Can Humans Tell? A Dual-Axis Study of Human Perception of LLM-Generated News*

Alexander Loth†
alexander.loth@stud.fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Martin Kappes
kappes@fra-uas.de
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany

Marc-Oliver Pahl
marc-oliver.pahl@imt-atlantique.fr
IMT Atlantique, UMR IRISA, Chaire Cyber CNI
Rennes, France

arXiv:2604.03755v2 [cs.CY] 29 Oct 2026

Abstract

Can humans tell whether a news article was written by a person or a large language model (LLM)? We investigate this question using JudgeGPT, a study platform that independently measures source attribution (human vs. machine) and authenticity judgment (legitimate vs. fake) on continuous scales. From 2,318 judgments collected from 1,054 participants across content generated by six LLMs, we report five findings: (1) participants cannot reliably distinguish machine-generated from human-written text ($p > .05$, Welch's t-test); (2) this inability holds across all tested models, including open-weight models with as few as 7B parameters; (3) self-reported domain expertise predicts judgment accuracy ($r = .35$, $p < .001$) whereas political orientation does not ($r = -.10$, n.s.); (4) clustering reveals distinct response strategies ("Skeptics" vs. "Believers"); and (5) accuracy degrades after approximately 30 sequential evaluations due to cognitive fatigue. The answer, in short, is no: humans cannot reliably tell. These results indicate that user-side detection is not a viable defense and motivate system-level countermeasures such as cryptographic content provenance.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems** → **Web mining**; • **Security and privacy** → *Social aspects of security and privacy*.

Keywords

LLM-Generated Text, Human Perception, Misinformation Detection, Dual-Axis Assessment, Fake News, Cognitive Fatigue, Content Provenance, Web Trust

1 Introduction

Text has long been the web's default trust carrier, yet LLMs now produce synthetic text that humans rate as no less credible than human-written content [3, 7]. As false and misleading content already spreads faster than corrections online [22], the ability to generate convincing fabrications at marginal cost amplifies what Wardle and Derakhshan term "information disorder" [23] across an already fragile ecosystem [4, 8, 21].

Automated detection approaches face a moving-target problem: classifiers trained on one model generation degrade rapidly on the next [2, 20]. A complementary line of research therefore focuses on the human side, investigating cognitive inoculation [6, 19] and digital literacy interventions [9]. Designing such interventions, however, requires empirical data on *how* and *where* human judgment fails.

Building on an earlier survey of the research landscape [11], we developed JudgeGPT [13], a web-based study platform with two methodological innovations. First, it decouples *authenticity* (legitimate vs. fake) from *source* (human vs. machine) via independent continuous scales, avoiding the conflation inherent in binary real/fake tasks. Second, stimuli are generated by RogueGPT [13, 14], a controlled multi-model framework that enables systematic comparison across LLM families. This poster reports five empirical findings from 2,318 judgments and discusses their implications for web system design.

2 Related Work

Research on AI-generated misinformation spans two complementary directions. On the detection side, neural classifiers such as Grover [24] and more recent transformer-based detectors [2] initially achieved high accuracy but degrade as models evolve [4]; comprehensive surveys confirm this trend [25]. On the human side, controlled experiments consistently show that participants perform near chance when distinguishing machine-generated from human-written text [3, 5]. Kreps et al. demonstrated that GPT-generated political messages are rated as credible as journalist-written ones [7], while Su et al. showed that detection difficulty increases with model scale [20].

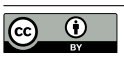
Most human-subject studies, however, employ binary classification tasks (real vs. fake) that conflate source and veracity: a machine-generated paraphrase of a true story is not misinformation, yet binary framing treats it as such. Our dual-axis method addresses this limitation. We additionally build on work in psychological inoculation [6, 19] and analytic thinking as a predictor of misinformation resilience [18], while connecting to emerging provenance standards (C2PA) as system-level alternatives [16].

