataset. Furthermore, we tested whether model performance may depend on model sizes. Therefore, we first fine-tuned the small-sized version of FLAN-T5 in every scenario, and then we repeated the same experiments in every scenario with the base-sized version, providing a response for Hypothesis 4.

To test Hypothesis 1b, i.e., to test the advantage of our approach when compared to classical machine learning models, we decided to compare the results with two benchmarks:

i) A basic approach consisting of a bag-of-words (BoW) encoder plus a logistic regression classifier (Lin et al., 2023), following the experimental procedure of Scenario 1;
ii) A literature baseline based on previous studies providing accuracy metrics on the same datasets using a machine learning or a deep learning approach (Capuozzo et al., 2020; Ilias et al., 2022; Kleinberg & Verschuere, 2021). For the Opinion dataset, characterized by opinions on five different topics per subject, we compared our results to the performance obtained in (Capuozzo et al., 2020) with respect to their "within-topic" experiments because our approach is equivalent to theirs, with the only difference that we addressed all the topics in one model.

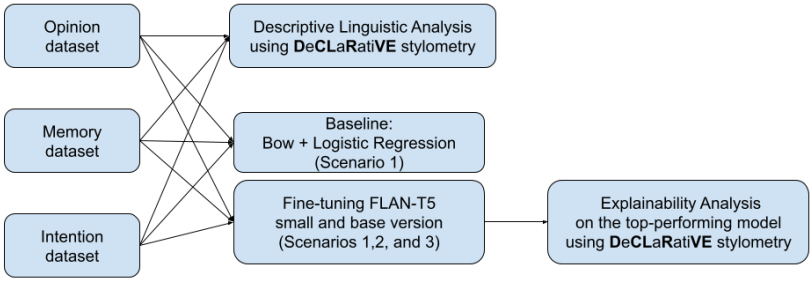
As a final step, we conducted an explainability analysis to investigate the differences in linguistic style between the truthful and deceptive statements that were correctly classified and misclassified by the model. This procedure aimed to provide a response to Hypothesis 5b, i.e., whether the model takes into account the linguistic style of statements for its final predictions. To achieve this result, we employed the **DeCLaRatiVE** stylometric analysis. In Figure 3, we provided a flow chart of the whole experimental setup.

2.5 Fine-tuning strategy

Fine-tuning of LLMs consists of adapting a pre-trained language model to a specific task by further training the model on task-specific data,

thereby enhancing its ability to generate contextually relevant and coherent text in line with the desired task objectives (Chung et al., 2022).

FIGURE 3. Visual illustration of the whole experimental setup.



Note. The Opinion, Memory, and Intention dataset underwent Descriptive Linguistic Analysis using **DeCLaRatiVE** stylometry. A baseline model consisting of Bag of Words (BoW) and Logistic Regression (Scenario 1) was also established for the three datasets. Then, the FLAN-T5 model in small and base versions was fine-tuned across Scenarios 1, 2, and 3. Finally, an Explainability Analysis was conducted on the top-performing model using **DeCLaRatiVE** stylometry to interpret the results

We fine-tuned FLAN-T5 in its small and base sizes using the three datasets and following the experimental setup described above. We approached the lie-detection task as a binary classification problem, given that the three datasets comprised raw texts associated with a binary label, specifically instances classified as truthful or deceptive.

To the best of our knowledge, no fine-tuning strategy is available in the literature for this novel downstream NLP task. Therefore, our strategy followed an adaptation of Hugginface's guidelines on fine-tuning an LLM for translation. Specifically, we chose the same optimization strategy used to pre-train the original model and the same loss function.

Notably, the classification task between deceptive and truthful statements has never been performed during the FLAN-T5 pre-training phase, nor is it included in any of the tasks the model has been pre-trained on. Therefore, we performed the same experiments, described in the Experimental set-up section, multiple times with different learning rate values (i.e., 1e-3, 1e-4, 1e-5), and we finally chose the configuration shown in Table 4, which yielded the best performance in terms of accuracy. All experiments and runs of the three scenarios were conducted on Google Colaboratory Pro+ using their NVIDIA A100 Tensor Core GPU.

TABLE 4. FLAN-T5 hyperparameters configuration for the small- and base-sized versions.

Model	Hyperparameter	Value
FLAN-T5 small	Learning rate	5e-4
	Weight decay coefficient	0.01
	Batch size	2
	Number of Epochs	3
FLAN-T5 base	Learning rate	5e-5
	Weight decay coefficient	0.01
	Batch size	2
	Number of Epochs	3

Note. The initial learning rate for every scenario was 5e-4 for the small model and 5e-5 for the base model. This choice was motivated by preliminary experiment results, with the smaller model, but not the base model, generally performing better with higher learning rates. The weight decay coefficient was set to 0.01 in all models and Scenarios. The batch size was set to 2 for computational reasons, specifically to avoid running out of available memory, even though it is known that a larger batch size usually leads to better performance. Finally, the number of epochs was set to 3 after preliminary experiments showing the maximum test accuracy after the third epoch without overfitting.

2.6 Statistical Procedure for Descriptive Linguistic Analysis

After applying the **DeCLaRatiVE** stylometry technique, we obtained a stylistic vector of 26 linguistic features for each text of the three datasets.

In order to assess the significance of the observed differences between the groups, a permutation t-test was employed (Moore, 1999). This non-parametric method involves pooling all observations and then randomly redistributing them into two groups, preserving the original group sizes. The test statistic of interest (i.e., the difference in means) is then computed for these permuted groups. By repeating this process thousands of times (i.e., n=10,000), we generated a test statistic distribution under the null hypothesis of no difference between the groups. The observed test statistic from the actual data was then compared to this distribution to compute a p-value, indicating the likelihood of observing such a difference if the null hypothesis was true. The advantage of using a permutation t-test is that no assumption about the distribution of data is needed. This analysis was conducted in Python using SciPy and the Pingouin library.

For the Memory and Intention dataset, we computed a permutation t-test (n=10,000) for independent samples for the 26 linguistic features to outline significant differences between the truthful and deceptive texts.

For the Opinion dataset, our analysis proceeded as follows. Firstly, we computed the **DeCLaRatiVE** stylometry technique for all the subjects' opinions. This resulted in a 2500 (opinions) x 26 (linguistic features) matrix. Then, since each subject provided five opinions (half truthful and half deceptive), we averaged the stylistic vector separately for the truthful and deceptive sets of opinions. This procedure allowed us to obtain two different averaged stylistic vectors for the same subject, one for the truthful opinions and one for the deceptive opinions. Importantly, this averaging process enabled us to obtain results that are independent of the topic (e.g., abortion or cannabis legalization) and the stance taken by the subject (e.g., in favor of or against that particular topic). Finally, we validated the statistical significance of these differences by conducting a paired sample permutation test (n=10,000). Results for each dataset were corrected for multiple comparisons with Holm-Bonferroni correction.

The effect size was expressed by Common Language Effect Size (CLES) with a confidence interval of 95 % (95 % CI), which is a measure of effect size that is meant to be more intuitive in its understanding by providing the probability that a specific linguistic feature, in a picked-at-random truthful statement, will have a higher score than in a picked-at-random deceptive one (McGraw & Wong, 1992). The null value for the CLES is the chance level at 0.5 (in a probability range from 0 to 1) and indicates that, when sampled, one group will be greater than the other, with equal chance. Cohen's d effect size with 95 % CI was also computed to add interpretation.

2.7 Statistical Procedure for Explainability Analysis

To examine whether the linguistic style of the input statements exerted an influence on the resulting output of the model and to provide explanations for the wrong classification outputs, we applied a **DeCLaRatiVE** stylometric analysis of statements correctly classified and misclassified by the top-performing model identified in Scenario 3 (Flan-T5 base)

To this aim, during each iteration of cross-validation, we paired the sentences belonging to the test set and their actual labels with the labels predicted by the model. After the cross-validation ended, for each of the ten folds and for each of the 26 linguistic features of the sentences that composed the test set for that fold, we performed a non-parametric permutation t-test for independent samples (n=10,000) for the following comparison of interest:

a) truthful statements misclassified as deceptive (False Negatives), with deceptive statements misclassified as truthful (False Positives);
b) statements correctly classified as deceptive (True Negatives) vs. truthful statements misclassified as deceptive (False Negatives);
c) statements correctly classified as truthful (True Positives) vs. deceptive statements misclassified as truthful (False Positives).
d) truthful vs. deceptive statements correctly classified by the model (True Positives vs. True Negatives).

To compute the effect size, we computed the average of the CLES and Cohen's d effect size scores with their respective 95 % CI obtained from each fold.

2.8 Data and code availability.

For the Opinion dataset, we obtained full access after contacting the corresponding author. The Memory dataset is downloadable at the link: https://www.microsoft.com/en-us/download/details.aspx?id=105291. The Intention dataset is publicly available at the link: https://osf.io/45z7e/.

All the Colab Notebooks to perform linguistic analysis on the three datasets, fine-tune the model in the three Scenarios, and conduct explainability analysis are available at https://github.com/robecoder/VerbalLieDetectionWithLLM.git.

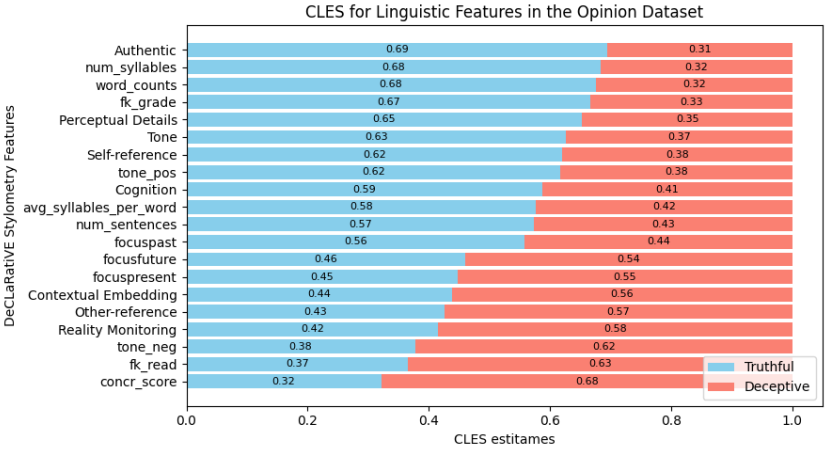
3. Results

3.1 Descriptive Linguistic Analysis

This section outlines the results of the descriptive linguistic analysis in terms of **DeCLaRatiVE** stylometric analysis to compare the three datasets on linguistic features.

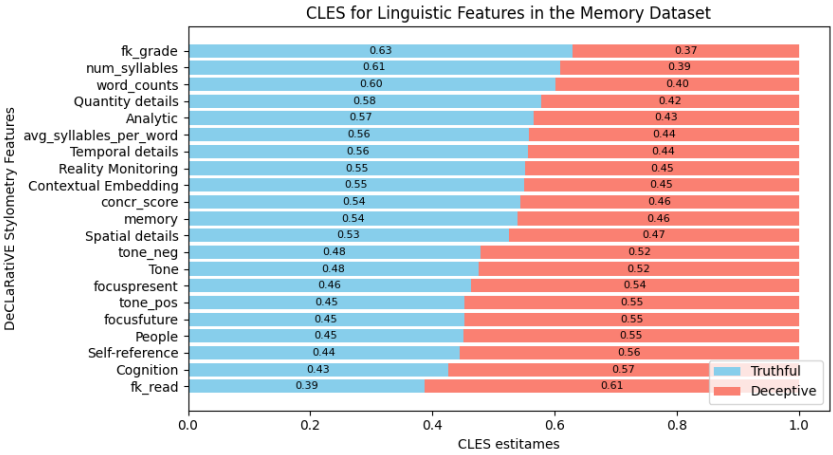
For the three datasets, Figures 4, 5, and 6 show the differences in the number, the type, the magnitude of the CLES effect size, and the direction of the effect for the linguistic features that survived post-hoc corrections. To make an example of these differences, the concreteness score of words ('concr_score') presented the largest CLES within the Intention dataset towards the truthful statements (Fig. 6), while in the Opinion dataset, it showed the largest CLES towards the deceptive statements (Fig. 4).

FIGURE 4. Horizontal stacked bar chart presenting the Common Language Effect Size (CLES) estimates for the significant linguistic features that survived post-hoc corrections in the Opinion dataset.



Note. The CLES estimates represent the probability (ranging from 0 to 1) of finding a specific linguistic feature in truthful opinions (sky blue) than in deceptive ones (salmon). The CLES for truthful opinions are sorted in descending order, while the CLES for deceptive opinions are sorted in ascending order.

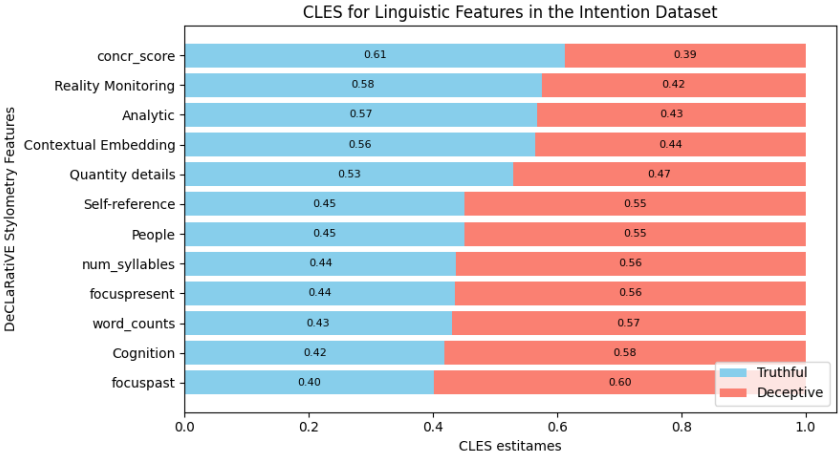
FIGURE 5. Horizontal stacked bar chart presenting the Common Language Effect Size (CLES) estimates for the significant linguistic features that survived post-hoc corrections in the Memory dataset.



Note. The CLES estimates represent the probability (ranging from 0 to 1) of finding a specific linguistic feature in truthful memories (sky blue) than in deceptive ones (salmon). The

CLES for truthful memories are sorted in descending order, while the CLES for deceptive memories are sorted in ascending order.

FIGURE 6. Horizontal stacked bar chart presenting the Common Language Effect Size (CLES) estimates for the significant linguistic features that survived post-hoc corrections in the Intention dataset.



Note. The CLES estimates represent the probability (ranging from 0 to 1) of finding a specific linguistic feature in truthful intentions (sky blue) than in deceptive ones (salmon). The CLES for truthful intentions are sorted in descending order, while the CLES for deceptive intentions are sorted in ascending order.

Overall, the Intentions dataset displayed fewer significant differences in linguistic features among truthful and deceptive statements than the Opinion and Memory datasets. In Table S5 (SM), we reported, for all the linguistic features and the three datasets, all the statistics, corrected p-values, effect-size scores (expressed as CLES and Cohen's D with 95% CI), and the direction of the effect for all linguistic features and the three datasets.

3.2 Performance on the Lie-Detection classification task

This section presents the performance, in terms of average accuracy (and standard deviation) of the 10-folds, on the test sets after the last epoch of the small and base models in all the Scenarios.

Scenario 1.

In Table 5 are depicted the test accuracies for the FLAN-T5 model, categorized by dataset and model size in Scenario 1. In each case, the base

model, on average, outperformed the small model, with the Memory dataset showing the largest improvement of 4% and the Intention dataset showing just a 0.06% increase in average accuracy. These results indicate that the larger model size generally leads to improved performance across the three datasets, with higher accuracy observed in the base version.

TABLE 5. Test accuracy of the FLAN-T5 models in Scenarios 1 and 3 for the three datasets.

Model	Opinion	Memory	Intention
Bag-of-words baseline	76.16 ± 2.9 %	57.57 ± 7.66 %	67.07 ± 3.18 %
Literature baseline	65.16 ± 5.7 %	-	69.00 [63; 74] %
			69.86 ± 2.34 %
			70.61 ±2.58 %
Flan-T5 small - Scenario 1	80.64 ± 2.03 %	76.87 ± 2.06 %	71.46 ± 3.65 %
Flan-T5 base - Scenario 1	82.60 ± 3.01 %	**80.61 ± 1.41** %	71.52 ± 2.21 %
Flan-T5 small - Scenario 3	79 ± 2.11 %	75.67 ± 1.90 %	69.32 ± 3.75 %
Flan-T5 base - Scenario 3	**82.72 ± 2.39** %	79.87 ± 1.60 %	**72.25 ± 2.86** %

Note. Reported values are means ± standard deviation of the 10 folds. Best results per evaluation metric are in bold. The literature baseline for the Opinion dataset refers to the average accuracy and standard deviation from all within-topic accuracies from FastText Embedding + Transformer (Capuozzo et al., 2020). The literature baseline for the Intention dataset refers to the accuracy from Vanilla Random Forest using LIWC features (confidence interval in square brackets) (Kleinberg & Verschuere, 2021), the averaged accuracy and standard deviation from RoBERTa + Transformers + Co-Attention model and BERT + co-attention model (Ilias et al., 2022), respectively.

Scenario 2.

This scenario aimed to investigate our fine-tuned LLM's generalization capability across different deception domains. As presented in Table 6, the test accuracy for the three experiments in this scenario significantly dropped to the chance level, showing that the model, in any case, was able to learn a general rule to detect lies coming from different contexts.

Scenario 3.

In Scenario 3, we tested the accuracy of the FLAN-T5 small and base versions on the aggregated Opinion, Memory, and Intention datasets. The small-sized FLAN-T5 achieved an average test accuracy of 75.45% (*SD*=1.6), while the base-sized FLAN-T5 exhibited a higher average test

accuracy of 79.31% (*SD*=1.3). In other words, the base-sized model outperformed the small model by approximately four percentage points.

TABLE 6. Test accuracy of FLAN-5 Models in Scenario 2 (three combinations of train sets).

Train set	Test set	Model Size	Test Accuracy
		Flan-T5 small	55.37
Opinion + Memory	Intention	Flan-T5 base	55.67
		Flan-T5 small	55.37
Opinion + Intention	Memory	Flan-T5 base	54.23
		Flan-T5 small	53.12
Memory + Intention	Opinion	Flan-T5 base	49.40

Note. The performance comparison is conducted among the small and base versions of the FLAN-T5 model in three combinations of training sets: opinion + memory, opinion + intention, and memory + intention.

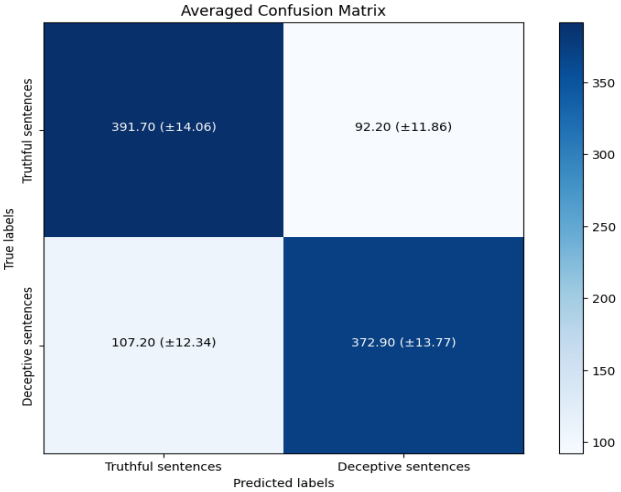
Results in Table 5 show the disaggregated performance on individual datasets between the small and base FLAN-T5 models in Scenario 3, with a comparison to their counterparts in Scenario 1. These comparisons show that FLAN-T5-small in Scenario 3 exhibited worse performance than in Scenario 1. Instead, in Scenario 3, the base model barely outperformed its counterparts of Scenario 1 on the Opinion and Intention datasets by less than 1% and slightly underperformed its counterpart of Scenario 1 on the Memory dataset.

We identified the top-performing model as the FLAN-T5 base in Scenario 3 because of its higher accuracy in the overall performance. The averaged confusion matrix of the 10 folds for this model is depicted in Figure 7. Notably, in each case, we were able to outperform both the bag-of-words + logistic regression classifier baseline and the performance achieved on the same datasets in previous studies (Capuozzo et al., 2020; Ilias et al., 2022; Kleinberg & Verschuere, 2021).

3.3 Explainability Analysis

This section aims to gain a deeper understanding of the top-performing model identified in Scenario 3 (FLAN-T5 base) through a **DeCLaRatiVE** stylometric analysis of statements correctly classified and misclassified by the model. The purpose of this analysis was to examine whether the linguistic style of the input statements exerted an influence on the resulting output of the model and to provide explanations for the wrong classification outputs. For this analysis, we compared:

FIGURE 7. Averaged confusion matrix of the top-performing model identified as FLAN-T5 base in Scenario 3.



Note. In each square, the results obtained represent the average (and standard deviation) from the test set of each iteration of the 10-fold cross-validation.

a) truthful statements misclassified as deceptive (False Negatives), with deceptive statements misclassified as truthful (False Positives);
b) statements correctly classified as deceptive (True Negatives) vs. truthful statements misclassified as deceptive (False Negatives);
c) statements correctly classified as truthful (True Positives) vs. deceptive statements misclassified as truthful (False Positives).
d) truthful vs. deceptive statements correctly classified by the model (True Positives vs. True Negatives).

The statistically significant features reported survived post-hoc correction for multiple comparisons in each fold. Overall, for comparison a), b), and c), we observed no statistically significant differences ($p < 0.05$) in any linguistic features for most of the splits, with the only exception of:

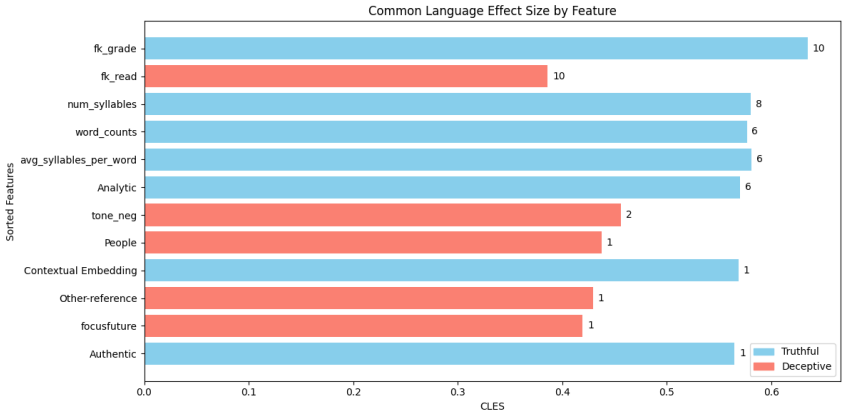
1) 'fk_read' in fold 1 (t=5.30, p=0.04, $CLES$ = 0.63 [0.55, 0.71], d=0.46 [0.18, 0.75]) and 'Reality Monitoring' in fold 6 (t=4.74, p=0.047, $CLES$=0.62 [0.54, 0.70], d=0.46 [0.17, 0.75]) for the a) comparison;
2) 'Reality Monitoring'in fold 6 (t=-3.39, p=0.04, $CLES$=0.40 [0.34, 0.46], d=-0.34 [-0.55, -0.13]) and 'Reality Monitoring' (t=-3.16,

p=0.04, CLES = 0.41 [0.34,0.47], d = -0.34 [-0.56, -0.12]) and 'Contextual Embedding' (t=-2.11, p=0.01, $CLES$=0.39 [0.33, 0.45], d=-0.42 [-0.63, -0.2]) in fold 7 the b) comparison;

3) 'num_syllables'(t=76.87, p=0.01, $CLES$=0.64 [0.57, 0.7], d=0.46 [0.27, 0.7]) and 'word_counts'(t=59.63, p=0.01, $CLES$=0.64 [0.57, 0.71], d=0.46 [0.21, 0.7]) in fold 9 for the c) comparison.

Conversely, for the d) comparison, several significant features emerged in all the folds and survived corrections for multiple comparisons. Figure 8 depicts the CLES effect size scores of linguistic features, sorted according to the number of times they were found to be significant among the ten folds. The top six features in Fig. 8 represented a cluster of linguistic features related to the Cognitive Load framework.

FIGURE 8. Linguistic features in Truthful and Deceptive statements that were accurately classified by FLAN-T5 base in Scenario 3.



Note. The bar plot shows the averaged Common Language Effect Size among the ten folds of linguistic features that survived post-hoc corrections. Linguistic features are sorted in descending order according to the number of times they were found to be significant among the 10 folds (displayed at the side of each bar). Linguistic features higher on average in truthful texts are shown in sky blue, while those higher on average in deceptive texts are shown in salmon.

4. Discussion

At the time of writing and to the best of our knowledge, this is the first study involving the use of an LLM for a lie-detection task. LLMs are Transformer-based models trained on large corpora of text that have proven to generate coherent text in human natural language and have extreme flexibility in a wide range of NLP tasks (Zhao et al., 2023). In addition, these models can be further fine-tuned on specific tasks using

smaller task-specific datasets, achieving state-of-the-art results (Zhao et al., 2023). In the present Chapter, we investigated the efficacy of a Large Language Model (LLM), specifically FLAN-T5 in its small and base versions, in learning and generalizing the intrinsic linguistic representation of deception across different contexts. To accomplish this, we employed three datasets encompassing genuine or fabricated statements regarding personal opinions, autobiographical experiences, and future intentions.

4.1 Descriptive Linguistic Analysis

Descriptive linguistic analysis was performed to compare the three datasets on linguistic features by exploring the differences in the **DeCLaRatiVE** style, i.e., analyzing 26 linguistic features extracted from the psychological frameworks of Distancing, Cognitive Load, Reality monitoring, and Verifiability approach. This analysis aimed to test Hypothesis 5a, which postulates a variation in the linguistic style that differentiates truthful from deceptive statements across varying contexts (i.e., personal opinions vs. autobiographical memories vs. future intentions). The results from this analysis confirmed our hypothesis, showing that the linguistic features exhibiting statistically significant differences between truthful and deceptive statements indeed varied across datasets. This variation was observed in terms of the total number and type of features, the magnitude of the effect size (from very small to medium), and the direction of the effect. In the following paragraphs, the interpretation of the significant linguistic features of each dataset will be discussed.

Opinions

After analyzing truthful and deceptive opinions using the **DeCLaRatiVE** stylometry, different linguistic features - related to the theoretical frameworks of CL, RM, and Distancing - were found to be significant. In line with the CL framework, we observed that truthful opinions were characterized by greater complexity, verbosity, and more authenticity in linguistic style (Hauch et al., 2015; Vrij et al., 2015). For features related to the RM framework, truthful opinions were characterized by a lesser number of concrete words and a greater number of cognitive words, as also previously shown (Mihalcea & Strapparava, 2009); in contrast, deceptive opinions showed higher scores in the concreteness of words, contextual details, and reality monitoring. These differences may reflect, on one side, the reasoning processes that truth-tellers engage in evaluating the pros and cons of abstract and controversial concepts (e.g., abortion), while for deceivers, it may be indicative of difficulty in abstraction, resulting in fake opinions that sound more grounded in reality. Finally, in line with previous literature on distancing framework (Hancock et al.,

2007; Newman et al., 2003) and deceptive opinions (Mihalcea & Strapparava, 2009; Pérez-Rosas & Mihalcea, 2015), deceivers utilized more other-related word classes ('Other-reference') and fewer self-related words ('Self-reference'), confirming that individuals may tend to avoid personal involvement when expressing deceptive statements.

Memories

Following the analysis of truthful and deceptive narratives of autobiographical memories through **DeCLaRatiVE** stylometry, various linguistic features associated with the theoretical frameworks of CL, RM, VA, and Distancing were found to be significant. As for opinions, according to the CL framework, truthful narratives of autobiographical memories exhibited higher levels of complexity and verbosity and appeared to be more analytical in style (Hauch et al., 2015; Vrij et al., 2015). In accordance with the RM framework (Amado et al., 2016; Gancedo et al., 2021; Johnson & Raye, 1981; S. Sporer, 1997; S. L. Sporer, 2004), posing that truthful memory accounts tend to reflect the perceptual processes involved during the experience of the event, while fabricated accounts are constructed through cognitive operations, we found genuine memories exhibiting higher scores in memory-related words and the number of words associated with spatial and temporal information ('Contextual Embedding'), as well as an overall higher RM score. Conversely, we found deceptive memories showing higher scores in words related to cognitive processes (e.g., reasoning, insight, causation). Furthermore, in line with Kleinberg's truthful concreteness hypothesis (Kleinberg et al., 2019), truthful memories were overall characterized by words with higher scores of concreteness. Along with the VA, truthful memories contained more verifiable details, as indicated by the greater number of named entities about times and locations (Kleinberg, Mozes, et al., 2018; Kleinberg, van der Toolen, et al., 2018). Notably, we found this effect although participants lied in a low-stakes scenario. However, deceptive memories were unexpectedly characterized by a higher number of self-references and named entities of 'People'. This result is in contrast with previous literature on the distancing framework (Hauch et al., 2015; Newman et al., 2003). One possible explanation of this significant but small effect is that liars may try to increase their credibility by fostering a sense of social connection.

Intentions

Upon examining truthful and deceptive statements of future intentions through **DeCLaRatiVE** stylometry, several linguistic features were

found to be significant. Our findings are consistent with previous research claiming that genuine intentions contain more 'how-utterances', i.e., indicators of careful planning and concrete descriptions of activities. In contrast, false intentions are characterized by 'why-utterances', i.e., explanations and reasons for why someone planned an activity or for doing something in a certain way (Kleinberg, van der Toolen, et al., 2018). Indeed, we found that true intentions were more likely to provide concrete and distinct information about the intended action, grounding their statements in real-world experiences and providing temporal and spatial references. Additionally, true intentions were characterized by a more analytical style and a greater presence of numerical entities. In contrast, false intentions exhibited a higher number of cognitive words and expressions and were temporally oriented toward the present and past. Furthermore, we found evidence in line with the claim that liars may over-prepare their statements (Kleinberg, van der Toolen, et al., 2018), as indicated by higher verbosity. Finally, in contrast with the distancing framework (Hauch et al., 2015; Newman et al., 2003), we found a significantly higher proportion of self-references and mentions of people in deceptive statements. However, the effect size for this finding is small. As for deceptive memories, one possible interpretation is that liars may attempt to appear more credible by creating a sense of social connection.

4.2 Lie Detection Task

In order to test the feasibility of fine-tuning a FLAN-T5 model for a lie detection task, we developed three scenarios.

In Scenario 1, we tested whether fine-tuning LLMs can effectively classify the veracity of short statements based on raw texts with performance highly above the chance level (Hypothesis 1a). To this aim, we fine-tuned FLAN-T5 in its small version to perform lie detection as a classification task. We repeated this procedure for the three datasets (i.e., opinions vs. memories vs. intentions). This fine-tuning process yielded promising results confirming our hypothesis, with an average accuracy of 80.64% (*SD*=2.03) for the Opinion dataset, 76.87% (*SD*=2.06) for the Memory dataset, and 71.46% (*SD*=3.65) for the Intention dataset.

In Scenario 2, we tested whether fine-tuning an LLM on deceptive narratives enables the model to detect new types of deception (Hypothesis 2). To verify this hypothesis, we fine-tuned FLAN-T5 (small version) on two datasets and tested on the third one (e.g., train: opinion + memory; test: intention). Our findings show that the model performed at chance level in all three combinations of this Scenario, suggesting that there are no universal rules the model can learn to distinguish truthful from deceptive statements, enabling a generalization of the task across different

contexts. Indeed, as shown in the Descriptive Linguistic Analysis section, the three datasets differed significantly in terms of the content and the linguistic style by which truthful and deceptive narratives are delivered. Therefore, the model struggled to identify a specific pattern of linguistic deception and appeared to engage in domain-specific learning, tailoring its classification capabilities to that specific domain of deception.

In Scenario 3, we tested whether fine-tuning an LLM on a multiple-context dataset enables the model to obtain successful predictions on a multi-context test set (hypothesis 3). To achieve this aim, we fine-tuned and tested FLAN-T5 (small version) with the three aggregated datasets (i.e., opinion + memory + intention). The small-sized FLAN-T5 achieved an average accuracy of 75.45% (st. dev. ± 1.6). Additionally, the disaggregated performance on individual datasets compared to their counterpart in Scenario 1 exhibited solely a small decrease in accuracy (around 1%). These findings confirmed our hypothesis, providing evidence of LLMs' ability to generalize when fine-tuned and tested on a multi-context dataset, in contrast to previous empirical evidence showing a decline in performance in machine learning models on the same scenarios (Hernández-Castañeda et al., 2016; Mihalcea & Strapparava, 2009; Pérez-Rosas & Mihalcea, 2014).

To test whether the model performance increases when employing larger models (Hypothesis 4), we repeated the same experiments in Scenarios 1, 2, and 3 with the base version of FLAN-T5. In Scenario 1, we found that the base version of FLAN-T5 provided higher accuracy than the small version. In Scenario 3, the base version of the model achieved an average accuracy of 79.31% (*SD*=1.3), outperforming the small model by approximately four percentage points. Additionally, this increase in the general accuracy did not compromise the performance on any individual dataset when compared to what was achieved by the smaller model or by the FLAN-T5 base in Scenario 1. In contrast, the base version of FLAN-T5 in Scenario 2 still obtained performance around the chance level. On one hand, the findings obtained from the base model in Scenarios 1 and 3 confirmed the hypothesis that the model size does influence the performance, likely because a bigger model is able to learn a better representation of linguistic patterns of genuine and deceptive narratives. Specifically, in Scenario 3, the FLAN-T5 base, with its larger size, possessed the capability to comprehend and integrate the features of the three distinct datasets altogether, thereby maintaining consistent performance across all individual datasets. In contrast, the smaller FLAN-T5 in Scenario 3 seemed to relinquish certain specialized abilities that are beneficial for specific datasets to classify deception across different contexts. On the other hand, findings from Scenarios 2 and 3 (with small and base FLAN-T5) showed that LLMs, despite having acquired a comprehensive

understanding of language patterns, still require exposure to prior examples to accurately classify deceptive texts within different domains. Given the results achieved, we highlight the importance of a diversified dataset to achieve a generalized good performance. We also consider crucial the balance between the diversity of the dataset and the size of the LLM, suggesting that the more diverse the dataset, the larger the model required to achieve higher-level accuracy.

Lastly, to test whether our approach outperforms classical machine learning and deep learning approaches in verbal lie detection (Hypothesis 1b), we compared the results obtained from FLAN-T5 in its small and base versions with the performance of a simpler baseline of a logistic regressor based on BoW embedding (Lin et al., 2023) and Transformer models previously employed in the literature on the Opinion (Capuozzo et al., 2020) and Intention datasets (Ilias et al., 2022; Kleinberg & Verschuere, 2021). Specifically, when comparing the Memory dataset to the logistic regression baseline, there was a 32% increase in performance. This improvement might be attributed to the longer and more complex nature of the stories in the Memory dataset, which challenges the effectiveness of more straightforward methods like logistic regression based on BoW in a lie detection task. In contrast, LLMs already possess a robust language representation; thus, fine-tuning LLMs leverages this representation, tailoring their NLP proficiency specifically for a lie detection task, yielding higher accuracy. The performance gained by fine-tuning LLMs was less pronounced for the Opinion and Intention datasets. For the Opinion dataset, this could be due to the relative ease of classification in these datasets, where simpler models can already achieve good performance, leaving a smaller margin for improvement. Nonetheless, the difference between our approach and the baselines is not negligible. In the Opinion dataset, we outperformed the literature baseline of a Transformer model trained from scratch by 17% accuracy and surpassed our logistic regression baseline by six percentage points. For the Intention dataset, our approach showed a 5-percentage point improvement over the logistic regression baseline and around 1-2% improvement over the best literature baseline. Notably, the best literature baseline for the Intention dataset (averaged accuracy: 70.61 ±2.58 %) used a similar approach to ours in terms of the type of model used, involving a Transformer-based model (BERT + Co-attention), which may explain the narrower performance gap.

Besides the differences in performance, the main advantage of our approach is its simplicity and flexibility compared to those used in previous studies (Capuozzo et al., 2020; Ilias et al., 2022; Kleinberg & Verschuere, 2021). Fine-tuning an LLM leverages an existing encoding procedure of language that effortlessly handles any type of statement,

unlike logistic regression based on BoW or training a new Transformer-based model from scratch. Taking all these aspects together, fine-tuning LLMs resulted in being more advantageous in terms of feasibility, flexibility, and performance accuracy.

4.3 Explainability Analysis

To improve the explainability of the performance collected, we investigated whether the linguistic style that characterizes truthful and deceptive narratives could have a role in the model's final predictions (Hypothesis 5b). For this aim, we applied a **D**e**CL**a**R**ati**VE** stylometric analysis on statements that were correctly classified and misclassified by the top-performing model identified in Scenario 3 (i.e., FLAN-T5 base).

