Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild

Anonymous Author(s)

Abstract

As foundation models (FMs) approach human-level fluency, distinguishing synthetic from organic content has become a key challenge for Trustworthy Web Intelligence. This paper presents **[EvalTool]** and **[GenTool]**, a dual-axis framework decoupling "authenticity" from "attribution" to investigate the causal mechanisms of human susceptibility. Analyzing 918 evaluations across five FMs (including GPT-4 and Llama-2), we employ Structural Causal Models (SCMs) to isolate the determinants of detection accuracy. Contrary to partisan narratives, we find that political orientation exerts negligible causal influence ($r = -0.10$). Instead, "fake news familiarity" serves as a mediator ($r = 0.35$), suggesting that exposure functions as adversarial training for human discriminators. We identify a "fluency trap" where GPT-4 outputs (HumanMachineScore: 0.20) bypass Source Monitoring mechanisms, rendering them indistinguishable from human text. These findings indicate that "pre-bunking" interventions must target cognitive source monitoring rather than demographic segmentation to ensure trustworthy information ecosystems.

Keywords

trustworthy foundation models, truth-default theory, human susceptibility, disinformation, hallucinations, causal analysis, pre-bunking, web intelligence

1 Introduction

The rapid proliferation of foundation models, particularly large language models (LLMs), presents both opportunities and challenges for web intelligence. Han et al. [6] define trustworthy machine learning as encompassing robustness, privacy, and causal integrity—properties that extend beyond model performance to the entire data-model-human pipeline. As these models improve, distinguishing between human-written and machine-generated content has become a central challenge for maintaining trustworthy information ecosystems online [anonymous]. Recent work has examined the causal reasoning capabilities of LLMs [3] and their tendency to hallucinate [15], yet how users perceive and evaluate foundation model outputs "in the wild" remains underexplored.

Building trustworthy AI systems requires understanding not only the technical reliability of models but also the causal factors that influence human susceptibility to AI-generated misinformation and disinformation. Yao et al. [29] frame web data as inherently "imperfect," subject to noise and annotation errors; we extend this insight by treating human perception errors as a form of imperfect annotation that reveals latent variables governing trust. Following Liu et al. [16], we approach the analysis of human judgment patterns as a method of causal discovery—identifying the hidden factors that mediate susceptibility.

To address this gap, we developed **[EvalTool]** [anonymous] ([anonymized]), an interactive, open-source platform for systematically studying human perception of AI-generated content, and **[GenTool]** [anonymous] ([anonymized]), a complementary content generation engine that produces diverse news fragments using multiple foundation models. Together, these tools enable investigation of causal relationships between user characteristics, content properties, and judgment accuracy.

Our research addresses the following questions with implications for trustworthy foundation models:

- **Causal Factors in Susceptibility:** What demographic and experiential factors causally influence users' ability to distinguish foundation model outputs from human-written content?
- **Model-Specific Vulnerabilities:** Do different foundation models vary in their ability to produce human-like content, and what are the implications for trustworthy AI deployment?
- **Learning and Adaptation:** Can exposure to AI-generated content improve detection abilities over time, suggesting potential for "pre-bunking" interventions?

1.1 Contributions

This paper offers three contributions:

- **[EvalTool] and [GenTool] Platforms:** Open-source tools for continuous data collection on human perception of foundation model outputs, featuring dual-axis assessment (authenticity and source attribution) with multilingual support.
- **Causal Analysis of Human Perception:** Empirical findings from 154 participants and 918 evaluations revealing how fake news familiarity, political orientation, and other factors causally influence detection accuracy.
- **Model Comparison Insights:** Evidence that GPT-4 produces more human-like text than other models tested, with implications for trustworthy foundation model development and deployment.

1.2 Paper Structure

Section 2 situates our work within related research. Section 3 describes the [EvalTool] and [GenTool] platforms. Section 4 reports our primary findings. Section 5 discusses implications for trustworthy foundation models and future research directions.

2 Related Work

Research on trustworthy foundation models intersects with several established areas that inform our approach.

2.1 Trustworthy AI and Foundation Models

The challenge of building trustworthy AI systems has received attention as foundation models become widespread [6]. Weidinger et al. [27] provide a taxonomy of LLM risks including misinformation

generation and human overreliance on fluent outputs. Key concerns include model reliability under noisy or adversarial conditions [7], the causal reasoning capabilities of LLMs [3], and understanding why models hallucinate [15]. While Chi et al. [3] demonstrate that LLMs lack "Level 2" causal reasoning, our data reveals that humans falsely attribute this capability to them due to output fluency—a phenomenon Bender et al. [1] term "stochastic parrots," where coherent-sounding text masks the absence of genuine understanding. These technical dimensions of trustworthiness are complemented by the need to understand human factors—specifically, how users perceive and interact with AI-generated content in web environments [29].