3 Methodology

JudgeGPT is a web-based study platform for collecting human perception data on AI-generated text [13]. Its design was informed by a systematic gap analysis of prior work [11].

Dual-Axis Assessment. For each news fragment, participants provide three independent ratings on continuous 0–100 sliders (Figure 1): (1) *Source judgment* (0 = certainly machine, 100 = certainly human), (2) *Authenticity judgment* (0 = certainly fake, 100 = certainly legitimate), and (3) *Topic familiarity*. Continuous scales were

*Published in the Companion Publication of the 2026 18th ACM Web Science Conference (WebSci Companion '26), pp. 34–39. DOI: 10.1145/3795513.3807431.
†Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.

chosen over Likert items to capture degree of certainty and to enable parametric analyses including Pearson correlation and clustering.

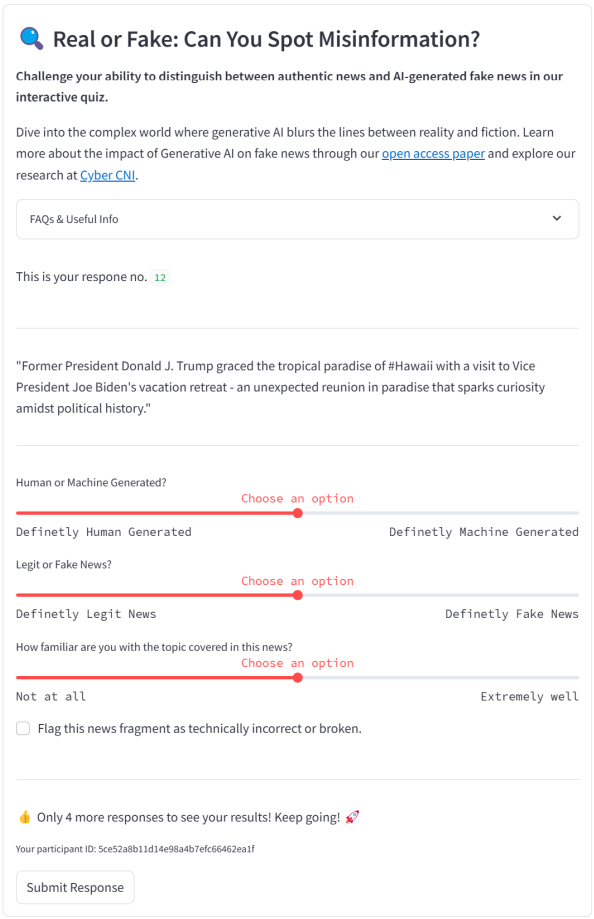


Figure 1: The JudgeGPT interface. Participants rate each fragment on three continuous axes: source attribution, authenticity, and topic familiarity.

Stimulus Design. Stimuli follow a 2×2 factorial design crossing *origin* (human, machine) with *veracity* (real, fake). Machine-generated fragments are produced by RogueGPT [13, 14], which orchestrates six LLMs (GPT-4 [1], GPT-3.5, GPT-4o, LLaMA-2 13B, Gemma 7B, Mistral 7B) with persona-based prompting. Human-sourced fragments are drawn from established news outlets and known misinformation databases. AI-generated texts are grounded in real news topics and verified for factual consistency to avoid conflating perception of style with detection of hallucinations. The stimulus set is intentionally skewed toward machine-origin fragments (~98%), with human-origin items serving as calibration anchors. This design choice reflects the study's focus on *within-AI* variation (across models) rather than *human-vs.-AI* comparison; participants are not informed of the base rate, and the near-chance detection results (Finding 1) hold when analyzed on the human-origin subset alone.

Data Collection. Participants provide informed consent and complete a demographic questionnaire (age, education, political orientation, AI familiarity) before evaluating a sequence of fragments. Each participant evaluates between 5 and 87 fragments (median = 12, IQR = 6–22), with presentation order randomized and model assignment balanced across participants. The platform records the three slider ratings, response latency, and an anonymous participant identifier for linking judgments to covariates. The sample skews toward educated European demographics (68% university-educated, 74% European); we discuss generalizability limitations in Section 5.

