The Indistinguishability Threshold: Measuring Cognitive Vulnerabilities to AI-Generated Disinformation

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

Abstract

The proliferation of Generative AI has precipitated an "Authenticity Crisis," where synthetic media effectively breaches the "Indistinguishability Threshold" of human perception. This dissertation presents an integrated Design Science framework to quantify this systemic vulnerability across nine chapters. We introduce three open-source artifacts: RogueGPT, a parametric stimulus engine for generating controlled synthetic news; JudgeGPT, a dual-axis perception platform measuring both authenticity and credibility judgments; and Origin Lens, a mobile application implementing C2PA cryptographic provenance verification. Our empirical pipeline reveals four core findings: (1) human detection accuracy for AI-generated news converges to random chance (~50%), independent of age or digital literacy—challenging the "digital native" assumption; (2) a significant "Fatigue Effect" causes detection capacity to degrade under volume, supporting the "Firehose of Falsehood" strategy; (3) a measurable "Liar's Dividend" where authentic content is dismissed as AI-generated; and (4) expert practitioners express skepticism toward detection tools and preference for provenance-based standards [25]. Building on the CSET RICHDATA and DISARM frameworks, we validate an adapted "Disinformation Kill Chain" using empirical perception data to identify stage-specific countermeasures. Our findings establish that mitigation must shift from "post-delivery" fact-checking to "pre-delivery" provenance and "post-exposure" cognitive inoculation. This work connects technical AI capabilities with human factors research, contributing evidence-based strategies for platform governance in the Web Science domain.

CCS Concepts

• **Social and professional topics** → **Misinformation**; • **Information systems** → *Web mining*; • **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Security and privacy** → *Social aspects of security and privacy*.

Keywords

generative AI, disinformation, indistinguishability threshold, authenticity crisis, cognitive vulnerabilities, synthetic media, human perception, content provenance, C2PA, LLMs, web science



This work is licensed under a Creative Commons Attribution 4.0 International License.
WebSci '26, Braunschweig, Germany
© 2026 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/26/05
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The digital information ecosystem is undergoing a *phase transition*, driven by the proliferation of Large Language Models (LLMs) capable of generating text that is structurally, stylistically, and semantically indistinguishable from human-authored content [20]. This shift marks the end of the "Fake News" era—characterized by low-quality, manually produced clickbait—and the beginning of the "Synthetic Media" era. In this new paradigm, the primary challenge is no longer merely verifying the *accuracy* of information, but verifying its *origin*.

The scale of the threat is substantial. Research has demonstrated that false information spreads faster and farther than accurate information on social networks [38], and the psychological mechanisms that make humans susceptible to misinformation—including cognitive laziness and partisan bias [31]—remain unchanged even as generation capabilities advance. Meanwhile, automated detection systems that once showed promise (e.g., GLTR [7], DetectGPT [28]) face an arms race with increasingly sophisticated generators.

This research addresses the systemic risks posed by this transition. We introduce the concept of the Indistinguishability Threshold—the point at which human detection accuracy converges to random chance. We argue that once this threshold is crossed, traditional "content-based" detection strategies become futile, necessitating a shift toward "provenance-based" governance [11]. Through a Design Science Research (DSR) framework [10], we present an empirical pipeline comprising three open-source artifacts: RogueGPT (a parametric stimulus engine), JudgeGPT (a dual-axis perception sensor), and Origin Lens (a C2PA provenance verification tool).

The research contributes: (1) a formalized definition of the Indistinguishability Threshold grounded in computational theory, (2) empirical evidence quantifying human cognitive vulnerabilities to synthetic text, and (3) the "Disinformation Kill Chain"—a systemic mitigation framework. This interdisciplinary approach addresses the Web Science focus on the interplay between technology and society [36, 40].

2 Related Work

2.1 Theoretical Foundations

The study of online misinformation has evolved from early taxonomies of "information disorder" [40] to empirically grounded frameworks. Lazer et al. [16] established a cross-disciplinary research agenda for the science of fake news, while Vosoughi et al. [38] demonstrated that false information spreads faster and farther than accurate information on social networks. Recent work in the Web Science community has examined the linguistic underpinnings of misinformation spread within echo chambers [39]. The psychological mechanisms underlying susceptibility—cognitive laziness,

partisan bias, and continued influence effects—are synthesized in Pennycook and Rand [31] and in a comprehensive review by Ecker et al. [6], who identify the drivers of misinformation belief and its resistance to correction.