In the misclassified sample, truthful and deceptive statements did not differ significantly for any linguistic feature extracted with the **D**e**CL**a**R**ati**VE** stylometry technique. The only exception was fold 1, which showed significant differences in the text's readability score. No significant differences were detected in each fold in linguistic features between deceptive statements that were correctly classified as deceptive (True Negatives) and truthful statements that were misclassified as deceptive (False Negatives), with the exception of Reality Monitoring in folds 6 and 7 and Contextual Embedding score in fold 7. Finally, truthful statements that were correctly classified as truthful (True Positives) and deceptive statements that were misclassified as truthful (False Positives) exhibited no significant differences, except for the number of syllables and the number of words in fold 9. We argue that the observation of significant differences in selected linguistic features across specific folds is more indicative that these findings may not be generalizable and are likely influenced by the particular fold under analysis. When taken together, most of the analyzed folds showed a substantial overlap in linguistic style. Consequently, the model might have exhibited poor classification performance for those statements because, while deceptive, they showed a linguistic style resembling truthful statements and vice versa.

In contrast, correctly classified statements displayed several significant differences between truthful and deceptive statements. Notably, the top six linguistic features in Fig. 8 resulted in statistical significance in at least 6 out of 10 folds. The fact that we found a consistent pattern of linguistic features in correctly classified statements but not in misclassified statements provides evidence for our hypothesis, suggesting that the linguistic style of statements does have a role in the model's final predictions. More in detail, the top-six linguistic features depicted in Fig. 8 represent a cluster of linguistic cues associated with the CL framework (Vrij et al., 2015), specifically low-level features related to the length, complexity,

and analytical style of the texts that may have enabled the distinction between truthful and deceptive statements. The fact that linguistic cues of CL survived among the several features available in a mixed dataset of utterances reflecting opinions, memories, and intentions raises the question of whether CL cues may be more generalizable than other cues that are, in contrast, more specific to a particular type of deception.

4.4 Limitations and future outlooks

Despite the demonstrated success of our model, three significant limitations impact the ecological validity of our findings and their practical application in real-life scenarios.

The first notable limitation pertains to the narrow focus of our study, which concentrated solely on lie detection within three specific contexts: personal opinions, autobiographical memories, and future intentions. This restricted scope limits the possibility of accurately classifying deceptive texts within different domains. A second limitation is that we exclusively considered datasets developed in experimental setups designed to collect genuine and completely fabricated narratives. However, individuals frequently employ embedded lies in real-life scenarios, in which substantial portions of their narratives are true, rather than fabricating an entirely fictitious story. Finally, the datasets employed in this study were collected in experimental low-stakes scenarios where participants had low incentives to lie and appear credible. Because of all the above issues, the application of our model in real-life contexts may be limited, and caution is advised when interpreting the results in such situations.

The limitations addressed in this study underscore the need for future research to expand the applicability and generalizability of lie-detection models for real-life settings. Future works may explore the inclusion of new datasets, trying different LLMs (e.g., the most recent GPT-4), different sizes (e.g., FLAN-T5 XXL version), and different fine-tuning strategies to investigate the variance in performance within a lie-detection task. Furthermore, our fine-tuning approach completely erased the previous capabilities possessed by the model; therefore, future works should also focus on new fine-tuning strategies that do not compromise the model's original capabilities.

Conclusion

In the present Chapter, we investigated the efficacy of a Large Language Model (LLM), specifically FLAN-T5 in its small and base versions, in

learning and generalizing the intrinsic linguistic representation of deception across three different contexts.

To this aim, we first tested whether fine-tuning an LLM is a valid procedure to detect deception from raw texts above chance level and outperform the classical machine learning and deep learning approaches. We found that fine-tuning FLAN-T5 on one context is a valid procedure to obtain a state-of-the-art accuracy, as proved by the fact that this procedure outperformed the baseline model (BoW + logistic regression) and previous works that applied machine and deep learning techniques on the same contexts (Capuozzo et al., 2020; Ilias et al., 2022; Kleinberg & Verschuere, 2021; Lin et al., 2023).

Second, we wanted to investigate whether fine-tuning an LLM on deceptive narratives enables the model to also detect new types of deceptive narratives. Findings from Scenario 2 disconfirm this hypothesis, suggesting that the model requires previous examples of different deceptive narratives to provide adequate accuracy in this classification task.

Third, we investigated whether it is possible to successfully fine-tune an LLM on a multiple-context dataset. Results from Scenario 3 confirm that a fine-tuned LLM may provide adequate accuracy in detecting deception from different contexts. We also found that fine-tuning on multiple datasets can increase the performance with respect to when fine-tuned on a single context.

Furthermore, we hypothesized that the model performance may depend on the model size, given that the larger the model, the better the model forms its inner representation of language. Results from Scenarios 1 and 3 confirmed that the base-sized model of FLAN-T5 provides higher accuracy than the small-sized version.

Finally, with our experiments, we introduced the **DeCLaRatiVE** stylometry technique, a new theory-based stylometric approach to investigate deception in texts from four psychological frameworks in one shot (Distancing, Cognitive Load, Reality Monitoring, and Verifiability Approach). We employed the **DeCLaRatiVE** stylometry technique to compare three datasets on linguistic features, and we found that fabricated statements from different contexts exhibit different linguistic cues of deception. We also employed the **DeCLaRatiVE** stylometry technique to conduct an explainability analysis and investigate whether the linguistic style by which truthful or deceptive narratives are delivered is a feature that the model takes into account for its final prediction. To this aim, we compared correctly classified and misclassified statements by the top-performing model (FLAN-T5 base in Scenario 3), finding that correctly classified statements share linguistic features related to the cognitive

load theory. In contrast, truthful and deceptive misclassified statements do not present significant differences in linguistic style.

Altogether, the findings in this Chapter, together with findings in Chapter 2, have highlighted the potential of resorting to computational methods stemming from natural language processing and machine learning for the coding of statements and the prediction of verbal deception detection. However, these approaches have been tested so far on fabricated statements, which involve producing a completely made-up statement. Therefore, in the next chapter (Chapter 4), we will investigate to what extent the opportunities highlighted so far also apply to a more nuanced and ecological form of deception, also known as embedded lies.

References

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *European Journal of Psychology Applied to Legal Context*, *7*(1), 3–12. https://doi.org/10.1016/J.EJPAL.2014.11.002

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/J.IJCHP.2016.01.002

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/S15327957PSPR1003_2

Bond, G. D., Holman, R. D., Eggert, J. A. L., Speller, L. F., Garcia, O. N., Mejia, S. C., Mcinnes, K. W., Ceniceros, E. C., & Rustige, R. (2017). 'Lyin' Ted', 'crooked hillary', and 'Deceptive Donald': Language of lies in the 2016 US presidential debates. *Appl. Cognit. Psychol.*, *31*(6), 668–677. https://doi.org/10.1002/acp.3376

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology*, *19*(3), 313–329. https://doi.org/10.1002/ACP.1087

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. (2020). DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th language resources and evaluation conference* (Vol. 12, pp. 1423-1430).

Chen, H. (2011). *Dark Web: Exploring and Mining the Dark Side of the Web*. 1–2. https://doi.org/10.1109/EISIC.2011.78

Chen, X., Hao, P., Chandramouli, R., & Subbalakshmi, K. P. (2011). Authorship similarity detection from email messages. *Lecture Notes in*

Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *6871 LNAI*, 375–386. https://doi.org/10.1007/978-3-642-23199-5_28

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). *Scaling Instruction-Finetuned Language Models*. http://arxiv.org/abs/2210.11416

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.*, *52*(1), 1–4. https://doi.org/10.1002/pra2.2015.145052010082

Daelemans, W. (2013). Explanation in computational stylometry. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7817 LNCS*(PART 2), 451–462. https://doi.org/10.1007/978-3-642-37256-8_37

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Fornaciari, T., & Poesio, M. (2012). *DeCour: a corpus of DEceptive statements in Italian COURts* (pp. 1585–1590). http://www.lrec-conf.org/proceedings/lrec2012/pdf/377_Paper.pdf

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in Italian court cases. *Artif. Intell. Law*, *21*(3), 303–340. https://doi.org/10.1007/s10506-013-9140-4

Fornaciari, T., & Poesio, M. (2014). Identifying fake Amazon reviews as learning from crowds. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, 279–287. https://doi.org/10.3115/V1/E14-1030

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/EJPALC2021A10

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Process.*, *45*(1), 1–23. https://doi.org/10.1080/01638530701739181

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues

to Deception. *Personality and Social Psychology Review*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., & Flores, J. J. G. (2016). Cross-domain deception detection using support vector networks. *Soft Comput.*, *21*(3), 585–595. https://doi.org/10.1007/s00500-016-2409-2

Ilias, L., Soldner, F., & Kleinberg, B. (2022). *Explainable Verbal Deception Detection using Transformers*. https://arxiv.org/abs/2210.03080v1

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018). Using Named Entities for Computer-Automated Verbal Deception Detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., Nahari, G., Arntz, A., & Verschuere, B. (2017). An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/COLLABRA.80

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, *32*(3), 354–366. https://doi.org/10.1002/ACP.3407

Kleinberg, B., van der Vegt, I., Arntz, A., & Verschuere, B. (2019). *Detecting deceptive communication through linguistic concreteness*. https://doi.org/10.31234/OSF.IO/P3QJH

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica*, *213*. https://doi.org/10.1016/J.ACTPSY.2020.103250

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Commun. Monogr.*, *66*(2), 125–144. https://doi.org/10.1080/03637759909376468

Levitan, S. I., Maredia, A., & Hirschberg, J. (2018). Linguistic cues to deception and perceived deception in interview dialogues. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 1941–1950. https://doi.org/10.18653/V1/N18-1176

Lin, Y. C., Chen, S. A., Liu, J. J., & Lin, C. J. (2023). Linear Classifier: An Often-Forgotten Baseline for Text Classification. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *2*, 1876–1888. https://doi.org/10.18653/v1/2023.acl-short.160

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law*, *11*(1), 99–122. https://doi.org/10.1080/10683160410001726356

Mbaziira, A. V., & Jones, J. H. (2017). Hybrid text-based deception models for native and non-native English cybercriminal networks. *ACM International Conference Proceeding Series*, *Part F130280*, 23–27. https://doi.org/10.1145/3093241.3093280

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychol. Bull.*, *111*(2), 361. https://doi.org/10.1037/0033-2909.111.2.361

Mihalcea, R., & Strapparava, C. (2009). The lie detector. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, 309. https://doi.org/10.3115/1667583.1667679

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Scientific Reports*, *8*(1). https://doi.org/10.1038/S41598-018-20462-6

Moore, J. H. (1999). Bootstrapping, permutation testing and the method of surrogate data. *Phys. Med. Biol.*, *44*(6), L11. https://doi.org/10.1088/0031-9155/44/6/101

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*(2), 227–239. https://doi.org/10.1111/J.2044-8333.2012.02069.X

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 309–319.

Pennebaker, J. W., Botth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic Inquiry and Word Count: LIWC2015. In *Austin, TX: Pennebaker Conglomerates*.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 59–66. https://doi.org/10.1145/2818346.2820758

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 3391–3401.

Pérez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural deception detection. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, *2*, 440–445. https://doi.org/10.3115/V1/P14-2072

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 1120–1125. https://doi.org/10.18653/V1/D15-1133

Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2020). Beyond Modelling: Understanding Mental Disorders in Online Social Media. *Advances in Information Retrieval*, *12035*, 296. https://doi.org/10.1007/978-3-030-45439-5_20

Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proc. Natl. Acad. Sci.*, *119*(45), e2211715119. https://doi.org/10.1073/pnas.2211715119

Sarzynska-Wawer, J., Pawlak, A., Szymanowska, J., Hanusz, K., & Wawer, A. (2023). Truth or lie: Exploring the language of deception. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281179

Schutte, M., Bogaard, G., Mac Giolla, E., Warmelink, L., Kleinberg, B., & Verschuere, B. (2021). *Man versus Machine: Comparing manual with LIWC coding of perceptual and contextual details for verbal lie detection*. https://doi.org/10.31234/OSF.IO/CTH58

Solà-Sales, S., Alzetta, C., Moret-Tatay, C., & Dell'Orletta, F. (2023). Analysing Deception in Witness Memory through Linguistic Styles in Spontaneous Language. *Brain Sciences*, *13*(2). https://doi.org/10.3390/BRAINSCI13020317

Sporer, S. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Appl. Cognit. Psychol.*, *11*(5), 373–397.

Sporer, S. L. (2004). Reality monitoring and detection of deception. *The Detection of Deception in Forensic Contexts*, 64–102. https://doi.org/10.1017/CBO9780511490071.004

Street, C. N. H., & Masip, J. (2015). The source of the truth bias: Heuristic processing? *Scandinavian Journal of Psychology*, *56*(3), 254–263. https://doi.org/10.1111/SJOP.12204

Tomas, F., Dodier, O., & Demarchi, S. (2022). Computational Measures of Deceptive Language: Prospects and Issues. *Frontiers in Communication*, *7*. https://doi.org/10.3389/FCOMM.2022.792378

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour*, *7*(5), 718–728. https://doi.org/10.1038/S41562-023-01556-2

Vrij, A., & Fisher, R. P. (2016). Which lie detection tools are ready for use in the criminal justice system? *Journal of Applied Research in Memory and Cognition*, *5*(3), 302–307. https://doi.org/10.1016/j.jarmac.2016.06.014

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/LCRP.12088

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal Lie Detection: Its Past, Present and Future. *Brain Sciences*, *12*(12). https://doi.org/10.3390/BRAINSCI12121644

Vrij, A., & Nahari, G. (2019). The Verifiability Approach. *Evidence-Based Investigative Interviewing*, 116–133. https://doi.org/10.4324/9781315160276-7

Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. (2014). A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas Psychol.*, *34*(1), 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A Survey of Large Language Models. *ArXiv Preprint ArXiv:2303.18223*. https://arxiv.org/abs/2303.18223v16

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, *13*(1), 81–106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6F

Supplementary Materials

Dataset description

The **Deceptive Opinions dataset** (Capuozzo et al., 2020) consists of 5000 opinions about highly controversial issues such as abortion, cannabis legalization, gay marriage, euthanasia, and policy on migrants. The dataset was collected from two samples in the US and Italy through Amazon Mechanical Turk, and participants were asked to provide both truthful and deceptive opinions in their native language (i.e., English or Italian). The experimental paradigm employed a ground-truth approach in which participants were instructed to provide both truthful and deceptive opinions for half of the topics, ensuring counterbalancing of the proportion of truthful and deceptive opinions for each topic. For our study, we selected opinions collected from the English (US) sample for a total of 2500 opinions from 500 participants.

The **Hippocorpus dataset** (Sap et al., 2022) is a collection of stories gathered through three stages on Amazon Mechanical Turk. The first stage involved workers writing a story about a memorable event they experienced in the past six months. The second stage involved workers from the first stage retelling their stories after 3-6 months. In the third stage, new workers were assigned a subset of story topics from the first stage and instructed to imagine a complete narrative based on that topic. They were then asked to write down the story as if they had personally experienced it. After writing, workers completed a questionnaire about the personal significance of their stories. The dataset contains 6,854 stories: 2,779 recalled stories, 2,756 imagined stories, and 1,319 retold stories. For the aim of this study, we included only recalled and imagined stories. Additionally, 11 stories were excluded due to missing data, and 19 stories containing a number of words below 2.5 standard deviations from the average (narratives with fewer than 24.34 words) were removed. We considered these stories to have too few words for analysis and indicative of a lack of engagement among the Turkers. Therefore, the final sample of stories employed for the study was 5,506.

The **Intention dataset** (Kleinberg & Verschuere, 2021) is a collection of statements about participants' most significant non-work-related activity recruited via Prolific Academic. Participants were instructed to provide convincing answers to two brief questions:

- Q1. "Please describe your activity as specifically as possible.";
- Q2. "Which information can you give us to reassure us that you are telling the truth?"

The activity described was required to be specific and not a continuous or daily activity occurring within the next seven days, with a defined start and end time.

Participants were randomly assigned to either the truthful or deceptive condition. In the deceptive condition, participants were assigned matched activities from the truthful condition. The dataset contains 1640 statements (857 deceptive and 783 truthful) with two answers per participant. For the aim of this study, we selected only statements from Q1, for a total of 1640 statements.

Vocabulary Uniqueness

Vocabulary uniqueness was computed by applying Jaccard's index to the truthful and deceptive vocabulary sets for each dataset, as in Ríssola et al. (2020) and Ilias et al. (2022). Jaccard's index is a measure of **similarity between two sample sets**. In the context of text analysis, we used Jaccard's index to compare the vocabulary used in truthful and deceptive statements, as in the following equation:

$$J = (setTrue \cap setFalse) / (setTrue \cup setFalse) \quad (1)$$

To ensure the accuracy of our Jaccard's index calculations, we implemented a series of preprocessing steps on the text data using the Python library SpaCy. These steps included converting all text to lowercase, applying tokenization to segment the text into individual words, removing stop words, and lemmatizing the remaining words. This preprocessing was necessary to focus solely on content words when computing Jaccard's index.

Subsequently, we transformed the preprocessed text samples for each group (truthful vs. deceptive) into sets of unique words utilizing the split() and explode() methods available in the Pandas library. This approach allowed us to construct sets comprising all the distinct words present in the truthful and deceptive text samples. Jaccard's index was derived by calculating the intersection (common words) and union (total words) of these two sets. The resulting index ranges from 0, indicating a completely different vocabulary between the two sets, to 1, indicating a completely identical vocabulary between the two sets. This index served as a measure of similarity or overlap between the word choices of truthful and deceptive statements within the respective datasets.

DeCLaRatiVE STYLOMETRY

Using several Natural Language Processing (NLP) techniques, we computed **D**e**CL**a**R**ati**VE** stylometry to a) describe linguistic differences in truthful and deceptive statements in the three datasets and b) conduct explainability analysis by exploring the linguistic style of sentences the model correctly classified and misclassified in order to understand whether the style of those sentences was a relevant feature for the model to generate its predictions. **D**e**CL**a**R**ati**VE** stylometry consisted of the extraction process of 26 linguistic features among the psychological frameworks of **D**istancing (Newman et al., 2003), **C**ognitive **L**oad (Vrij et al., 2015), **R**eality monitoring (Johnson & Raye, 1981), and **V**erifiability **A**pproach (Nahari et al., 2012; Vrij & Nahari, 2019). The extraction process for each framework is well described in the paragraphs below.

Cognitive Load:

Previous research has associated statistics about length and readability of the text with complexity and has employed them to study deception along with the cognitive load framework (see Hauch et al. (2015) for a meta-analysis of studies of linguistic cues for deception).

Using the Python library TEXTSTAT, we automatically computed several statistics from raw texts:

- number of sentences (**num_sentences**),
- number of words (**num_words**),
- number of syllables (**num_syllables**),
- the average number of syllables per word (**avg_syllables_per_word**),
- the Flesch-Kincaid Grade Level (**fk_grade**),
- the Flesch Reading-Ease Level (**fk_read**)

The Flesch-Kincaid Grade Level and the Flesch Reading-Ease Level were used to determine the difficulty of understanding a passage in English. Both tests utilize the same basic measures of word length and sentence length, but differ in their weighting factors.

The results of these tests are inversely correlated: a text with a high score on the Reading Ease test will typically have a lower score on the Grade-Level test. Higher scores on the Flesch Reading-Ease test indicate that the

material is easier to read, while lower scores indicate more difficult passages. The Flesch reading-ease[7] score is computed as in equation 1):

1) $206.835 - 1.015(total\ words/total\ sentences) - 84.6(total\ syllables/total\ words)$

The Flesch-Kincaid Grade Level Formula[8] produces a score corresponding to a U.S. grade level, providing an intuitive index of the readability level of texts. The index may be interpreted as the years of education typically required to understand the text. The grade level is computed as in equation 2):

2) $0.39(total\ words/total\ sentences) + 11.8(total\ syllables/total\ words) - 15.59$

Reality Monitoring and Distancing Framework

LINGUISTIC INQUIRY AND WORD COUNT (LIWC)

Linguistic Inquiry and Word Count (LIWC) is the gold standard software for analyzing word usage (Boyd et al., 2022). Given a text, it calculates the percentage of total words corresponding to more than 100 categories in the dictionary related to different psychosocial dimensions, which have been validated by human evaluators using rigorous procedures. A detailed description of LIWC-22 functioning and categories is reported in (Tausczik et al., 2010). Using the English dictionary, we scored each text along with all the categories present in LIWC-22.

LIWC scoring was computed on tokenized text using the English dictionary. The selection of the LIWC categories was guided by previous research on computerized verbal lie-detection (Newman et al., 2003; Sap et al., 2022) and a recent meta-analysis (Hauch et al., 2015). Therefore, we employed LIWC to investigate the presence of verbal cues related to the Reality Monitoring and Distancing frameworks. Additionally, we computed summary scores about the analytical style, the authenticity, and the tone of the text, as well as indices about the writer's temporal orientation.

[7] Flesch, R. How to write plain English. University of Canterbury. Available at http://www.mang. canterbury. ac. nz/writing_guide/writing/flesch.shtml. [Retrieved 5 February 2016]. (1979)

[8] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch*. (1975).

The selected and compounded linguistic features of interest are listed below:

- Summary statistics:
    - **'Analytic':** It describes the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns, also known as Analytic Thinking. People low in Analytical Thinking tend to write and think using more intuitive and personal language. Language scoring high in Analytic Thinking tends to be rewarded in academic settings and correlates with grades and reasoning skills. Language scoring low in Analytic Thinking tends to be viewed as less cold and rigid and more friendly and personable.
    - **'Authentic':** It describes the degree to which a person is self-monitoring while speaking or writing. Examples of texts that score low in authenticity include prepared texts (i.e., speeches written ahead of time) and texts where a person is being socially cautious. Examples of highly authentic texts tend to be spontaneous conversations between close friends or political leaders with little to no social inhibitions.
    - **'Tone':** Although LIWC-22 includes both positive and negative tone dimensions, the Tone variable puts the two dimensions into a single summary variable. The algorithm is built so that the more positive the tone, the higher the score. Scores below 50 suggest a more negative emotional tone.
- Basic Dictionary:
    - **'tone_pos':** provides the percentage score of words related to a positive sentiment (rather than positive emotion per se). It includes words related to positive emotions (e.g., happy, joy) and words related to those emotions (e.g., birthday, beautiful).
    - **'tone_neg':** provides the percentage score of words related to a negative sentiment (rather than negative emotion per se). It includes words related to negative emotions (e.g., sad, angry) and words related to those (e.g., kill, funeral).
    - **'Cognition':** It is the overarching dimension that reflects different ways people think or refer to their thinking. It includes the subcategories of all-or-none thinking,

different cognitive processes (i.e., insight, causation, discrepancy, tentative, certitude, differentiation), and memory.
        - **'memory'**: (e.g., remember, forget) reflects people's references and attention to their memories, beliefs about memory, and the processes of recall and forgetting.
- Writer's temporal orientation:
    - **'focuspast'**: refers to the use of past tense verbs and adverbs related to the past in language.
    - **'focuspresent'**: refers to the use of present tense verbs and adverbs related to the present in language.
    - **'focusfuture'**: refers to the use of future tense verbs and adverbs related to the future in language.

The following scores were computed by feature-engineering the available LIWC features according to theories on lie detection:

- To investigate the use of personal pronouns along with the **Distancing framework,** we computed two metrics (Newman et al., 2003):
    - **Self-reference:** computed as the sum of LIWC categories: 'i' + 'we.'
    - **Other-reference:** computed as the sum of LIWC categories: 'shehe' + 'they' + 'you.'
- To investigate the **Reality Monitoring framework**, we computed the following metrics below, following the same procedure as in Bond et al. (2017), Bond & Lee (2005), Kleinberg et al. (2017):
    - **Contextual Embedding**: computed as the sum of LIWC categories: 'space' + 'motion' + 'time.'
    - **Perceptual Details:** computed as the sum of LIWC categories: 'attention' + 'visual' + 'auditory' + 'feeling.'
    - **Reality Monitoring:** computed as the sum of LIWC categories: Contextual Embedding + Perceptual Details + 'Affect' - 'Cognition.'

CONCRETENESS SCORE

Another way to investigate the assumption of Reality Monitoring is through the linguistic concreteness of words used in truthful and deceptive texts. Kleinberg et al. (2019) have already investigated the concreteness of words in the framework of lie detection, postulating the *truthful concreteness hypothesis*, which states that truthful accounts are typically

characterized by specific, concrete, and situationally embedded information, while deceptive and fabricated statements tend to contain more abstract and less concrete information.

To determine the average level of concreteness for each statement, we utilized the concreteness annotation dataset developed by Brysbaert et al. (2014) (download is available in the Supplementary Materials here: https://link.springer.com/article/10.3758/s13428-013-0403-5#Sec10). This dataset involved around 40.000 English word lemmas, which were scored by a large group of human annotators using a five-point Likert scale, ranging from 1 (abstract) to 5 (concrete).

As for the computation of the Jaccard's index, we implemented a series of preprocessing steps on the text data using the Python library SpaCy to ensure the accuracy of our ***concreteness score*** calculation. These steps included converting all text to lowercase, applying tokenization to segment the text into individual words, removing stop words, and lemmatizing the remaining words. This process resulted in a list of lemmatized content words. Lemmatization was a necessary step to ensure a consistent overlap between our list of content words and those in the annotated concreteness dataset.

To compute the average concreteness score for each text, we cross-referenced the content words with the annotated concreteness dataset, assigning the respective concreteness scores when a match was found. Finally, we computed our dependent variable 'concr_score' by averaging the concreteness scores of all content words in the list. Higher values of the 'concr_score' indicated a greater degree of concreteness in the language employed within the statements.

Verifiability Approach

The verifiability approach in verbal lie detection suggests that truthful statements are more likely to be verifiable than false or deceptive statements, as liars avoid mentioning details that could be verified with independent evidence to conceal their deception (Nahari et al., 2012; Vrij & Nahari, 2019). Verifiable details may be represented by activities involving or witnessed by identified individuals, documented through video or photographic evidence, or leaving digital or physical traces (e.g., phone calls or receipts) (Nahari et al., 2012; Vrij & Nahari, 2019).

Automatically extracting **verifiable details** by using **named-entity recognition** (NER) has been proven to be effective for the detection of deception in hotel reviews (Kleinberg, van der Toolen, et al., 2018) as well as in participants' intentions on their weekend plans.

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique that deals with identifying and extracting information from text (so-called named entities) and classifying them into predefined categories, such as persons, locations, organizations, time, and many more. Using SpaCy, a Python library for NLP, we automatically extracted unique named entities from each raw statement through a Transformer algorithm for the English language (*en_core_web_trf*, https://spacy.io/models/en#en_core_web_trf). Table S3 shows the list of all entities available in SpaCy with their descriptions and some examples.

TABLE S3. List of the labels, brief descriptions, and a few examples of the extracted named-entities with the Python library SpaCy.

Entity	Description	Example
DATE	Absolute or relative dates or periods	"December 25, 2022", "10th August 1998", "Yesterday"
TIME	Times smaller than a day	"2:30 PM", "9 o'clock", "morning"
GPE	Countries, cities, states	"United States", "Paris", "Tokyo"
LOC	Non-GPE locations, mountain ranges, bodies of water	"Central Park", "Mount Everest", "Amazon Rainforest"
PERSON	People, including fictional	"Steve Jobs", "Emma Johnson", "Harry Potter"
ORDINAL	"first", "second", etc.	"First", "Third", "Tenth"
ORG	Companies, agencies, institutions, etc.	"Google", "Apple Inc.", "United Nations"
QUANTITY	Measurements, as of weight or distance	"10 kilograms", "5 liters", "100 meters"
WORK_OF_ART	Titles of books, songs, etc.	"Mona Lisa", "Hamlet", "Gone with the Wind"
PRODUCT	Objects, vehicles, foods, etc. (not services)	"iPhone", "Coca-Cola", "Nike shoes"
CARDINAL	Numerals that do not fall under another type	"Five", "Twenty", "One hundred"
NORP	Nationalities or religious or political groups	"American", "Muslim", "Republican"
MONEY	Monetary values, including unit	"$10", "€50", "¥1000"
LANGUAGE	Any named language	"English", "Spanish", "French"

Entity	Description	Example
FAC	Buildings, airports, highways, bridges, etc.	"Eiffel Tower", "White House", "Golden Gate Bridge"
EVENT	Named hurricanes, battles, wars, sports events, etc.	"Olympic Games", "Wedding ceremony", "Concert"
PERCENT	Percentage, including "%"	"50%", "10.5%", "75.2%"
LAW	Named documents made into laws.	"Constitution", "Copyright Act", "Traffic laws"

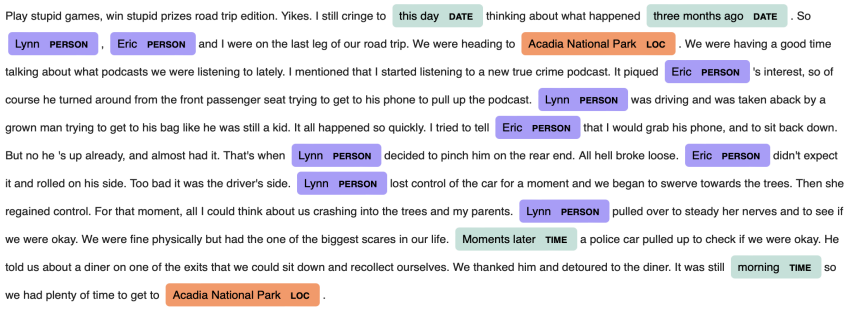
Table S4 shows the list of the combined entities along with the Verifiability approach.

TABLE S4. New linguistic features derived after grouping named entities.

Label	Grouped Entities
People	PERSON
Temporal details	DATE + TIME + EVENT
Spatial details	GPE + LOC + FAC
Quantity details	PERCENT + MONEY + QUANTITY + CARDINAL + ORDINAL

The example depicted in Figure S1 is a common way to represent text with annotated named-entities.

FIGURE S1. Narrative of an autobiographical event annotated with named entities using Named-Entities Recognition in SpaCy.



Note. In this narrative, entities related to named people (e.g., Lynn and Eric), location (e.g., Acadia National Park), dates (e.g., this day, three months ago), and times (e.g., moments later, morning) were automatically detected.

Descriptive Linguistic Analysis

In Table S5, we provided the results of the descriptive linguistic analysis conducted on the three datasets using the **DeCLaRatiVE** stylometry.

TABLE S5. Linguistic differences in the **DeCLaRatiVE** features between truthful and deceptive statements for each dataset.

Linguistic Feature	Dataset	Permutation t-statistic	Corrected *p*-value	CLES (95 % CI)	Cohen's *d* (95% CI)	Direction
num_sentences	Opinion	0.264	0.006**	0.57 (0.55, 0.59)	0.19 (0.1, 0.28)	Truthful
	Memory	0.138	0.365	0.52 (0.5, 0.53)	0.04 (-0.02, 0.09)	Truthful
	Intention	-0.092	1	0.47 (0.44, 0.5)	-0.06 (-0.15, 0.04)	Deceptive
word_counts	Opinion	15.508	0.005**	0.68 (0.66, 0.7)	0.57 (0.48, 0.67)	Truthful
	Memory	29.268	0.005**	0.6 (0.59, 0.62)	0.32 (0.27, 0.37)	Truthful
	Intention	-6.509	0.005**	0.43 (0.4, 0.46)	-0.22 (-0.31, -0.12)	Deceptive
num_syllables	Opinion	23.182	0.005**	0.68 (0.66, 0.7)	0.61 (0.51, 0.7)	Truthful
	Memory	41.277	0.005**	0.61 (0.6, 0.62)	0.35 (0.3, 0.41)	Truthful
	Intention	-7.5	0.005**	0.44 (0.41, 0.47)	-0.20 (-0.3, -0.1)	Deceptive
avg_syllabes_per_word	Opinion	0.027	0.005**	0.58 (0.55, 0.6)	0.27 (0.18, 0.36)	Truthful
	Memory	0.015	0.005**	0.56 (0.54, 0.57)	0.21 (0.16, 0.27)	Truthful
	Intention	0.003	1	0.50 (0.47, 0.53)	0.03 (-0.07, 0.13)	Truthful
fk_grade	Opinion	1.395	0.005**	0.67 (0.65, 0.69)	0.56 (0.47, 0.66)	Truthful
	Memory	0.863	0.005**	0.63 (0.61, 0.64)	0.46 (0.4, 0.51)	Truthful
	Intention	-0.465	0.094	0.46 (0.43, 0.49)	-0.13 (-0.23, -0.03)	Deceptive
fk_read	Opinion	-5.095	0.005**	0.37 (0.34, 0.39)	-0.48 (-0.57, -0.39)	Deceptive
	Memory	-3.133	0.005**	0.39 (0.37, 0.4)	-0.39 (-0.45, -0.34)	Deceptive
	Intention	0.984	1	0.52 (0.49, 0.55)	0.08 (-0.02, 0.17)	Truthful

Linguistic Feature	Dataset	Permutation t-statistic	Corrected *p*-value	CLES (95 % CI)	Cohen's *d* (95% CI)	Direction
Analytic	Opinion	0.624	0.623	0.50 (0.47, 0.53)	0.03 (-0.06, 0.12)	Truthful
	Memory	5.139	0.005**	0.57 (0.55, 0.58)	0.23 (0.18, 0.29)	Truthful
	Intention	5.859	0.005**	0.57 (0.54, 0.59)	0.21 (0.12, 0.31)	Truthful
Authentic	Opinion	14.984	0.005**	0.69 (0.67, 0.72)	0.70 (0.6, 0.8)	Truthful
	Memory	0.605	0.415	0.5 (0.49, 0.52)	0.02 (-0.03, 0.07)	Truthful
	Intention	3.329	0.449	0.54 (0.51, 0.56)	0.10 (0.01, 0.2)	Truthful
Tone	Opinion	8.191	0.005**	0.63 (0.59, 0.66)	0.41 (0.32, 0.51)	Truthful
	Memory	-2.545	0.015*	0.48 (0.46, 0.49)	-0.08 (-0.13, -0.03)	Deceptive
	Intention	-4.505	0.073	0.47 (0.44, 0.49)	-0.14 (-0.24, -0.05)	Deceptive
tone_pos	Opinion	0.633	0.005**	0.62 (0.58, 0.65)	0.37 (0.28, 0.46)	Truthful
	Memory	-0.334	0.005**	0.45 (0.44, 0.47)	-0.16 (-0.21, -0.1)	Deceptive
	Intention	-0.307	0.664	0.46 (0.44, 0.49)	-0.09 (-0.19, 0.01)	Deceptive
tone_neg	Opinion	-1.138	0.005**	0.38 (0.34, 0.41)	-0.42 (-0.51, -0.33)	Deceptive
	Memory	-0.1	0.029*	0.48 (0.47, 0.49)	-0.08 (-0.13, -0.02)	Deceptive
	Intention	-0.01	1	0.48 (0.47, 0.5)	-0.01 (-0.11, 0.09)	Deceptive
Cognition	Opinion	1.201	0.005**	0.59 (0.55, 0.62)	0.26 (0.17, 0.35)	Truthful
	Memory	-0.991	0.005**	0.43 (0.41, 0.44)	-0.27 (-0.32, -0.22)	Deceptive
	Intention	-1.175	0.005**	0.42 (0.39, 0.45)	-0.25 (-0.35, -0.16)	Deceptive
memory	Opinion	-0.003	0.595	0.50 (0.49, 0.51)	-0.03 (-0.12, 0.05)	Deceptive
	Memory	0.054	0.005**	0.54 (0.53, 0.55)	0.12 (0.07, 0.17)	Truthful
	Intention	-0.013	1	0.50 (0.49, 0.5)	-0.05 (-0.15, 0.05)	Deceptive

Linguistic Feature	Dataset	Permutation t-statistic	Corrected *p*-value	CLES (95 % CI)	Cohen's *d* (95% CI)	Direction
focuspast	Opinion	0.23	0.048*	0.56 (0.52, 0.59)	0.18 (0.09, 0.27)	Truthful
	Memory	0.213	0.064	0.51 (0.5, 0.53)	0.07 (0.01, 0.12)	Truthful
	Intention	-0.835	0.005**	0.40 (0.38, 0.43)	-0.36 (-0.46, -0.26)	Deceptive
focuspresent	Opinion	-0.489	0.029*	0.45 (0.42, 0.48)	-0.20 (-0.29, -0.11)	Deceptive
	Memory	-0.226	0.005**	0.46 (0.45, 0.48)	-0.13 (-0.18, -0.08)	Deceptive
	Intention	-0.605	0.005**	0.44 (0.41, 0.46)	-0.20 (-0.3, -0.1)	Deceptive
focusfuture	Opinion	-0.264	0.008**	0.46 (0.43, 0.49)	-0.21 (-0.3, -0.12)	Deceptive
	Memory	-0.228	0.005**	0.45 (0.44, 0.47)	-0.2 (-0.25, -0.14)	Deceptive
	Intention	0.107	1	0.51 (0.48, 0.54)	0.03 (-0.07, 0.12)	Truthful
Self-reference	Opinion	1.052	0.005**	0.62 (0.59, 0.65)	0.41 (0.32, 0.5)	Truthful
	Memory	-0.536	0.005**	0.44 (0.43, 0.46)	-0.18 (-0.23, -0.13)	Deceptive
	Intention	-0.572	0.039*	0.45 (0.42, 0.48)	-0.15 (-0.25, -0.05)	Deceptive
Other-reference	Opinion	-0.654	0.005**	0.43 (0.39, 0.46)	-0.29 (-0.38, -0.2)	Deceptive
	Memory	-0.128	0.342	0.49 (0.47, 0.5)	-0.04 (-0.1, 0.01)	Deceptive
	Intention	-0.42	0.048*	0.45 (0.43, 0.48)	-0.14 (-0.24, -0.05)	Deceptive
Perceptual details	Opinion	0.469	0.005**	0.65 (0.62, 0.69)	0.45 (0.36, 0.54)	Truthful
	Memory	-0.089	0.083	0.48 (0.47, 0.5)	-0.06 (-0.11, -0.01)	Deceptive
	Intention	-0.203	1	0.46 (0.44, 0.49)	-0.07 (-0.17, -0.03)	Deceptive
Contextual Embedding	Opinion	-0.736	0.007**	0.44 (0.4, 0.47)	-0.22 (-0.31, -0.13)	Deceptive
	Memory	0.729	0.005**	0.55 (0.54, 0.57)	0.19 (0.14, 0.24)	Truthful
	Intention	1.84	0.005**	0.56 (0.54, 0.59)	0.24 (0.15, 0.34)	Truthful

Linguistic Feature	Dataset	Permutation t-statistic	Corrected *p*-value	CLES (95 % CI)	Cohen's *d* (95% CI)	Direction
Reality Monitoring	Opinion	-1.994	0.005**	0.42 (0.38, 0.45)	-0.27 (-0.36, -0.18)	Deceptive
	Memory	1.217	0.005**	0.55 (0.54, 0.57)	0.2 (0.14, 0.25)	Truthful
	Intention	2.474	0.005**	0.58 (0.55, 0.6)	0.26 (0.17, 0.36)	Truthful
Concreteness score	Opinion	-0.112	0.005**	0.32 (0.29, 0.35)	-0.60 (-0.69, -0.5)	Deceptive
	Memory	0.037	0.005**	0.54 (0.53, 0.56)	0.17 (0.11, 0.22)	Truthful
	Intention	0.125	0.005**	0.61 (0.59, 0.64)	0.39 (0.29, 0.49)	Truthful
People	Opinion	-0.038	0.111	0.49 (0.48, 0.51)	-0.10 (-0.19, -0.01)	Deceptive
	Memory	-0.001	0.005**	0.45 (0.44, 0.46)	-0.23 (-0.28, -0.17)	Deceptive
	Intention	-0.335	0.005**	0.45 (0.43, 0.47)	-0.24 (-0.34, -0.14)	Deceptive
Temporal details	Opinion	0.011	0.663	0.52 (0.5, 0.55)	0.03 (-0.06, 0.12)	Truthful
	Memory	0.002	0.005**	0.56 (0.54, 0.57)	0.2 (0.15, 0.25)	Truthful
	Intention	0.105	1	0.50 (0.47, 0.53)	0.03 (-0.06, 0.13)	Truthful
Spatial details	Opinion	0.003	0.934	0.51 (0.49, 0.54)	0.01 (-0.08, 0.09)	Truthful
	Memory	0.001	0.005**	0.53 (0.51, 0.54)	0.09 (0.04, 0.15)	Truthful
	Intention	-0.096	1	0.47 (0.45, 0.49)	-0.05 (-0.15, 0.04)	Deceptive
Quantity details	Opinion	0.022	0.448	0.52 (0.5, 0.55)	0.05 (-0.04, 0.14)	Truthful
	Memory	0.002	0.005**	0.58 (0.56, 0.59)	0.29 (0.23, 0.34)	Truthful
	Intention	0.352	0.005**	0.53 (0.51, 0.55)	0.22 (0.12, 0.31)	Truthful

Note. The table reports the descriptive linguistic analysis from the **DeCLaRatiVE** stylometry for each dataset (Opinion, Memory, Intention), statistic of the permutation *t*-test, *p*-values after Holm-Bonferroni correction (* $p < .05$, ** $p < .01$, *** $p < .001$), effect size (Common language effect size and Cohen's *d*) with 95 % confidence intervals, and direction of the effect (truthful vs. deceptive).