4 Findings

The dataset comprises 2,318 judgments from 1,054 unique participants. Table 1 provides an overview; each finding is detailed below.

Table 1: Summary of empirical findings (N = 2,318 judgments from 1,054 participants; median 12 items per participant, range 5–87).

Finding	Statistic	Implication
F1: Undetectable	$t(2316) = 0.87$, $p = .38$	Intuition fails as defense
F2: Model-agnostic	All 6 LLMs at $\bar{x} \approx 0.50$	Low barrier to deceptive text
F3: Expertise > Politics	$r = .35$ vs. $r = -.10$	Literacy-based interventions
F4: User personas	$k = 2$ clusters (sil. = .41)	Adaptive designs needed
F5: Fatigue	Decline after ~30 items	Sustained vigilance infeasible

Finding 1: Machine-Generated Text Is Not Detectable by Participants. Figure 2 shows near-complete overlap between human- and machine-generated content on both axes. A Welch's t-test yields no significant difference in source scores between conditions ($t(2316) = 0.87$, $p = .38$, Cohen's $d = 0.04$), confirming that participants' intuitions do not reliably discriminate AI-generated from human-written text.

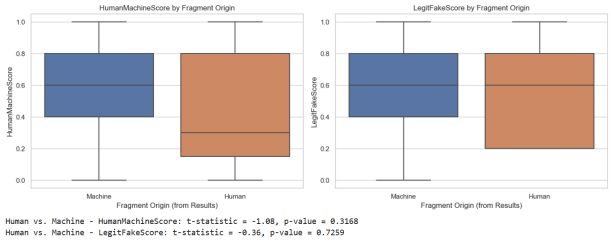


Figure 2: Source score (left) and authenticity score (right) distributions for machine- vs. human-written fragments. Overlapping distributions and a non-significant t-test indicate participants cannot distinguish the two conditions.

Finding 2: Detection Failure Is Model-Agnostic. Figure 3 disaggregates source scores by generating model. Mean scores for all six LLMs fall within the 0.44–0.55 range, clustering around the 0.5 chance level. A one-way ANOVA reveals no significant between-model effect ($F(5, 2349) = 1.92$, $p = .09$). Open-weight models with as few as 7B parameters produce text rated no differently from GPT-4o output, indicating that the capability to generate human-indistinguishable text is no longer restricted to frontier models.

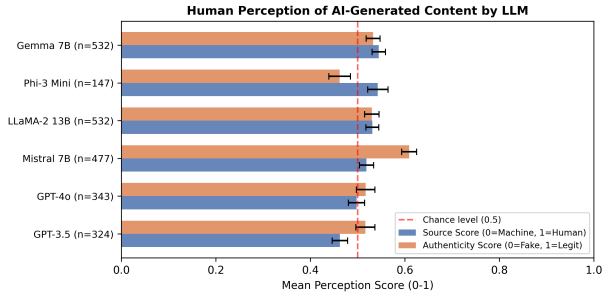


Figure 3: Mean source and authenticity scores per LLM (± SE). The dashed line marks chance level (0.5). No model is reliably identified as machine-generated.

Finding 3: Domain Expertise Predicts Accuracy; Political Orientation Does Not. Pearson correlations (Figure 4) show that self-reported fake news familiarity is positively associated with source judgment accuracy ($r = .35$, $p < .001$) and authenticity accuracy ($r = .29$, $p < .001$). Political extremity shows only a weak, non-significant negative correlation ($r = -.10$, $p = .12$). While the political orientation measure is coarse (a single self-report item) and the stimulus set was not designed to control for politically sensitive content, the pattern suggests that learned analytical skills may be a stronger predictor of detection performance than ideological predisposition. A dedicated study with politically balanced stimuli would be needed to draw causal conclusions about the role of political orientation.