2.2 Cognitive Frameworks for AI Perception

Our analysis draws on established cognitive science frameworks. The Source Monitoring Framework [12] explains how humans use heuristic cues (detail, fluency) rather than verified source information when attributing text origins—we term this the "Fluency Trap." Truth-Default Theory, recently applied to AI [18], suggests that AI systems exhibit truth bias, which may create "validity illusions" for users who assume AI outputs are accurate; fluent foundation model outputs may thus be extiteroding the truth-default that humans apply to textual communication. Finally, dual-process theory [13] explains why detection accuracy degrades over time: sustained analytical effort (System 2) depletes cognitive resources, leading participants to rely on faster but less accurate heuristics.

2.3 Human Perception and Fake News Detection

The intersection of generative AI and misinformation presents a dual challenge: while LLMs can assist in detection, they also lower barriers to fake news generation [anonymous]. Automated fake news detection has employed techniques from multimodal architectures [26] to transformer-based systems like GROVER [30]. The RAID benchmark [5] provides the current standard for evaluating machine-generated text detectors, demonstrating that automated systems struggle with adversarial attacks and domain shifts. Our work complements RAID by providing *human-centric* metrics: while RAID measures algorithmic detection, [EvalTool] measures human detection—both are necessary for a complete picture of trustworthiness in web intelligence.

Studies have investigated how demographic factors influence credibility judgments [2, 25, 28] and factors influencing trust in machine-generated content [9]. Clark et al. [4] found that human evaluation of generated text is unreliable, with evaluators often unable to distinguish GPT-3 outputs from human writing. Ippolito et al. [11] demonstrated that automatic detection is easiest precisely when humans are fooled, suggesting complementary roles for algorithmic and human detection. The MIST test [17] measures misinformation susceptibility using psychometric analysis. However, most work focuses on binary classification (real vs. fake), missing the nuanced spectrum of human perception.

2.4 Pre-bunking and Psychological Inoculation

Our finding on learning effects connects to the literature on psychological inoculation against misinformation. Roozenbeek and van der Linden [22, 24] have demonstrated that gamified "pre-bunking" interventions can improve resilience against misinformation on social media. The Debunking Handbook [14] synthesizes evidence on effective correction strategies, emphasizing that prebunking is more effective than debunking. Google Jigsaw's collaboration on EU election pre-bunking campaigns [23] demonstrates real-world applications of this research for web intelligence. We position [EvalTool] as a *dynamic pre-bunking environment*: beyond passive measurement, the platform's iterative exposure to AI-generated content may itself serve as a form of cognitive inoculation.

Table 1 positions [EvalTool] relative to existing approaches.

Tool	Human Eval	Dual-Axis	Multi-LLM
GROVER [30]	✗	✗	✗
MIST [17]	✓	✗	✗
FakeGPT [8]	✗	✗	✗
[EvalTool] (Ours)	✓	✓	✓

Table 1: [EvalTool] provides dual-axis human evaluation across multiple foundation models.

[EvalTool] extends prior work by: (1) capturing nuanced perceptions using continuous scales for both authenticity and source attribution; (2) enabling demographic analysis to identify causal factors in AI content perception; (3) comparing multiple foundation models to assess relative "human-likeness"; and (4) supporting multilingual evaluation for cross-cultural insights.

3 [EvalTool] and [GenTool] Platforms

Our dual-system framework combines **[EvalTool]** [anonymous] ([anonymized]), an interactive platform for evaluating human perception of content authenticity and source, with **[GenTool]** [anonymous] ([anonymized]), a generative engine producing news fragments using multiple foundation models. Together, these platforms form the empirical foundation for our ongoing research program on human-AI interaction in the context of misinformation [anonymous].

3.1 [EvalTool]: Human Evaluation Platform

[EvalTool] implements a Streamlit-based web interface that presents news fragments and captures user evaluations along two continuous dimensions:

- **Perceived Authenticity (LegitFakeScore):** From "Completely Fake" to "Completely Legitimate" — capturing how believable the content appears
- **Source Attribution (HumanMachineScore):** From "Definitely Human-Written" to "Definitely Machine-Generated" — capturing perceived authorship

This dual-axis approach enables analysis of whether humans can distinguish AI-generated content (source attribution) independently of whether they find it credible (authenticity assessment). The platform supports four languages (English, German, French, Spanish) with automatic browser language detection. After obtaining informed consent, participants provide demographic information (age, gender, education, political orientation, fake news familiarity) before evaluating news fragments. The system logs response times and interaction patterns alongside judgments, enabling temporal analysis of learning and fatigue effects.