The "Authenticity Crisis" describes the erosion of heuristic cues humans use to evaluate trust [3]. As LLMs exhibit emergent capabilities at scale [2], linguistic artifacts previously relied upon for detection disappear, leading to *Generalized Distrust*: users disengage from the information ecosystem entirely. Vaccari and Chadwick [37] showed that even brief exposure to synthetic political video increases uncertainty and reduces trust in news on social media. The *Liar's Dividend* [3, 34] compounds this problem: bad actors weaponize public awareness of AI content to dismiss authentic evidence. The "Firehose of Falsehood" model [29] describes how high-volume dissemination overwhelms cognitive defenses, forcing users from analytical "System 2" to heuristic "System 1" processing [13].

2.2 Detection and Provenance

Automated detection tools (GLTR [7], DetectGPT [28]) degrade across model generations [41]. Ippolito et al. [12] found that automatic detection is easiest precisely when humans are most deceived, highlighting the complementary failure modes of human and machine classifiers. Groh et al. [9] demonstrated that while individual humans perform near chance at detecting deepfakes, machine-informed crowds achieve significantly higher accuracy—suggesting hybrid human-AI approaches. These findings motivate a shift from detection-based to provenance-based approaches [4, 11]. Countermeasures drawing on inoculation theory [33] and critical digital literacy [14] offer complementary "pre-bunking" strategies.

2.3 Prior Work in This Dissertation

This symposium paper draws on a body of published work by the author. The author's engagement with generative AI predates the current wave: *Decisively Digital* [17] examined NLP and GPT's emerging potential to reshape digital content, a subsequent book chapter [18] explored machine learning and synthetic data methods in risk controlling, while *KI für Content Creation* [19] explicitly addressed the ethical risks of generative AI—including disinformation, deepfakes, and the imperative for media literacy. An initial survey [20] then mapped the dual role of generative AI in both producing and detecting disinformation. Subsequent work examined how LLM-generated misinformation affects digital ecosystems at scale [24], and established a causal framework for analyzing human susceptibility to foundation model hallucinations [23]. A complementary expert perception survey [25] revealed practitioner skepticism toward detection tools and preference for provenance standards. A dual-axis empirical study [21] quantified human detection accuracy for LLM-generated news across demographic and cognitive dimensions. On the mitigation side, Origin Lens [26, 27] implements privacy-first cryptographic image provenance verification on mobile devices, reclaiming trust on the AI-mediated web. The present paper synthesizes these contributions into a unified dissertation framework.

3 Research Questions and Methodology

The dissertation addresses one primary research question supported by four sub-questions:

Primary RQ: To what extent does the ecosystem of GenAI-driven disinformation—from generation to human perception—reveal new vulnerabilities and require new mitigation frameworks?

- **RQ1 (Generation):** How can modern LLMs be systematically operationalized for synthetic news generation, and what design parameters influence output quality?
- **RQ2 (Public Perception):** How do non-expert humans perceive and judge AI-generated text, and which demographic, cognitive, and temporal factors predict detection accuracy?
- **RQ3 (Expert Perception):** How do domain experts assess the threat landscape and mitigation strategies, and where do their assessments converge with or diverge from empirical public perception data?
- **RQ4 (Mitigation):** Can the Disinformation Kill Chain be empirically validated using human perception data, and what stage-specific countermeasures does this validation reveal?

3.1 Design Science Research Framework

The methodology follows the Design Science Research (DSR) paradigm [10, 30], structured around three iterative cycles. The *Relevance Cycle* connects the research to societal needs for understanding synthetic media threats. The *Design Cycle* involves building and evaluating the software artifacts. The *Rigor Cycle* grounds the work in established theories of persuasion, cognition, and AI safety.