Chapter 4

When lies are mostly truthful: examining embedded lies through computational text analysis

This chapter is based on: Loconte, R., & Kleinberg, B. (2025). Examining embedded lies through computational text analysis. *Scientific Reports*, *15*(1), 26482. https://doi.org/10.1038/s41598-025-11327-w

Abstract

Verbal deception detection research relies on narratives and commonly assumes statements as truthful or deceptive. A more realistic perspective acknowledges that the veracity of statements exists on a continuum, with truthful and deceptive parts being embedded within the same statement. However, research on embedded lies has been lagging behind. We collected a novel dataset of 2,088 truthful and deceptive statements with annotated embedded lies. Using a counterbalanced within-subjects design, participants provided two versions of an autobiographical event. One was described truthfully, and the other one deceptively by including embedded lies. Participants later highlighted those embedded lies and judged them on lie centrality, deceptiveness, and source. We show that a fine-tuned language model (Llama-3-8B) can classify truthful statements and those containing embedded lies significantly above the chance level (64% accuracy). Individual differences, linguistic properties, and explainability analysis suggest that the challenge of moving the dial towards embedded lies stems from their resemblance to truthful statements. Typical deceptive statements consisted of 2/3 truthful information and 1/3 embedded lies, largely derived from past personal experiences and with minimal linguistic differences from their truthful counterparts. We present this dataset as a novel resource to address this challenge and foster research on embedded lies in verbal deception detection.

Keywords: Deception, Embedded Lies, Lying Profile, Natural Language Processing, Individual Differences.

1. Introduction

Everyone engages in some form of deception daily (Verigin et al., 2019). Rather than fabricating entirely false accounts, however, most individuals tend to combine elements of truth with elements of falsehood (Markowitz, 2024). This deception strategy is known as the embedding of lies. Embedded lies present a distinctive challenge in deception research and remain a largely under-investigated phenomenon.

1.1 Verbal deception detection

Research on verbal deception detection has often been focused on methods using manual coding, which involve training human judges to evaluate statements based on predefined verbal cues. One of the most common and widely applied techniques in real-world settings is the Criteria-Based Content Analysis (CBCA, Amado et al., 2016). CBCA was originally developed to evaluate children's testimonies on alleged sexual abuse cases and is now used to assess the credibility of testimonies in legal contexts. CBCA requires a human to identify and score a narrative on specific verbal cues, such as the amount of detail, unexpected complications, or spontaneous corrections, that truth-tellers are more likely to exhibit than deceivers. Another widely investigated technique in research is Reality Monitoring (RM, Johnson & Raye, 1981; Sporer, 2004), which distinguishes between truth and lies by focusing on the richness of sensory and contextual details provided by the speaker. Truth-tellers are thought to provide more vivid and detailed sensory information than deceivers, who typically rely on fabricated or imagined events. Building on the RM, the Verifiability Approach (VA, Nahari et al., 2012) capitalises on the tendency of truth-tellers to provide more verifiable details compared to lie-tellers, who avoid that because it could expose their deceit. While these methods have promise (Amado et al., 2016; Gancedo et al., 2021; Palena et al., 2021; Verschuere et al., 2021), they are more time-consuming and reliant on the expertise of practitioners than automated procedures with computational models, limiting their scalability (Fitzpatrick et al., 2015).

1.2 Computer-automated verbal deception detection

Recent advances in Natural Language Processing (NLP), often combined with methods from machine learning (ML), have introduced automated methods for detecting deception, enhancing both scalability and objectivity. NLP techniques allow the representation of textual data in a numerical vector form, with different levels of granularity. For instance,

the Linguistic Inquiry and Word Count (LIWC, Boyd et al., 2022) computes the frequency of words that pertain to psychological, social, and emotional dimensions (e.g., cognitive words, affective words, social words, etc); part-of-speech (POS) tagging informs on the shallow syntactic text structure by automatically assigning grammatical categories (e.g., nouns, verbs, pronouns) to words; named-entity recognition (NER) identifies and labels proper nouns and piece of information into predefined broader categories (i.e., *25.12.2024* into DATES, *Central Park* into LOCATIONS, *Google* into ORGANIZATIONS); *n*-gram models represents the text into frequency patterns of tokens; and embeddings use a vectorial numerical representation of words or statements that preserves the semantic and contextual relationship, allowing similar items to have closer representations in a multi-dimensional space.

Early approaches in computer-automated verbal deception detection (e.g., Hauch et al., 2015; Kleinberg et al., 2017; Mihalcea & Strapparava, 2009; Ott et al., 2011) first extract information from textual data and then use supervised machine learning to train models that use these extracted variables to derive a truthfulness judgment (see Constancio et al., 2023; Fitzpatrick et al., 2015; Hauch et al., 2015) for an extensive overview on the topic). For example, a previous study detected deceptive opinions by training a naïve Bayes and a support vector machine classifier on n-grams reaching 70.8% and 70.1% accuracy (Mihalcea & Strapparava, 2009). Likewise, another study detected deceptive opinions after training a support vector machine classifier on a combination of n-grams and LIWC features, reaching 89.8% accuracy (Ott et al., 2011). Finally, other scholars extracted the proportion of unique named-entities and found a significant discriminative power of 0.67 and 0.65 in detecting positive and negative deceptive hotel reviews, respectively (Kleinberg et al., 2017).

In contrast, more recent approaches go beyond feature extraction and use recent advances with Transformer-based models (Vaswani et al., 2017) that already incorporate a semantic representation of texts. For example, a previous study employed a Bidirectional Encoder Representations for Transformers (BERT,(Devlin et al., 2018) to incorporate contextual embeddings with attention-based mechanisms and detected deceptive utterances in a dataset of transcripts of criminal proceedings hearings, reaching 71.61% accuracy (DeCour, Fornaciari & Poesio, 2012). Another study fine-tuned a large language model (i.e., FLAN-T5, (Chung et al., 2022) to classify deceptive statements in three datasets encompassing personal opinions, autobiographical events, and future intentions, reaching 79.31% accuracy (Loconte et al., 2023).

This increasing sophistication of NLP techniques paves the way for future research to further refine methods and address more complex challenges in automated deception detection, such as the detection of embedded lies.

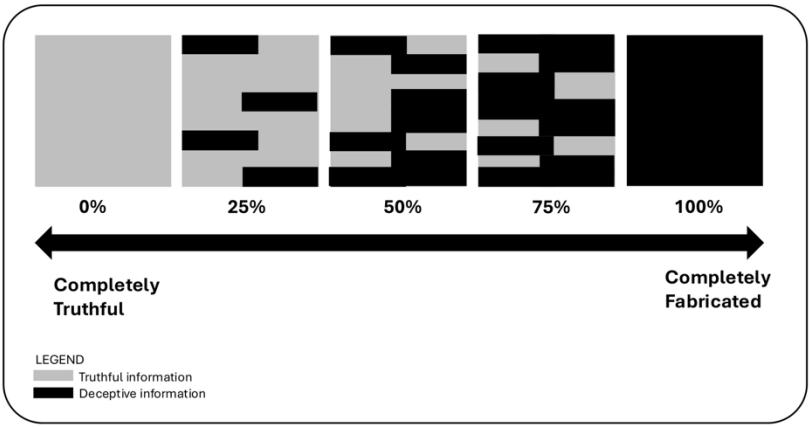
1.3 Embedded deception

The concept of embedded lies emerged when researchers started to ask lie-tellers about their lying strategies. Across a range of different contexts and studies, lie-tellers were found to frequently draw on past experiences to make their lies more believable (Bell & DePaulo, 1996; DePaulo et al., 2003; Hartwig et al., 2007; Leins et al., 2017; Wang et al., 2004). For instance, in two studies, 67% and 86% of liars, respectively, chose to construct their deceptive statements by incorporating elements of previously experienced events (Leins et al., 2013). Similarly, in another research, over 20% of deceptive statements consisted of truthful information (Nahari et al., 2012).

Despite being widely acknowledged within the field of deception detection, only a limited number of studies have directed their attention towards the phenomenon of embedded lies. Most research has focused on fabrication, conceptualising fabricated stories as entirely false and resulting in a dichotomous view of deception as either completely deceptive or not deceptive at all. A study typical of this perspective research requires a between-subjects design where participants engage in or view a mock crime event and are then allocated to one of two conditions: truth-tellers, in which they recollect exactly what they watched or experienced, and lie-tellers, in which they fabricate the event. For example, in a previous study, truth-tellers played a game during a staged event with a confederate, while lie-tellers, who did not participate in the event, were instructed to steal £10 from a wallet and then fabricated their involvement in the staged activities during a subsequent interview (Vrij et al., 2008). In other experiments, researchers implemented a matched-pairs design to match participants with the specific content of false statements (Kleinberg & Verschuere, 2021; Monaro et al., 2022; Sap et al., 2022). For example, participants in the honest condition were asked to tell the researcher about their past holidays, whereas, in the deceptive condition, participants were instructed to pretend to have experienced that same holiday (Monaro et al., 2022).

However, a more nuanced and realistic perspective acknowledges that a deceptive statement exists on a continuum where truthful and deceptive parts are embedded within the same statement (see Fig. 1). We, hence, adopted this framework in this Chapter to address the challenge of moving the dial towards embedded lies.

FIGURE 1 Graphical representation of the deception continuum framework.(Monaro et al., 2022)



Note. Deception is embedded into truthful statements in a continuous range from 0% (= no lies are present) to 100% (= the whole statement is made up). Levels in between represent various degrees of deception in the form of embedded lies.

1.4 Detecting embedded lies

While embedded deception has been acknowledged as a problem by many (Bell & DePaulo, 1996; DePaulo et al., 2003; Hartwig et al., 2007; Leins et al., 2013, 2017; Nahari et al., 2012; Wang et al., 2004), only a few studies reported on the nature and detection of embedded lies.

Using manual coding methods, two studies investigated embedded lies as fabricated statements contained within otherwise truthful statements (Verigin, Meijer, & Vrij, 2020; Verigin, Meijer, Vrij, et al., 2020). The first found that lies embedded in a fabricated statement were not qualitatively different from lies embedded in an otherwise truthful statement, suggesting that verbal credibility assessment tools may be robust against the embedding of lies. The second was a follow-up study employing the within-statement baseline comparison (Vrij, 2016), which is a strategy based on the idea that deception can be better detected if compared with a truthful baseline. However, the results showed that this strategy was not effective in improving the deception detection rate.

Another work investigated embedded lies to test whether specific manually annotated verbal cues, such as complications, common knowledge details, and self-handicapping strategies, varied across different amounts of lying in statements (Caso et al., 2023). The study found only

a significant difference in the number of complications, with a larger difference between truthful and outright deceptive statements compared to truthful and embedded lies.

Another work challenged the dichotomy between bald-faced lies and bald-faced truths (Markowitz, 2024). It was hypothesized that both truthful and deceptive narratives draw from a common pool of information so that lies and truths are not mutually exclusive but rather coexist within the same accounts. In this study, participants were randomly allocated to the truthful or deceptive condition and were tasked with writing opinions about their friends. They were then instructed to indicate within their statements the specific parts that were deceptive and those that were truthful. The results suggested that truthful narratives inherently include a certain proportion of embedded lies (37.06% of embedded lies in deceptive texts compared to 29.21% in truthful texts). While deceptive narratives contained a significantly higher rate of fabrications than truthful narratives, it is noteworthy that even the latter was aplenty with fabrications.

The few studies to date have not yet addressed two important challenges in lie detection research. Firstly, measuring embedded lies is inherently complex, with few studies employing within-subjects designs that control for individual differences and statement topics. Secondly, the lack of granular analytical methods hampered the detection of embedded lies, which are harder to identify than general deception. These methodological limitations have hindered the exploration of embedded lies, leaving them underrepresented in the literature despite their significance. Additionally, previous scholars have mentioned the importance of individual differences (e.g., demographic factors, personality traits, cognitive styles, and emotional states) in engaging in deceptive behaviour and in the type and dynamic of the deception involved (Halevy et al., 2014; C. L. Hart et al., 2020; Jones & Paulhus, 2017; Kashy & DePaulo, 1996; Levine et al., 2013; Serota et al., 2010; Serota & Levine, 2015; Weiss & Feldman, 2006); for a recent and complete review, see (Semrad et al., 2019). In the context of embedded lies, only one study explored whether and how personality (i.e., dark triad traits, Paulhus & Williams, 2002) and demographic factors (i.e., age, gender, ethnicity, and political ideology) influence this specific form of deception (Markowitz, 2024). However, no significant differences emerged from this specific study. Hence, with respect to individual differences, embedded lies represent an even more unexplored phenomenon. We, therefore, aim to connect these two streams of research in this Chapter by also promoting the investigation of individual differences in embedded lies.

1.5 The current study

This Chapter aims to help bridge the gap between deception practice and research by focusing explicitly on embedded lies. Prior work has usually employed between-subject or matched-pairs designs to study deception intended as fully fabricated accounts. Further, the majority of deception work relies on relatively small datasets (Kleinberg et al., 2019) and manual procedures (e.g., Vrij et al., 2022). Embedded lies also need further investigation in terms of individual differences, with only one study focusing on demographic and individual traits affecting embedded deception (Markowitz, 2024). We seek to address these limitations. First, we present a new dataset of embedded lies collected in a within-subjects experimental design that is sufficiently large to conduct meaningful computational analysis, including predictive modelling. Second, we enrich the scope of the dataset beyond the narratives and provide data at the individual level, allowing for analyses of individual differences in verbal deception behaviour. Third, we utilize automated approaches to retrieve variables from the narrative data using NLP methods and further resort to supervised machine learning to train models in detecting embedded deception.

2. Materials and Method

2.1 Participants

We recruited a total of 1058 participants[9] fluent in English from the general population through the online participant pool Prolific. Each participant provided informed consent before taking part in the Qualtrics-administered experimental task. Participation in the study was reimbursed 2$ upon experiment completion. Eight participants who did not follow the instructions (i.e., repeated the same phrase in multiple boxes) or provided non-sensical completions to the open-answer fields (e.g., writing random characters to fill the box) were removed for analysis. Eight participants from the subset of participants that freely recall a memorable event (after selecting the option "None of them") were removed because provided a title story that was too long (i.e., with a number of words

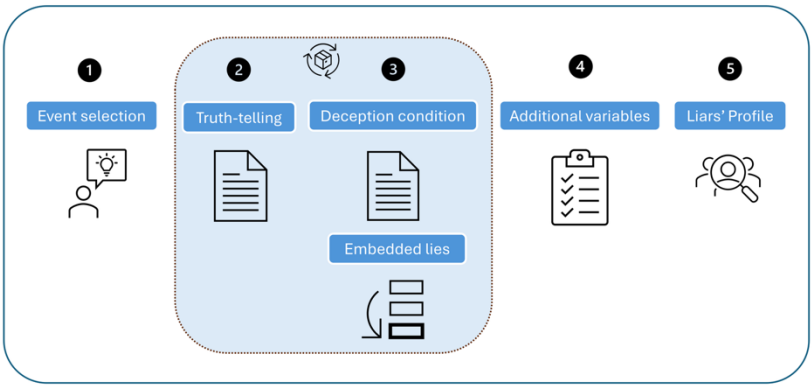
[9] An a priori power analysis for a small effect with a power of 0.90 (Cohen's d=0.20, α=.05, two-tailed) resulted in a sample size of 265 participants. Since we aimed to present a dataset adequate for computational analyses, we collected significantly more data.

higher than two standard deviations from the average) and were basically anticipating the main story in the wrong section. The final sample consisted of 1042 participants (58.23% females, 41.17% males, 0.19% preferred not to say, 0.38% expired data or removed consent on Prolific). The mean age was 30.32 years (SD=9.35, range: 18-105).

2.2 Experimental task

A previous study found that truthful statements may contain deceptive parts (Markowitz, 2024). However, we argue that truthful statements may also be, by definition, completely truthful, and those that are partially deceptive might be residuals from research design. Therefore, in this study, we developed an experimental task that followed a different perspective, considering fabrication on a continuum from 0 (fully truthful statements) to 100 (fully deceptive statements). The experiment was conducted in a counterbalanced within-subjects design (Fig. 2) where half of the participants performed the truth-telling task and then the deceptive task, followed by the embedded lie selection and rating. The other half performed the deceptive task first, followed by the embedded lie selection and rating, and then the truth-telling task. Additionally, we collected demographic variables and participants' lying attitudes to advance the understanding of individual differences in embedded lies.

FIGURE 2. Experimental procedure adopted in the experiment.



Note. The order of tasks two and three was counterbalanced across participants.

Step 1: Event selection

The experiment started with the event selection task. Participants were provided with a list of eleven pre-selected autobiographical events that they might have experienced in the past 24 months. The events were deemed relevant for lying in the subsequent deceptive writing task. After participants selected all of the events that they had experienced themselves in the past 24 months, they were randomly assigned to one of them and answered five questions about the event with the aim of collecting the following memory-related variables:

i) time: "how long ago did the event happen?" through a multiple-choice question with 25 options (from <1 months to 24 months);
ii) recollection: "how often do you think or talk about this event?" on a 5-point scale (1=never; 5=always);
iii) importance: "how important is this event to you?" on a 5-point scale (1=not important at all; 5=extremely important);
iv) accuracy: "how well do you remember this event?" on a 5-point scale (1=not well at all; 5=extremely well);
v) valence: "how would you rate this event in emotional terms?" on a 5-point scale (-1=extremely negative; 1=extremely positive).

The assigned event served as the basis for the remainder of the task. If participants did not experience any of the events in the list, they were instructed to choose the option "*none of them*". They were then asked to think about a positive or negative event occurring in the last 24 months that was memorable, emotional, and that directly involved them and were asked to provide a short title for the event. The 18.91% (n=197) of participants chose the "*none of them*" option.

We focused on autobiographical events that were deemed relevant for lying to mirror real-world scenarios and enhance the practical application of our research findings in improving the accuracy of deception detection.

Step 2: Truth-telling task

In the truth-telling task, participants were required to write an accurate and truthful account of the event in question. They were also asked to use correct spelling and grammar and were reminded not to use AI assistants in the writing task. Copy-pasting was blocked to prevent it. This task required a minimum of 3000 characters to move to the next phase of the experiment. Full instructions are provided in Supplementary Materials (SM) 1.

Step 3a: Deception task

For the deception task, participants were provided with a context relevant to lying and instructed to write a deceptive account of the selected event by incorporating false information. Specifically, participants were told to write an alternative version of the selected event in order to get a specific advantage from lying. Participants were also warned not to make up a statement about a new event and not to mention in any way that they were lying.

The list of contexts, matched for each event, and the number of participants allocated are provided in Table 1. Other than the deceptive instructions, the general writing instructions were identical to the truthful task. To motivate participants to do their best, they were informed about the chance of winning an extra 50£ if their statement was considered credible by the experts. In reality, all participants were included in the draw and the payment was distributed to a randomly selected participant after the data collection concluded. Full instructions are provided in SM 1.

TABLE 1. List of events, contexts for lying, and number (percentages) of participants allocated to that event.

Events	Context for lying	No. of participants allocated (%)
A job interview for your dream job	Inflate your past experiences to get the job	**160 (15.36%)**
Being hospitalized and undergoing surgery	Exaggerate some side effects to receive extra compensation from your health insurance	**70 (6.67%)**
Being involved in a car accident	Increase the claimed amount of damage you received to get more money	**47 (4.51%)**
Causing a car accident	Describe the event so that it's not your fault	**15 (1.44%)**
Cheating on an exam	Describing how you passed the exam, given that you cannot admit that you cheated	**48 (4.61%)**
Cheating on your partner	Convince your partner that you didn't cheat on them	**36 (3.45%)**
Ending a long romantic relationship	Pretend that you just had an argument with your partner	**152 (14.59%)**
Getting a speeding fine	Pretend it wasn't you driving the car that day	**62 (5.95%)**
Getting fired	Pretend that you just had a bad day at work	**34 (3.26%)**

Events	Context for lying	No. of participants allocated (%)
Missing a deadline at work because of bad organization	Find excuses that allow you not to appear forgetful or disorganised	**97 (9.31%)**
None of them	-	**197 (18.91%)**
Taking the bus/train without the ticket	Convince the ticket inspector that they shouldn't fine you for not having the ticket	**124 (11.90%)**

Step 3b: Embedded lies

Once participants had written the deceptive account, participants were instructed to copy and paste words, phrases, or sentences from their deceptive statements into a maximum of 20 text boxes (similar to Markowitz, 2024). For each word or phrase that was copy-pasted, participants rated the deceptiveness (*"how deceptive was this detail for your whole lie statement?")* and centrality *("how central was this detail for your whole lie statement?"*) of each embedded lie on a 5-point scale (1=not at all deceptive/central to 5=extremely deceptive/central).

Through a multiple-choice question, participants provided the source on which they relied for the embedded lie. The source options were based on liars' relying on their past experiences or cognitive processes (i.e., from memory, imagination, and planning). The following source options were provided: 1) you connected the detail to a past personal experience; 2) you saw a similar event happen to someone else and used that as a basis for the detail; 3) you derived the detail from a story another person told you, or from a book, or a movie; 4) you imagined the detail without any specific memory or experience; 5) you used planned, future activities as a reference.

To account for individual variability (i.e., participants copy-pasting a single word vs. multiple phrases or sentences), the number of embedded lies was also standardized for each subject by computing the ratio between the number of words provided in the 20 boxes and the total number of words in their deceptive text. The standardized number of embedded lies ranged from 0 to 1.

Step 4: Additional variables

Once the two writing tasks were completed, participants rated the following additional variables on a 5-point scale (1=completely disagree; 5=completely agree): i) difficulty of the task (i.e., "*I found the task was difficult*"); ii) clarity of instructions (i.e., "*I found the instructions were clear*");

iii) motivation of telling the truth (i.e., "*I was motivated to provide a convincing truthful statement*"); iv) motivation of lying (i.e., "*I was motivated to provide a convincing deceptive statement*").

Step 5: Liars' profile

To measure potential individual differences in participants' lying attitudes, the lying profile questionnaire (Makowski et al., 2023) was administered. The lying profile questionnaire measured dispositional traits of deception and was composed of 16 items grouped into four factors: frequency of lying (frequency); ability to lie (ability); negative attitude towards lying (negativity); and positive attitudes toward lying depending on the context (contextuality). Since participants may be prone to mask their lying attitude, the Balanced Inventory of Socially Desirable Responding Short Form (BIDR, (C. M. Hart et al., 2015) was also administered and used to correct the lying profile scores for potential effects of social desirability. The BIDR was a 16-item questionnaire which measured two main dimensions of social desirability: 1) self-deception enhancement (SDE): the unconscious tendency of individuals to provide honest but positively biased self-reports to protect self-esteem; 2) impression management (IM): the habitual and conscious tendency of individuals to present themselves of a favourable public image. We report results on both the raw lying profile scores as well as the ones after correcting for the BIDR scores. The correction procedure was conducted by fitting a general linear model that regressed out the SDE and IM scores from each lying profile factor.

2.3 Textual analysis of narrative data

Linguistic Inquiry and Word Count analysis

The Linguistic Inquiry and Word Count (LIWC, Boyd et al., 2022; Pennebaker et al., 2015) software is the gold standard for analysing word usage and semantics in texts across more than 100 features by calculating the percentage of total words corresponding to each category using validated dictionaries of words associated with psychosocial dimensions. Specifically, the English dictionary (version LIWC-22) was employed for this analysis, and 118 features were extracted from tokenized text.

DeCLaRatiVE stylometry

The **DeCLaRatiVE** stylometry approach (Loconte et al., 2023) subsumes 26 linguistic variables derived from four theoretical lines in verbal deception research: Distancing (Newman et al., 2003), Cognitive Load

(Monaro et al., 2018; Vrij et al., 2015), Reality Monitoring (Johnson & Raye, 1981; S. Sporer, 1997), and VErifiability Approach (Nahari et al., 2012; Palena et al., 2021). Linguistic variables associated with the cognitive load, such as text length, readability, and complexity, were computed using the Python library TEXTSTAT. Those related to the Distancing and RM framework were computed using LIWC-22 features extracted from tokenized text. RM was also investigated through linguistic concreteness by cross-referencing an annotated dataset (Brysbaert et al., 2014) with the content words in our dataset and averaging the final scores per statement. The preprocessing steps to derive content words from statements were tokenization, conversion to lowercase, stop-word removal, and lemmatization and were run with the SpaCy library in Python. Finally, verifiable details were extracted as entities with the named-entity recognition (NER) model available on the SpaCy library (en_core_web_trf, https://spacy.io/models/en#en_core_web_trf). A full list of the 26 linguistic variables, along with a brief description, is presented in Table 2 (see Loconte et al., 2023, for a more detailed understanding of the approach).

TABLE 2. List and short description of the 26 linguistic features pertaining to the **DeCLaRatiVE** Stylometry technique.

Label	Description
num_sentences	Total number of sentences
num_words	Total number of words
num_syllables	Total number of syllables
avg_syllabes_per_word	Average number of syllables per word
fk_grade	Index of the grade level required to understand the text
fk_read	Index of the readability of the text
Analytic	LIWC summary statistic analyzing the style of the text in term of analytical thinking (0 - 100)
Authentic	LIWC summary statistic analyzing the style of the text in term of authenticity (0 - 100)
Tone	Standardized difference (0-100) of 'tone_pos' - 'tone_neg'
tone_pos	Percentage of words related to a positive sentiment (LIWC dictionary)
tone_neg	Percentage of words related to a negative sentiment (LIWC dictionary)
Cognition	Percentage of words related to semantic domains of cognitive processes (LIWC dictionary)

Label	Description
memory	Percentage of words related to semantic domains of memory/forgetting (LIWC dictionary)
focuspast	Percentage of verbs and adverbs related to the past (LIWC dictionary)
focuspresent	Percentage of verbs and adverbs related to the present (LIWC dictionary)
focusfuture	Percentage of verbs and adverbs related to the future (LIWC dictionary)
Self-reference	Sum of LIWC categories 'i' + 'we'
Other-reference	Sum of LIWC categories 'shehe' + 'they' + 'you'
Perceptual details	Sum of LIWC categories 'attention' + 'visual' + 'auditory'+ 'feeling'
Contextual Embedding	Sum of LIWC categories 'space' + 'motion' + 'time'
Reality Monitoring	Sum of Perceptual details + Contextual Embedding + Affect - Cognition
Concreteness score	Mean of concreteness score of words
People	Unique named-entities related to people: e.g., 'Mary', 'Paul', 'Adam'
Temporal details	Unique named-entities related to time: e.g., 'Monday', '2:30 PM', 'Christmas'
Spatial details	Unique named-entities related to space: e.g., 'airport', 'Tokyo', 'Central park'
Quantity details	Unique named-entities related to quantities: e.g., '20%', '5 $', 'first', 'ten', '100 meters'

n-gram differentiation

Using the *n*-gram differentiation test (as in Mozes et al., 2021), we compared the frequencies of unigrams, bigrams, and trigrams in truthful and deceptive statements within each event. This comparison was made using a signed rank sum test approach. Ties in ranks were fixed by averaging random ranks in 500 iterations. Statements were first pre-processed using SpaCy library in Python by removing stop words and lemmatising the remaining words. Only *n*-grams that appeared in at least 5% of all documents were included in the analysis. The effect size used for the frequency comparisons was *r*, which ranged from -1.0 to 1.0.

2.4 Machine-learning classification

To investigate whether deceptive statements with embedded lies can be distinguished from truthful statements, we performed a document classification task using different state-of-the-art ML approaches. The models included both traditional and advanced architectures. Specifically, four Random Forest (RF) classifiers were trained on Bag of Words (BOW) representations (Ignatow & Mihalcea, 2017), LIWC variables (Boyd et al., 2022), **DeCLaRatiVE** variables (Loconte et al., 2023), and GPT-embedding representations[10], respectively. Additionally, we tested the performance of different fine-tuned language models, such as distilBERT (Sanh et al., 2019), FLAN-T5 base (Chung et al., 2022), and Llama-3-8B (Sanh et al., 2019). We finally explored the performance of a deception language model from a previous study (Loconte et al., 2023), which was a FLAN-T5 base model fine-tuned on three large datasets of deception with 79.31% (± 1.3) accuracy. Language models (i.e., distilBERT, FLAN-T5, and Llama-3) were trained using the HuggingFace library and the Google Colaboratory Pro + interface with the A100 Tensor Core GPU. Cross-validation was performed to ensure robust evaluation. Specifically, RF models were trained using 10-fold nested cross-validation, while language models were fine-tuned with 5-fold cross-validation to optimize computational costs. Classification performance was assessed in terms of overall accuracy (Formula 1), as well as precision, recall, and F1-score by condition (truthful vs deceptive). Specifically, for each condition, precision measured the proportion of positive predictions that were actually positive (Formula 2); recall measured the proportion of actual positives that were correctly classified as positives (Formula 3); and F1 was the harmonic mean of precision and recall (Formula 4). Details of each model and the training procedure are reported in SM 4.

(1) $Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$

(2) $Precision = \frac{T_P}{T_P + F_P}$

(3) $Recall = \frac{T_P}{T_P + F_N}$

(4) $F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

[10] https://platform.openai.com/docs/guides/embeddings

Annotations: T_P = True positives, T_N = True negatives, F_P = False positives, and F_N = False negatives.

2.5 Analysis Plan

We first looked at the subject level to examine characteristics of the reported embedded lies, such as their frequency, source, deceptiveness, and centrality for a deceptive account. Second, we examined individual differences related to demographic variables and lying profiles. Furthermore, we assessed linguistic differences between the narratives by using the LIWC and **DeCLaRatiVE** approach. From the LIWC, we obtained for each subject 118 variables; from the **DeCLaRatiVE** stylometry technique, we obtained 26 variables. A within-subject permutation t-test with 9,999 permutations (Moore, 1999) was employed to test for statistical differences in these variables by statement veracity (truthful vs deceptive). Results from multiple comparisons were corrected using Bonferroni correction. Truthful and deceptive statements were also analysed in terms of *n*-grams by using the *n*-grams differentiation test. These analyses were conducted in R using the *MKinfer* and *effectsize* libraries. Finally, state-of-the-art machine learning approaches were employed in a classification task to differentiate truthful from deceptive statements with embedded lies.

2.6 Transparency statement

The study was approved by the Ethics Review Board of the Tilburg School of Social and Behavioral Sciences (Reference Number: TSB_RP1442). Data collection reported in this Chapter was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants prior to their involvement in the study, and they were subsequently debriefed once their participation had been completed. Data and scripts used to run the experiments are available at https://osf.io/jzrvh.

3. Results

3.1 Corpus descriptives

We collected a corpus of 2084 truthful and deceptive statements, collectively, across 11 events deemed relevant for lying[11]. We found deceptive statements (M=7.13, SD=4.65) containing a significantly higher average number of sentences than truthful statements (M=6.77, SD=4.48), $t_{(9999)}$= -0.37, p=.003, d =-0.08 [-0.14, -0.03]. Likewise, the number of words was, on average, significantly higher in deceptive (M=145.29, SD=83.65) than in truthful statements (M=131.74, SD=78.49), $t_{(9999)}$=-13.55, p<.001, d=-0.17 [-0.22, -0.12]. However, these findings might be an artefact of the instructions, as participants were instructed to add details to appear deceptive and achieve a specific goal, resulting in producing longer statements.

3.2 Embedded lies

Embedded lies included an average of 5.03 lies per text with an average number of words of 46.27 (SD=42, $Median$=35, see Table 3).

TABLE 3. Descriptive statistics of participants' responses in variables associated with embedded lies (M, SD, Median).

	Embedded lies		
	M	SD	Median
Words	46.27	42.23	35
Absolute no. of embedded lies	5.03	3.35	4
Standardized no. of embedded lies	0.32	0.20	0.29
Deceptiveness	3.94	0.79	4
Centrality	3.55	0.82	3.59

The average ratio between the number of words in the annotated embedded lies and the respective deceptive statement was 0.32 (SD=0.20, $Median$=0.29). The average of embedded lies for each event is reported in Table 4S (SM 2). Using a 5-point scale, embedded lies were rated as moderately deceptive (M=3.94, SD=0.79, $Median$=4) and central to the overall narrative (M=3.55, SD=0.82, $Median$=3.59). Further, 35.86% of embedded lies (n=1881) relied on personal past experiences that involved participants directly and 10.41% (n=546) indirectly; 33.86% of embedded lies (n=1776) relied on participants' imagination, while 14.95% (n=784) on

[11] Descriptive statistics of the variables associated with the events are reported in Table 3S in SM 2.

others' experiences and only 4.92% (n=258) on personal future plans. An example of subjects' responses is provided in Box 1. Correlational analysis between variables associated with embedded lies, lying profile and BIDR scales is provided in SM 2.