Finding 4: Clustering Reveals Distinct Response Strategies. K-means clustering ($k = 2$, selected via silhouette analysis, silhouette coefficient = .41) on participant-level mean scores identifies two groups (Figure 5): "Skeptics," who assign low trust across all content regardless of origin, and "Believers," who maintain high baseline trust. These distinct response strategies imply that uniform interventions will be suboptimal.

Finding 5: Initial Learning Effect Followed by Cognitive Fatigue. A rolling-window analysis of sequential judgments (Figure 6) reveals improved accuracy during the first 15–20 evaluations, consistent with a learning effect. Beyond approximately 30 items, accuracy declines and participants increasingly default to "fake" classifications, indicating cognitive fatigue. This temporal pattern limits the practical duration of detection-based interventions.

5 Discussion and Implications

Three design implications follow from our findings:

- **Provenance over detection.** The failure of human judgment across all tested models (F1, F2) implies that detection cannot be offloaded to users. Cryptographic provenance frameworks such as C2PA, which attach verifiable content histories at the point of creation, offer a more scalable alternative [16, 17], complemented by source-level credibility signals [12].
- **Bounded inoculation.** The learning effect (F5, first 15–20 items) supports the efficacy of inoculation-based interventions [6, 19], but the subsequent fatigue effect constrains their practical deployment to short, focused sessions rather than continuous monitoring.
- **Persona-aware design.** The Skeptic/Believer distinction (F4) suggests that web platforms should adapt trust indicators to user disposition rather than applying uniform labels. The expertise effect (F3) further implies that interventions building analytical skill may be more productive than those targeting partisan bias, though this finding should be validated with politically balanced stimuli in future work.

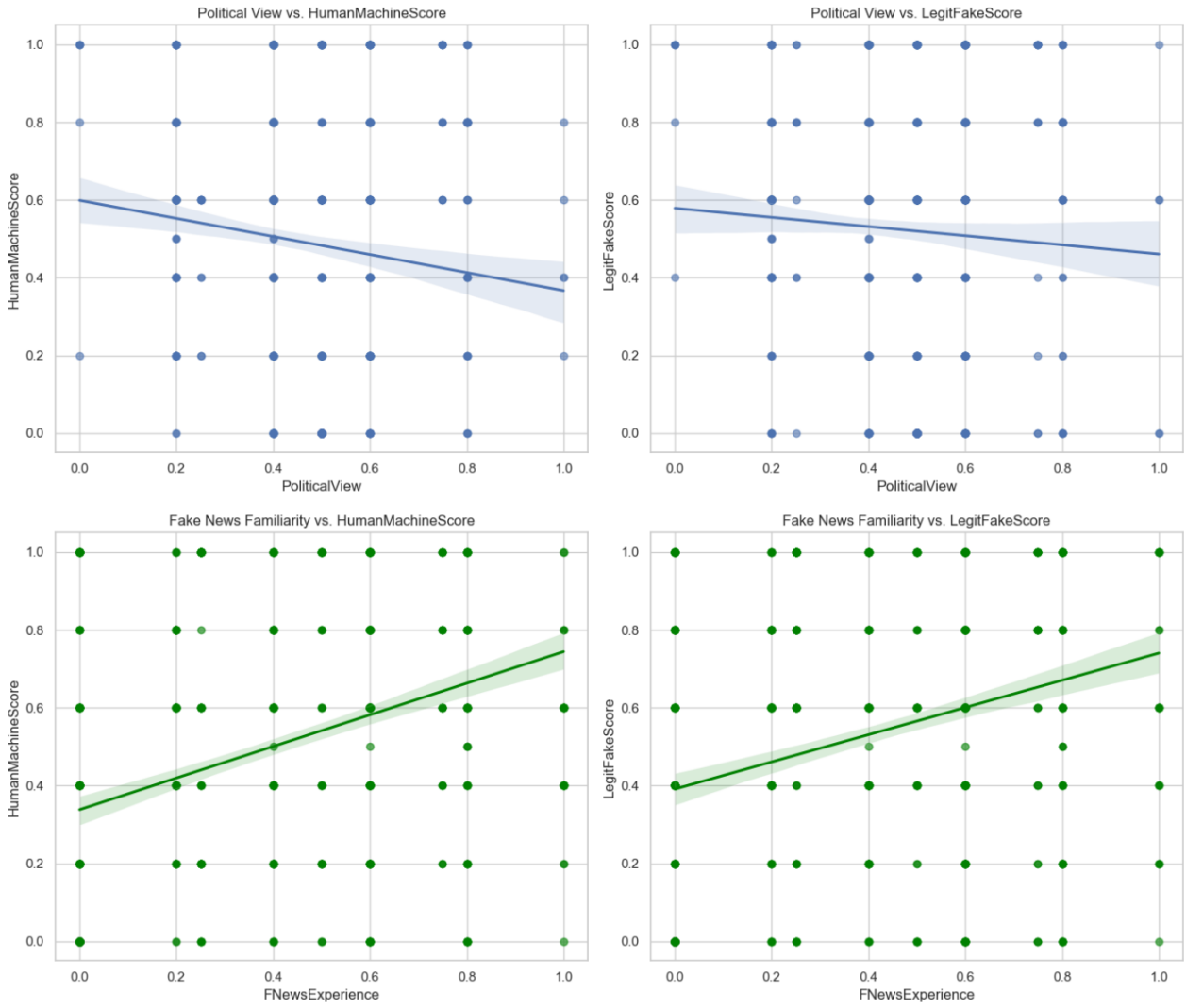
These findings are part of a broader empirical program on AI-driven disinformation encompassing generation infrastructure [14], human perception [13], expert assessment [15], provenance-based countermeasures [16, 17], domain credibility infrastructure [12], and a unified dissertation framework formalizing the "Indistinguishability Threshold" [10].