3.2 [GenTool]: Content Generation Engine

[GenTool] generates news fragments using multiple foundation models, enabling comparative analysis of how different models vary in producing human-like content. The currently supported models include:

- **OpenAI:** GPT-4, GPT-4o, GPT-3.5-turbo
- **Meta:** Llama-2-13b-chat
- **Google:** Gemma-7b-it
- **Mistral AI:** Mistral-7b-instruct
- **Microsoft:** Phi-3-mini-4k-instruct

This multi-model approach informs trustworthy AI deployment: if certain models consistently evade human detection, they pose greater risks for misinformation propagation on the web.

3.3 Data Collection and Metrics

The platform collects three types of data for causal analysis:

(1) **Participant demographics:** Age, gender, education level, political orientation, fake news familiarity, newspaper subscriptions, and geographic location — enabling analysis of susceptibility factors
(2) **Fragment metadata:** Source (human or AI), topic, language, generating model, and veracity (based on factual accuracy) — enabling model comparison
(3) **Assessment metrics:** Continuous slider positions for both dimensions, response times, and interaction patterns — enabling behavioral analysis

As of December 2025, the dataset includes 154 validated participants and 918 evaluations across 2301 news fragments in four languages.

4 Findings

This section presents findings from 918 valid participant responses, examining causal factors influencing human perception of foundation model outputs and comparative model performance.

4.1 From Correlation to Structural Causal Models

While correlational analysis provides initial insights, we frame our findings within a Structural Causal Model (SCM) [10, 20] to make explicit causal claims. We propose a Directed Acyclic Graph (DAG) with the following structure:

- T **(Treatment):** Model Architecture (e.g., GPT-4 vs. Mistral vs. Llama-2 [19])
- M **(Mediator):** Textual Features (perceived fluency, coherence, stylistic markers)
- C **(Confounders):** User Demographics (fake news familiarity, age, education, political orientation)
- Y **(Outcome):** Source Attribution (HumanMachineScore)

This formalization enables claims using Pearl's *do*-calculus: "Controlling for political orientation (C), the direct causal effect of Model Architecture (T) on detection failure (Y) is mediated by Perceived Fluency (M)." Formally, $P(Y|do(T))$ captures the interventional effect of deploying different foundation models on human susceptibility. This language aligns with the causal inference focus of the workshop [6, 16].

4.2 Causal Factors: Fake News Familiarity and Political Orientation

We analyzed correlations between participant characteristics (confounders C) and judgment accuracy (Figure 1):

- **Fake news familiarity predicts accuracy:** Moderate positive correlations exist between fake news familiarity and both source attribution ($r = 0.35$) and legitimacy assessment ($r = 0.29$). This suggests that prior experience with misinformation improves detection capabilities, consistent with findings from our larger reproduction study [anonymous].
- **Political orientation shows weak effects:** Only weak negative correlations exist between political extremity and judgment accuracy ($r = -0.10$ for source, $r = -0.05$ for legitimacy), suggesting political orientation is not a primary causal factor—consistent with Pennycook and Rand's [21] finding that analytical thinking, not partisan bias, predicts fake news susceptibility.

These findings have implications for trustworthy AI: interventions targeting fake news awareness may be more effective than those targeting political groups.

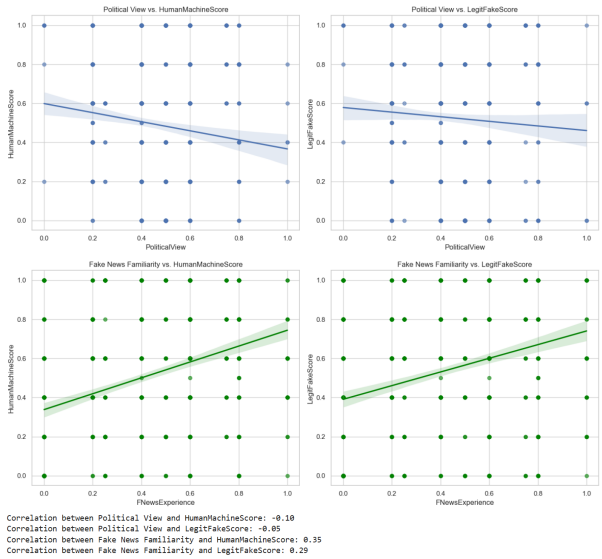

Figure 1: Correlations between participant characteristics and judgment scores. Fake news familiarity shows stronger predictive power than political orientation.

4.3 Foundation Model Comparison: Perceived Humanness

A notable finding for trustworthy foundation models is the variation in how "human-like" different models appear to users. Table 2 shows the average HumanMachineScore by model (lower = more human-like).

Result: GPT-4 outputs received lower scores (0.20) than all other models (next best: 0.44), indicating GPT-4 produces text that participants perceive as more human-like. This finding has implications

for foundation model trustworthiness: as models become more capable, human detection becomes less reliable. Strikingly, when comparing human-written and AI-generated fragments in aggregate (Figure 2), independent t-tests found no significant mean differences between groups for either HumanMachineScore or LegitFakeScore—participants could not reliably distinguish AI content from human content at the population level.