BOX 1. Example of a statement provided by participants during the task

EVENT: Being involved in a car accident	**EVENT: Being involved in a car accident** **INSTRUCTIONS: lie about the event to increase the claimed amount of damage you received to get more money**
«I was driving home after getting my dog from her sitter. My mom was sitting next to me to keep company to the dog, when we got met with a lot of traffic. So we were advancing quite slowly towards our destination when we come across this intersection, where on the right the cars have a STOP sign. This guy, very old, probably in his 60s, doesn't stop and continues moving towards us. I stomped on the break, but it wasn't in time, and the car crashed against our side. It had been years since I was involved in anything of the sort, so while I was pretty sure it was not my fault, I was shaking the entire time I was dealing with the men to fill out the paperwork».	*«I was driving home after I got my dog from her sitter. My mom and dog were sitting in the passenger seat,* ***my dog likes to ride on the ground between my moms legs****. We came across quite a bit of traffic and were moving slowly towards our destination. As I move through this intersection, where cars on the right have a STOP sign (so they have to stop, and I have priority), this guy thats at least 60 years old, completely ignores the sign and advances towards us at quite a speed. Because there was traffic in front of me, I* ***could nothing but watch*** *as the car crashed into us, directly on the passengers side.* ***My mom was thrown to my side****, kept in place only by the seat belt, and* ***her leg was pretty badly hurt****, cause* ***she used her body to protect our doggy****. This experience is clear in my mind because after I got off* ***I had to have a fight with the other driver because he was incapable of acknowledging fault****».*

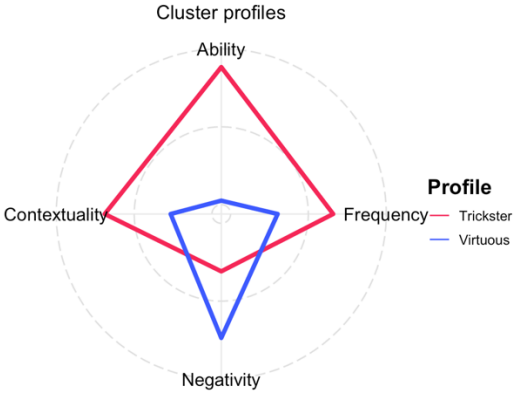
Note. On the left side is a sample truthful statement from a participant about being involved in a car accident. On the right side, the same participant provided the deceptive statements about the same event following the given instructions. In **bold**, the embedded lies identified by the participant.

3.3 Individual differences

We investigated individual differences in the absolute and standardized number of embedded lies, deceptiveness, and centrality scores. Regarding demographic factors (see also SM 2), we found a gender difference for the average deceptiveness scores ($diff$=0.11 ± 0.05, p=0.03, d=0.14 [0.02, 0.26]), with females (M=3.98, SD=0.79) reporting higher values than

males (M=3.88, SD=0.77). As for age, we found a small, significant positive correlation between age and deceptiveness (rho=0.075, S=172823388, p=0.015).

FIGURE 3. Radar plot of the average values at the four lying profile factors in the trickster and virtuous cluster.



Note. The scores at the lying profile factors are corrected for Social Desirability.

Furthermore, we investigated the presence of subpopulations of liars by a cluster analysis of participants' scores in the four-factor lying profile questionnaire (Makowski et al., 2023). Lying profile scores were first adjusted for social desirability[12]. We then followed the procedure in Makowski et al. (2023) to cluster participants (see SM 3). Our dataset was deemed suitable for clustering (Hopkins' H=0.25)[13]. The method agreement procedure supported the existence of two clusters, as indicated by ten methods out of 29 (34.48%). After applying the k-means clustering algorithm, the two clusters accounted for 31.97% of the total variance of

[12] The correction procedure employed a Generalized Linear Model approach to regress out the scores of each lying profile factor (i.e., LIE_Ability, LIE_Contextuality, LIE_Frequency, LIE_Negativity) for social desirability effects (i.e., SDE and IM). The adjusted scores were calculated using the *adjust* function from the *datawizard* package in Rstudio.

[13] The Hopkins' H statistic checks whether the data is appropriate for clustering. We can reject the null hypothesis and conclude that the dataset is significantly clusterable when H < 0.5. A value for H lower than 0.25 indicates a clustering tendency at the 90% confidence level (Lawson & Jurs, 1990). The H statistic was computed using the *check_clusterstructure* function from the *performance* package in Rstudio.

the original data. The first cluster (44.72% of the sample) was characterized by participants with very low reported lying ability, low levels of frequency and contextuality, and strong negative attitudes towards lying; the second cluster (55.28% of the sample) was characterized by people with higher levels of contextuality and frequency of deception, very high levels of ability and low levels of negative attitudes towards lying (Fig. 3). Following the original work (Makowski et al., 2023), we labelled the first cluster as the virtuous and the second as the trickster cluster. To test the validity of this two-cluster solution, we trained a logistic regression that used, as features, the adjusted scores of the four scales of the lying profile questionnaire and, as a predicted variable, the labels obtained from the cluster analyses (as in Bambini et al., 2022). We obtained an almost perfect classification (*accuracy*=0.99). This result supported the validity of our two-cluster solution, confirming that the labels associated with each participant were not randomly assigned but actually reflected an inherently different pattern of responding. However, no significant differences were found in any dependent variable in the two groups (see Table 6S in SM 3).

3.4 Textual analysis of narrative data

Tables 4 and 5 suggest that a few linguistic indicators were significantly indicative of deception, albeit often with small effect sizes. LIWC variables associated with deceptive statements pertained mainly to using emotional and social words and references (i.e., social words, social references, pronouns and personal pronouns, social behaviour, language of status and leadership; Table 4). In contrast, LIWC features associated with truthfulness included mainly words associated with memory (i.e., remember, forget, remind) and numbers.

When we conducted the analysis by event, significant differences emerged for some LIWC variables by statement veracity for four events (i.e., Being hospitalized and undergoing surgery, Ending a long romantic relationship, Getting a speeding fine, and Taking the bus/train without the ticket).

TABLE 4. Effect sizes (and CIs) of significant LIWC features for the entire dataset and specific events.

Topic	LIWC feature	LIWC interpretation	LIWC example words	Cohen's *d*	Adjusted CI	Direction
Overall	Social	Social words	Argue, boyfriend, chat	-0.22	-0.33, -0.11	D > T
	WC	Total word counts	-	-0.19	-0.30, -0.08	D > T

Topic	LIWC feature	LIWC interpretation	LIWC example words	Cohen's *d*	Adjusted CI	Direction
	Period	Number of periods	.	0.19	0.07, 0.29	T > D
	socrefs	Social references	you, we, he, she	-0.18	-0.29, -0.07	D > T
	shehe	Third singular personal pronouns	she, he, her, his	-0.18	-0.29, -0.07	D > T
	ppron	Personal pronouns	I, you, my, me	-0.17	-0.28, -0.06	D > T
	memory	Memory words	remember, forget, remind	0.17	0.06, 0.28	T > D
	socbehav	Social behavior words	said, love, say, care	-0.15	-0.25, -0.04	D > T
	male	Male references	he, his, him, man	-0.14	-0.25, -0.03	D > T
	number	Numbers	one, two, first, once	0.14	0.03, 0.25	T > D
	emo_anger	Emotion of anger	hate, mad, angry, frustr*	-0.14	-0.25, -0.03	D > T
	pronoun	Pronouns	I, you, that, it	-0.13	-0.24, -0.02	D > T
	det	Determiners	the, at, that, mine	-0.13	-0.23, -0.02	D > T
	Clout	Language of leadership, status	-	-0.11	-0.22, -0.001	D > T
Being hospitalized and undergoing surgery	Tone	Emotional tone	-	0.49	0.04, 0.94	T > D
	Period	Number of periods	.	0.48	0.03, 0.92	T > D
	WC	Total word count	-	-0.47	-0.92, -0.02	D > T
	power	Words of power	own, order, allow, power	-0.45	-0.89, -0.002	D > T
Ending a long romantic relationship	emo_anger	Emotion of anger	hate, mad, angry, frustr*	-0.41	-0.71, -0.12	D > T
	conflict	Conflict words	fight, kill, killed, attack	-0.36	-0.66, -0.06	D > T
	ppron	Personal pronouns	I, you, my, me	-0.34	-0.63, -0.04	D > T

Topic	LIWC feature	LIWC interpretation	LIWC example words	Cohen's *d*	Adjusted CI	Direction
	socbehav	Social behavior words	said, love, say, care	-0.32	-0.61, -0.02	D > T
Getting a speeding fine	Social	Social words	Argue, boyfriend, chat	-0.75	-1.26, -0.24	D > T
	socrefs	Social references	you, we, he, she	-0.70	-1.20, -0.19	D > T
	shehe	Third singular personal pronouns	she, he, her, his	-0.68	-1.18, -0.18	D > T
	Clout	Language of leadership, status	-	-0.56	-1.04, -0.07	D > T
Taking the bus/train without the ticket	Social	Social words	Argue, boyfriend, chat	-0.65	-1.00, -0.30	D > T
	shehe	Third singular personal pronouns	she, he, her, his	-0.57	-0.92, -0.23	D > T
	socrefs	Social references	you, we, he, she	-0.56	-0.90, -0.22	D > T
	male	Male references	he, his, him, man	-0.54	-0.88, -0.20	D > T
	socbehav	Social behavior words	said, love, say, care	-0.50	-0.84, -0.16	D > T
	comm	Communication words	said, say, tell, thank*	-0.46	-0.79, -0.13	D > T
	ppron	Personal pronouns	I, you, my, me	-0.41	-0.74, -0.08	D > T
	pronoun	Pronouns	I, you, that, it	-0.41	-0.74, -0.07	D > T
	Cognition	Words of Cognition	know, think, but, if	0.37	0.04, 0.70	T > D
	WC	Total word count	-	-0.35	-0.68, -0.03	D > T
	tentat	Words of tentativeness	if, or, any, something	0.34	0.01, 0.67	T > D
	visual	Visual words	see, lool, eye*, saw	-0.33	-0.65, -0.002	D > T

Note. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. Linguistic features are sorted by the absolute value of the effect size magnitude for each event. For the direction of the effect, T = truthful and D = deceptive.

For **DeCLaRatiVE** linguistic features (Table 5), only a few of them were significantly indicative of deception in five out of eleven events. When testing the whole dataset, the only significant features for deceptive statements were references to others, the number of words and number of syllables. In contrast, significant features for truthful statements were memory-related words and temporal details.

TABLE 5. Effect sizes (and CIs) of significant **DeCLaRatiVE** features for the entire dataset and specific events.

Topic	DeCLaRatiVE feature	Cohen's *d*	Adjusted CI	Direction
Overall	Other-reference	-0.19	-0.29, -0.10	D > T
	num_syllables	-0.19	-0.29, -0.09	D > T
	num_words	-0.19	-0.28, -0.09	D > T
	memory	0.17	0.07, 0.27	T > D
	Temporal details	0.10	0.0004, 0.19	T > D
A job interview for your dream job	memory	0.25	0.003, 0.51	T > D
Being hospitalized and undergoing surgery	Tone	0.49	0.09, 0.88	T > D
	num_words	-0.47	-0.87, -0.08	D > T
	num_syllables	-0.47	-0.86, -0.08	D > T
Getting a speeding fine	Other-reference	-0.66	-1.10, -0.22	D > T
Missing a deadline at work because of bad organization	Other-reference	-0.35	-0.67, -0.02	D > T
	memory	0.33	0.006, 0.66	T > D
Taking the bus without the train ticket	Other-reference	-0.52	-0.82, -0.22	T > D
	Cognition	0.37	0.08, 0.66	T > D
	num_words	-0.35	-0.64, -0.06	D > T
	num_syllables	-0.35	-0.64, -0.07	D > T

Note. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. Linguistic features are sorted by the absolute value of the effect size magnitude for each event. For the direction of the effect, T = truthful and D = deceptive.

Finally, the *n*-gram differentiation analysis (Table 6) revealed how deceptive statements with embedded lies may appear very similar to their truthful counterparts, resulting in few or no significant

differences in word usage. This result highlights the reasons why detecting embedded lies is a hard task.

TABLE 6. Effect sizes (*r*) and CIs of significant n-grams for specific events after using the n-grams differentiation test.

Event	*n*-gram	r	Adjusted CI	Direction
Taking the bus/train without the train ticket	tell	-0.20	-0.37, -0.02	D > T
	ticket	-0.18	-0.23, -0.14	D > T
	time	-0.14	-0.26, -0.01	D > T
Ending a long romantic relationship	relationship	-0.07	0.001, 0.13	T > D
Missing a deadline at work because of bad organisation	time	0.10	0.004, 0.20	T > D
Cheating on your partner	feel	0.33	0.06, 0.60	T> D
Being hospitalized and undergoing surgery	pain	-0.22	-0.38, -0.06	D > T
	surgery	-0.14	-0.22, -0.05	D > T
Getting fired	fire	0.25	0.09, 0.40	T > D
Getting a speeding fine	speed	0.10	0.005, 0.20	T > D
Cheating on an exam	study	-0.28	-0.43, -0.13	D > T
	answer	0.19	0.01, 0.36	T > D
Causing a car accident	drive	-0.21	-0.38, -0.03	D > T

Note. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. *N*-grams are sorted by effect size after comparing truthful and deceptive statements for each event. For the direction of the effect, T = truthful and D = deceptive.

3.5 Predictive modelling performance

We trained different machine learning and language models to distinguish deceptive statements with embedded lies from truthful ones. Table 7 shows that all models could classify statements better than the chance level (with $p < .01$ after running an exact binomial test), but the highest performance reached 64% accuracy after fine-tuning a Llama-3 model.

3.6 Exploratory explainability analysis

To add interpretations to the achieved performance, we conducted an explainability analysis on the Llama-3 and deception language model. We computed Spearman's rank correlations between the deceptive class probabilities and the absolute and standardized number of embedded

lies, deceptiveness, and centrality scores (Table 7S in SM 4). There was a significant positive correlation between the class probability of deceptiveness and the absolute (*rho*=.10, *S*=170216978, $p < .01$) and standardized number of embedded lies (*rho*=.10, *S*=170565831, $p = .001$). For the deception language model, we found a significant positive correlation between the absolute number of embedded lies and class probability (*rho*=.09, *S*=171758230, *p*=.004). Finally, only for the Llama-3 model, we found correct classifications having a significantly higher amount of absolute number of embedded lies (*M*=5.31, *SD*=3.39) compared to incorrect ones (*M*=4.43, *SD*=2.83), *d*=0.27 [0.14, 0.40]. Similarly, a standardized number of embedded lies was significantly higher in correctly classified statements (*M*=0.34, *SD*=0.21) with respect to incorrect ones (*M*=0.29, *SD*=0.19), *d*=0.22 [0.09, 0.35] (see Table 8S in SM 4). These findings suggest that the more a statement is fabricated, namely, the greater the number of embedded lies within an otherwise truthful statement, the higher the probability of a language model to accurately and confidently predict the class of that statement.

TABLE 7. Classification performance of predictive models.

		Truthful			Deceptive		
Model	Accuracy	Precision	Recall	F1	Precision	Recall	F1
BOW + RF	.55 (.03)	.55 (.03)	.53 (.04)	.54 (.03)	.55 (.03)	.56 (.04)	.55 (.03)
LIWC + RF	.57 (.03)	.58 (.03)	.55 (.03)	.57 (.03)	.57 (.03)	.60 (.05)	.58 (.04)
DeCLaRatiVE + RF	.56 (.02)	.56 (.02)	.55 (.06)	.55 (.04)	.56 (.02)	.56 (.06)	.56 (.03)
GPT-embeddings + RF	.62 (.03)	.62 (.03)	.62 (.05)	.62 (.03)	.62 (.03)	.62 (.05)	.62 (.03)
distilBERT	.60 (.02)	.64 (.05)	.51 (.19)	.55 (.10)	.60 (.05)	.69 (.16)	.63 (.05)
Fine-tuned FLAN-T5 base	.60 (.02)	.60 (.06)	.57 (.03)	.59 (.03)	.59 (.04)	.63 (.04)	.61 (.01)
Fine-tuned Llama-3-8B	**.64 (.04)**	**.67 (.05)**	**.55 (.13)**	**.60 (.08)**	**.62 (.05)**	**.73 (.10)**	**.67 (.05)**
Deception language model	.56	.54	.76	.63	.60	.35	.44

Note. The values refer to the average performance after performing cross-validation. In brackets, the standard deviation is reported. The deception language model was only employed to predict the class in our dataset; therefore, no cross-validation was performed. All models were significantly better than the chance level with $p < .01$. In bold is the performance of the best model.
Abbreviations: BOW = bag of words; RF = random forest; LIWC = Linguistic inquiry and word count.

4. Discussion

4.1 Moving forward on embedded lies

In this Chapter, we sought to spark renewed research interest in verbal deception detection and to move the dial towards embedded lies. With this aim, we presented a dataset of 2084 statements (i.e., truthful vs. deceptive with embedded lies) about eleven categories of autobiographical events deemed relevant for lying. We focused on autobiographical memories because of their relevance in forensic contexts, where the credibility of witnesses' and suspects' statements is assessed and often centred on autobiographical events. Additionally, this new dataset was collected in a within-subjects design, providing data at the statement and the individual level. Specifically, it provides granular information on the statement level, including annotations and ratings of embedded lies and memory-related measures about each event (e.g., how in the past, how frequently it is remembered, how important it is, etc.) and demographic data and personality-related measures, such as attitudes towards lying and social desirability, at the individual level. We believe this resource might be valuable in fostering psychological research on linguistic, contextual, and individual differences associated with embedded lies.

4.2 The nature of embedded lies

Our findings suggest that participants used, on average, five embedded lies in their statements to achieve a predefined deception goal. About 1/3 of the length of deceptive statements were embedded lies. Similar figures are reported in previous research(Markowitz, 2024) for embedded lies in faked opinions about friends (37%). As for the source of embedded lies, most participants relied, whether directly or indirectly, on their personal experiences (46.27%), while a smaller percentage used their imagination (33.86%) or drew from others' experiences (14.95%). This finding supports the notion that liars often integrate elements of truth into their lies to enhance plausibility, making the detection of deception more difficult (Markowitz, 2024). A realistic deceptive statement (i.e., one with embedded lies rather than a full-blown deceptive narrative) can thus be typified as one that consists of about 2/3 of truthful information and 1/3 of embedded lies, which are most likely to be derived from personal experience.

4.3 Individual differences in embedded lies

We further investigated individual differences in the nature of embedded lies. We found gender playing a role in how individuals self-rated the deceptiveness of their embedded lies, with females scoring higher in deceptiveness than males, albeit with small effect sizes. Age also played a role, with older participants being more openly deceptive in their statements.

In terms of lying attitude, the results of the cluster analysis were slightly different from the original paper (Makowski et al., 2023). We identified only two, rather than three, clusters of liars that resembled the original *virtuous* and *trickster* clusters. Specifically, the virtuous cluster was mainly characterised by a strong aversion to deception, while the tricksters tended to lie more frequently, to perceive themselves as good liars, and to adapt their lying behaviour to the context. However, despite this clear distinction, no significant differences were reflected in their behaviour and, specifically, in the absolute and standardised number of embedded lies, as well as in their deceptiveness and centrality scores. A possible explanation for why the difference in the lying attitude was not reflected in the lying behaviour (i.e., in the number of embedded lies) might be that all participants were instructed to write the statement deceptively by adding embedded lies, and this might have reduced the variability in their responses.

4.4 Textual properties of embedded lies

In addition to individual differences among liars, we examined linguistic properties of embedded lies by leveraging automated NLP techniques. Linguistic analysis using psycholinguistic variables and a deception-specific set of variables (**DeCLaRatiVE**) revealed few differences between truthful statements and those with embedded lies, with small effect sizes. Deceptive statements contained a larger proportion of social references, while truthful statements tended to include more references to memory processes. Similarly, the **DeCLaRatiVE** analysis suggested that deceptive statements contained more references to other people, which is in line with the distancing framework (Newman et al., 2003) and a higher number of words and syllables, which may reflect the experimental instructions that encouraged participants to add more details to achieve their deceptive goals. Conversely, truthful statements contained more memory-related words and temporal details, which is more in line with the Reality Monitoring framework (Johnson & Raye, 1981; S. L. Sporer, 2004).

When we zoomed in on the event level, we found significant differences in LIWC variables only in four out of eleven events and in **DeCLaRatiVE** variables in five out of eleven events. Altogether, these findings suggest that while there are some discernible differences between truthful and deceptive statements, these differences are often subtle and context-dependent. This is also in line with previous studies showing that truthful statements do not necessarily contain more details than embedded lies (Markowitz, 2024; Verigin, Meijer, Vrij, et al., 2020).

A term frequency analysis of n-grams underscored the difficulty of detecting deception through word usage when embedded lies are involved. In nine out of eleven events, we found negligible effects, with only one or two significant *n*-grams per event (e.g., "pain" and "surgery" as significant n-grams in deceptive statements for the event "Being hospitalized and undergoing surgery") and with small effect size, highlighting the subtle nature of embedded lies. This supports previous findings that verbal detection remains challenging, particularly when lies are carefully embedded within otherwise truthful narratives (Markowitz, 2024; Vrij et al., 2010). In addition, this overlap can be attributed to the within-subject design employed for this study, which eliminated any potential linguistic confounders derived from having different participants write about the same task under two conditions (honest vs. deceptive), typical of between-subjects studies.

4.5 Detecting embedded lies

By collecting a dataset that was sufficiently large to perform predictive modelling, we resorted to simpler supervised approaches based on machine learning models trained on extracted features and on state-of-the-art language models to classify statements as completely truthful or with embedded lies. The results showed that embedded lies present a significant challenge for deception detection due to their incorporation of truthful elements. Specifically, the highest performance of a language model with competitive capabilities (Llama-3.1-8B, Grattafiori et al., 2024), which we fine-tuned for this specific task, reached 64% accuracy. The result from our Llama model was in line with commonly reported performances in previous research (Fornaciari & Poesio, 2012b; Kleinberg & Verschuere, 2021; Rubin & Conroy, 2012, 2011). Notably, a language model published in a previous study (Loconte et al., 2023), which reported an accuracy of 79.31% in detecting fabricated statements across different contexts, dropped to 56% accuracy when applied to our study. An explanation for that drop could be attributed to overfitting. Related works have shown, in fact, that deception classifiers drop remarkably when tested on new samples (Kleinberg et al., 2018). However, we argue this was not the case. In the original study, the detection rate (i.e., the

recall) for truthful (81%) and deceptive statements (78%) was balanced. In contrast, in our study, the deception language model showed a recall of 76% for truthful statements, similar to that of the original study, but a remarkable drop to 35% specifically for deceptive statements. This drop indicates that the struggle was mainly in the detection of embedded lies, which were often misclassified as truthful statements (here: 65% of embedded lies were misclassified as truthful, vs. 22% in the original study). If it were a matter of overfitting, we would have also expected a remarkable drop in the recall of truthful statements. However, this decline was not observed, which indicates that while the deception model was able to resort to what was learnt during the training phase to classify new samples of truthful statements correctly, it was unable to do so for the deceptive ones. We argue this was attributed to the fact that the deception involved was different (here: embedded lies vs. fabrication in the original study). Moreover, the explainability analyses on the Llama-3 and deception language model provided further evidence for the notion that the more nuanced the embedded lies are, the harder they are to detect.

Finally, when employing other common approaches, typically employed to detect deception (i.e., ML models trained on BOW representation, LIWC features, and embeddings), performance was significantly better than chance – albeit reaching just 55% to 62% accuracy. Other fine-tuned language models (here: distilBERT and FLAN-T5 base) were no more effective in performing the task. Altogether, these findings indicate that the challenge in identifying embedded lies stems from their resemblance to truthful statements and, as the degree of fabrication increases, the classification process becomes more straightforward.

4.6 Limitations and future outlooks

Despite the study's aim to overcome known limitations related to deception detection research (e.g., focus on fabrication, use of between-subjects designs, and small sample sizes), it comes with its own limitations in methodology and findings.

Regarding methodology, embedded lies were both self-reported and self-annotated by the participants, leading to subjective interpretations of what constitutes an embedded lie. This subjectivity could reduce the consistency and reliability of the data. In our analysis, we standardized the number of embedded lies by computing the ratio of words in embedded lies to the total number of words in the deceptive statement to ensure that the results were not influenced by individual interpretations of what a unit of embedded lie was. We recommend future researchers adopting this or other forms of standardization (even during the data collection

process) to ensure consistency. Second, while we recruited a Prolific sample that was sufficiently large to conduct meaningful computational analysis, these findings should be replicated with laboratory experiments where participants are in contact with the interviewer and can offer a longer verbal narrative, instead of a short written account. Additionally, previous research showed that more proactive interviewing techniques, such as the strategic use of evidence, the use of unexpected questions, or the Reality Interview (see Vrij et al., 2022, for an overview of these approaches), increase differences between truth-tellers and lie-tellers, enhancing deception detection rates. Therefore, further investigation on the detection of embedded lies using these interviewing approaches is needed, as they might promise higher accuracy rates. Third, while the dataset covers eleven distinct events, focusing the investigation on individual events, it may result in smaller sample sizes, limiting the statistical power and the ability to conduct predictive analyses within specific events. Finally - and in contrast to the study design employed by Markowitz (2024) - we conceptualised truthful statements as entirely truthful, while deceptive statements were situated on a continuum ranging from embedded lies to completely fabricated statements. While it is reasonable that individuals may occasionally offer partial truths, it is also feasible to convey completely truthful statements. Consequently, we opted to narrow our focus of investigation by contrasting completely truthful statements with varying degrees of embedded lies. A potential avenue for future research could involve incorporating partial truths, as Markowitz (2024) did in his design, or alternatively, having three versions of the statement: truthful, embedded lies, and fully deceptive.

Regarding findings, we focused on predictive modelling, with the task being conceptualized as a classification task (i.e., whether a statement is truthful or contains embedded lies). However, future studies can go beyond this binary classification and conceptualize the task as a regression task where ML models quantify the extent of deception (e.g., the number of embedded lies) in a given statement. Additionally, future studies might focus on a sequence classification task to predict how and where lies are embedded within truthful narratives. Another obvious limitation is that the models tested were not compared with truth/lie judgments of untrained humans. Although meta-analytical evidence indicates that untrained judges perform close to the chance level (Bond & DePaulo, 2006), this finding is worth replication when lying involves embedded lies. Further, previous theoretical frameworks of deception and theories relying on manual coding, such as the use-the-best heuristic (Verschuere et al., 2023), the verifiability approach (Nahari et al., 2012), as well as the role of complications, common knowledge details, or self-handicapping

strategies (Caso et al., 2023), should be tested on this new dataset of deception to provide novel insights on what works on embedded lies.

Conclusion

In this Chapter, we addressed the first challenge for automated verbal deception detection: the robustness against a more nuanced form of deception. To this aim, we presented a novel dataset as a resource to encourage research on embedded lies in verbal deception detection. The analysis of individual differences and linguistic properties, as well as the results from predictive modelling and explainability analysis, highlighted how the unique challenge in detecting embedded lies stems from their mixed nature and resemblance to truthful statements.

In the next Chapter, we will address another challenge for automated verbal deception detection: human adoption or aversion toward algorithmic predictions. We examined the opportunities of automated methods for verbal deception detection for data coding and prediction in Chapters 2 and 3; however, especially for the case of automated predictions, little is known about the extent to which humans rely on such predictions. This question will be further investigated in the next Chapter.

References

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/J.IJCHP.2016.01.002

Bambini, V., Frau, F., Bischetti, L., Cuoco, F., Bechi, M., Buonocore, M., Agostoni, G., Ferri, I., Sapienza, J., Martini, F., Spangaro, M., Bigai, G., Cocchi, F., Cavallaro, R., & Bosia, M. (2022). Deconstructing heterogeneity in schizophrenia through language: a semi-automated linguistic analysis and data-driven clustering approach. *Schizophrenia 2022 8:1*, *8*(1), 102-. https://doi.org/10.1038/s41537-022-00306-z

Bell, K. L., & DePaulo, B. M. (1996). Liking and lying. *Basic and Applied Social Psychology*, *18*(3), 243–266. https://doi.org/10.1207/S15324834BASP1803_1

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personal. Soc. Psychol. Rev.*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app .

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Caso, L., Cavagnis, L., Vrij, A., & Palena, N. (2023). Cues to deception: can complications, common knowledge details, and self-handicapping strategies discriminate between truths, embedded lies and outright lies in an Italian-speaking sample? *Frontiers in Psychology*, *14*. https://doi.org/10.3389/FPSYG.2023.1128194

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). *Scaling Instruction-Finetuned Language Models*. http://arxiv.org/abs/2210.11416

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*,

18(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Fitzpatrick, E., Bachenko, J., & Fornaciari, T. (2015). *Automatic Detection of Verbal Deception*. https://doi.org/10.1007/978-3-031-02158-9

Fornaciari, T., & Poesio, M. (2012a). *DeCour: a corpus of DEceptive statements in Italian COURts* (pp. 1585–1590). http://www.lrec-conf.org/proceedings/lrec2012/pdf/377_Paper.pdf

Fornaciari, T., & Poesio, M. (2012b). *On the Use of Homogenous Sets of Subjects in Deceptive Language Analysis*. 39–47. https://aclanthology.org/W12-0406.pdf

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/EJPALC2021A10

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models*. https://arxiv.org/abs/2407.21783v3

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being Honest about Dishonesty: Correlating Self-Reports and Actual Lying. *Human Communication Research*, *40*(1), 54–72. https://doi.org/10.1111/HCRE.12019

Hart, C. L., Lemon, R., Curtis, D. A., & Griffith, J. D. (2020). Personality Traits Associated with Various Forms of Lying. *Psychological Studies*, *65*(3), 239–246. https://doi.org/10.1007/S12646-020-00563-X/TABLES/4

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding Short Form (BIDR-16). *SAGE Open*, *5*(4). https://doi.org/10.1177/2158244015621113

Hartwig, M., Granhag, P. A., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime & Law*, *13*(2), 213–227. https://doi.org/10.1080/10683160600750264

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Ignatow, G., & Mihalcea, R. (2017). Text Mining: A Guidebook for the Social Sciences. *Text Mining: A Guidebook for the Social Sciences*. https://doi.org/10.4135/9781483399782

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*(1), 67–85. https://doi.org/10.1037/0033-295X.88.1.67

Jones, D. N., & Paulhus, D. L. (2017). Duplicity among the Dark Triad: Three faces of deceit. *Journal of Personality and Social Psychology*, *113*(2), 329–342. https://doi.org/10.1037/PSPP0000139

Kashy, D. A., & DePaulo, B. M. (1996). Who Lies? *Journal of Personality and Social Psychology*, *70*(5), 1037–1051. https://doi.org/10.1037/0022-3514.70.5.1037

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/JOURNAL.PONE.0220228

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2017). Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, *32*(3), 354–366. https://doi.org/10.1002/ACP.3407

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica*, *213*. https://doi.org/10.1016/J.ACTPSY.2020.103250

Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information & Computer Sciences*, *30*(1), 36–41. https://doi.org/10.1021/CI00065A010

Leins, D. A., Fisher, R. P., & Ross, S. J. (2013). Exploring liars' strategies for creating deceptive reports. *Legal and Criminological Psychology*, *18*(1), 141–151. https://doi.org/10.1111/J.2044-8333.2011.02041.X

Leins, D. A., Zimmerman, L. A., & Polander, E. N. (2017). Observers' real-time sensitivity to deception in naturalistic interviews. *Journal of Police and Criminal Psychology*, *32*(4), 319–330. https://doi.org/10.1007/S11896-017-9224-2

Levine, T. R., Serota, K. B., Carey, F., & Messer, D. (2013). Teenagers Lie a Lot: A Further Investigation into the Prevalence of Lying. *Communication Research Reports*, *30*(3), 211–220. https://doi.org/10.1080/08824096.2013.806254

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. *Scientific Reports 2023 13:1*, *13*(1), 1–19. https://doi.org/10.1038/s41598-023-50214-0

Makowski, D., Pham, T., Lau, Z. J., Raine, A., & Chen, S. H. A. (2023). The structure of deception: Validation of the lying profile questionnaire. *Current Psychology*, *42*(5), 4001–4016. https://doi.org/10.1007/S12144-021-01760-1

Markowitz, D. M. (2024). Deconstructing deception: Frequency, communicator characteristics, and linguistic features of embeddedness. *Applied Cognitive Psychology*, *38*(3), e4215. https://doi.org/10.1002/ACP.4215

Mihalcea, R., & Strapparava, C. (2009). The lie detector. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, 309. https://doi.org/10.3115/1667583.1667679

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Scientific Reports*, *8*(1). https://doi.org/10.1038/S41598-018-20462-6

Monaro, M., Maldera, S., Scarpazza, C., Sartori, G., & Navarin, N. (2022). Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. *Computers in Human Behavior*, *127*. https://doi.org/10.1016/J.CHB.2021.107063

Moore, J. H. (1999). Bootstrapping, permutation testing and the method of surrogate data. *Phys. Med. Biol.*, *44*(6), L11. https://doi.org/10.1088/0031-9155/44/6/101

Mozes, M., van der Vegt, I., & Kleinberg, B. (2021). A repeated-measures study on emotional responses after a year in the pandemic. *Scientific Reports 2021 11:1*, *11*(1), 23114-. https://doi.org/10.1038/s41598-021-02414-9

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*(2), 227–239. https://doi.org/10.1111/J.2044-8333.2012.02069.X

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 309–319.

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2021). The Verifiability Approach: A Meta-Analysis. *Journal of Applied Research in Memory and Cognition*, *10*(1), 155–166. https://doi.org/10.1016/J.JARMAC.2020.09.001

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Pennebaker, J. W., Botth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic Inquiry and Word Count: LIWC2015. In *Austin, TX: Pennebaker Conglomerates*.

Rubin, V. L., & Conroy, N. (2012). Discerning truth from deception: Human judgments and automation efforts. *First Monday*, *17*(3). https://doi.org/10.5210/FM.V17I3.3933

Rubin, V. L., & Conroy, N. J. (2011). Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4. https://doi.org/10.1002/MEET.2011.14504801098

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/abs/1910.01108v4

Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proc. Natl. Acad. Sci.*, *119*(45), e2211715119. https://doi.org/10.1073/pnas.2211715119

Semrad, M., Scott-Parker, B., & Nagel, M. (2019). Personality traits of a good liar: A systematic review of the literature. *Personality and Individual Differences*, *147*, 306–316. https://doi.org/10.1016/J.PAID.2019.05.007

Serota, K. B., & Levine, T. R. (2015). A Few Prolific Liars. *Journal of Language and Social Psychology*, *34*(2), 138–157. https://doi.org/10.1177/0261927X14528804

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The Prevalence of Lying in America: Three Studies of Self-Reported Lies. *Human Communication Research*, *36*(1), 2–25. https://doi.org/10.1111/J.1468-2958.2009.01366.X

Sporer, S. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Appl. Cognit. Psychol.*, *11*(5), 373–397.

Sporer, S. L. (2004). Reality monitoring and detection of deception. *The Detection of Deception in Forensic Contexts*, 64–102. https://doi.org/10.1017/CBO9780511490071.004

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. 1. https://arxiv.org/abs/1706.03762v7

Verigin, B. L., Meijer, E. H., Bogaard, G., & Vrij, A. (2019). Lie prevalence, lie characteristics and strategies of self-reported good liars. *PLoS ONE*, *14*(12). https://doi.org/10.1371/JOURNAL.PONE.0225566

Verigin, B. L., Meijer, E. H., & Vrij, A. (2020). A within-statement baseline comparison for detecting lies. *Psychiatry, Psychology, and Law : An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *28*(1), 94–103. https://doi.org/10.1080/13218719.2020.1767712

Verigin, B. L., Meijer, E. H., Vrij, A., & Zauzig, L. (2020). The interaction of truthful and deceptive information. *Psychology, Crime & Law*, *26*(4), 367–383. https://doi.org/10.1080/1068316X.2019.1669596

Verschuere, B., Bogaard, G., & Meijer, E. (2021). Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis. *Applied Cognitive Psychology*, *35*(2), 374–384. https://doi.org/10.1002/ACP.3775

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 2023 7:5*, *7*(5), 718–728. https://doi.org/10.1038/s41562-023-01556-2

Vrij, A. (2016). Baselining as a Lie Detection Method. *Applied Cognitive Psychology*, *30*(6), 1112–1119. https://doi.org/10.1002/ACP.3288

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/JIP.82

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal Lie Detection: Its Past, Present and Future. *Brain Sciences*, *12*(12). https://doi.org/10.3390/BRAINSCI12121644

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, Supplement*, *11*(3), 89–121. https://doi.org/10.1023/B:GRUP.0000021838.66662.0c

Wang, G., Chen, H., & Atabakhsh, H. (2004). Criminal identity deception and deception detection in law enforcement. *Group Decision and Negotiation*, *13*(2), 111–127.