3.2 Research Artifacts

Following Hevner's DSR guidelines [10], we developed a closed-loop experimental pipeline comprising four complementary open-source artifacts that together address the full synthetic media lifecycle:

RogueGPT: The Parametric Stimulus Engine.[1] This tool wraps multiple LLM backends (GPT-4, Llama, Mistral) and provides a research interface for varying generation parameters: *political leaning*, *emotional intensity*, *journalistic style*, *perplexity*, and *burstiness*. The system implements adversarial prompt engineering techniques [42] to probe model safety boundaries, storing each fragment with full metadata for reproducibility.

JudgeGPT: The Socio-Technical Sensor.[2] This web-based platform presents participants with a stream of mixed human-authored and synthetic content. Its core methodological contribution is the Dual-Axis Assessment: (1) *Authenticity Axis*—"Do you believe this was written by a Human or AI?" and (2) *Credibility Axis*—"Do you believe the information is True or False?" This separation enables quantification of two key metrics: the *Deception Rate* (AI-generated content incorrectly rated as human-authored) and the *Liar's Dividend Rate* (authentic human content incorrectly dismissed as AI-generated). These metrics capture both successful disinformation and the second-order effect of generalized skepticism.

Origin Lens: The Provenance Verification Tool.[3] This mobile application performs on-device cryptographic verification of C2PA

[1] https://github.com/aloth/RogueGPT
[2] https://github.com/aloth/JudgeGPT
[3] https://github.com/aloth/origin-lens

Content Credentials [4], validates signature chains, and detects AI-generated content via multiple signals (C2PA, EXIF, SynthID). It serves as both a research probe for studying verification tool adoption and a practical C2PA countermeasure.

CRED-1: The Domain Credibility Dataset.[4] This open dataset [22] aggregates multi-signal credibility assessments for 2,672 news domains, combining fact-checker ratings, domain age, WHOIS transparency, TLS configuration, and content signals. CRED-1 enables automated "pre-bunking" by providing machine-readable credibility baselines at the source level—complementing content-level detection with source-level provenance.

Together, these four artifacts form an integrated research pipeline: RogueGPT generates the threat, JudgeGPT measures the vulnerability, Origin Lens provides the countermeasure, and CRED-1 supplies the credibility infrastructure—enabling study of the synthetic media lifecycle from generation to mitigation.

4 Results

Borrowing from computational indistinguishability in cryptography [8] and geo-indistinguishability in differential privacy [1], we formalize the *Indistinguishability Threshold* as the point at which a human discriminator D can no longer reliably distinguish between human-authored text H and AI-generated text A:

$$P(D(x) = 1 \mid x \in H) \approx P(D(x) = 1 \mid x \in A) \quad (1)$$

For short-form news (tweets, headlines), aggregate detection accuracy is 52%—statistically indistinguishable from random guessing ($p > 0.05$). The Indistinguishability Threshold has been crossed for this content class, consistent with recent findings [5, 15]. Once crossed, detection-based defenses become unreliable, motivating a shift toward provenance verification [11].

4.1 The Fatigue Effect and Cognitive Load

We observed a statistically significant logarithmic decay in detection accuracy relative to session duration ($r = -0.42$, $p < 0.01$). As cognitive load increases, participants abandon analytical "System 2" verification [13] and rely on fluency heuristics—which favor the grammatical perfection of LLMs. This provides empirical support for the "Firehose of Falsehood" strategy [29] in the AI era: high-volume attacks exploit finite human cognitive resources.

4.2 Debunking the "Digital Native" Myth

Contrary to the influential "Digital Native" hypothesis [32], our data reveals no significant correlation between age and detection accuracy. Participants aged 18–24 performed no better than those aged 60+. This challenges the assumption that "growing up with the web" confers immunity to synthetic media, and carries policy implications: current "Digital Literacy" curricula should evolve into "Synthetic Media Literacy" programs emphasizing provenance verification rather than content analysis [14].

4.3 The Liar's Dividend Operationalized

Our Dual-Axis Assessment captures the *Liar's Dividend* [3, 34]—the strategic benefit bad actors derive when authentic evidence can be dismissed as "AI-generated." Additionally, qualitative data reveals *Generalized Distrust*: participants experiencing repeated detection failure enter "Epistemic Paralysis," distrusting *all* content regardless of provenance.