Limitations. The participant sample skews toward educated European demographics (68% university-educated, 74% European), limiting generalizability to broader populations. The controlled evaluation format does not capture the social and contextual cues present in real-world social media encounters. The stimulus set is intentionally skewed toward machine-generated content (∼98%); while this design serves the study's focus on within-AI variation across models, a balanced design would strengthen claims about human-origin anchoring effects and control for potential base-rate adaptation by participants. The political orientation finding (F3) is preliminary: the single-item measure and lack of politically controlled stimuli preclude causal claims. Future work will address these limitations through a longitudinal browser-extension study with a more diverse participant pool and balanced stimulus design.

6 Conclusion

Can humans tell? Our dual-axis study of 2,318 judgments across six LLM families provides a clear empirical answer: they cannot. Machine-generated text is indistinguishable from human writing regardless of model size or family, domain expertise predicts detection accuracy more strongly than political orientation, participants adopt distinct trust strategies, and cognitive fatigue limits sustained detection. These findings support a shift from user-level detection toward system-level countermeasures, including content provenance, adaptive trust indicators, and bounded inoculation interventions.

Data Availability. All datasets are archived on Zenodo under restricted academic access: RogueGPT stimulus corpus (DOI: 10.5281/zenodo.18703138), JudgeGPT human perception data (DOI: 10.5281/zenodo.18703385), and the CRED-1 domain credibility dataset (DOI:




```
Correlation between Political View and HumanMachineScore: -0.10
Correlation between Political View and LegitFakeScore: -0.05
Correlation between Fake News Familiarity and HumanMachineScore: 0.35
Correlation between Fake News Familiarity and LegitFakeScore: 0.29
```

Figure 4: Relationships between participant covariates and judgment scores. Top: political view (weak slope). Bottom: fake news familiarity (positive slope). Regression lines with 95% CI shown.

10.5281/zenodo.19355308). The JudgeGPT platform[1] and RogueGPT framework[2] are open-source.