Weiss, B., & Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews. *Journal of Applied Social Psychology*, *36*(4), 1070–1086. https://doi.org/10.1111/J.0021-9029.2006.00055.X

Supplementary Material - 1

Instructions

TABLE 1S. Full instructions provided to participants when the order of presentations of conditions was first truthful and then deceptive.

Truthful condition	Deceptive condition
Your task is to write about the event **"Being involved in a car accident"** *twice. For now, provide a* **completely truthful statement**. *Then, you are going to write an alternative version of the same event following additional instructions. This means that* **for now** *you should provide a* ***detailed, truthful account of that event.***	*In the previous task, you wrote about* **"Being involved in a car accident"**
Make sure to use correct spelling and grammar and separate your sentences with punctuation.	*This time you have to* ***lie about the event*** *you just described in order to obtain a benefit or avoid a loss. Specifically, you have to write an alternative version of the story about* **"Being involved in a car accident"** *in which you are deceptive to* **increase the claimed amount of damage you received to get more money.**
Describe what happened, who was involved, where and when it took place, and why it was memorable to you. Your statement should be at least 300 characters.	*Specifically,* **the event should essentially be the same, but you have to fabricate details**. *Afterward, we are going to ask you in which specific part of the story you lied and how.*
IMPORTANT: *We are aware that AI-assistant tools (e.g., ChatGPT) are increasingly used for tasks on Prolific. Please do not use it for this task. We seek to understand how humans write these statements. If you feel unable to do the task, please leave this spot open for others.*	*Now,* **write the deceptive version of your story about "Being involved in a car accident".**
Write your statement about the event **"Being involved in a car accident"** *here:"*	*Your statement should be at least 300 characters. Please don't mention in any way that you are lying in this statement.*
	IMPORTANT: *Try to be as convincing as possible. A researcher who is an expert in verbal-lie detection will evaluate your statement. If the experimenter would consider your statement as credible, you will have the chance to win an* **extra 50£ compensation** *by participating in a draw.*
	We are aware that AI-assistant tools (e.g., ChatGPT) are increasingly used for tasks on Prolific. Please do not use it for this task. We seek to understand how humans write these statements. If you feel unable to do the task, please leave this spot open for others.

TABLE 2S. Full instructions provided to participants when the order of presentations of conditions was first deceptive and then truthful.

Deceptive condition	Truthful condition
You are going to write about the event **"Being involved in a car accident"** *twice.* *Now you have to* ***lie about the event*** *in order to obtain a benefit or avoid a loss. Specifically, you have to write an alternative version of the story about* **"Being involved in a car accident"** *in which you are deceptive to* **increase the claimed amount of damage you received to get more money.** *Specifically,* **the event should essentially be the same but you have to fabricate details**. *Afterward, we are going to ask you in which specific part of the story you lied and how.* *Now,* **write the deceptive version of your story about "Being involved in a car accident" using the fabrication strategy.** *Your statement should be at least 300 characters. Please don't mention in any way that you are lying in this statement.* **IMPORTANT:** *Try to be as credible as possible because the experimenter is an expert in verbal-lie detection and will read your statement to evaluate it as credible or not. If the experimenter would consider your statement as credible, you will have the chance to win an* **extra 50£ compensation** *by participating in a draw.* *We are aware that AI-assistant tools (e.g., ChatGPT) are increasingly used for tasks on Prolific. Please do not use it for this task. We seek to understand how humans write these statements. If you feel unable to do the task, please leave this spot open for others.*	*In the first task we asked you to lie about your experience with* **"Being involved in a car accident".** *Now we ask you to provide the* **truthful version of the same story.** *This means you should provide a* **complete truthful statement** *of what happened, without omitting relevant information or adding made-up details, as you may have done before!* *Make sure to use correct spelling and grammar and separate your sentences with punctuation. Your statement should be at least 300 characters.* **IMPORTANT:** *We are aware that AI-assistant tools (e.g., ChatGPT) are increasingly used for tasks on Prolific. Please do not use it for this task. We seek to understand how humans write these statements. If you feel unable to do the task, please leave this spot open for others.* *How did things really turn out? Now, re-write the truthful version of your statement here.* *Describe what happened, who was involved, where and when it took place, and why it was memorable to you.*

Supplementary Material - 2

Descriptive statistics

Here we report the descriptive statistics (*M, SD,* Median) of the memory-related variables associated with the events, such as the time elapsed since the event occurred (in months), frequency of recollection, importance, accuracy of the recollection, and emotional valence (Table 3S).

TABLE 3S. Descriptive statistics (M, SD, Median) of participants' responses on anchored scales regarding the time elapsed since the event occurred (in months), frequency of recollection, importance, accuracy of the recollection, and emotional valence of the event.

	M	*SD*	*Median*	*Range*
Time				
"*how long ago did the event happen?*" (in months)	8.80	6.69	7	0 - 24
Recollection				
"*how often do you think or talk about this event?*"	2.58	1.09	2	1 - 5
Importance "*how important is this event to you?*"	3.28	1.34	4	1 - 5
Accuracy				
"*how well do you remember this event?*"	4.04	0.92	4	1 - 5
Valence "*how would rate this event in emotional terms?*"	-0.19	0.7	-0.5	-1 - +1

Embedded lies per event

Table 4S reports the descriptives (mean and standard deviation) of the absolute and standardized number of embedded lies split by events.

TABLE 4S. Average absolute and standardized number of embedded lies per event.

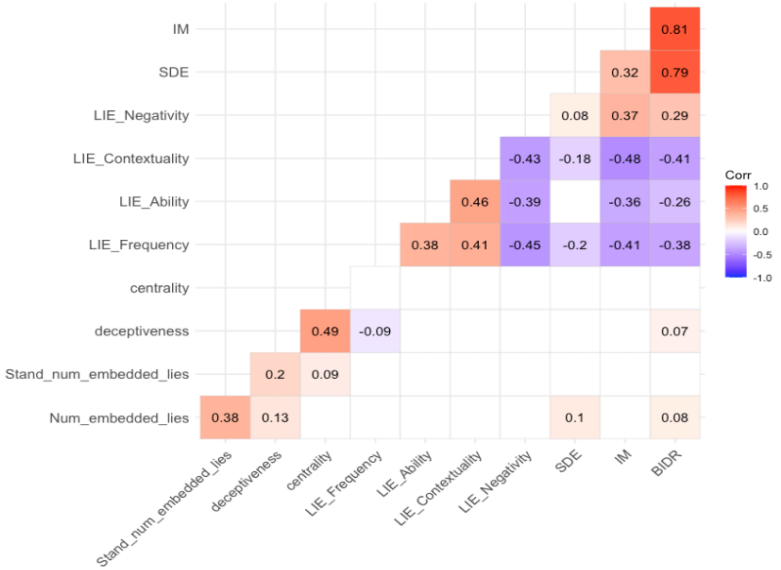
Events	Absolute number of embedded lies	Standardized number of embedded lies
A job interview for your dream job	5.53 (3.68)	0.34 (0.20)
Being hospitalized and undergoing surgery	5.91 (3.87)	0.30 (0.18)
Being involved in a car accident	4.89 (2.36)	0.31 (0.18)
Causing a car accident	5.20 (3.59)	0.22 (0.16)
Cheating on an exam	4.19 (3.37)	0.28 (0.15)
Cheating on your partner	6.75 (4.33)	0.31 (0.17)
Ending a long romantic relationship	5.66 (3.38)	0.33 (0.23)
Getting a speeding fine	4.63 (2.08)	0.34 (0.21)
Getting fired	5.38 (3.46)	0.36 (0.22)
Missing a deadline at work because of bad organization	4.97 (3.27)	0.34 (0.21)
None of them	4.50 (2.92)	0.30 (0.19)
Taking the bus/train without the ticket	4.00 (2.20)	0.35 (0.19)

Note. Standard deviation is reported in brackets.

Correlational analysis

In Figure 1S we show the significant Spearman's rank correlations between the lying profile and BIDR scales and the dependent variables associated with embedded lies.

FIGURE 1S. Correlation matrix.



Note. Only significant correlations with $p < .05$ that survived FDR correction for multiple comparisons are reported.

Abbreviations:
Num_embedded_lies = number of embedded lies
Stan_num_embedded_lie = standardized number of embedded lies
BIDR = Balanced Inventory of Social Desirability Responding scale
SDE = self-deception enhancement subscale of BIDR
IM = impression management subscale of BIDR

Individual differences

After excluding participants who preferred not to express their gender, expired, or revoked their consent to show gender-related data ($n = 6$), we tested for gender differences in the variables of interest. Specifically, results from a permutation t-test ($n_{perm} = 9.999$) revealed a significant difference between males ($M = 4.68$, $SD = 2.29$) and females ($M = 5.26$, $SD = 3.39$) in the absolute number of embedded lies ($diff = 0.58 \pm 0.20$, $p = 0.003$, $d = 0.18$ [0.06, 0.31]) but not in the standardized number of embedded lies ($diff = 0.01 \pm 0.01$, $p = 0.41$, $d = 0.05$ [-0.07, 0.17]). Therefore, these findings suggest that the difference in gender was mainly driven by statements' length.

There was another gender difference for the average deceptiveness scores ($diff = 0.11 \pm 0.05$, $p = 0.03$, $d = 0.14$ [0.02, 0.26]), with females ($M = 3.98$, $SD = 0.79$) reporting higher values than males ($M = 3.88$, $SD = 0.77$), but not for the average centrality scores ($diff = 0.01 \pm 0.05$, $p = 0.85$, $d = 0.01$ [-0.11, 0.14]). Using Spearman's rank correlations, we found only a small but significant positive correlation between age and deceptiveness ($rho = 0.075$, $S = 172823388$, $p = 0.015$). However, we found no significant correlation between age and the absolute number of embedded lies, the standardized number of embedded lies, as well as age and centrality scores (Table 5S).

TABLE 5S. Spearman's rank correlations between age and the dependent variables.

Variables	*rho*	*S*	*p*	Significance ($p < .05$)
Age - Absolute no. of embedded lies	-.046	178303408	.137	No
Age - Standardized no. of embedded lies	.002	186940849	.99	No
Age - Deceptiveness	.075	172823388	.015	Yes
Age - Centrality	.03	180713689	.28	No

Supplementary Material - 3

Clusters of liars

Following the original procedure in Makowski et al., (2023), we investigated the presence of subpopulations of liars by clustering participants' scores at the four-factor lying profile questionnaire. The only difference with the original procedure was that we used the lying profile scores corrected for social desirability. The correction procedure employed a Generalized Linear Model (GLM) approach to regress out the scores of each lying profile factor (i.e., LIE_Ability, LIE_Contextuality, LIE_Frequency, LIE_Negativity) for social desirability effects (i.e., SDE and IM). The adjusted scores were calculated using the *adjust* function from the *datawizard* package in Rstudio. In our dataset, the agreement method procedure suggested an optimal solution with two clusters and a second solution with three clusters. This final two-cluster solution, that we reported in the main text, reflected the *trickster* and the *virtuous* clusters from the original study.

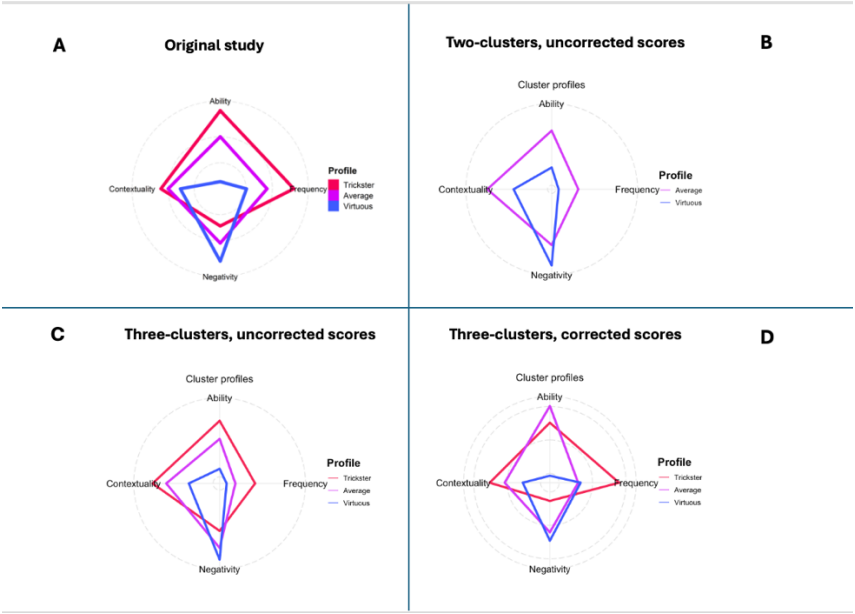
However, with the aim of replicating the findings of the original study, we also applied the k-means algorithm to compute the three-cluster solution. As shown in Fig. 2S Panel D, the obtained clusters were very different from the original ones (see Fig. 2S Panel A). Specifically, we found a group of participants with very low self-reported lying ability, frequency and contextuality, and strong negative emotions and moral attitudes associated with lying, that closely resembled the *virtuous* cluster (40.88% of the sample). A second group of participants (26.20%), which should reflect the *Average* cluster in the original paper, actually, showed average levels of frequency and contextuality, higher levels of negative attitudes and, unexpectedly, extremely high levels of ability. This different score distribution makes this group less analogous to the original one. The third group, which should reflect the *trickster* (32.92%) in the original paper, was composed by people showing very low levels of negativity, extremely high levels of frequency and contextuality and high levels of ability (but less high than in the original paper).

To investigate whether the correction procedure altered the clustering output, we replicated the same analytical procedure as in Makowski et al., (2023), but this time using the raw lying profile scores. The dataset was deemed suitable for clustering (Hopkins' H = 0.27). However, the method agreement procedure again supported the existence of 2 clusters, as indicated by 12 methods out of 29 (41.38%). For replication issues, we applied the k-means clustering algorithm and compared the results

when obtaining two and three clusters, respectively. The former accounted for 37.76% of the total variance of the original data, and the latter accounted for 48.98%.

When grouping participants into two clusters (Fig. 2S, Panel B), we found a group of people resembling the *virtuous* cluster in the original study and a group of people resembling the *Average* cluster, but with a lower level of reported frequency. In contrast, when participants were grouped into three clusters (Fig. 2S, Panel C), we obtained a more similar output, yet we failed to fully replicate the original findings due to the frequency scores being overall lower than in the original study.

FIGURE 2S. Radar plot of the average values of the four lying profile factors in the two- and three-cluster solutions.



Here we report in Table 6S the results of the paired permutation t-test on the differences between trickster and virtuous in the absolute and standardized number of embedded lies, deceptiveness, and centrality scores. No significant differences were found.

TABLE 6S. Embedded lies and associated measures between trickster and virtuous profiles.

	M (*SD*)		*diff* (*SD*)	*p*	*d*	95% CI
	Trickster	Virtuous				
Absolute no. of embedded lies	4.91 (3.16)	5.19 (3.35)	-0.29 (0.20)	.15	-0.09	-0.21, 0.03
Standardized no. of embedded lies	0.32 (0.19)	0.33 (0.21)	-0.003 (0.01)	.78	-0.02	-0.14, 0.10
Deceptiveness	3.94 (0.81)	3.94 (0.75)	-0.01 (0.05)	.99	-0.002	-0.12, 0.12
Centrality	3.57 (0.80)	3.53 (0.84)	0.04 (0.05)	.46	0.05	-0.08, 0.17

Note. The table reports means (*M*) and standard deviations (*SD*) for each variable. The mean differences (*diff*), *p*-values, Cohen's *d* effect size, and 95% confidence intervals (CI) for the differences come from a permutation t-test with 9.999 permutations.

Supplementary Material - 4

Machine-learning classification

State-of-the-art ML models were employed in a classification task to distinguish truthful from deceptive statements with embedded lies. We adopted two approaches: a simpler approach where a Random Forest (RF) model was trained on extracted linguistic features, and a more sophisticated approach that involved fine-tuning pre-trained language models.

Random forest models

The simpler approach consisted of training four Random Forest (RF) models in a binary classification task using as features a Bag of Words representations (BOW, Ignatow & Mihalcea, 2017), LIWC variables (Boyd et al., 2022), **DeCLaRatiVE** variables (Loconte et al., 2023), and GPT-embeddings[14]. RF is an ensemble learning technique that leverages multiple decision trees in the training phase to then select as a final output the most frequent prediction from the individual trees. Below, we provide the details on how we proceeded for the feature extraction:

- **BOW:** we used a bag-of-words representation of unigrams, bigrams, and trigrams to train the model. The bow model was applied on preprocessed text. Preprocessing included lowercasing, lemmatization and the removal of stop-words. We then included only *n*-grams that were present at least 5% of times across documents to exclude rare-words. This bow representation consisted of a vector of length 158.
- **LIWC:** we used the LIWC-22 software to extract 117 syntactic and semantic features from raw text. All features were included in the training phase.
- **DeCLaRatiVE:** we followed the procedure described in Loconte et al., (2023), and we extracted 26 linguistic features associated with four theoretical frameworks of deception (i.e., Distancing, Cognitive Load, Reality Monitoring, and Verifiability approach). All features were employed to train the model.

[14] https://platform.openai.com/docs/guides/embeddings/

- **GPT-embeddings:** the embedding representation was extracting using the OpenAI embeddings models[15]. Specifically, we employed the *text-embedding-3-large* model and, following OpenAI guidelines, we specified a vectorial dimension of 256 using the *dimension* parameter. This allowed us to have a meaningful statement representation without falling in the dimensionality curse (i.e., when the number of features exceeds the number of observations).

The training-test procedure employed a nested cross-validation framework. Specifically, it consisted of an inner loop repeated across 10 folds for hyperparameter optimization and an outer loop across 10 folds for model performance evaluation. The hyperparameter optimization was conducted through Grid Search. Once, the best hyperparameter combination was identified in the inner loop, it was then used to train the model on the entire training set in the outer loop. Models' performance was evaluated in terms of accuracy, precision, recall and F1 score.

Fine-tuning language models

The value of leveraging language models lies in two key areas: the robust numerical representation of natural language learned during the pre-training phase and the ability to adapt the model to a downstream task with minimal fine-tuning of the parameters in the final layer(s), without altering the underlying architecture. Fine-tuning can be accomplished through further training on task-specific data, which improves the model's capacity to generate coherent and contextually relevant text that aligns with the target task.

To assess models' performance in a robust manner we conducted a 5-fold cross-validation, ensuring that both truthful and deceptive statements from the same participants were either present in training test or in test set. This procedure was employed to avoid information leakage and biased performance metrics. Models' performance was assessed in terms of accuracy, precision, recall and F1 score. For our analysis, we tested the performance of fine-tuned versions of distilBERT (Sanh et al., 2019), FLAN-T5 base (Chung et al., 2022), and Llama-3-8B (Grattafiori et al., 2024). All language models, with the exception of the deception language model, were freely available through Hugginface platform.

DistilBERT is a smaller, faster, and cheaper version of the original BERT base model. It was trained by distillation meaning that it was trained to

[15] https://platform.openai.com/docs/guides/embeddings/embedding-models

predict the same probabilities as the original BERT model (https://huggingface.co/docs/transformers/en/model_doc/distilbert). In the present study, distilBERT was fine-tuned using the following configuration of parameters: learning rate 5e-5; weight decay coefficient: 0.01; batch size: 32; number of epochs: 3.

FLAN-T5 is a text-to-text general model developed by Google researchers and capable of solving many NLP task, such as sentiment analysis, question answering, and machine translation (https://huggingface.co/docs/transformers/model_doc/flan-t5). Among the several versions available we employed the FLAN-T5 base, which was fine-tuned with the following configuration: learning rate 5e-5; weight decay coefficient: 0.01; batch size: 2; number of epochs: 3.

Llama-3 model is the most refined version of Llama models, i.e., an open-source collection of foundation language models, developed by Meta AI (https://huggingface.co/docs/transformers/en/model_doc/llama3). This generation of Llama models demonstrated state-of-the-art performance on a wide range of benchmarks and showed improved reasoning. We employed the version with eight billion of parameters (Llama-3-8B), which was fine-tuned with a quantized low rank optimization (QLoRA) procedure and the following configuration: learning rate 1e-4; weight decay coefficient: 0.01; batch size: 2; number of epochs: 3.

The deception language model is a fine-tuned version of a FLAN-T5 base model to classify deceptive statements. In the original study, the deception language model was fine-tuned in three datasets encompassing 2500 personal opinions, 5506 autobiographical memories, and 1640 future intentions, reaching 79.31% accuracy (Loconte et al., 2023). For this study, the deception model was employed as it is to predict deception in our dataset without further fine-tuning.

Exploratory explainability analysis

Here, we report the exploratory explainability analysis we conducted on the Llama-3 model and deception language model in terms of correlations between deception class probabilities and embedded lies-dependent variables (Table 7S) and differences between correct and incorrect classifications in those dependent variables (Table 8S).

TABLE 7S. Spearman's rank correlations between class probability for deceptive statements and the dependent variables in the Llama-3 and deception language model.

Model	Variables	*rho*	*S*	*p*
Llama-3-8B model	Absolute no. of embedded lies	.10	170216978	.0009*
	Standardized no. of embedded lies	.10	170565831	.001*
	Deceptiveness	.05	180010090	.10
	Centrality	.05	180335919	.11
Deception language model	Absolute no. of embedded lies	.09	171758230	.004*
	Standardized no. of embedded lies	.01	185868739	.645
	Deceptiveness	-.02	191376556	.630
	Centrality	-.03	193262992	.421

Note. Positive correlations mean that the class probability of being deceptive (range 0.5 – 1.0) is higher when the dependent variable of interest is higher.
* $p < .01$

TABLE 8S. Embedded lies and associated measures between correct and incorrect classifications in the Llama-3-8B and deception language model.

Model	Variables	*M* (*SD*)		*diff* (*SD*)	*p*	*d*	95% CI
		Correct	Incorrect				
Llama 3-8B model	Absolute no. of embedded lies	5.31 (3.39)	4.43 (2.83)	0.88 (0.22)	.0001*	0.27	0.14, 0.40
	Standardized no. of embedded lies	0.34 (0.21)	0.29 (0.19)	0.04 (0.01)	.0005*	0.22	0.09, 0.35
	Deceptiveness	3.95 (0.80)	3.92 (0.76)	0.03 (0.05)	.6238	0.03	0.10, 0.16
	Centrality	3.57 (0.81)	3.50 (0.85)	0.07 (0.05)	.2212	0.08	-0.05, 0.21
Deception language model	Absolute no. of embedded lies	4.90 (3.30)	5.10 (3.22)	-0.20 (0.21)	.3475	-0.06	-0.19, 0.07
	Standardized no. of embedded lies	0.33 (0.21)	0.32 (0.20)	0.02 (0.01)	.2192	0.08	-0.05, 0.21
	Deceptiveness	3.93 (0.83)	3.94 (0.77)	-0.02 (0.05)	.7446	-0.02	-0.15, 0.11
	Centrality	3.58 (0.87)	3.53 (0.80)	0.05 (0.05)	.4156	0.06	-0.07, 0.18

Note. The table reports means (*M*) and standard deviations (*SD*) for each variable. The mean differences (*diff*), *p*-values, Cohen's *d* effect size, and 95% confidence intervals (CI) for the differences come from a permutation t-test with 9.999 permutations.

* $p < .01$

Chapter 5

Humans incorrectly reject confident accusatory AI judgments

This chapter is based on: Loconte, R., Monaro, M., Pietrini, P., Verschuere, B., & Kleinberg, B. (2026). Humans incorrectly reject confident accusatory AI judgments. *Computers in Human Behavior*, 109019. https://doi.org/10.1016/j.chb.2026.109019

Abstract

Automated verbal deception detection using methods from Artificial Intelligence (AI) has been shown to outperform humans in disentangling lies from truths. Research suggests that transparency and interpretability of computational methods tend to increase human acceptance of using AI to support decisions. However, the extent to which humans accept AI judgments for deception detection remains unclear. We experimentally examined how an AI model's accuracy (i.e., its overall performance in deception detection) and confidence (i.e., the model's uncertainty in single-statement predictions) influence human adoption of the model's judgments. Participants (n=373) were presented with veracity judgments of an AI model with high or low overall accuracy and various degrees of prediction confidence. The results showed that humans followed predictions from a highly accurate model more than from a less accurate one. Interestingly, the more confident the model, the more people deviated from it, especially if the model predicted deception. We also found that human interaction with algorithmic predictions either worsened the machine's performance or was ineffective. While this human aversion to accept highly confident algorithmic predictions was partly explained by participants' tendency to overestimate humans' deception detection abilities, we also discuss how truth-default theory and the social costs of accusing someone of lying help explain the findings.

Keywords: deception, verbal deception detection, human-AI interaction, decision-making, AI explainability

1. Introduction

Detecting deception remains one of the most challenging tasks for researchers and practitioners. Traditional methods often rely on subjective cues or limited behavioral indicators (DePaulo et al., 2003; Hartwig & Bond, 2011), making consistent and accurate detection difficult (Bond & DePaulo, 2006; Hartwig & Bond, 2014). Recent advances in artificial intelligence (AI) offer promising tools to enhance deception detection by analyzing complex patterns in language (Constancio et al., 2023; Hauch et al., 2015), improving both accuracy and scalability.

However, in high-stakes domains, such as in the legal domain, where outcomes directly impact individuals' lives, delegating decisions to machines is strongly disapproved, underscoring the need for including humans in the loop to oversee machine limitations (Kotsoglou & Oswald, 2020; Orsini et al., 2025; van Dijck, 2022). This remarks the need for more research on how humans can benefit from the AI assistance while retaining responsibility for their final decisions and, at the same time, understanding under which conditions AI-based judgments on the credibility of statements are endorsed or rejected by human decision-makers.

1.1 Human deception detection

Decades of research suggest that humans are *truth-biased* and tend to judge statements as truthful more frequently and to be more accurate in detecting truthful statements than deceptive ones. This truth bias is explained by the truth-default theory (TDT; Levine, 2014). According to TDT, people generally assume honesty in communication because deception is quite rare, and its occurrence is limited to a few prolific liars (Levine, 2014; Serota et al., 2010). Indeed, deception is a strategic act that typically occurs only when truthfulness obstructs personal goals (Levine, Kim, & Blair, 2010; Levine, Kim, & Hamel, 2010). As a consequence, people become suspicious only in case of strong triggers that disrupt this default truth assumption (e.g., hidden goals, inconsistencies, or third-party warnings).

Another robust finding is that humans are poor at detecting deception. More specifically, meta-analytical evidence indicates that laypeople's performance in discriminating between deceptive and truthful statements is no better than chance level (Hartwig & Bond, 2011, 2014). Additionally, expertise in the field (e.g., being a police officer dealing with potentially deceptive suspects) does not significantly improve the detection rate (Bond & DePaulo, 2006). In contrast, when relying on evidence-based tools (e.g., Criteria-Based Content Analysis, Reality Monitoring),

interviewing strategies (e.g., Strategic Use of Evidence, Imposing cognitive load), and heuristics (e.g., detailedness), human deception detection can be around 70% (Amado et al., 2016; Gancedo et al., 2021; Hartwig et al., 2014; Sporer et al., 2025; Verschuere et al., 2023; Vrij et al., 2008).

However, these traditional deception detection methods require either a one-on-one interaction, especially for strategic interviewing, or human involvement in scoring verbal transcripts for content analysis, making scalability expensive in terms of time and resources (Kleinberg et al., 2019b). Moreover, 70% accuracy is often considered low and unsatisfactory when applied to real-life and sensitive settings (Kleinberg et al., 2019a). A potential avenue for overcoming these limitations lies in the utilization of computational approaches to estimate the credibility of statements.

1.2 Computational approaches

Advancements in computational methods for deception detection have been made possible thanks to progress in computers, machine learning (ML), and natural language processing (NLP), enhancing both scalability and objectivity.

For example, focusing on the well-established knowledge that "lying takes time" (Suchotzki et al., 2017), early computer-aided approaches investigated deception detection through the analysis of mouse (Monaro et al., 2017) and keyboard dynamics (Monaro et al., 2018), while more recent research trained ML models on response latencies and error rates to unexpected questions (Melis et al., 2024). On the other hand, studies focusing on computer vision developed ML models trained at detecting deception through the analysis of facial expressions (see Delmas et al., 2024, for a review of studies).

Among all the computational approaches available, those based on textual data – combining techniques from NLP and ML models - are the most useful because they can be applied to automate manual credibility assessment methods (e.g., Reality Monitoring, Verifiability Approach) and can be used in combination with interviewing techniques. For example, one study employed the proportion of unique named-entities (NER) as a proxy of verifiable details to detect positive and negative deceptive hotel reviews, reaching an area under the curve (AUC) of 0.67 and 0.65, respectively (Kleinberg et al., 2017). Another study relied on a pretrained language model (i.e., Bidirectional Encoder Representations from Transformers - BERT) to detect deceptive utterances in a dataset of transcripts of trial hearings, reaching 71.61% accuracy (Pérez-Rosas et al., 2015). Fi-

nally, another work showed that ML algorithms significantly outperformed naïve (accuracy= 54.7%) and expert judges (accuracy= 59.4%) in detecting deception from transcripts of interviews with unexpected questions when trained both on theory-led (accuracy=69.4%) and data-driven features (accuracy=77.3%) (Loconte et al., 2025).

However, computational approaches for credibility assessment are not yet applied in real-life situations. Previous research has shown that these models show a broad range of accuracy (from 60% to 90%) and have limited generalisability to different domains (Kleinberg et al., 2019; Loconte et al., 2023; Velutharambath & Klinger, 2023). Nevertheless, we might imagine a near future where these limitations might be easily overcome and new language models on deception become available. If this is the case, a key problem with automated verbal deception detection is the inherent difficulty of relying exclusively on AI-based predictions in high-stakes domains. In fact, in sensitive contexts, such as forensic settings, delegating decisions to machines has been strongly disapproved, while including humans in the loop has been encouraged (Kotsoglou & Oswald, 2020; Orsini et al., 2025; van Dijck, 2022). Therefore, further research is needed on how humans can be effectively integrated into the loop to benefit from AI assistance while retaining responsibility. In fact, while hybrid decision-making has already been commonly used in online content moderation (Jhaver et al., 2019)and is becoming more popular in medical diagnoses (Bulten et al., 2020), it remains understudied in the context of deception detection.

1.3 Hybrid decision-making in deception detection

Hybrid decision making consists of integrating human oversight into AI predictions, allowing decision-makers to leverage AI's analytical strengths and scalability while maintaining responsibility. The rationale behind hybrid decision-making is that AI-based judgments are provided to the human decision-makers to obtain an overall better performance than either mode in isolation.

In the context of deception detection, previous research investigated hybrid decision-making by integrating supervised ML with human judgment and comparing the combined performance to that of each modality in isolation (Kleinberg & Verschuere, 2021). While ML alone achieved a classification accuracy of 69%, human involvement, by fully overruling or adjusting within given boundaries the AI-based predictions, did not significantly improve the deception detection performance and brought the accuracy back to chance level (Kleinberg & Verschuere, 2021).

Related work has built on that framework and examined how the availability of a lie-detection algorithm influences human judgment (von Schenk et al., 2024). Participants either received no algorithmic aid (control condition), were always shown AI predictions (forced condition), or could choose to access these predictions (choice condition). Notably, algorithmic availability altered social dynamics: participants were less willing to accuse others without AI support, but those who actively sought and relied on algorithmic predictions (choice condition) were more likely to follow accusatory predictions, compared to those who were passively exposed in the forced condition.

Together, these findings challenge the assumption that human-AI collaboration inherently improves decision-making and underscore the complexity of designing effective hybrid systems. In fact, integrating AI into human decision-making is not merely a technical problem but also a psychological and social challenge. One common limitation in these two studies is that both lacked manipulation of the information provided to participants about the model's overall performance (i.e., accuracy) and uncertainty scores (i.e., confidence) for individual predictions. Indeed, previous studies found accuracy and confidence as key drivers of human trust in AI models. It was found that higher AI stated and observed accuracy reliably increased user reliance (Alufaisan et al., 2021; He et al., 2023; Yin et al., 2019). Similarly, AI confidence levels influenced human self-confidence (Li et al., 2024; Zhang et al., 2020), with alignment effects persisting beyond the interaction (Zhang et al., 2020). This leaves open the question of whether more fine-grained information—such as information on the model's accuracy and prediction confidence— affects hybrid decision-making on deception detection in a more nuanced way. In other terms, we need to understand under which conditions automated judgments on the veracity of statements are endorsed or rejected by human decision-makers, and whether this interplay is beneficial for improved verbal deception detection.

1.4 The present study

The current study adds to research on hybrid human-machine deception detection by explicitly incorporating and manipulating information about (1) the overall accuracy of the AI model, (2) the AI uncertainty for individual predictions, and (3) the final veracity judgment of the AI model. To this aim, we developed an experimental task in which participants made veracity judgments about statements presented in contexts relevant to lying and were provided with the predictions of a fictitious AI-based classifier for deception detection.

We conducted a 2 (Accuracy: low=54% vs high=89%) by 5 (Confidence: indecisive, poorly confident, moderately confident, confident, very confident) by 2 (Classification: truthful vs deceptive) mixed-design experiment. The experimental design allows us to isolate the effects of AI model characteristics on human deception judgments when interacting with the model. We hypothesized a two-way interaction between Accuracy and Confidence: for a very accurate model (=89%), humans deviate less from AI predictions when these are made with high confidence compared to low confidence predictions; for a low-accuracy model, confidence does not exert any effect in human deviations, as the underlying predictions are inherently inaccurate regardless of confidence level.

A second research question entails whether information on the model's accuracy and confidence can improve verbal deception detection in the hybrid modality than with the AI mode in isolation. To reply to this research question, we relied on the receiver operating characteristic (ROC) framework to test the diagnostic power of AI- and human AI-assisted judgments and compared the areas under the curve (AUC) to seek potential differences under the two accuracy conditions.

2. Materials and Methods

2.1 Participants

We recruited 493 participants fluent in English from the general population through Prolific16. Each participant provided informed consent. Participants were excluded if they failed at least one of two attention checks. The first one consisted of asking participants to position two sliders at designated points during the task (i.e., -20 and 33). However, since a few participants provided feedback and reported that the slider might have shifted slightly after submission, we were more lenient and retained participants whose responses deviated by only ± 2 points, excluding n=66 participants. The second attention check consisted of asking participants, after the task, to recall the model's accuracy in a multiple-

[16] An a priori power analysis was conducted using G*Power (Faul et al., 2007) to establish the required number of participants. The results indicated that a sample size of 454 is sufficiently large to achieve a statistical power (1-β) of 0.95 in a mixed design with repeated measures (no. measurements = 10) involving two experimental conditions, given a significance level of $\alpha = 0.01$, a small effect size (0.1), low correlation between repeated measures (0.1), and a non-sphericity correction $\varepsilon = 0.1$.

choice question with six answer options, excluding n=52 participants. After data collection, we checked for the presence of any outliers for the time taken to complete the task. Participants were considered outliers if they completed the task in a time that was two standard deviations lower than the average value, but any outlier was detected. Finally, we checked for careless responding consisting of more than three consecutive identical responses (using the *careless* package in R; Johnson, 2005). Two participants were removed as they provided four and eight consecutive identical responses, respectively. The final remaining sample consisted of 373 participants (52.28% females, 47.72% males) with a mean age of 39.42 years (SD=13.61, range: 18-72).

2.2 Study design

The experimental design was 2 (Accuracy: low vs high, between-subjects) by 5 (Confidence: indecisive vs poorly confident vs moderately confident vs confident vs very confident, within-subjects) by 2 (Classification: truthful vs deceptive, within-subjects). Specifically, participants were randomly assigned to one of the two Accuracy conditions (low=54% vs. high=89%). Accuracy conditions were manipulated in the task by showing participants a number of correct predictions equivalent to the model's accuracy rate (e.g., the model with 89% accuracy showed 89% of the time a correct prediction). Additionally, for correct predictions, the veracity and confidence condition of the statement was also kept. In contrast, when a prediction was incorrect, the confidence range and value remained randomly assigned but flipped. Confidence was manipulated within-subjects, with participants evaluating ten statements (half truthful and half deceptive) paired with twice five different confidence levels (i.e., *indecisive, poorly confident, moderately confident, confident, very confident*). Statements were randomly paired with confidence levels and then with a random value from a specific confidence range (for truthful statements: *indecisive* = 0–9, *poorly confident* = 10–19, *moderately confident* = 20–29, *confident* = 30–39, *very confident* = 40–50; for deceptive statements: *indecisive* = 0 – -9; *poorly confident* = -10 – -19; *moderately confident* = -20 – -29; *confident* = -30 – -39; *very confident* = -40 – -50). Classification, i.e., whether the AI model predicted that statement as truthful or deceptive, was inherently manipulated within subjects. Positive values of confidence represented predictions of truthfulness, while negative values represented predictions of deception. Values equal to zero were excluded from the analysis.