4.4 The Automated GenAI Kill Chain

Building on the RICHDATA framework [35] and the DISARM taxonomy, we develop the "GenAI Disinformation Kill Chain" (Table 1). Unlike the conceptual frameworks it builds on, our Kill Chain is empirically validated: each stage is mapped to quantitative findings from the public perception study (Chapter 6 of the dissertation) and qualitative evidence from the expert survey (Chapter 7). The full integrated evaluation (Chapter 8) demonstrates that early-phase interventions yield the highest expected impact.

Table 1: The GenAI Disinformation Kill Chain

Stage	GenAI Adaptation	Mitigation Strategy
Recon	Auto-Profiling: LLMs psychograph targets at scale	Privacy-preserving ad tech
Weaponize	Generative Injection: Personalized narratives	Watermarking/C2PA provenance
Deliver	Agentic Swarms: Autonomous dialogue agents	Behavioral bot detection
Exploit	Cognitive Saturation: Fatigue Effect	Friction/throttling for System 2
Impact	Epistemic Collapse: Generalized distrust	Pre-bunking/inoculation games

This framework prioritizes early-phase interventions—model governance and platform detection—over post-hoc fact-checking, which our data suggests is cognitively unsustainable at scale. The "Pre-bunking" strategy draws on inoculation theory [33], which has shown effectiveness in building resistance to manipulation techniques.

5 Dissertation Structure and Future Work

The dissertation comprises nine chapters organized around the four research questions. After establishing the theoretical foundations (Chapters 1–3), we present the generation methodology via RogueGPT (Chapter 4, RQ1), the perception measurement platform JudgeGPT (Chapter 5), quantitative public perception analysis revealing the four core findings (Chapter 6, RQ2), qualitative expert perceptions from a survey of over 500 practitioners (Chapter 7, RQ3), and an integrated evaluation through the empirically validated Disinformation Kill Chain (Chapter 8, RQ4). Origin Lens serves as the mitigation validation artifact throughout, demonstrating that cryptographic provenance standards (C2PA) provide a viable countermeasure once the Indistinguishability Threshold is crossed.

Our findings establish that the "Authenticity Crisis" is not a speculative future risk but an empirically measurable present condition. Human cognition, regardless of digital literacy, is insufficient to reliably detect current-generation LLM output. The primary cost of AI disinformation is not the spread of lies, but the erosion of truth—the Liar's Dividend effect where authentic news is distrusted. This motivates a shift in Web Science from "content moderation" (detecting the fake) to "provenance assertions" (verifying the real).

[4] https://github.com/aloth/cred-1

Future Research Agenda. This dissertation will pursue: (1) *Multimodal Extension*—expanding to audio/video deepfakes; (2) *Cross-Cultural Studies*—identifying universal vs. culture-specific vulnerabilities; (3) *Inoculation Interventions*—testing "Pre-bunking" games [33]; (4) *Provenance Adoption*—studying barriers to C2PA adoption via Origin Lens; and (5) *CRED-1 Refinement*—expanding the domain credibility dataset with additional signals and evaluating its effectiveness for automated pre-bunking at scale.

By providing validated tools and frameworks, this research contributes evidence-based strategies for managing the risks of generative AI in digital information ecosystems. All artifacts are open-source to enable replication and extension by the research community.

References

[1] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential Privacy for Location-Based Systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. 901–914.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[3] Robert Chesney and Danielle Keats Citron. 2019. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review* 107 (2019), 1753–1820.

[4] Coalition for Content Provenance and Authenticity (C2PA). 2024. C2PA Technical Specification v2.1. https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

[5] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12763–12771.

[6] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The Psychological Drivers of Misinformation Belief and Its Resistance to Correction. *Nature Reviews Psychology* 1, 1 (2022), 13–29. doi:10.1038/s44159-021-00006-y

[7] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 111–116.

[8] Shafi Goldwasser and Silvio Micali. 1984. Probabilistic Encryption. *J. Comput. System Sci.* 28, 2 (1984), 270–299.

[9] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119. doi:10.1073/pnas.2110013119

[10] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design Science in Information Systems Research. *MIS Quarterly* 28, 1 (2004), 75–105.