Acknowledgments

We thank all participants of the JudgeGPT study.

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[2] Canyu Chen and Kai Shu. 2024. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *AI Magazine* 45, 3 (2024), 354–368. doi:10.1002/aaai.12188

[3] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 7282–7296. doi:10.18653/v1/2021.acl-long.565

[4] Emilio Ferrara. 2024. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science* 7, 1 (2024), 549–569. doi:10.1007/s42001-024-00250-1

[1] https://github.com/aloth/JudgeGPT
[2] https://github.com/aloth/RogueGPT



Figure 5: Pairplot of participant-level mean scores colored by cluster. The two groups show clearly separated distributions on both axes.

[5] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human Heuristics for AI-Generated Language Are Flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120. doi:10.1073/pnas.2208839120

[6] Anastasia Kozyreva, Sam Wineburg, Stephan Lewandowsky, and Ralph Hertwig. 2023. Critical Ignoring as a Core Competence for Digital Citizens. *Current Directions in Psychological Science* 32, 1 (2023), 81–88. doi:10.1177/09637214221121570

[7] Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117. doi:10.1017/XPS.2020.37

[8] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998

[9] Stephan Lewandowsky and Sander van der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* 32, 2 (2021), 348–384. doi:10.1080/10463283.2021.1876983

[10] Alexander Loth. 2026. The Indistinguishability Threshold: Measuring Cognitive Vulnerabilities to AI-Generated Disinformation. In *Companion Publication of the 2026 18th ACM Web Science Conference* (Braunschweig, Germany). ACM, New York, NY, USA, 79–82. doi:10.1145/3795513.3807421 PhD Symposium.



Figure 6: Rolling mean of judgment scores across sequential evaluations. Scores improve initially before declining after ~30 items, consistent with a learning-then-fatigue pattern.

[11] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or curse? A survey on the Impact of Generative AI on Fake News. arXiv:2404.03021 [cs.CL]

[12] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. CRED-1: An Open Multi-Signal Domain Credibility Dataset for Automated Pre-Bunking of Online Misinformation. SSRN preprint. doi:10.2139/ssrn.6448466

[13] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild. In *Companion Proceedings of the ACM Web Conference 2026* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 1125–1132. doi:10.1145/3774905.3795832

[14] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In *Companion Proceedings of the ACM Web Conference 2026* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 295–302. doi:10.1145/3774905.3795471

[15] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In *Companion Proceedings of the ACM Web Conference 2026* (Dubai, United Arab Emirates). ACM, New York, NY, USA, 980–988. doi:10.1145/3774905.3795484

[16] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: A Privacy-First Mobile Framework for Cryptographic Image Provenance and AI Detection. arXiv:2602.03423 [cs.CR] Accepted at the FIIR workshop, The Web Conference 2026; the workshop was subsequently cancelled.

[17] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In *Companion Publication of the 2026 18th ACM Web Science Conference* (Braunschweig, Germany). ACM, New York, NY, USA, 1–5. doi:10.1145/3795513.3806658

[18] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* 25, 5 (2021), 388–402. doi:10.1016/j.tics.2021.02.007

[19] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological Inoculation Improves Resilience Against Misinformation on Social Media. *Science Advances* 8, 34 (2022), eabo6254. doi:10.1126/sciadv.abo6254

[20] Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting Fake News Detection to the Era of Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (Mexico City, Mexico), Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, 1473–1490. doi:10.18653/v1/2024.findings-naacl.95

[21] Edson C Tandoc, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news": A typology of scholarly definitions. *Digital journalism* 6, 2 (2018), 137–153. doi:10.1080/21670811.2017.1360143

[22] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (2018), 1146–1151. doi:10.1126/science.aap9559

[23] Claire Wardle and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making.* Report DGI(2017)09. Council of Europe, Strasbourg. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[24] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

[25] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5, Article 109 (sep 2020), 40 pages. doi:10.1145/3395046