2.3 Stimuli

Statements were selected from a subset of a validated dataset of past autobiographical experiences deemed relevant for lying (Loconte et al., 2025). The deception involved in the original dataset was in the form of embedded lies. Specifically, we selected a total of 406 statements (203 truthful and 203 deceptive) about i) being hospitalised and undergoing surgery, ii) being involved in a car accident, iii) causing a car accident, and iv) missing a deadline at work because of bad organisation (see Table 1).

The instruction for the main task reads as follows:

> 📖*In this experiment, you will read twelve short statements about past experiences that are either truthful or lies.*
> *Your task is to guess whether each statement is truthful* ✅*or a lie* 🤔
> *These statements were randomly selected from a larger dataset containing both truthful and deceptive statements.*
> *To help you with your task, we provide you with the predictions of a lie detection algorithm based on artificial intelligence (AI)* 🤖*.*
> *You'll see an example on the next pages.*
>
> ***Please note that you should read the statements carefully, as after the task, you will also have to take a quick quiz. The quiz serves to validate your participation.***
>
> *Please read the following statement carefully:*
>
> ***Topic:*** *Being hospitalised and undergoing surgery*
>
> ***Context:*** *They might have lied to exaggerate some side effects and receive extra compensation from the health insurance.*
>
> ***Statement 1****: I was hospitalised for almost three months. I was sick, like really sick and doctors were unable to identify what is the problem. They kept me in hospital for almost a month just to keep on doing blood test and taking samples on my body, until they found out I'm diagnosed with lupus. It was a hard journey, I stated attending Albert Luthuli hospital for Rheumatologist, Ngwelezane hospital for Dermatologist, until I get better.*

The experiment was implemented in a web app designed with the Python-based library *Streamlit* (v.1.45.0; screenshots of the task available here: https://osf.io/28avp/files/4ta5x).

TABLE 1. Context for lying, verbatim example of a statement, total number of statements, and average number of words (and SD) per topic.

Topic	Context for lying	Example statement	No. of statements in topic	Average no. of words (and SD)
Being hospitalised and undergoing surgery	The author of the statement might have lied to exaggerate some side effects to receive extra compensation from the health insurance	Six months ago I underwent planned surgery to remove a very large cyst from my scalp and several medium sized cysts. The surgery took longer than planned and as a result I was very unwell afterwards. It took me some time to come round whilst in recovery. The effects of general anaesthetic and anti-emetic drugs caused me to be violently sick, I could not even keep water down. I was rushed to another hospital and admitted to a ward for observation and to administer fluids via a drip. I waited four years for this surgery, during this time my details were lost and several appointments cancelled due to strikes and other issues. If I had received the surgery earlier, the procedure would not have taken so long and I would not have been so unwell after the surgery.	138	157 (95.4)
Being involved in a car accident	The author of the statement might have lied to increase the claimed amount of damage received to later get more money from their insurance	Last month, I was in a car accident exiting a mall. The street wasn't all that busy and it was a fairly calm day. I arrived at the four way intersection to a red light and stopped accordingly. When the light turned green, I proceeded to make my turn when a car from the adjacent side ignored their red light and continued passing through and hit my car in the rear. When I exited to confront the individual, they sped away.	94	144 (76.7)
Causing a car accident	The author of the statement might have lied to describe the event in a way	A morning i was driving my car to work and i was in a hurry because i was going to late. As i driving i reached traffic lights that was at that moment red. When the traffic lights became green i start moving again and all of a sudden the car in front of me	30	169 (88.3)

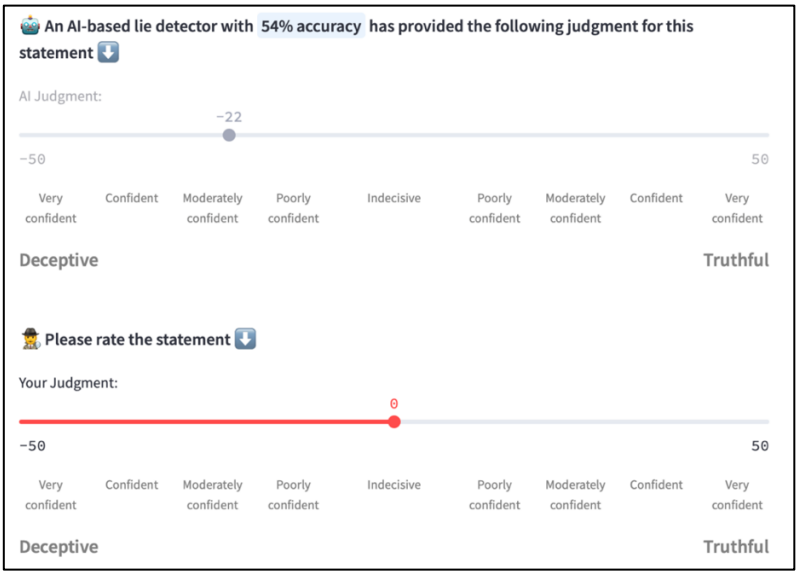
Topic	Context for lying	Example statement	No. of statements in topic	Average no. of words (and SD)
	so that it's not their fault.	stopped withour a reason, causing me to collide with it. The woman driving the car, came out of the car and she was yelling at me saying that i caued the accident. But the fault was on her, because she stopped suddenly without a reason and all of the other vehicles were moving and also as i noticed she was talking on her cellphone. In my point of view she shouldn't using her phone while driving and be more cautious about the traffic lights and the vehicles behind her.		
Missing a deadline at work because of bad organisation	The author of the statement might have lied to find excuses that allow them not to appear forgetful or disorganised.	I missed a timeline of the important work, because of bad organisation of my colleague. I have been trying my best, but they were putting the task on hold since the time has passed and now they are trying to convince everybody that was not their fault, but mine. I have repeated many times that the task is important to finish just before 2nd half of May and now there is a lot of to catch up, but they preferred to be passive and take a lot of brakes causing the delay in the job accolated to our team. Now everyone has to work on the task, but our client is losing his patience and judge the whole company by the laziness of the teammates.	144	127 (65.5)

Note. Example statements are reported *verbatim*

2.4 Procedure

Participants read twelve statements (i.e., ten experimental statements and two attention checks) and were instructed to 1) act as a judge, 2) read each assigned statement paired with a fictitious AI judgment, and 3) decide whether the statement was truthful or deceptive. The AI judgment was displayed in the form of a slider with a continuous scale from -50 to +50 and 0 as a midpoint. Specifically, values from -50 to 0 represented deceptive statements, and values from 0 to +50 represented truthful statements. Values closer to the extremes of the slider (i.e., -50 or 50) represented very high levels of confidence, and values closer to 0 represented indecisive judgments. This range was intended to mimic class probabilities, which here was defined as *confidence* to ensure naïve understanding. Participants provided their judgment using another slider with the same scale. In this way, we measured participants' judgments and uncertainty in one shot. This similarity across sliders also ensured participants a fair understanding of what confidence means (Figure 1). The slider of each judgment started at the indecisive midpoint of 0, and participants could freely move it in any direction.

FIGURE 1. Illustration of participants' and AI's sliders for judgments.



Note. The upper part of the figure depicts the AI judgment in the form of a slider, providing information about the overall veracity judgment plus the confidence

level. The lower part of the picture depicts the participant's slider that will be used to provide the final judgment.

The average duration of each trial (i.e., reading the statement and making a judgment) was 49.03 seconds (*SD*=28.97, *Median*=43.72). After performing the task, participants rated on a 10-point scale their level of motivation (1=Not at all,10=Very much), difficulty of the task (1=Very easy, 10=Very difficult), familiarity with machine learning models (1=Not familiar at all, 10=Very familiar), and their performance in the deception detection task (1=Algorithm performance is better, 10=My performance is better) as well as that of the average human compared to the AI model (1=Algorithm performance is better, 10=Human performance is better). After the experiment ended, participants were thanked and debriefed about the fictitious nature of the AI model. Participants were remunerated with GBP 1.80. The duration of the task was, on average, 21.46 minutes (*SD*=51.94, *Median*=15.82).

2.5 Analysis plan

The main and preregistered analyses are reported in the *Confirmatory analyses* paragraph of the *Results* section. For the main analysis, we rescaled the judgments from -50 - +50 to 0-100 to have all values in the positive range. The dependent variable was computed as the difference between human judgment and AI judgment ($\Delta y = y_{human} - y_{AI}$), where positive values indicated that humans judged the statement as more truthful than the AI, and negative values indicated that humans judged the statement as more deceptive. We used a linear mixed model to account for the nested structure of the data, as each participant provided multiple judgments and each statement was judged multiple times. We determined the optimal model structure by comparing three nested linear mixed-effects models in predicting Δy (see Supplementary Materials, SM 1). Based on the results, Δy was modelled as a function of the main effects and full interactions of Accuracy (low = 54% vs. high = 89%), Confidence (five levels: indecisive, poorly confident, moderately confident, confident, very confident), and Classification (truthful vs. deceptive) as fixed effects. The model included one random intercept for participants and one for statements, to account for individual- and stimulus-related differences in baseline deviation.

As a robustness check, we re-run the same analysis by including as covariates participants' scores on i) motivation, ii) difficulty of the task, iii) familiarity with ML models, and iv) goodness of their performance and that of the v) average human compared to AI. All analyses were conducted in R, using the *lme4* and *lmerTest* libraries. The alpha

level for statistical significance was 0.01, and effect sizes are reported with 99% confidence intervals.

2.6 Exploratory analysis

Given that one-fourth (n=120) of the sample was excluded from the analysis due to failing attention and statistical checks, we re-run our main analysis on the whole sample as a robustness check.

Moreover, we also analyzed the data from a different perspective by examining absolute values of human deviations ($|\Delta y| = |y_{AI} - y_{human}|$), namely deviations from model predictions irrespective of whether participants judged statements as more truthful or more deceptive. Put differently, the analysis focused on the magnitude of the deviation rather than the direction of the deviation.

Finally, since findings show that participants tend to deviate from AI predictions under specific conditions, we tested whether these human corrections improve detection between deceptive and truthful statements. With this aim, we used the ROC framework, which evaluates the diagnostic power of a continuous response variable (here: the human judgment) by plotting sensitivity against 1-specificity across all possible thresholds. The area under the curve (AUC) quantifies this performance, ranging from 0 (=perfectly inaccurate) to 1 (=perfectly accurate), with 0.50 indicating chance-level discrimination. Differences in AUC between conditions were tested using the DeLong test (DeLong et al., 1988).

2.7 Ethics and transparency statement

The experiment was designed in accordance with the Declaration of Helsinki and was approved by the local ERB (code: TSB_RP1450). The study design, methods, hypotheses, and confirmatory analyses were preregistered at: https://aspredicted.org/cxy6-t3sq.pdf. All data, materials, and the code to reproduce the analysis are available at https://osf.io/28avp/overview.

3. Results

3.1 Preliminary analysis

Descriptive statistics of participants' deviation from AI scores (Δy) across conditions are reported in Table S1.

A permutation *t*-test with 9,999 permutations indicated no difference between the high vs low accuracy condition in i) the motivation to perform the task well, ii) the perceived difficulty of the task, and iii) the level of ML familiarity (Table 2). However, participants in the low-accuracy condition scored significantly higher on believing that AI performs worse than the average human compared to those in the high-accuracy condition, d=0.31 [99% CI: 0.04, 0.58]. Participants in the low-accuracy condition also scored significantly higher on believing that their own judgments were better than AI judgments compared to the high accuracy condition, d=0.29 [0.02, 0.56].

After computing participant-level averages for deviations (Δy) and absolute deviations ($|\Delta y|$), associations with self-reported covariates using Spearman rank correlations were examined. Results indicated a small positive correlation between the average deviation (Δy) from AI predictions and participants' belief that the average human performs better than the AI model ($r_s(371)=0.18$, $p=0.002$). Small positive correlations were also found between the **average absolute deviation ($|\Delta y|$)** and participants' perception of outperforming the model ($r_s(371)=0.34$, $p<0.001$) and individuals' beliefs that the average human performs better than the AI model ($r_s(371)=.29$, $p<0.001$). Correlations between the remaining covariates (i.e., motivation, difficulty, and ML familiarity) were not found to be significant (see Figure S1 in SM 2 for the correlation matrix). This suggests that participants who perceived the average human, or themselves, as more capable in spotting deception than the AI model tended to display larger deviations from the algorithmic predictions.

TABLE 2. Mean, standard deviation, and Cohen's *d* of the difference between covariates across accuracy conditions.

	Accuracy low		Accuracy high		
	Mean	*SD*	Mean	*SD*	*d* (99% CI)
Motivation	8.75	1.43	8.66	1.53	0.06 (-0.21; 0.33)
Difficulty	3.65	2.79	3.55	2.83	0.03 (-0.23; 0.30)
ML familiarity	6.11	2.47	6.01	2.47	0.04 (-0.23, 0.31)
AI vs Average Human	6.67	1.84	6.06	2.14	0.31 (0.04, 0.58) *
AI vs Yourself	6.51	1.81	5.95	2.01	0.29 (0.02, 0.56) *

Abbreviations: SD = *Standard deviation;* CI = Confidence intervals.
* $p < .01$

3.2 Confirmatory analyses

For the main analysis, a linear mixed-effects model was fitted with random intercepts for participants and statements (see Table S2 in SM3). Using the *anova( )* function, type III F-tests were conducted to investigate the main and interaction effects of the categorical predictors.

The 2 (Accuracy: low vs. high) by 5 (Confidence: indecisive vs poorly confident vs moderately confident vs confident vs very confident) by 2 (Classification: truthful vs deceptive) ANOVA showed significant main effects of Classification, a significant two-way interaction effect of Accuracy by Classification and Confidence by Classification, and a significant three-way interaction effect of Accuracy by Confidence by Classification (Table 3).

TABLE 3. Type III Analysis of Variance for the Linear Mixed Model predicting human deviation.

Effect	Sum of Sq	Mean Sq	Num DF	Den DF	*F*-value	η^2 (99% CI)
Accuracy	9	9	1	377.5	0.02	0.00 (0.00, 0.00)
Confidence	928	232	4	180.0	0.46	0.01 (0.00, 0.05)
Classification	514,132	514,132	1	3,425.6	**1026.21****	0.23 (0.20, 0.26)
Accuracy × Confidence	2,594	649	4	3,241.6	1.29	0.00 (0.00, 0.01)
Accuracy × Classification	20,655	20,655	1	3,426.1	**41.23****	0.01 (0.00, 0.02)
Confidence × Classification	287,802	71,951	4	3,426.4	**143.61****	0.14 (0.12, 0.017)
Accuracy × Confidence × Classification	8,370	2,092	4	3,426.4	**4.18***	0.01 (0.00, 0.01)

Note. For this model, the bobyqa optimiser was used to fit the model. Results are based on Satterthwaite's approximation for degrees of freedom. In **bold** are reported significant *F*-values.
Abbreviations: Num DF= number of degrees of freedom; Sq = squares
* $p < .01$, ** $p < .001$

The main effect of Classification suggested that the human deviation from model predictions was significantly higher when the model predicted deceptive (M=19.40, SE=0.76) than when it predicted truthful (M=-5.25, SE=0.77), d=0.55 [0.50, 0.59].

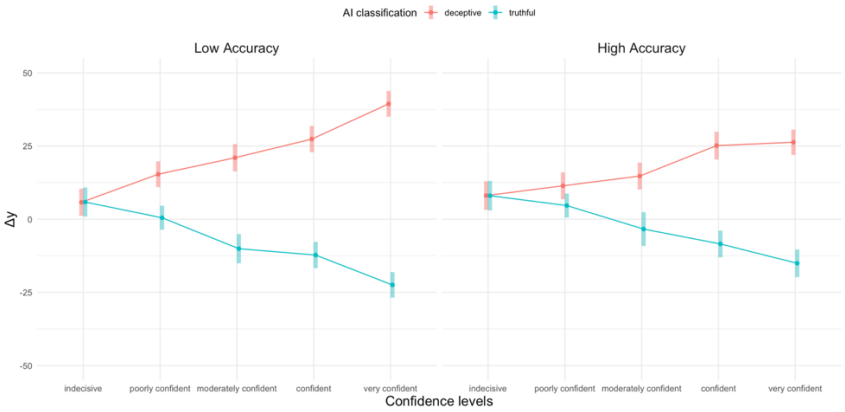
For the two-way interaction of Accuracy by Classification, the human deviation from the model's predictions of deceptiveness was significantly higher for the low-accuracy model (M=21.80, SE=0.97) than for the

high-accuracy one (M=16.99, SE=0.99), d=0.13 [0.04, 0.22], moving the AI judgment towards more truthful but with a greater extent for the low-accuracy model. Similarly, the average deviation from the model's predictions of truthfulness was significantly more pronounced in the low-accuracy condition (M=-7.79, SE=0.98) than in the high-accuracy condition (M=-2.71, SE=1.02), d=-0.13 [-0.22, -0.05]. The negative deviation values reflect participants changing the initial AI judgment towards more deceptive, but did so to a greater extent in the high-accuracy than in the low-accuracy.

For the two-way interaction of Confidence by Classification, human deviation from AI predictions of deceptiveness increased significantly with each step up in the confidence range, except between the indecisive and poorly confident, and poorly confident and moderately confident range, suggesting that the higher the model's confidence, the higher the human deviation from AI judgments of deception. Similarly, for AI predictions of truthfulness, deviations were significantly more pronounced across steps up in the confidence range, except between indecisive and poorly confident, and between moderately confident and confident (see Table S3 for pairwise contrasts in SM).

Most importantly, we found a significant three-way interaction between Accuracy, Confidence, and Classification, which we unpacked by Accuracy (Figure 2).

FIGURE 2. Average values of deviation Δy across Accuracy (low vs. high), Confidence (indecisive vs. poorly confident vs. moderately confident vs. confident vs. very confident), and Classification (deceptive vs. truthful) conditions.



Note. Error bars represent 99% CI.

Low accuracy model

Under the low-accuracy condition, the Confidence by Classification ANOVA indicated a significant main effect of Classification and a significant interaction effect. For the main effect of Classification, $F(1, 1812.3)=696.31$, $p<.001$, $\eta^2=0.28$ [0.23, 0.32], human deviation was significantly higher for AI predictions of deceptiveness ($M=21.96$, $SE=1.03$) than for predictions of truthfulness ($M=-7.73$, $SE=1.05$), $d=0.62$ [0.55, 0.69], indicating that humans tended to disagree more from AI accusations of deception.

The significant interaction between Confidence and Classification, $F(4, 1832.56)=84.25$, $p<0.001$, $\eta^2=0.16$ [0.12, 0.19]. revealed that human deviation from AI predictions of deceptiveness increased significantly with each step up in the confidence range, except between *poorly confident* and *moderately confident*, and between *moderately confident* and *confident* levels. For AI predictions of truthfulness, deviation was more pronounced within steps up in confidence levels, except between *indecisive* and *poorly confident*, and between *moderately confident* and *confident*. These findings suggest that, for both types of AI predictions (i.e., truthful and deceptive), human deviation from the initial AI judgment tends to grow as the AI expresses greater confidence (see Table S4 for pairwise contrasts in SM), although in different directions, namely judging deceptive statements as more truthful and truthful statements as more deceptive.

High accuracy model

Similarly to the low accuracy condition, the Confidence by Classification ANOVA indicated a significant main effect of Classification and a significant interaction effect for the high-accuracy condition. Concerning Classification, $F(1, 1605.6)=364.58$, $p<.001$, $\eta2=0.19$ [0.14, 0.23], humans deviated more from AI predictions of deceptiveness ($M=17.15$, $SE=0.95$) than from predictions of truthfulness ($M=-2.79$, $SE=0.98$), $d=0.48$ [0.41, 0.54].

The interaction between Confidence and Classification, $F(4, 1611.0)=58.76$, $p<.001$, $\eta^2=0.13$ [0.09, 0.17] revealed a more nuanced pattern of human deviation across confidence levels. When the AI classified statements as deceptive, deviations increased significantly only at higher levels of confidence (i.e., confident and very confident) compared to lower levels. Yet, the difference between confident and very confident was not significant. In contrast, when the AI classified statements as truthful, deviations remained more constant, with significant differences emerging at least every two steps up the confidence scale (e.g., moderately confident was significantly different from very confident but not

from confident). These results suggest that, under high-accuracy conditions, only high levels of confidence significantly influence human deviations from AI predictions of deceptiveness, while deviations from truthful AI predictions increase more gradually across confidence levels (see Table S5 in SM).

3.3 Exploratory analysis

Robustness check

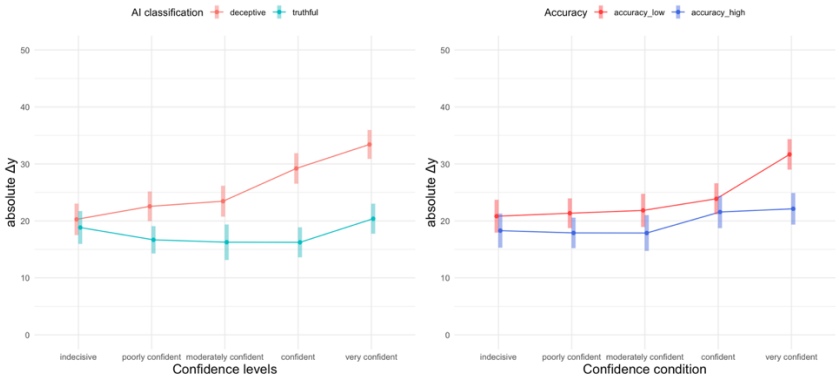
As a robustness check, the same analysis was re-run on a new model that included participants' scores of covariates and on a new model including the participants who failed the attention checks (n=120) and were previously excluded from analysis. Our findings were unaffected by the inclusion of the covariates (see Table S6). However, when previously excluded participants were included in the analysis, the three-way interaction effect was no longer significant (see Table S7). This suggests that, while the observed effects were not driven by individual differences in the covariates, participants' attention during the task was important.

Magnitude of deviation

Full analyses of the magnitude of human deviation (i.e., irrespective of whether participants judged statements as more truthful or more deceptive) are provided in SM6, while a summary of important results is reported below.

A first relevant finding was that, regardless of the condition, the magnitude of deviation was always significantly greater than zero (see Figure 3). Moreover, we observed a significant interaction between AI classification and confidence level, $F(4, 3463.4)=13.13$, $p<.001$, $\eta^2=0.01$ [0.01, 0.03]. As illustrated in Figure 3 (left panel), deviations from model predictions of truthfulness did not vary significantly across confidence levels (see Table S11). In contrast, participants' magnitude of deviations significantly increased with increases in confidence levels for predictions of deceptiveness (see Table S11). Finally, we found a significant interaction between confidence level and model accuracy, $F(4, 3293.6)=5.27$, $p<.001$, $\eta^2=0.006$ [0.00, 0.01] (see Figure 3, right panel). Between the two conditions, a significant difference in magnitude was observed only for *very confident* predictions stemming from the low-accuracy model than from the high-accuracy one. Relatedly, no differences were found for the high-accuracy condition within confidence levels, while, for the low-accuracy condition, the average magnitude at the *very confident* level was significantly greater than those at the remaining lower levels (see Table S12).

FIGURE 3. Average values of absolute deviation between i) Classification and Confidence (left panel) and ii) Accuracy and Confidence (right panel).



Note. Error bars represent 99% CI.

This pattern of findings confirms that participants tend to reject more over-confident predictions, especially those stemming from a low-accuracy model and those predicting deception.

Human performance in detecting deception

Participants guided by the high-accuracy model achieved a detection performance that was significantly higher than chance (AUC=0.76 [0.73; 0.79]) and significantly higher than participants' performance under the low-accuracy model condition (AUC=0.57 [0.54; 0.60]; DeLong test: $D(3692.5)=11.23$, $p<.001$). Yet, this detection rate was significantly lower than that of the highly accurate model (AUC=0.90 [0.88; 0.92]; DeLong test for two correlated ROC curves: $z=-12.42$, $p<.001$), suggesting that human interaction with the model's predictions worsens detection performance compared to relying entirely on the AI model.

As for participants' performance under the low-accuracy model condition (AUC=0.57 [0.54;0.60]), this was not significantly better than the model's performance alone (AUC=0.54 [0.51,0.57]), $z=-1.82$, $p=.06$.

This pattern of findings suggests that participants were influenced by model predictions in both conditions, with human performance mirroring that of the corresponding model, yet without overcoming the model's detection rate.

4. Discussion

This study investigated hybrid deception detection by exploring under which conditions algorithmic judgments on statement veracity are endorsed or rejected by human decision-makers. Specifically, this study focused on the role of the model's accuracy and uncertainty for individual predictions. Using statements about past events deemed relevant for lying (Loconte & Kleinberg, 2025), participants were provided with predictions from a fictitious AI-based deception classifier and were instructed to act as a judge and evaluate the veracity of statements. The experimental task was designed to isolate the effects of the model's characteristics (i.e., accuracy and confidence) on human deviations from the model's judgments. We predicted a two-way interaction effect between the model's accuracy and confidence, so that, only for the highly accurate model, but not the lowly accurate one, human deviation significantly decreases as the model's confidence increases. In other words, for a very accurate model, higher confidence in predictions produces greater endorsement of the model's predictions by humans. However, this hypothesis was only partially supported.

4.1 Main findings

Contrary to our expectations, humans tended to reject AI judgments more when these were made with high confidence. That rejection was more pronounced when the prediction indicated *deception,* compared to predictions indicating *truthfulness,* and when stemming from a low-accuracy model than from a very accurate one. As for the direction of the deviation, the more the AI was confident in predicting *deception*, the more participants leaned their judgment towards less deceptive (hence, as more truthful). In contrast, the more the model predicted *truthfully* with greater confidence, the more participants leaned towards less truthful (hence, more deceptive).

These findings are partially in contrast with previous studies showing that accuracy and confidence are key drivers of human trust, with higher levels of AI accuracy and confidence being associated with higher levels of trust in AI models (Alufaisan et al., 2021; He et al., 2023; Li et al., 2024, 2025; Yin et al., 2019). While our findings are consistent with existing literature on AI-accuracy, suggesting that an increase in the stated model accuracy leads to a greater degree of human trust in these models (He et al., 2023; Yin et al., 2019), we observed a different direction for the effect of confidence (Li et al., 2024, 2025). Specifically, we observed greater deviations from *very confident* predictions than from low-confident ones. Furthermore, we did not replicate a full truth-bias effect that was found in previous studies on hybrid deception detection (Kleinberg &

Verschuere, 2021; Levine, 2014; von Schenk et al., 2024). While participants tended to judge deceptive statements as more truthful, especially when predicted with higher levels of confidence, this was not mirrored for truthful statements, which were judged as less truthful under the same high-confidence conditions in our experiment.

Finally, our findings add to the emerging work on hybrid deception detection. Others report that AI availability can increase the accusation rate in deception settings (von Schenk et al., 2024). In our work, we show that providing users with transparent information about the model's performance and uncertainty can discourage them from blindly following an accusatory AI judgment. Confidence cues, in particular, play a critical role in this sense: when users learn that an algorithm displays high confidence in a deceptive judgment, they are less likely to accept it.

We argue that there are two possible explanations for this high-confidence aversion effect. First, humans' beliefs about their own ability in detecting deception, or that of the average human, could explain why participants tended to disagree more with AI predictions. Recent studies have suggested that the type of task assigned to an AI influences the extent to which humans rely on its output (Thorp et al., 2025): when performing objective tasks (e.g., counting dots in an image), people conform more to AI, while, in subjective tasks (e.g., interpreting emotions in an abstract image), they align more with responses of other humans (Riva et al., 2022). Being deception detection somehow subjective, it is likely that humans tend to rely more on their judgment than on that of an AI model. Furthermore, by overestimating human ability in spotting deception, individuals may perceive the high-confident model's judgments as overconfidence, thus raising skepticism toward AI predictions. Prior research already showed that laypeople and practitioners who strongly believed in the wrong cues for deception detection (e.g., behavioral and non-verbal cues such as gaze aversion) were also less inclined to rely on the most relevant ones (i.e., verbal cues). Similarly, individuals who hold strong beliefs about the average human performance in detecting deception might be less inclined to trust an AI model, regardless of its stated level of accuracy.

The distrust in AI possibly also explains why we did not replicate a full truth-bias effect, as participants who relied more on their capabilities might have disagreed with AI judgments regardless of the classification label. However, since these associations are small, other factors may contribute to explaining these findings. A second explanation, through the lens of Truth Default Theory (TDT; Levine, 2014), is that participants were more comfortable making judgments with moderate levels of confidence. According to TDT, people generally assume honesty unless they have reasons to suspect otherwise. Highly confident AI accusations of

deception may have lacked sufficient justification to prevail over this truth-default status, thus favoring a more cautious stance, more aligned with their expectation of truthfulness, and acknowledging the social costs of falsely accusing someone of lying.

4.2 Humans-AI performance in deception detection

This study also tested whether human oversight on machine predictions was beneficial or simply worsened the algorithmic performance. We found participants exposed to predictions from a low-accuracy model performed only slightly better than chance (accuracy: 0.57), but without significantly improving the accuracy of the model itself. In contrast, participants guided by the high-accuracy model, despite showing a detection rate that was significantly better than chance (accuracy: 0.76), performed significantly lower than the high-accuracy model (accuracy: 0.90). Notably, had participants under the latter condition relied entirely on the algorithmic predictions, their performance would have closely matched the 89% accuracy. Moreover, this drop suggests that participants' modifications were not just a mere deviation in the confidence level but represent a substantial modification in the classification itself. In other words, participants deviated from AI judgments to an extent that reversed the classification judgment from *deceptive* to *truthful*, and vice versa.

These findings align with previous research suggesting that the human interplay with algorithmic predictions either worsens machine performance or is simply ineffective (Kleinberg & Verschuere, 2021; von Schenk et al., 2024). What we added to the previous body of research is that adding a transparency and interpretability layer on the model's accuracy and confidence is not sufficient to achieve a hybrid-performance that is better than that of the model in isolation. One possible interpretation for this finding is that the nature of the task itself might be a critical factor. Prior meta-analytical evidence showed that human-AI collaboration tends to be beneficial in tasks where humans already outperform AI, whereas impairments are observed when AI alone outperforms humans (Vaccaro et al., 2024). Given that humans are known to be poor at deception detection (Bond & DePaulo, 2006; DePaulo et al., 2003; Hartwig & Bond, 2011), it is plausible that, for this task, a hybrid modality is unlikely to exceed the performance of an AI model in isolation.

4.3 Limitations and future research

A few limitations are worth mentioning for this study. First, our study entails the fact that the model's predictions were always available to participants. In related work (von Schenk et al., 2024), some participants always received predictions from a lie-detection algorithm (i.e., forced condition) before making their own judgments, while others could choose to request such predictions (i.e., choice condition), and others performed the task without any AI aid (i.e., control condition). Their findings show that model availability increases the accusation rate, especially among those who actively seek such algorithmic predictions. Our study adds more nuance to these findings by showing that including a transparency layer on the model's performance and uncertainty can discourage users from blindly following an accusatory AI judgment. However, unlike von Schenck et al. (2024), our study explored the role of transparency on trust exclusively under a *forced condition*, namely, with participants always having access to the AI predictions. Future studies should, instead, investigate the role of transparency in human trust by also including conditions where participants are free to choose whether to access or not to AI predictions (i.e., choice condition).

Second, we run our experiment in a non-incentivised task, and we did not provide any information about possible consequences that the author of the statements would have experienced if accused of lying. It might be that by adding this information, participants could have been even more averse towards over-confident AI predictions. Future research may try to investigate the role of accuracy and confidence in an incentivized task that also incorporates the possibility to choose whether or not participants want to seek AI predictions.

Third, in our study, humans were not aware of their inherent limited capability for deception detection, and, indeed, we found that greater deviations positively correlated with this erroneous belief that their own performance, or that of the average human, is usually better than that of an AI model. Future research should explore whether and how informing users about their inherent limitations in deception detection changes their trust in AI predictions. Finally, while we informed participants on the model's accuracy and predicting confidence, the role of the model's explainability was not explored in this study. **Explainability** has been proposed as a mechanism to improve trust and decision-making by providing users with insights into the rationale behind predictions and by potentially enabling better understanding and error correction (Alufaisan et al., 2021). However, while some studies from other domains support the benefits of explainability (Balasubramaniam et al., 2023; Oswald et al., 2018; Sadeghi et al., 2024), meta-analytic evidence

suggests that its impact on performance may not exceed that of AI predictions alone (Schemmer et al., 2022). These mixed findings highlight the need for further research into how and whether explainability can improve human trust in AI, specifically for deception detection.

Conclusion

This study investigated how the accuracy and confidence of a fictitious AI model for deception detection influence human reliance on AI-based veracity judgments. Human judges refrained from relying on highly confident AI predictions, especially when accusing someone of deception or stemming from a low-accuracy model. Human aversion toward AI predictions was partially explained by participants' tendency to overestimate their own deception detection capabilities. In line with the truth-default theory, such aversion also reflected users' cautiousness in deception accusations, given the social costs of falsely accusing someone of lying. This caution, however, comes at the expense of reducing individuals' own ability to detect deception effectively.