[11] Luis-Daniel Ibáñez, Elena Simperl, Fabien Gandon, and Henry Story. 2017. Redecentralizing the Web with Distributed Ledgers. *IEEE Intelligent Systems* 32, 1 (2017), 92–95.

[12] Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic Detection of Generated Text is Easiest When Humans Are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1808–1822. doi:10.18653/v1/2020.acl-main.164

[13] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[14] 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), 39–43. doi:10.1177/09637214221121570

[15] 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.

[16] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998

[17] Alexander Loth. 2021. *Decisively Digital: From Creating a Culture to Designing Strategy*. Wiley, Hoboken, NJ.

[18] Alexander Loth. 2023. Erweiterung des Machine Learning-gestützten Risikocontrollings mit Synthetic Control. In *Data Driven Controlling: Data Analytics und KI kennen und nutzen*. Haufe Lexware Verlag, 219–230.

[19] Alexander Loth. 2024. *KI für Content Creation: Texte, Bilder, Audio erstellen mit ChatGPT & Co.* mitp, Frechen.

[20] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or Curse? A Survey on the Impact of Generative AI on Fake News. *arXiv preprint arXiv:2404.03021* (2024). https://arxiv.org/abs/2404.03021

[21] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Can Humans Tell? A Dual-Axis Study of Human Perception of LLM-Generated News. In *Proceedings of the 18th ACM Web Science Conference 2026 (WebSci '26)*. ACM, New York, NY, USA.

[22] 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. (2026). doi:10.2139/ssrn.6448466 Preprint available at SSRN.

[23] 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 (WWW '26 Companion)*. ACM, New York, NY, USA. doi:10.1145/3774905.3795832 Also available as arXiv:2601.22871.

[24] 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 (WWW '26 Companion)*. ACM, New York, NY, USA. doi:10.1145/3774905.3795471 Also available as arXiv:2601.21963.

[25] 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 (WWW '26 Companion)*. ACM, New York, NY, USA. doi:10.1145/3774905.3795484 Also available as arXiv:2602.02100.

[26] 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. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*. ACM, New York, NY, USA. Also available as arXiv:2602.03423.

[27] 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 *Proceedings of the 18th ACM Web Science Conference 2026 (WebSci '26)*. ACM, New York, NY, USA.

[28] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

[29] Christopher Paul and Miriam Matthews. 2016. *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. Technical Report PE-198-OSD. RAND Corporation. https://www.rand.org/pubs/perspectives/PE198.html

[30] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 3 (2007), 45–77.

[31] Gordon Pennycook and David G. Rand. 2021. The Psychology of Fake News. *Trends in Cognitive Sciences* 25, 5 (2021), 388–402.

[32] Marc Prensky. 2001. Digital Natives, Digital Immigrants. *On the Horizon* 9, 5 (2001), 1–6.

[33] Jon Roozenbeek and Sander van der Linden. 2019. Fake News Game Confers Psychological Resistance Against Online Misinformation. *Palgrave Communications* 5, 65 (2019).

[34] Kaylyn Jackson Schiff, Daniel S. Schiff, and Natália S. Bueno. 2024. The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability? *American Political Science Review* (2024).

[35] Katerina Sedova, Christine McNeill, Aurora Johnson, Aditi Joshi, and Ido Wulkan. 2021. *AI and the Future of Disinformation Campaigns*. Technical Report. Center for Security and Emerging Technology (CSET), Georgetown University.

[36] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[37] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (2020). doi:10.1177/2056305120903408

[38] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (2018), 1146–1151.

[39] Xinyu Wang, Jiayi Li, and Sarah Rajtmajer. 2024. Inside the Echo Chamber: Linguistic Underpinnings of Misinformation on Twitter. In *Proceedings of the 16th ACM Web Science Conference (WebSci '24)*. ACM, New York, NY, USA, 31–41. doi:10.1145/3614419.3644009

[40] Claire Wardle and Hossein Derakhshan. 2017. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Technical Report DGI(2017)09. Council of Europe.

[41] 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.

[42] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. In *arXiv preprint arXiv:2307.15043.* https://arxiv.org/abs/2307.15043