References

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does Explainable Artificial Intelligence Improve Human Decision-Making? *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *8A*, 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *Int. J. Clin. Health Psychol.*, *16*(2), 201–210. https://doi.org/10.1016/j.ijchp.2016.01.002

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197. https://doi.org/10.1016/J.INFSOF.2023.107197

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/S15327957PSPR1003_2

Bulten, W., Balkenhol, M., Belinga, J. J. A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., Pereira, G., Roy, P., Saile, G., Salles, P., Schaafsma, E., Tschui, J., Vos, A. M., Delahunt, B., Samaratunga, H., … Litjens, G. (2020). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology 2020 34:3*, *34*(3), 660–671. https://doi.org/10.1038/s41379-020-0640-y

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

Delmas, H., Denault, V., Burgoon, J. K., & Dunbar, N. E. (2024). A Review of Automatic Lie Detection from Facial Features. *Journal of Nonverbal Behavior 2024 48:1*, *48*(1), 93–136. https://doi.org/10.1007/S10919-024-00451-2

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837. https://doi.org/10.2307/2531595

DePaulo, B. M., Malone, B. E., Lindsay, J. J., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality monitoring: A meta-analytical review for forensic practice. *Eur. J. Psychol. Appl. Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/ejpalc2021a10

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/A0023589

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661–676. https://doi.org/10.1002/ACP.3052

Hartwig, M., Granhag, P. A., & Luke, T. (2014). Strategic Use of Evidence During Investigative Interviews: The State of the Science. *Credibility Assessment: Scientific Research and Applications*, 1–36. https://doi.org/10.1016/B978-0-12-394433-7.00001-4

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personal. Soc. Psychol. Rev.*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

He, G. ; Buijsman, S. ; & Gadiraju, U. (2023). *How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System*. *7*(CSCW2), 29. https://doi.org/10.1145/3610067

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-Machine Collaboration for Content Regulation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *26*(5). https://doi.org/10.1145/3338243

Kleinberg, B., Arntz, A., & Verschuere, B. (2019a). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/JOURNAL.PONE.0220228

Kleinberg, B., Arntz, A., & Verschuere, B. (2019b). Detecting deceptive intentions: Possibilities for large-scale applications. *The Palgrave Handbook of Deceptive Communication*, 403–427. https://doi.org/10.1007/978-3-319-96334-1_21/TABLES/3

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2017). Using named entities for computer-automated verbal deception detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica*, *213*. https://doi.org/10.1016/J.ACTPSY.2020.103250

Kotsoglou, K. N., & Oswald, M. (2020). The long arm of the algorithm? Automated Facial Recognition as evidence and trigger for police intervention. *Forensic Science International: Synergy*, *2*, 86. https://doi.org/10.1016/J.FSISYN.2020.01.002

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Levine, T. R., Kim, R. K., & Blair, J. P. (2010). (In)accuracy at Detecting True and False Confessions and Denials: An Initial Test of a Projected Motive Model of Veracity Judgments. *Human Communication Research*, *36*(1), 82–102. https://doi.org/10.1111/J.1468-2958.2009.01369.X

Levine, T. R., Kim, R. K., & Hamel, L. M. (2010). People Lie for a Reason: Three Experiments Documenting the Principle of Veracity. *Communication Research Reports*, *27*(4), 271–285. https://doi.org/10.1080/08824096.2010.496334

Li, J., Yang, Y., Liao, Q. V., Zhang, J., & Lee, Y. C. (2025). As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. *Conference on Human Factors in Computing Systems - Proceedings* . https://dl.acm.org/doi/10.1145/3706598.3713336

Li, J., Yang, Y., Zhang, R., Liao, Q. V., Song, T., Xu, Z., & Lee, Y. (2024). Understanding the Effects of Miscalibrated AI Confidence on User Trust, Reliance, and Decision Efficacy. *Preprint at* https://doi.org/10.48550/arXiv.2402.07632

Loconte, R., Battaglini, C., Maldera, S., Pietrini, P., Sartori, G., Navarin, N., & Monaro, M. (2025). Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Processing. *Journal of Language and Social Psychology*. https://doi.org/10.1177/0261927X251316883

Loconte, R., & Kleinberg, B. (2025). Examining embedded lies through computational text analysis. *Scientific Reports 2025 15:1*, *15*(1), 1–16. https://doi.org/10.1038/s41598-025-11327-w

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. *Scientific Reports 2023 13:1*, *13*(1), 1–19. https://doi.org/10.1038/s41598-023-50214-0

Melis, G., Ursino, M., Scarpazza, C., Zangrossi, A., & Sartori, G. (2024). Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions. *Scientific Reports*, *14*(1). https://doi.org/10.1038/S41598-024-63156-Y

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Sci Rep*, *8*(1), 1976. https://doi.org/10.1038/s41598-018-20462-6

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, *12*(5). https://doi.org/10.1371/JOURNAL.PONE.0177851

Orsini, F., Cioffi, A., Cipolloni, L., Bibbò, R., Montana, A., De Simone, S., & Cecannecchia, C. (2025). The application of artificial intelligence in forensic pathology: a systematic literature review. *Frontiers in Medicine*, *12*, 1583743. https://doi.org/10.3389/FMED.2025.1583743

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223–250. https://doi.org/10.1080/13600834.2018.1458455

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015). Deception detection using real-life trial data. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, 59–66. https://doi.org/10.1145/2818346.2820758

Riva, P., Aureli, N., & Silvestrini, F. (2022). Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica*, *229*, 103681. https://doi.org/10.1016/J.ACTPSY.2022.103681

Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/J.COMPELECENG.2024.109370

Schemmer, M., Hemmer, P., Nitsche, M., Kuhl, N., & Vossing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence

in Human-AI Decision-Making. *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The Prevalence of Lying in America: Three Studies of Self-Reported Lies. *Human Communication Research*, *36*(1), 2–25. https://doi.org/10.1111/J.1468-2958.2009.01366.X

Sporer, S. L., Hauch, V., Masip, J., & Martschuk, N. (2025). A Meta-Analysis of Field Studies on Criteria-Based Content Analysis. *European Psychologist*. https://doi.org/10.1027/1016-9040/A000561

Suchotzki, K., Verschuere, B., Bockstaele, B. Van, Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/BUL0000087

Thorp, S. O., Slupphaug, S., Rimol, L. M., Lervik, S., Kristoffer, S., Hoel, B., & Grassini, S. (2025). Conformity towards humans versus AI in different task domains: the type of task matters. *Journal of Psychology and AI*, *1*(1). https://doi.org/10.1080/29974100.2025.2540762

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour 2024 8:12*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

van Dijck, G. (2022). Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy and Research*, *28*(3), 407–423. https://doi.org/10.1007/s10610-022-09516-8

Velutharambath, A., & Klinger, R. (2023). UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 39–51. https://doi.org/10.18653/V1/2023.WASSA-1.5

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 2023 7:5*, *7*(5), 718–728. https://doi.org/10.1038/s41562-023-01556-2

von Schenk, A., Klockmann, V., Bonnefon, J. F., Rahwan, I., & Köbis, N. (2024). Lie detection algorithms disrupt the social dynamics of accusation behavior. *IScience*, *27*(7), 110201. https://doi.org/10.1016/j.isci.2024.110201

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/JIP.82

Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300509

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

Supplementary Material

1. Optimal linear-mixed model structure

To determine the optimal model structure, we compared three nested linear mixed-effects models predicting the absolute deviation from AI judgments.

Model 1 included fixed effects for the two-way interaction between Accuracy and Confidence, considered AI Classification as a covariate, and included random intercepts for participants and statements. Model 2 added a three-way interaction among the predictors. Model 3 included the same fixed effects as Model 2 but added random slopes for Confidence by participant to account for individual differences in how participants adjust their deviation across confidence levels.

1) $\Delta y = Accuracy * Confidence + Classification +$
$(1 \mid \text{Participant\_id}) + (1 \mid \text{Statement\_id})$

2) $\Delta y = Accuracy * Confidence * Classification +$
$(1 \mid \text{Participant\_id})) + (1 \mid \text{Statement\_id})$

3) $\Delta y = Accuracy * Confidence * Classification + (1 +$
$\text{Confidence} \mid \text{Participant\_id}) + (1 \mid \text{Statement\_id})$

A likelihood ratio test comparing the models (refitted using maximum likelihood) showed that Model 2 provided a significantly better fit than Model 1, $\chi^2(9) = 589.41$, $p < .001$, and Model 3 did not significantly outperform Model 2, $\chi^2(14) = 12.35$, $p = .578$.

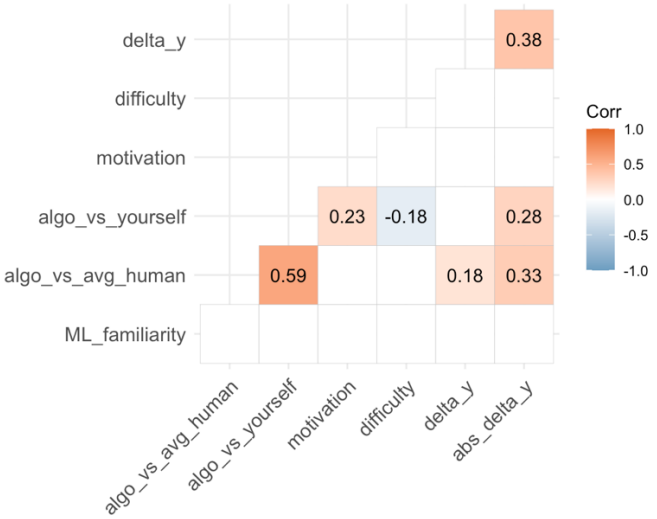
Based on these results, we selected Model 2 as the final model, indicating that the inclusion of a three-way interaction and a random intercept for participants and statements was enough to significantly improve the model fit.

2. Preliminary Analysis

TABLE S1. Mean, standard deviation, and median values of participants' deviation from AI (Δy) across conditions.

		Accuracy low			Accuracy high		
Classification	Confidence	*M*	*SD*	*Median*	*M*	*SD*	*Median*
Deceptive	Indecisive	5.83	24.75	8	7.85	21.80	6
	Poorly Confident	15.87	26.31	15.5	11.57	23.51	6
	Moderately Confident	21.01	25.82	15	14.9	23.43	10
	Confident	27.21	28.07	20	25.16	26.15	19.5
	Very Confident	39.5	29.90	31	26.37	26.45	20
Truthful	Indecisive	6.07	23.09	12	8.20	19.54	9.5
	Poorly Confident	0.12	22.55	6	4.55	18.91	9
	Moderately Confident	-10.15	22.06	-3	-3.68	20.37	1.5
	Confident	-12.07	22.73	-6	-8.45	22.0	-1.5
	Very Confident	-22.61	25.17	-15	-15.06	20.88	-9

FIGURE S1. Correlation matrix of Spearman correlations between the participant's average deviation and absolute deviation and covariates of motivation, difficulty, machine learning familiarity, belief in performing better than the AI, and belief that the average human performs better than the AI.



Note: Only significant correlations at $p < .01$ after *false discovery rate* correction for multiple comparisons are displayed in the correlation matrix.
Abbreviations:
(1) abs_delta_y = average absolute human deviation at the participant-level (dependent variable); (2) delta_y = average human deviation at the participant-level (dependent variable); (3) algo_vs_avg_human = covariate of belief that average human's performance is better than that of the AI; (4) algo_vs_yourself = covariate of belief that individual's performance is better than that of the AI; (5) difficulty = covariate of difficulty of the the task; (6) ML_familarity = covariate of familiarity with machine learning models; (7) motivation = covariate of motivation in performing the task well.

3. Confirmatory Analysis

TABLE S2. Variance, Standard Deviation (SD), and adjusted Intraclass Correlation Coefficient (ICC) for Random Effects from Model 2.

Random Effects			
Group	Variance	SD	ICC
Participant_id	39.44	6.28	0.07
Statement_id	35.91	5.99	0.06
Residual	501.00	22.38	-

TABLE S3. Pairwise contrasts for Confidence range by Classification.

Pairwise contrast	AI Classification = deceptive		AI Classification = truthful	
	Diff (SE)	*d* (99% CI)	*Diff* (*SE*)	*d* (99% CI)
indecisive – poorly confident	-6.36 (2.14)	-0.14 (-0.27, -0.02)	3.86 (2.13)	0.09 -(0.04, 0.22)
indecisive – moderately confident	-11.14 (2.27) **	-0.27 (-0.42, -0.13)	13.10 (2.43) **	0.27 (0.14, 0.40)
indecisive – confident	-19.01 (2.21) **	-0.45 (-0.58, -0.31)	16.84 (2.23) **	0.38 (0.25, 0.52)
indecisive – very confident	-26.30 (2.15) **	-0.63 (-0.78, -0.49)	24.93 (2.22) **	0.55 (0.42, 0.69)
poorly confident – moderately confident	-4.78 (2.18)	-0.12 (-0.26, -0.02)	9.23 (2.25) **	0.21 (0.08, 0.34)
poorly confident – confident	-12.65 (2.11) **	-0.30 (-0.43, -0.17)	12.98 (2.03) **	0.34 (0.20, 0.48)
poorly confident – very confident	-19.95 (2.05) **	-0.48 (-0.61, -0.35)	21.06 (2.01) **	0.53 (0.39, 0.67)
moderately confident – confident	-7.87 (2.24) **	-0.20 (-0.34, -0.05)	3.74 (2.35)	0.08 (-0.05, 0.21)
moderately confident – very confident	-15.16 (2.18) **	-0.39 (-0.54, -0.24)	11.83 (2.33) **	0.25 (0.12, 0.38)
confident – very confident	-7.29 (2.11) *	-0.18 (-0.31, -0.04)	8.09 (2.12) **	0.20 (0.06, 0.33)

Note. *M* = estimated marginal means; *SE* = standard error; *d* = Cohen's *d*, *CI* = confidence interval.
* $p < .011$
** $p < .001$

TABLE S4. Pairwise contrasts for Confidence range by Classification under the low-accuracy condition.

	Low-accuracy condition			
Pairwise contrast	AI Classification = deceptive		AI Classification = truthful	
	Diff (*SE*)	*d* (99% CI)	*Diff* (*SE*)	*d* (99% CI)
indecisive – poorly confident	-9.60 (2.90) *	-0.16 (-0.28, -0.03)	4.37 (2.92)	0.07 (-0.05, 0.19)
indecisive – moderately confident	-15.47 (3.06) **	-0.27 (-0.41, -0.13)	14.62 (3.24) **	0.22 (0.09, 0.35)
indecisive – confident	-20.79 (2.96) **	-0.35 (-0.49, -0.22)	17.12 (3.06) **	0.27 (0.14, 0.40)
indecisive – very confident	-33.44 (2.91) **	-0.57 (-0.70, -0.43)	26.83 (3.02) **	0.42 (0.30, 0.55)
poorly confident – moderately confident	-5.86 (2.96)	-0.10 (-0.23, 0.03)	10.24 (2.98) *	0.17 (0.04, 0.31)
poorly confident – confident	-11.19 (2.85) *	-0.19 (-0.31, -0.06)	12.75 (2.78) **	0.23 (0.10, 0.36)
poorly confident – very confident	-23.84 (2.81) **	-0.40 (-0.52, -0.27)	22.46 (2.73) **	0.40 (0.27, 0.54)
moderately confident – confident	-5.32 (3.01)	-0.09 (-0.23, 0.04)	2.50 (3.11)	0.04 (-0.09, 0.17)
moderately confident – very confident	-17.97 (2.97) **	-0.32 (-0.46, -0.18)	12.21 (3.07) **	0.20 (0.07, 0.33)
confident – very confident	-12.65 (2.87) **	-0.22 (-0.35, -0.09)	9.71 (2.88) *	0.17 (0.04, 0.30)

Note. Diff = estimated difference of marginal means; *SE* = standard error; *d* = Cohen's *d*, *CI* = confidence interval.
* $p < .01$
** $p < .001$

TABLE S5. Pairwise contrasts for Confidence range by Classification under the high-accuracy condition.

	High-accuracy condition			
Pairwise contrast	AI Classification = deceptive		AI Classification = truthful	
	Diff (*SE*)	*d* (99% CI)	*Diff* (*SE*)	*d* (99% CI)
indecisive – poorly confident	-3.42 (2.68)	-0.06 (-0.18, 0.06)	3.31 (2.60)	0.06 (-0.06, 0.19)
indecisive – moderately confident	-6.75 (2.77)	-0.13 (-0.27, 0.01)	11.30 (3.06) *	0.17 (0.05, 0.29)
indecisive – confident	-17.21 (2.74) **	-0.31 (-0.44, -0.18)	16.42 (2.72) **	0.30 (0.17, 0.43)
indecisive – very confident	-18.69 (2.64) **	-0.36 (-0.49, -0.22)	22.61 (2.74) **	0.38 (0.26, 0.51)
poorly confident – moderately confident	-3.32 (2.65)	-0.07 (-0.20, 0.07)	7.99 (2.85)	0.13 (0.01, 0.25)
poorly confident – confident	-13.79 (2.62) **	-0.25 (-0.37, -0.13)	13.12 (2.47) **	0.27 (0.14, 0.41)
poorly confident – very confident	-15.27 (2.51) **	-0.29 (-0.42, -0.17)	19.30 (2.49) **	0.38 (0.25, 0.51)
moderately confident – confident	-10.46 (2.72) *	-0.21 (-0.35, -0.07)	5.12 (2.95)	0.08 (-0.04, 0.20)
moderately confident – very confident	-11.95 (2.61) **	-0.26 (-0.41, -0.11)	11.31 (2.97) *	0.17 (0.05, 0.29)
confident – very confident	-1.48 (2.58)	-0.03 (-0.16, 0.10)	6.19 (2.62)	0.12 (-0.01, 0.24)

Note. *Diff* = estimated difference of marginal means; *SE* = standard error; *d* = Cohen's *d*, *CI* = confidence interval.
* $p < .01$
** $p < .001$

4. Robustness check by including covariates

Covariates:

- Motivation: "*How much were you motivated to perform well?*" (0=Not at all, 10=Very much)
- Difficulty: "*How difficult did you find the study?*" (0=Very easy, 10=Very difficult)
- ML familiarity: "*How familiar are you with AI-based algorithms?*" (0=Not familiar at all, 5=Neutral, 10=Very familiar)
- AI vs Average human: "*How good do you think the **average human performance** is compared to the performance of the AI-based lie detector in predicting whether a statement is true or false?*" (0=Algorithm's performance is better, 5=Equal, 10=Human's performance is better)
- AI vs Yourself: "*How good do you think **your performance** is compared to the performance of the AI-based lie detector in predicting whether a statement is true or false?*" (0=Algorithm's performance is better, 5=Equal, 10=My performance is better)

TABLE S6. Type III Analysis of Variance for the Linear Mixed Model predicting human deviation.

Effect	Covariates	Sum of Sq	Mean Sq	Num DF	Den DF	*F*-value	η^2 (99% CI)
Accuracy	Excluded	9	9	1	377.5	0.02	0.00 (0.00, 0.00)
	Included	244	244	1	373.0	0.49	0.00 (0.00, 0.03)
Confidence	Excluded	928	232	4	180.0	0.46	0.01 (0.00, 0.05)
	Included	920	230	4	180.3	0.46	0.01 (0.00, 0.05)
Classification	Excluded	514,132	514,132	1	3,425.6	**1026.21****	0.23 (0.20, 0.26)
	Included	512,845	512,845	1	3,427.3	**1023.20****	0.23 (0.20, 0.26)
Accuracy × Confidence	Excluded	2,594	649	4	3,241.6	1.29	0.00 (0.00, 0.01)
	Included	2,669	667	4	3,242.5	1.22	0.00 (0.00, 0.01)
Accuracy × Classification	Excluded	20,655	20,655	1	3,426.1	**41.23****	0.01 (0.00, 0.02)
	Included	20,854	20,854	1	3,431.6	**41.61****	0.01 (0.00, 0.02)

Effect	Covariates	Sum of Sq	Mean Sq	Num DF	Den DF	*F*-value	η^2 (99% CI)
Confidence × Classification	Excluded	287,802	71,951	4	3,426.4	**143.61****	0.14 (0.12, 0.017)
	Included	287,726	71,931	4	3,460.5	**143.51****	0.14 (0.12, 0.017)
Accuracy × Confidence × Classification	Excluded	8,370	2,092	4	3,426.4	**4.18***	0.01 (0.00, 0.01)
	Included	8,470	2,118	4	3,468.6	**4.22***	0.01 (0.00, 0.01)

Note. For this model, the bobyqa optimiser was used to fit the model. Results are based on Satterthwaite's. In **bold** are reported significant *F*-values.
Abbreviations: Num DF= number of degrees of freedom; Sq = squares
* $p < .001$

5. Robustness check by including previously excluded participants

TABLE S7. Type III Analysis of Variance for the Linear Mixed Model predicting human deviation.

Effect	Sample	Sum of Sq	Mean Sq	Num DF	Den DF	*F*-value	η^2 (99% CI)
Accuracy	Reduced	9	9	1	377.5	0.02	0.00 (0.00, 0.00)
	Full	53	53	1	499.7	0.10	0.00 (0.00, 0.00)
Confidence	Reduced	928	232	4	180.0	0.46	0.01 (0.00, 0.05)
	Full	1730	432	4	182.2	0.85	0.02 (0.00, 0.08)
Classification	Reduced	514,132	514,132	1	3,425.6	**1026.21****	0.23 (0.20, 0.26)
	Full	749,367	749,367	1	4,591.6	**1470.48****	0.24 (0.22, 0.27)
Accuracy × Confidence	Reduced	2,594	649	4	3,241.6	1.29	0.00 (0.00, 0.01)
	Full	1,942	486	4	4,341.4	0.95	0.00 (0.00, 0.00)
Accuracy × Classification	Reduced	20,655	20,655	1	3,426.1	**41.23****	0.01 (0.00, 0.02)
	Full	13,602	13,602	1	4,596.6	**26.69****	0.01 (0.00, 0.01)
Confidence × Classification	Reduced	287,802	71,951	4	3,426.4	**143.61****	0.14 (0.12, 0.017)
	Full	393,315	98,329	4	4,624.5	**192.95****	0.14 (0.12, 0.017)
Accuracy × Confidence × Classification	Reduced	8,370	2,092	4	3,426.4	**4.18***	0.01 (0.00, 0.01)
	Full	5,866	1,466	4	4,627.2	2.88	0.00 (0.00, 0.01)

Note. For this model, the bobyqa optimiser was used to fit the model. Results are based on Satterthwaite's. *Reduced* sample refers to the sample after excluding participants who failed attention and statistical checks. *Full* sample refers to the whole sample of participants collected without any exclusion. In **bold** are reported significant *F*-values.
Abbreviations: Num DF= number of degrees of freedom; Sq = squares
* $p < .01$
** $p < .001$

6. Absolute values of human deviation

Model equation:

$|\Delta y| = Accuracy * Confidence * Classification + (1|\text{ Participant\_id})) + (1|\text{ Statement\_id})$

TABLE S8. Variance, Standard Deviation (SD), and adjusted Intraclass Correlation Coefficient (ICC) for Random Effects.

Random Effects			
Group	Variance	SD	ICC
Participant_id	45.84	6.77	0.13
Statement_id	2.05	1.43	0.01
Residual	318.11	17.84	-

TABLE S9. Type III Analysis of Variance for the Linear Mixed Model predicting absolute values of human deviation.

Effect	Sum of Sq	Mean Sq	Num DF	Den DF	*F*-value	η^2 (99% CI)
Accuracy	7,126	7,126	1	376.9	**22.40****	0.06 (0.01, 0.13)
Confidence	27,204	6,801	4	159.4	**21.38****	0.35 (0.19, 0.47)
Classification	56,292	56,292	1	3,440.4	**176.96****	0.05 (0.03, 0.07)
Accuracy × Confidence	6,712	1,678	4	3,293.6	**5.27****	0.00 (0.00, 0.01)
Accuracy × Classification	1,401	1,401	1	3,441.9	4.40	0.00 (0.00, 0.01)
Confidence × Classification	16,704	4,1761	4	3,463.4	**13.13****	0.01 (0.01, 0.03)
Accuracy × Confidence × Classification	591	148	4	3,470.4	0.46	0.00 (0.00, 0.00)

Note. For this model, the bobyqa optimiser was used to fit the model. Results are based on Satterthwaite's approximation for degrees of freedom. In **bold** are reported significant *F*-values.
Abbreviations: Num DF= number of degrees of freedom; Sq = squares
* $p < .01$, ** $p < .001$

TABLE S10. Pairwise contrasts for the main effects of accuracy, confidence, and classification.

Main effect	Pairwise contrast	*Diff.* (*SE*)	*d* (99% CI)
Accuracy	low-accuracy – high-accuracy	4.36 (0.92) **	0.24 (0.11, 0.38)
Confidence	indecisive – poorly confident	-0.06 (0.99)	-0.00 (-0.19, 0.18)
	indecisive – moderately confident	-0.30 (1.07)	-0.02 (-0.23, 0.18)
	indecisive – confident	-3.17 (1.02)	-0.24 (-0.44, -0.04)
	indecisive – very confident	-7.35 (1.01) **	-0.55 (-0.75, -0.34)
	poorly confident – moderately confident	-0.24 (1.01)	-0.02 (-0.22, 0.18)
	poorly confident – confident	-3.11 (0.95)	-0.25 (-0.45, -0.05)
	poorly confident – very confident	-7.28 (0.94) **	-0.58 (-0.78, -0.37)
	moderately confident – confident	-2.86 (1.04)	-0.23 (-0.44, -0.01)
	moderately confident – very confident	-7.04 (1.03) **	-0.55 (-0.77, -0.33)
	confident – very confident	-4.18 (0.97) **	-0.33 (-0.54, -.013)
Classification	deceptive – truthful	8.1 (0.61) **	0.23 (0.18, 0.27)

Abbreviations: *Diff.* = Estimated difference of marginal means; *SE* = standard error; *d* = Cohen's d; CI = confidence interval
** $p < .001$

TABLE S11. Pairwise contrasts for Confidence range by Classification.

Pairwise contrast	AI Classification = deceptive		AI Classification = truthful	
	Diff (*SE*)	*d* (99% CI)	*Diff* (*SE*)	*d (99% CI)*
indecisive – poorly confident	-2.28 (1.38)	-0.07 (-0.17, 0.04)	2.16 (1.36)	0.06 (-0.04, 0.17)
indecisive – moderately confident	-3.18 (1.41)	-0.11 (-0.23, 0.02)	2.58 (1.57)	0.07 (-0.04, 0.17)
indecisive – confident	-8.94 (1.40) **	-0.27 (-0.39, -0.16)	2.60 (1.43)	0.08 (-0.03, 0.18)
indecisive – very confident	-13.15 (1.36) **	-0.41 (-0.53, -0.30)	-1.55 (1.43)	-0.04 (-0.14, 0.06)
poorly confident – moderately confident	-0.90 (1.37)	-0.03 (-0.15, 0.09)	0.42 (1.44)	0.01 (-0.09, 0.12)
poorly confident – confident	-6.65 (1.36) **	-0.20 (-0.30, -0.09)	0.44 (1.28)	0.02 (-0.10, 0.13)
poorly confident – very confident	-10.87 (1.32) **	-0.33 (-0.43, -0.22)	-3.70 (1.29)	-0.12 (-0.22, -0.01)
moderately confident – confident	-5.75 (1.39) **	-0.20 (-0.32, -0.07)	0.02 (1.50)	0.00 (-0.11, 0.11)
moderately confident – very confident	-9.96 (1.35) **	-0.36 (-0.48, -0.23)	-4.12 (1.50)	-0.11 (-0.22, -0.01)
confident – very confident	-4.21 (1.34)	-0.13 (-0.24, -0.02)	-4.15 (1.35)	-0.13 (-0.23, -0.02)

Note. Diff = estimated difference of marginal means; *SE* = standard error; *d* = Cohen's *d*, *CI* = confidence interval.
** $p < .001$

TABLE S12. Pairwise contrasts for Confidence range by Accuracy condition.

Pairwise contrast	Low-accuracy condition		High-accuracy condition	
	Diff (*SE*)	*d* (99% CI)	*Diff* (*SE*)	*d* (99% CI)
indecisive – poorly confident	-0.52 (1.34)	-0.02 (-0.12, 0.09)	0.40 (1.38)	0.01 (-0.09, 0.11)
indecisive – moderately confident	-1.02 (1.43)	-0.03 (-0.15, 0.09)	0.41 (1.52)	0.01 (-0.10, 0.12)
indecisive – confident	-3.06 (1.37)	-0.10 (-0.21, 0.02)	-3.28 (1.43)	-0.09 (-0.20, 0.01)
indecisive – very confident	-10.85 (1.36) **	-0.34 (-0.46, -0.23)	-3.84 (1.41)	-0.11 (-0.21, -0.01)
poorly confident – moderately confident	-0.50 (1.35)	-0.02 (-0.14, 0.10)	0.02 (1.44)	0.00 (-0.11, 0.11)
poorly confident – confident	-2.54 (1.29)	-0.09 (-0.20, 0.03)	-3.67 (1.33)	-0.11 (-0.22, -0.01)
poorly confident – very confident	-10.33 (1.27) **	-0.35 (-0.46, -0.23)	-4.24 (1.31)	-0.13 (-0.24, -0.03)
moderately confident – confident	-2.04 (1.38)	-0.07 (-0.19, 0.05)	-3.69 (1.47)	-0.11 (-0.22, 0.00)
moderately confident – very confident	-9.83 (1.37) **	-0.34 (-0.47, -0.22)	-4.26 (1.46)	-0.12 (-0.23, -0.01)
confident – very confident	-7.79 (1.31) **	-0.27 (-0.38, -0.15)	-0.57 (1.36)	-0.02 (-0.13, 0.09)

Note. Diff = estimated difference of marginal means; *SE* = standard error; *d* = Cohen's *d*, *CI* = confidence interval.
** $p < .001$

Chapter 6

General discussion

1. Contribution of this thesis

Research at the intersection between psychology and computer science has shown the potential of computational methods stemming from machine learning (ML) and natural language processing (NLP) for the automated detection of verbal deception (Constancio et al., 2023; Hauch et al., 2015). However, further exploration of the topic was necessary to understand the extent to which we can rely on such automated methods. The present thesis aimed to address this research question by exploring key **opportunities** (Chapters 2 and 3) and **challenges** (Chapters 3, 4, and 5), among the many relevant issues, in automated verbal deception detection. The findings are discussed below with a focus on forensic contexts.

1.1 Opportunities for automated coding

The potential of automated verbal deception detection was first explored in terms of automated coding of statements (Chapters 2 and 3).

In this thesis, we could automatically extract several linguistic features, by resorting to NLP techniques, which we then linked to specific theoretical frameworks. In this manner, we could automate the most common manual approaches in psycholegal research. In one study, we employed this approach to mimic the Reality Monitoring (Johnson & Raye, 1981; Sporer, 1997, 2004) and Cognitive Load (Vrij et al., 2008) (see Chapter 2). In a second study, we developed a new stylometric technique, which we named **DeCLaRatiVE** (Chapter 3). **DeCLaRatiVE** is a theory-based stylometric technique that investigates 26 linguistic features associated with deception, grounded in the psychological frameworks of **D**istancing (Newman et al., 2003), **C**ognitive **L**oad (Vrij et al., 2008), **R**eality **M**onitoring (Johnson & Raye, 1981; Sporer, 1997, 2004), and **VE**rifiability Approach (Nahari et al., 2012; Vrij & Nahari, 2019). The extraction of these features combines statistical features of text, dictionary-based approaches (e.g., the Linguistic Inquiry Word Count - LIWC; Boyd et al., 2022; Tausczik et al., 2010), and machine-learning methods (i.e., named-entity recognition). **DeCLaRatiVE** was applied to extract and compare the verbal cues of deception that emerged from different contexts (e.g., personal opinions, autobiographical memories, and future intentions; see Chapter 3) and to compare more nuanced forms of deception (i.e., embedded lies) with their truthful counterparts (Chapter 4). In contrast to the time and effort that manual coding would have required, this technique enabled the analysis of verbal cues of deception from a multi-theoretical perspective in a single shot.

Human coding is, in psychological research, the standard practice for verbal deception detection. It involves training individuals to read and evaluate statements according to predefined criteria before providing a final judgment on veracity. For example, according to the Criteria-Based Content Analysis (Steller & Koehnken, 1989) and Reality Monitoring (Sporer, 2004, 1997), statements can be evaluated through specific criteria, such as the number of details, the plausibility of the story, or the vividness of the recollection. The presence of these criteria can be coded using frequency counts (i.e., counting the number of details), categorical measures (i.e., presence vs. absence), or rating scales (e.g., scoring from 1=none to 7=very much) (Gancedo et al., 2021). Similarly, along with the Verifiability approach (VA; Nahari et al., 2012; Vrij & Nahari, 2019), deceptive statements can be distinguished from truthful ones based on the number of verifiable details (e.g., information that can be double-checked by an independent witness, video footage, and other types of evidence). However, these procedures come with limitations that automated approaches can easily overcome. We present a close comparison between manual and automated approaches for coding statements in Table 1, underscoring the opportunities of automated methods against the limitations of manual approaches.

TABLE 1. Comparison between manual and automated coding in scalability, reliability, analysis of complex data, and learning.

Property	Manual coding	Automated coding
Scalability	Manual coding procedures (e.g., CBCA, RM, VA) require considerable effort and time and are, therefore, often validated on small sample sizes (Kleinberg et al., 2019).	Manual coding is both demanding and time-consuming, but not necessarily more accurate than automated methods (Hacking et al., 2023; Schutte et al., 2021; Szojka et al., 2025). Psychological research can rely on NLP techniques to automatically extract features from statements, thereby expanding investigations on a large scale and eliminating the need for validation on small sample sizes, which is typical of manual coding procedures.
Reliability	Manual coding procedures have been criticized for including criteria that are difficult to operationalize and interpret, thereby introducing bias and disagreements among raters (Hauch et al., 2017), raising	By coding linguistic features in an automated and replicable manner, we reduce concerns about raters' disagreement or differing interpretations of constructs, as the same statement

Property	Manual coding	Automated coding
	the possibility that different raters may code the same statement differently and, in the worst-case scenario, even reach divergent judgments regarding its veracity.	will always be coded in the same way, thereby enhancing objectivity and reproducibility.
Complexity of textual data	Manual approaches require human coders that focus on one or maximum a few criteria at a time, and rely on heuristics for their judgments (Verschuere et al., 2023).	Automated coding enables the possibility of extracting multiple features (see DeCLaRatiVE) and, by resorting to ML, those features can be combined to result in one single judgment.
Learning	Human learning is bounded, as it also requires time and efforts, and the scale of improvements is limited.	Automated approaches relying on ML models can learn from many data points (the more the better) and can be consistently updated with new data points to improve their predictions.

Taken together, these advantages highlight how automated and machine learning–based methods not only address the practical and methodological limitations of manual coding but also pave the way for more scalable, objective, and predictive approaches to deception detection.

1.2 Opportunities for automated prediction

The potential of automated verbal deception detection was also explored in terms of automated predictions (Chapters 2 and 3).

In four experiments, we closely compared the performance of naïve judges, forensic experts, theory-led, and data-driven models in detecting deception from transcripts of interviews with unexpected questions (Chapter 2). Naïve judges relied on gut feeling and common knowledge for their judgments. Forensic experts were trained in and employed Reality Monitoring for evaluating the credibility of statements. Previous meta-analytical evidence has shown that, using Reality Monitoring, deception can be detected with around 70% accuracy (Gancedo et al., 2021). Theory-led and data-driven models leveraged features extracted through NLP. The theories involved in the theory-led models were Reality Monitoring and Cognitive Load. For the data-driven, a selection of the most significant linguistic features was conducted. The results showed that both theory-led (accuracy=69.4%) and data-driven models (accuracy=77.3%) significantly outperformed naïve (accuracy=54.7%) and expert judges (accuracy=59.4%). While this finding was expected for naïve judges, who are known to be generally poor lie-detectors (Bond & DePaulo, 2006; Hartwig & Bond, 2011), it was somewhat unexpected for forensic experts, given previous research showing the effectiveness of

Reality Monitoring (Gancedo et al., 2021). From a practical perspective, this evidence suggests that computational methods may represent a valid alternative to aid forensic experts in those cases where they perform poorly, even when relying on well-established tools for credibility assessment.

In a following study, we built upon the findings on Chapter 2 by investigating two key aspects: 1) the efficacy of Large Language Models (LLMs) (i.e., small and base versions of FLAN-T5) in automated verbal deception detection and 2) the robustness of their performance across different domains (i.e., three datasets: personal opinions, autobiographical experiences, and future intentions; see Chapter 3). Unlike traditional statistical models that rely on preselected features, LLMs are models of human language that have been trained on extensive corpora and are able to generate texts that look indistinguishable to humans (Jakesch et al., 2023) and are able to perform well in a wide range of NLP tasks (Chung et al., 2022). One advantage of these models is their ability to be fine-tuned[17] for specific tasks (here: deception detection), while inherently computing rich text embeddings, thereby eliminating the need for manual feature engineering or selection. Instead, they process entire statements in input in the form of embeddings and, in the case of deception detection, adapt their learned language representations to try distinguishing between deceptive and truthful statements. At the time of writing and to the best of our knowledge, this was the first study trying to fine-tune an LLM for verbal deception detection. Our findings showed that fine-tuning LLMs on a single dataset yields state-of-the-art performance, outperforming a simpler baseline model (here: a bag-of-words model) and previous models on the same datasets (Capuozzo et al., 2020; Ilias et al., 2022; Kleinberg & Verschuere, 2021). However, while fine-tuning on one domain did not generalize to new domains, training on multiple domains improved robustness and further enhanced performance, reaching up to 79.31% accuracy.

We argue that the approach of fine-tuning LLM represents the state-of-the-art for automated verbal deception detection for three reasons. First, by training these models on a diversified dataset (i.e., including multiple domains), they achieve cross-domain generalizability. One important note is that the balance between LLM size and the number of domains is crucial, as our findings indicate that larger LLMs achieve higher performance, likely because greater dataset diversity requires a

[17] Importantly, this is true for early open-access LLMs, such as FLAN-T5 or Llama-2. In contrast, fine-tuning closed-source or very large models is often more challenging due to computational and accessibility constraints. In this latter case, prompt-tuning is more common.

larger model to reach higher accuracy. Second, once fine-tuned, these models can directly predict deception from raw texts, without requiring further preprocessing and feature extraction. This opens the door to real-life applications where deception models become available and accessible to a wider non-technical audience. Finally, given the aforementioned limitations of manual approaches to deception detection, such as difficulties in scalability and reliability, complexity of data, and limited human improvements through training, the approach of fine-tuning LLMs is by far a more robust and scalable solution.

1.3 Challenges for a universal model of deception

While the first chapters showcased the potential of automated methods for detecting deception, this thesis also explores their limitations and challenges. The two main challenges we explored for automated verbal deception detection entail i) their capability to generalize across different domains and types of deception, and ii) the extent of their adoption or aversion by a wider audience for real-life applications.

If we aim for a universal model of deception, such a model should be able to generalize its performance across different domains, such as deception in political opinions, fake alibis, and malicious intentions, as well as various deceptive strategies, including fabrications, omissions, and embedded lies.

Generalization across domains

Our findings on fine-tuning LLMs for deception detection (Chapter 3) showed that such models require training on datasets that encompass different domains to generalize across multiple domains; otherwise, performance drops to chance level. Previous studies have already emphasized this challenge (Velutharambath & Klinger, 2023), extended it to difficulties in generalizing across different cultural dimensions of individualism and collectivism (Papantoniou et al., 2022), and some have even questioned whether deception can actually be detected at all (Velutharambath et al., 2025). This highlights a critical challenge: domain-specific linguistic patterns have a strong influence on model predictions, making single-domain training insufficient for real-world applications where deceptive statements vary widely in context. To overcome this limitation, future research should prioritize the development and training of multi-domain models of deception.

However, this solution is not that straightforward. While training or fine-tuning language models on multi-domain datasets of deception can open new avenues for a foundation model of deception (e.g., as already

claimed for a foundation model of human cognition; Binz et al., 2025), it also raises a fundamental concern, as deception is inherently contextual (Levine, 2014). In fact, linguistic markers of deception vary not only across topics (Velutharambath & Klinger, 2023) but also across cultures (Papantoniou et al., 2022) and contexts (Markowitz & Hancock, 2019), making it difficult to expect any model to cover all possible scenarios for a definitive and universal solution.

Generalization across deceptive strategies

Findings in Chapters 2 and 3 highlight the potential of automated methods for predicting deception in fabricated narratives. However, fabrication is the most extreme form of deception and represents only one of the many deception strategies available. In fact, in the worst-case scenario, the most critical information is not even fabricated: it is simply omitted (Leal et al., 2020, 2023). To account for this limitation, we investigated deception from a more nuanced and realistic perspective, acknowledging that deception lies on a continuum where truthful and deceptive information merge, and communications span from being completely truthful to completely fabricated, or mixed with embedded lies (Bell & DePaulo, 1996; Hartwig et al., 2007; Leins et al., 2013; Nahari et al., 2012; Wang et al., 2004; see Chapter 4). Little research has already investigated how people's lies are embedded in overall truthful narratives (Caso et al., 2023; Verigin et al., 2020), and, as shown in our systematic review (Chapter 1), even less research has been conducted on the effectiveness of computational methods in detecting embedded lies (Markowitz, 2024). Therefore, we further addressed this knowledge gap by developing a new large dataset of embedded lies to understand whether what we know about fabrication actually translates to embedded lies (Chapter 4).

Our results showed that embedded lies represent a significant challenge for automated verbal deception detection (but also for deception detection in general), due to their incorporation of truthful information. We examined the nature of embedded lies and related individual differences and linguistic properties, finding that typical deceptive statements are overall truthful, with only 1/3 of embedded lies, mostly stemming from past personal experiences, and with minimal individual and linguistic differences from completely truthful narratives. By relying again on fine-tuning LLMs (here: Llama-3-8B), this time, we could only achieve 64% accuracy in classifying truthful from deceptive statements with embedded lies. Notably, when resorting to the fine-tuned FLAN-T5 (from Chapter 3), its performance dropped drastically to 56% (its accuracy in Chapter 3 was 79.31%), albeit without difficulties in predicting truthful

statements, but with evident problems in detecting embedded lies correctly. Explainability analysis revealed that classification becomes more straightforward as the number of embedded lies increases and deception approaches fabrication. These findings, altogether, indicate that some forms of deception are more challenging than others, and that computational models require adaptation to various deceptive strategies for real-life applications.

1.4 Challenges for human adoption of algorithmic predictions

In this thesis, we showed how automated methods may represent the state-of-the-art for verbal deception detection and should be taken as a valid alternative to aid human decision-makers in assessing the credibility of statements when traditional methods fall short (Chapters 2 and 3).

If we envision a near future where the inherent limitations of automated verbal deception detection are overcome and a definitive deception language model becomes widely available, we should already be researching and discussing how to calibrate human trust in AI predictions. In high-stakes contexts, such as in the forensic domain, human judges must remain central to decision-making, with AI serving only as a supportive tool (Kotsoglou & Oswald, 2020; Orsini et al., 2025; van Dijck, 2022). This highlights the importance of research on how humans can benefit from AI assistance while maintaining responsibility for their own final decisions. Put differently, it is crucial to understand the conditions under which AI-based judgments on statement veracity are accepted or rejected by human decision-makers, and whether this interplay enhances detection accuracy.

However, while hybrid decision-making was already investigated in online content moderation (Jhaver et al., 2019) and medical diagnoses (Bulten et al., 2020), it remained understudied in the context of deception detection, with only two studies available (Kleinberg & Verschuere, 2021; von Schenk et al., 2024). In Chapter 4, we built upon previous studies and addressed this topic by conducting an experiment that investigated how transparency about a model's overall accuracy and confidence in individual predictions influences human trust in AI decisions. Our results showed that humans tended to follow more predictions stemming from a highly accurate model, but also rejected more confident accusations of deception. Moreover, we found that human interaction with algorithmic predictions had no effect or even reduced the model's performance.

These findings, while suggesting a degree of human aversion to AI judgments, also indicate that the deployment of even highly accurate models

may encounter resistance in the general audience. Previous research highlights widespread concerns about AI replacing humans in high-stakes roles, such as those of judges or doctors, although these fears vary across cultures (Dong et al., 2024). More interestingly, these concerns seem to be influenced by perceived mismatches between psychological traits required for a role (e.g., human judge) and those attributed to AI (Dong et al., 2024). Alternatively, models' explainability, namely clarifying the rationale behind predictions, has been proposed as a mechanism to reduce concerns and improve trust in AI (Alufaisan et al., 2021), despite some mixed findings about its actual effectiveness (Balasubramaniam et al., 2023; Oswald et al., 2018; Sadeghi et al., 2024; Schemmer et al., 2022). Future research should, therefore, investigate further how to reduce humans' aversion to AI, either by discovering the psychological trait that humans believe an AI should possess to be a reliable deception detector or by studying how explanations behind the model's decision help foster trust.

2. Implications of this thesis

2.1 Implications for theories

Although this thesis did not aim to expand existing theories or define new ones, its findings highlight how deception theories are generally weak and underspecified, operating differently across different domains (Chapter 3) and deceptive strategies (Chapters 2, 3, and 4). Details on how our findings support or contradict existing theories of verbal deception are provided in the paragraphs below, along with considerations on how computational methods can also be employed as tools to refine these theories.

Distancing framework

The distancing framework posits that individuals may tend to engage in psychological distancing when expressing deceptive statements (Hancock et al., 2007; Newman et al., 2003) by using more other-references (e.g., we, you, they) than self-references (e.g., I, me, my). Across domains, these predictions were only partially supported. Our findings, in fact, align with previous research on deceptive opinions (Mihalcea & Strapparava, 2009; Pérez-Rosas & Mihalcea, 2015), which shows that deceivers use more other-references than self-references. However, they contrast with previous studies on distancing in past memories and future intentions (Hauch et al., 2015; Newman et al., 2003), as we found deceptive memories and intentions being more associated with a higher number of

self-references, suggesting that lie-tellers may have attempted to appear more credible by creating a sense of social connection. Finally, when shifting from fabrication to embedded lies, our findings still supported the distancing framework, as we found that deceptive statements with embedded lies contained more other-references than truthful statements.

Cognitive Load

The cognitive load approach posits that lie-tellers may engage in higher cognitive demand, given the effort required for lying (Vrij et al., 2008). However, these predictions were not consistent across domains and deceptive strategies in our findings. In fact, while deceptive opinions and past experiences were shorter and linguistically less complex (Hauch et al., 2015; Vrij et al., 2015), deceptive intentions were characterized by greater verbosity, suggesting that intentions may be over-prepared by lie-tellers (Kleinberg, van der Toolen, et al., 2018). Furthermore, when applying the cognitive load theory to embedded lies, we observed effects in the opposite direction: participants provided longer statements for embedded deception than for truthful statements. Although this pattern may partly reflect the experimental instructions (i.e., adding embedded lies to a statement in order to achieve a goal), it is also possible that cognitive load is more diagnostic of fabrication but loses its explanatory power for embedded lies, as mixing truthful and deceptive information may reduce cognitive demands, thereby eliminating the need for brevity.

Reality Monitoring

Reality Monitoring assumes that truthful accounts stem from episodic and autobiographical recollections and are therefore more characterized by perceptual details, contextual, and affective information, while fabricated accounts stem more from cognitive processes, such as inferences, reasoning, and imagination (Johnson & Raye, 1981; Sporer, 1997, 2004) Across domains, the evidence in favor of Reality Monitoring was mixed. For past experiences and future intentions, findings were largely consistent with Reality Monitoring predictions: truthful accounts exhibited higher RM scores and greater use of concrete, memory-related, spatial, and temporal information, whereas deceptive accounts relied more on abstract words and words reflecting cognitive processes (Amado et al., 2016; Gancedo et al., 2021; Johnson & Raye, 1981; Sporer, 1997, 2004). In contrast to previous expectations, deceptive opinions showed higher scores in the concreteness of words, contextual details, and reality monitoring. This different direction of effects was also previously shown (Mihalcea & Strapparava, 2009). These findings were interpreted as deceivers show difficulties in abstraction, resulting in fake opinions that

sound more grounded in reality. Finally, when switching from fabrication to embedded lies, only a few indicators of Reality Monitoring were found to be more predictive of truthfulness (i.e., memory-related words and temporal information) than embedded deception. This suggests that Reality Monitoring may be more diagnostic for fully fabricated accounts than for narratives where deceptive information is embedded within truthful information.

Verifiability approach

Along with the verifiability approach, lie-tellers face the dilemma of providing as many details as possible to be believed, while avoiding, as much as possible, providing verifiable details (Nahari et al., 2012). Across domains, our findings align with the verifiability approach and previous research for past experiences (Kleinberg, Mozes, et al., 2018; Kleinberg, van der Toolen, et al., 2018) and future intentions (Kleinberg, van der Toolen, et al., 2018), with truthful statements being characterized by more verifiable details (measured as the proportion of named entities) than deceptive statements. Interestingly, deceptive statements of past experiences and future intentions were also more characterized with references to people, which, as for the distancing framework, we interpreted as lie-tellers want to foster their social connections to appear as more truthful. However, no features related to verifiable details differentiated truthful from deceptive opinions. Finally, when moving from fabrication to embedded lies, no significant differences emerged between truthful statements and embedded deception in terms of verifiable details.

Truth-default theory

According to the Truth-default theory (TDT; Levine, 2014), individuals generally assume honesty in communication and become suspicious of deception only in those cases where a strong trigger disrupts this default truth assumption (e.g., hidden goals, inconsistencies, or third-party warnings). Findings from this thesis on hybrid deception detection (Chapter 4) strongly align with the TDT. We found that human judges tended to reject more confident algorithmic predictions of deception than confident predictions of truthfulness, suggesting that individuals do not abandon their truth-default status if such confident algorithmic accusations seem to lack clear evidence of deception. This also underscores that, in such cases, individuals tend to support a more thoughtful perspective that acknowledges the social costs of falsely accusing someone of lying and aligns with individuals' expectations of truthfulness. Overall, our findings extend the TDT to hybrid human–algorithm con-

texts, showing that the truth-default status applies not only to interpersonal communication but also to contexts involving the evaluation of algorithmic deception judgments.

General considerations for theories

The findings of this thesis, when taken together, only partially support existing theories on verbal cues to deception, highlighting the challenges of generalizing across different domains and deceptive strategies, suggesting that current theories are underspecified. For example, the Cognitive Load theory has been criticized for failing to specify the presumed cognitive mechanisms underlying deception and for not being a falsifiable theory (Neequaye, 2022). While this thesis did not aim to substantially advance theoretical developments, as this went beyond its primary scope, it underscores the need for future research to formalize theories that can generalize across diverse contexts and deceptive strategies.

In this regard, automated approaches may also play a role in theory development and refinement. For instance, in this thesis, we resorted to ML models to understand whether forensic experts fail to detect verbal deception due to inherent difficulties in combining multiple verbal cues or due to lack of informativeness of such criteria by comparing models trained on experts' judgments and models trained on features derived from NLP (Chapter 2). Additionally, explainability analysis on why one trained model may fail to predict deception across different domains (Chapter 3) revealed that some statement may be misclassified because they mimic the narrative style of truthful statements, with successful classifications occurring only when features related to cognitive load are significantly present in such statements. Similarly, explainability analysis was useful to understand that the challenge in predicting embedded lies stems from their similarity to truthful narratives (Chapter 4). Finally, we see opportunities in resorting to adversarial attacks to reveal how individuals think of and engage in deceptive narratives. Adversarial attacks, in NLP, are a deliberate yet subtle manipulation of the input text, often achieved through word substitutions, misspellings, or paraphrasing, that does not alter the original meaning for humans but effectively misleads or degrades the performance of a language model (Morris et al., 2020). Previous research already pointed out that by asking human participants to challenge model predictions by paraphrasing statements or substituting single words, research can gain a deeper understanding of how individuals think deception works and can employ dynamic settings where constant feedback from a classifier is provided, thereby holding the potential for new insights into the dynamics of deception and for refining existing theories (Kleinberg et al., 2025).

To conclude, these considerations underscore the pressing need for more robust and generalizable theories of verbal deception and encourage openness to resort to computational methods to help refine theoretical frameworks and achieve greater explanatory and predictive power.

2.2 Implications for practice

Practical implications of this thesis stem from using automated methods for coding statements on a large scale, as well as relying on trained ML models to predict deception when traditional methods fall short. In fact, in this thesis, we see automated approaches i) going beyond chance-level predictions, overcoming the performance of unaided naïve and expert judges (Aamodt & Custer, 2006; Bond & DePaulo, 2006; Hartwig & Bond, 2011); ii) being well-suited for predicting in-sample data and learning from multiple observations and features, thereby enabling continuous learning and the integration of different theoretical perspectives at the same time (see **DeCLaRatiVE**); iii) overcoming limitations related to manual approaches in terms of scalability, reliability, and continuous learning from complex data, thereby representing the state-of-the-art in verbal deception detection

However, despite these potential benefits, there are practical limitations that warrant mention. First, automated methods often show difficulties in generalizing out of their in-sample training, and in this thesis, we showed how they struggle in generalizing learning across different domains of deception. Additionally, deception is typically predicted at the statement level, but the detection at the embedded lie level remains particularly challenging. Second, it is important to acknowledge that humans are somehow averse to algorithmic predictions of deception, thereby posing a challenge to the effective integration of algorithmic predictions and human decision-makers for hybrid deception detection. Importantly, human-AI interaction often yields lower performance than the AI does in isolation in the domain of verbal deception detection. Furthermore, one note for the use of algorithmic prediction in forensic contexts concerns explainability. In such high-stakes settings, interpretable models should be preferred to highly complex, opaque systems, even when the latter offer relatively higher accuracy (Oswald et al., 2018). Explainability is, therefore, a fundamental requirement. However, some models, such as those relying on embeddings or deep architectures, are considered black boxes, as they often lack transparency, making it difficult to justify decisions in court or during investigative procedures (Garrett & Rudin, 2023). Conversely, we argue that models trained on theory-led features, as in the case of **DeCLaRatiVE**, provide greater practical value because they enable researchers and practitioners to explain which ver-

bal cues were detected and how these cues align with established theoretical frameworks, with the potential to foster human trust in the system and support accountability (Alufaisan et al., 2021).

In summary, while automated methods and ML models offer significant advantages in scalability, accuracy, and theoretical integration, their practical application requires careful consideration of limitations such as generalizability, explainability, and human aversion. Future progress in automated verbal deception detection will depend not only on improving algorithmic performance but also on fostering transparency and trust through interpretable models that can be effectively integrated into real-world decision-making contexts.

3. Limitations

Despite the contributions of this thesis, a few considerations about the constraints on the generalizability of our findings are worth mentioning.

First, as highlighted in our systematic review (Chapter 1), establishing a clear and verifiable ground truth is often a challenging task. In this thesis, we relied on a clear but non-verifiable ground truth level, meaning that participants were assigned to experimental conditions (here: truthful and deceptive), but without the possibility to verify what participants stated. Despite it being a common practice, this ground truth manipulation constrains the validity of current findings, as it mostly relies on participants' compliance with instructions rather than the factual veracity of statements. Previous studies already showed that participants may not only embed truthful information in deceptive statements, but also deceptive information in truthful ones (Markowitz, 2024; Verigin et al., 2020; Vrij et al., 2010). Therefore, future research should start testing theories and models of deception under experimental conditions where the ground truth can be clearly verified. For example, mock crime paradigms, video recollection tasks, or object description experiments enable researchers to systematically manipulate and verify deceptive information, as these settings allow for full control over who lies and who tells the truth, and ensure that every statement can be checked against a known reality. Once internal validity is established in such controlled contexts, the same predictions should then be tested in more naturalistic environments, such as real-life interviews or high-stakes situations, to assess external and ecological validity. This two-step approach would enhance the robustness and generalizability of findings and the scientific rigor of deception research.

Second, we acknowledge that our studies involved low-stakes scenarios where individuals participate in lab or online experiments and have

nothing to lose if they are caught as liars. In contrast, people may resort to deception in real-world scenarios because the stakes are high: for example, they deceive because otherwise they may lose money, reputation, or be sentenced to prison. However, low stakes are not necessarily a limitation. While it is true that in high-stakes contexts, individuals may prepare their deceptive stories more thoroughly (Porter & Ten Brinke, 2010), potentially reducing differences in cognitive load and diminishing detection accuracy (Walczyk et al., 2018), it is also true that higher stakes may simply amplify the small observed effects that are already seen in psychological research. For example, in research experiments, participants may freely exaggerate their claims and fabricate unrealistic alibis (e.g., stating that they have been abroad), whereas in real-life high-stakes scenarios, deceptive narratives tend to be more constrained to appear plausible (e.g., if they travelled, they should have a receipt for their flight ticket). This aligns with the verifiability approach (Nahari et al., 2012; Vrij & Nahari, 2019), which states that lie-tellers tend to provide fewer verifiable details, and is partly reflected in investigative contexts where law enforcement actively verifies suspects' statements, thereby limiting the feasibility of extreme fabrications and thereby suggesting that what is found in such low-stakes lab experiments may be even more effective in these naturalistic scenarios.

Third, this thesis involved some static experimental settings, such as typing a deceptive statement instead of producing it spontaneously in speech. While investigating static deception may be relevant for certain online contexts (e.g., deception in fake reviews), in other contexts, such as investigative interviews, deception should be studied with more dynamic, interactive approaches that mimic real-life scenarios. This is particularly important, given that the cognitive load theory (Vrij et al., 2008) predicts that lie-tellers will be more affected by a higher demand for lying. However, asking participants to write rather than engage in a conversation may reduce the cognitive costs of lying, as writing gives more time to plan and formulate responses, potentially masking important cues of deception (Derrick et al., 2013; Hauch et al., 2015). Additionally, stylometric techniques are most effective for extended statements rather than yes/no or very brief answers, limiting their applicability in situations where individuals may be reluctant to provide detailed and extensive responses. Future research should explore ways to overcome this constraint, such as developing hybrid models that combine stylometric features with other linguistic or behavioral cues, such as reaction times (Suchotzki et al., 2017), aggregating multiple short responses to approximate longer texts, or leveraging interviewing strategies that maximize differences between truth-tellers and lie-tellers, making resistance an important cue of deception (Vrij et al., 2008; Vrij & Granhag, 2012). These

strategies would broaden the applicability of stylometric approaches beyond contexts like online reviews, fake news, and other digital communications.

Finally, limitations on the reproducibility of findings pose a further concern. Proprietary tools and libraries are not always fully transparent and stable over time. For example, LIWC English dictionaries are not available to users but are embedded within the software, thereby limiting the possibility of scrutinizing and debating word-category assignments. Similarly, certain functions and models, such as those used for computing embeddings from closed-source LLMs like OpenAI's GPT models, are under constant refinement, beyond the researcher's control. All these factors introduce uncertainty and hinder the long-term replicability of findings. By relying on open-source tools and models (e.g., Llama, Mistral, and DeepSeek), future research can enhance the transparency, replicability, and reproducibility of findings, thereby aligning with the principles of open science.

4. Future outlooks

First, findings from our systematic review (Chapter 1) highlighted the need for more research on two key areas: 1) establishing a clearer and more verifiable ground truth by developing procedures that ensure the verifiability of statements, and 2) exploring forms of deception that extend beyond fabrication, including omissions and embedded lies, among others. Such developments would not only enhance the internal and external validity of deception research, but also fundamentally reshape how deception is conceptualized and studied. Establishing a clearer and verifiable ground truth would allow researchers to move beyond ambiguous or self-reported truthfulness, enabling rigorous testing of theoretical predictions and improving the reliability of automated detection systems. Similarly, expanding the scope beyond fabrication to include omissions, embedded lies, and mixed strategies would reflect the complexity of real-world deception, where lies rarely occur as complete fabrication. This broader perspective could lead to both more general theories of deception and more robust detection systems, capable of handling diverse and subtle forms of lying. Ultimately, these advances would bridge the gap between controlled laboratory studies and high-stakes applications, making deception research more scientifically robust and practically relevant.

Second, given the large body of available research, future research should prioritize meta-analyses to consolidate fragmented evidence and identify key moderators of detection accuracy. This is particularly im-

portant because, while numerous meta-analyses exist for manual approaches to deception detection, automated methods remain largely unexplored. In fact, the only meta-analysis on this topic (Hauch et al., 2015) was conducted 10 years ago and focused mostly on dictionary-based methods, leaving research with a knowledge gap on how modern automated approaches actually perform. Addressing this gap would enable researchers to quantify the overall detection accuracy of these methods and assess their robustness across various conditions. For example, meta-analyses could examine moderators, such as the impact of i) stakes (low vs. high), which affect the motivation of lie-tellers, ii) modality (written vs. transcribed), which have an impact on the cognitive load of individuals (Derrick et al., 2013; Hauch et al., 2015; Sporer, 2016), and iii) interaction level (static vs. dynamic), which determines whether deception occurs in isolated statements or in interactive exchanges, influencing both cue manifestation and detection strategies. By systematically analyzing these moderators, meta-analyses can reveal not only whether automated verbal deception detection works, but also under which conditions it succeeds or fails, providing the first truly comprehensive and generalizable picture of its practical applicability.

Third, mega-analysis of existing datasets represents another promising avenue, as it enables testing models across multiple domains and assessing the feasibility of developing a foundational deception model, a large-scale, multi-domain model that captures universal deception cues while adapting to context-specific nuances. Similar to what was claimed about foundation models of cognition (Binz et al., 2025), such a model would represent a major leap forward, enabling researchers and practitioners to move beyond fragmented, domain-limited approaches and toward a generalizable framework for automated deception detection. However, while such foundation models offer scalability and predictive power, they are typically opaque because they rely on neural networks and embeddings (see again Binz et al., 2025). This raises concerns for high-stakes applications where transparency and accountability are essential (Oswald et al., 2018). Therefore, future research should not only pursue accuracy but also prioritize interpretability. In this regard, theory-led models, based on psychologically grounded features, may offer practical advantages, as they enable practitioners to explain which verbal cues are detected and how they relate to established theories, albeit at the cost of slightly lower accuracy and constrained generalization. One potential solution to this stems from dynamic deception detection and lies in developing Bayesian models that, starting from a 50-50 chance, progressively update probabilities as evidence accumulates from existing knowledge and new information (e.g., as provided during interactive interviews), until sufficient confidence is reached to predict truthfulness or deception. This can be designed as an interactive human-AI system,

and it would represent a similar approach to that adopted by law enforcement in forensic settings, thereby satisfying the requirement that algorithmic decisions cannot and should not replace human decision-making (Oswald et al., 2018).

Additionally, in court trials, witnesses and suspects are asked to repeat the same story multiple times and to different people to assess the consistency of the statement, thus creating multiple versions (or paraphrases) of the same story (Fisher et al., 2009). However, previous research has shown that paraphrasing a deceptive statement using, for example, widely accessible language models (e.g., ChatGPT) is an effective strategy to mislead deception classifiers and appear as more truthful (Kleinberg et al., 2025). This suggests that future research should also focus on developing models of deception that are inherently robust against adversarial attacks, such as paraphrasing. This finding introduces a critical challenge: automated systems may be vulnerable to adversarial attacks that exploit linguistic variability. Unlike humans, who are unaffected by subtle shifts in meaning or emphasis, current classifiers often rely on surface-level features, making them easy to fool through paraphrasing or word substitutions. This vulnerability is particularly concerning given the growing accessibility of generative AI tools, which allow individuals to produce convincing adversarial attacks at scale. Future research should therefore focus on developing models that are inherently robust against such adversarial strategies, for example, by incorporating adversarial examples in the training phase to improve robustness against subtle textual variations. Addressing this issue is essential to ensure that automated deception detection remains reliable in high-stakes contexts such as legal proceedings, security screenings, and online fraud prevention.

Finally, a parallel research line should examine human trust in AI-assisted deception detection, an underexplored area with significant implications for deployment in high-stakes domains. This includes developing experiments that manipulate the level of explainability, user agency, and exposure to such models to investigate their influence on human acceptance of AI predictions, as well as employing quasi-experimental designs to control for the effects of cultural and individual differences on human trust. Understanding these dynamics is essential if automated systems become widely accessible and integrated into forensic or security practices.

References

Aamodt, M., & Custer, H. (2006). Who can best catch a liar? *Forensic Examiner*, *15(1)*, 6.

Amado, B. G., Arce, R., Fariña, F., & Vilariño, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. *International Journal of Clinical and Health Psychology*, *16*(2), 201–210. https://doi.org/10.1016/J.IJCHP.2016.01.002

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197. https://doi.org/10.1016/J.INFSOF.2023.107197

Bell, K. L., & DePaulo, B. M. (1996). Liking and lying. *Basic and Applied Social Psychology*, *18*(3), 243–266. https://doi.org/10.1207/S15324834BASP1803_1

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature 2025 644:8078*, *644*(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/S15327957PSPR1003_2

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. https://www.liwc.app

Bulten, W., Balkenhol, M., Belinga, J. J. A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., Pereira, G., Roy, P., Saile, G., Salles, P., Schaafsma, E., Tschui, J., Vos, A. M., Delahunt, B., Samaratunga, H., … Litjens, G. (2020). Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology 2020 34:3*, *34*(3), 660–671. https://doi.org/10.1038/s41379-020-0640-y

Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. (2020). DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th language resources and evaluation conference* (Vol. 12, pp. 1423-1430).

Caso, L., Cavagnis, L., Vrij, A., & Palena, N. (2023). Cues to deception: can complications, common knowledge details, and self-handicapping strategies discriminate between truths, embedded lies and outright lies in an Italian-speaking sample? *Frontiers in Psychology*, *14*. https://doi.org/10.3389/FPSYG.2023.1128194

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2022). *Scaling Instruction-Finetuned Language Models*. http://arxiv.org/abs/2210.11416

Constancio, A. S., Tsunoda, D. F., de Fátima Nunes Silva, H., da Silveira, J. M., & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/JOURNAL.PONE.0281323

Derrick, D. C., Meservy, T. O., Jenkins, J. L., Burgoon, J. K., & Nunamaker, J. F. (2013). Detecting Deceptive Chat-Based Communication Using Typing Behavior and Message Cues. *ACM Transactions on Management Information Systems (TMIS)*, *4*(2). https://doi.org/10.1145/2499962.2499967

Dong, M., Conway, J. R., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2024). Fears About Artificial Intelligence Across 20 Countries and Six Domains of Application. *American Psychologist*. https://doi.org/10.1037/AMP0001454

Fisher, R. P., Brewer, N., & Mitchell, G. (2009). The Relation between Consistency and Accuracy of Eyewitness Testimony: Legal versus Cognitive Explanations. *Handbook of Psychology of Investigative Interviewing: Current Developments and Future Directions*, 121–136. https://doi.org/10.1002/9780470747599.CH8

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality Monitoring: A Meta-analytical Review for Forensic Practice. *European Journal of Psychology Applied to Legal Context*, *13*(2), 99–110. https://doi.org/10.5093/EJPALC2021A10

Garrett, B. L., & Rudin, C. (2023). Interpretable algorithmic forensics. *Proceedings of the National Academy of Sciences*, *120*(41), e2301842120. https://doi.org/10.1073/PNAS.2301842120

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2007). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Process.*, *45*(1), 1–23. https://doi.org/10.1080/01638530701739181

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/A0023589

Hartwig, M., Granhag, P. A., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime & Law*, *13*(2), 213–227. https://doi.org/10.1080/10683160600750264

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*, *19*(4), 307–342. https://doi.org/10.1177/1088868314556539

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of Criteria-Based Content Analysis. *Psychological Assessment*, *29*(6), 819–834. https://doi.org/10.1037/PAS0000426

Ilias, L., Soldner, F., & Kleinberg, B. (2022). *Explainable Verbal Deception Detection using Transformers*. https://arxiv.org/abs/2210.03080v1

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(11). https://doi.org/10.1073/PNAS.2208839120

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-Machine Collaboration for Content Regulation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *26*(5). https://doi.org/10.1145/3338243

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychol. Rev.*, *88*(1), 67–85. https://doi.org/10.1037/0033-295x.88.1.67

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), e0220228. https://doi.org/10.1371/JOURNAL.PONE.0220228

Kleinberg, B., Loconte, R., & Verschuere, B. (2025). Effective faking of verbal deception detection with target-aligned adversarial attacks. *Legal and Criminological Psychology*, *00*, 1–24. https://doi.org/10.1111/LCRP.70001

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018). Using Named Entities for Computer-Automated Verbal Deception Detection. *Journal of Forensic Sciences*, *63*(3), 714–723. https://doi.org/10.1111/1556-4029.13645

Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, *32*(3), 354–366. https://doi.org/10.1002/ACP.3407

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta Psychologica*, *213*. https://doi.org/10.1016/J.ACTPSY.2020.103250

Kotsoglou, K. N., & Oswald, M. (2020). The long arm of the algorithm? Automated Facial Recognition as evidence and trigger for police intervention. *Forensic Science International: Synergy*, *2*, 86. https://doi.org/10.1016/J.FSISYN.2020.01.002

Leal, S., Vrij, A., Deeb, H., & Fisher, R. P. (2023). Interviewing to detect omission lies. *Applied Cognitive Psychology*, *37*(1), 26–41. https://doi.org/10.1002/ACP.4020

Leal, S., Vrij, A., Deeb, H., Hudson, C., Capuozzo, P., & Fisher, R. P. (2020). Verbal cues to deceit when lying through omitting information. *Legal and Criminological Psychology*, *25*(2), 278–294. https://doi.org/10.1111/LCRP.12180

Leins, D. A., Fisher, R. P., & Ross, S. J. (2013). Exploring liars' strategies for creating deceptive reports. *Legal and Criminological Psychology*, *18*(1), 141–151. https://doi.org/10.1111/J.2044-8333.2011.02041.X

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Markowitz, D. M. (2024). Deconstructing deception: Frequency, communicator characteristics, and linguistic features of embeddedness. *Applied Cognitive Psychology*, *38*(3), e4215. https://doi.org/10.1002/ACP.4215

Markowitz, D. M., & Hancock, J. T. (2019). Deception and language: The cOntextual Organization of Language and Deception (COLD)

framework. *The Palgrave Handbook of Deceptive Communication*, 193–212. https://doi.org/10.1007/978-3-319-96334-1_10

Mihalcea, R., & Strapparava, C. (2009). The lie detector. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, 309. https://doi.org/10.3115/1667583.1667679

Nahari, G., Vrij, A., & Fisher, R. P. (2012). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, *19*(2), 227–239. https://doi.org/10.1111/J.2044-8333.2012.02069.X

Neequaye, D. A. (2022). A Metascientific Empirical Review of Cognitive Load Lie Detection. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/COLLABRA.57508

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. *Law and Human Behavior*, *40*(4), 440–457. https://doi.org/10.1037/LHB0000193

Orsini, F., Cioffi, A., Cipolloni, L., Bibbò, R., Montana, A., De Simone, S., & Cecannecchia, C. (2025). The application of artificial intelligence in forensic pathology: a systematic literature review. *Frontiers in Medicine*, *12*, 1583743. https://doi.org/10.3389/FMED.2025.1583743

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223–250. https://doi.org/10.1080/13600834.2018.1458455

Papantoniou, K., Papadakos, P., Patkos, T., Flouris, G., Androutsopoulos, I., & Plexousakis, D. (2022). Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering*, *28*(5), 545–606. https://doi.org/10.1017/S1351324921000152

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *Conference Proceedings - EMNLP 2015: Conference*

on Empirical Methods in Natural Language Processing, 1120–1125. https://doi.org/10.18653/V1/D15-1133

Porter, S., & Ten Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology*, *15*(1), 57–75. https://doi.org/10.1348/135532509X433151

Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://doi.org/10.1016/J.COMPELECENG.2024.109370

Schemmer, M., Hemmer, P., Nitsche, M., Kuhl, N., & Vossing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128

Sporer, S. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Appl. Cognit. Psychol.*, *11*(5), 373–397.

Sporer, S. L. (2004). Reality monitoring and detection of deception. *The Detection of Deception in Forensic Contexts*, 64–102. https://doi.org/10.1017/CBO9780511490071.004

Sporer, S. L. (2016). Deception and cognitive load: Expanding our horizon with a working memory model. *Frontiers in Psychology*, *7*(APR), 172031. https://doi.org/10.3389/FPSYG.2016.00420

Steller, M., & Koehnken, G. (1989). Criteria-Based Content Analysis. *The Suggestibility of Children's Recollections*. https://doi.org/10.1037/T27704-000

Suchotzki, K., Verschuere, B., Bockstaele, B. Van, Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin*, *143*(4), 428–453. https://doi.org/10.1037/BUL0000087

Tausczik, Y., and, J. P.-J. of language, & 2010, undefined. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journals.Sagepub.Com*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

van Dijck, G. (2022). Predicting Recidivism Risk Meets AI Act. *European Journal on Criminal Policy and Research*, *28*(3), 407–423. https://doi.org/10.1007/S10610-022-09516-8

Velutharambath, A., & Klinger, R. (2023). UNIDECOR: A Unified Deception Corpus for Cross-Corpus Deception Detection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 39–51. https://doi.org/10.18653/V1/2023.WASSA-1.5

Verigin, B. L., Meijer, E. H., Vrij, A., & Zauzig, L. (2020). The interaction of truthful and deceptive information. *Psychology, Crime & Law*, *26*(4), 367–383. https://doi.org/10.1080/1068316X.2019.1669596

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., van Goor, T., Löwy, L. H. S., Appiah, O. K., & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 2023 7:5*, *7*(5), 718–728. https://doi.org/10.1038/s41562-023-01556-2

von Schenk, A., Klockmann, V., Bonnefon, J. F., Rahwan, I., & Köbis, N. (2024). Lie detection algorithms disrupt the social dynamics of accusation behavior. *IScience*, *27*(7), 110201. https://doi.org/10.1016/J.ISCI.2024.110201

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/JIP.82

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal Criminol. Psychol.*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117. https://doi.org/10.1016/J.JARMAC.2012.02.004

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest, Supplement*, *11*(3), 89–121. https://doi.org/10.1177/1529100610390861

Vrij, A., & Nahari, G. (2019). The Verifiability Approach. *Evidence-Based Investigative Interviewing*, 116–133. https://doi.org/10.4324/9781315160276-7

Walczyk, J. J., Sewell, N., & DiBenedetto, M. B. (2018). A review of approaches to detecting malingering in forensic contexts and promising cognitive load-inducing lie detection techniques. *Frontiers in Psychiatry*, *9*, 384915. https://doi.org/10.3389/fpsyt.2018.00700

Wang, G., Chen, H., & Atabakhsh, H. (2004). Criminal identity deception and deception detection in law enforcement. *Group Decision and Negotiation*, *13*(2), 111–127. https://doi.org/10.1023/B:GRUP.0000021838.66662.0c

Conclusion

The present thesis investigated the extent to which computational methods can be employed for automated verbal deception detection. To address this issue, this thesis explored key opportunities and challenges related to automated verbal deception detection, with findings holding both theoretical and practical implications.

First, opportunities were identified in the automated coding of statements, overcoming the practical and methodological limitations of manual coding, while also paving the way for more scalable, objective, and predictive approaches to deception detection. Second, opportunities were identified in the automated prediction of verbal deception, as models trained to detect fabricated statements have been found to perform above chance level and to overcome limitations of human judgments and manual approaches, becoming the state-of-the-art for verbal deception detection.

However, this thesis also identified important challenges. Current automated approaches for detecting verbal deception typically struggle to generalize across different domains and to more nuanced forms of deception, such as embedded lies, thereby undermining their potential application in out-of-sample data. Additionally, human aversion towards algorithmic predictions challenges the possibility of translating and integrating these models into real-life settings.

The theoretical implications of these findings are rooted in the acknowledgment that existing theories on deception are formulated too broadly, as their degree of support varies significantly depending on the context and the lying strategy employed. In other words, what appears consistent with a theory in one domain may be contradicted in another, signalling a lack of specificity and predictive power. While this thesis did not aim to refine deception theories, it highlights their limitations and provides considerations on how computational methods can also serve as tools for theory refinement.

Conversely, the practical implications of these findings stem from defining the key advantages of automated verbal deception detection, such as scalability, reliability, and accuracy, as well as its key challenges on generalizability, interpretability, and human aversion for real-life applications.

To conclude, this thesis showcases that while computational methods offer promising tools for verbal deception detection, their capabilities are

often overstated. Our findings show that these models can outperform human judgment under controlled conditions, yet their performance drops significantly when applied across domains or to nuanced forms of deception. Moreover, issues of interpretability and human trust remain unsolved, raising concerns about premature deployment in high-stakes contexts. These limitations highlight an important reality: methods from artificial intelligence are not a silver bullet for deception detection. Progress will require not only technical improvements but also rigorous validation, theoretical refinement, and careful consideration of ethical and social implications. Until these challenges are addressed, automated deception detection should be viewed as an exploratory solution or, at best, a complementary aid to human expertise.