Table 2: Average HumanMachineScore by foundation model (lower = more human-like).

Model	Avg. Score
`openai_gpt-4`	0.200
`mistralai_mistral-7b`	0.442
`openai_gpt-4o`	0.468
`google_gemma-7b`	0.499
`meta_llama-2-13b`	0.526
`openai_gpt-35-turbo`	0.548
`microsoft_Phi-3-mini`	0.580

4.4 Learning Effects and Behavioral Patterns

We examined whether participants' judgment criteria changed over the course of the experiment, possibly due to learning or fatigue. Figure 4 plots the average scores as a function of response order.

Learning Effects: The HumanMachineScore shows a downward trend early in sessions before stabilizing (Figure 4), suggesting participants improve at detecting AI text with exposure. This aligns with research on psychological inoculation against misinformation [24]. Our larger-scale analysis reveals asymmetric cognitive fatigue: fake news detection accuracy degrades under sustained exposure while AI source detection remains relatively stable [anonymous]. This fatigue effect is consistent with dual-process theory [13], where sustained System 2 effort leads to resource depletion. This learning effect suggests that targeted "pre-bunking" training could leverage natural learning mechanisms to improve detection capabilities.

Behavioral Clustering: Clustering analysis (Figure 3) revealed distinct "judging profiles" among participants:

- **Skeptical judges:** Consistently rated content as machine-generated and fake
- **Trusting judges:** Tended to rate content as human-written and legitimate—consistent with Truth-Default Theory applied to AI [18]
- **Variable judges:** Showed inconsistent patterns, possibly reflecting genuine uncertainty

These profiles were influenced by factors including response time patterns and fake news familiarity, suggesting that individual differences in cognitive style affect susceptibility to AI-generated misinformation.

Heterogeneity of Treatment Effect (HTE):. Following Yao et al. [29], we embrace the "imperfect data" inherent to web intelligence. The variation in human response times and the distinct "Skeptical vs. Trusting" clusters (Figure 3) are not merely noise—they represent *heterogeneity of treatment effect* (HTE). Different users exhibit differential susceptibility to different models: Trusting judges are particularly vulnerable to GPT-4's fluency, while Skeptical judges show relative immunity. Identifying *which users* are susceptible to *which models* constitutes a causal discovery task [16]. [EvalTool] provides the granular, user-level data necessary to estimate these heterogeneous effects—a prerequisite for *personalized AI safety interventions* that target specific cognitive profiles rather than broad demographic categories.

5 Conclusion

This research addresses the human element of Web Intelligence. Through our [EvalTool] platform, we collected 918 human evaluations across four languages and identified the specific cognitive and demographic factors that mediate susceptibility to AI-generated misinformation and disinformation.

Our analysis yielded three findings:

(1) **Fake news familiarity as causal mediator:** Prior experience with misinformation predicts detection accuracy ($r = 0.35$ for source attribution, $r = 0.29$ for legitimacy), while political orientation exerts negligible causal influence ($r = -0.10$). This suggests that media literacy interventions may be more effective than demographic targeting.
(2) **The GPT-4 "humanness gap":** GPT-4 produces text perceived as more human-like than other models (HumanMachineScore: 0.20 vs. 0.44+), quantifying the challenge for trustworthy web intelligence. This gap has implications for AI governance and deployment policies.
(3) **Learning effects and cognitive inoculation:** Participants exhibit learning over evaluation sessions, with distinct judgment profiles emerging. This aligns with psychological inoculation research [24] and suggests that "pre-bunking" training could leverage natural learning mechanisms to improve detection capabilities.

Toward Cognitive Inoculation: We align with the call by Han et al. [6] to view trustworthiness as a property spanning data, model, and human interaction. Our findings support a shift from *debunking* (post-hoc fact-checking) to *pre-bunking* (proactive inoculation).

Inoculation theory [22, 24] posits that exposing individuals to a weakened form of a persuasive argument (the "virus") triggers the production of counter-arguments (the "antibodies"), conferring resistance to future, stronger attacks. We position [EvalTool] as a *digital vaccine*: by exposing users to AI-generated text and providing immediate feedback (the "weakened dose"), we force users to update their Source Monitoring heuristics [12]. The learning curve observed in Figure 4 visualizes this "antibody production"—the strengthening of cognitive defenses through controlled exposure.

Policy Recommendations for Web Intelligence: Based on our findings, we propose three complementary interventions:

(1) **Mandatory Provenance (C2PA):** Since human detection of GPT-4 is failing (HumanMachineScore: 0.20), relying on "user vigilance" is a failed policy. We advocate for cryptographic watermarking standards (e.g., C2PA) to provide an external verification layer that does not depend on fallible human perception.

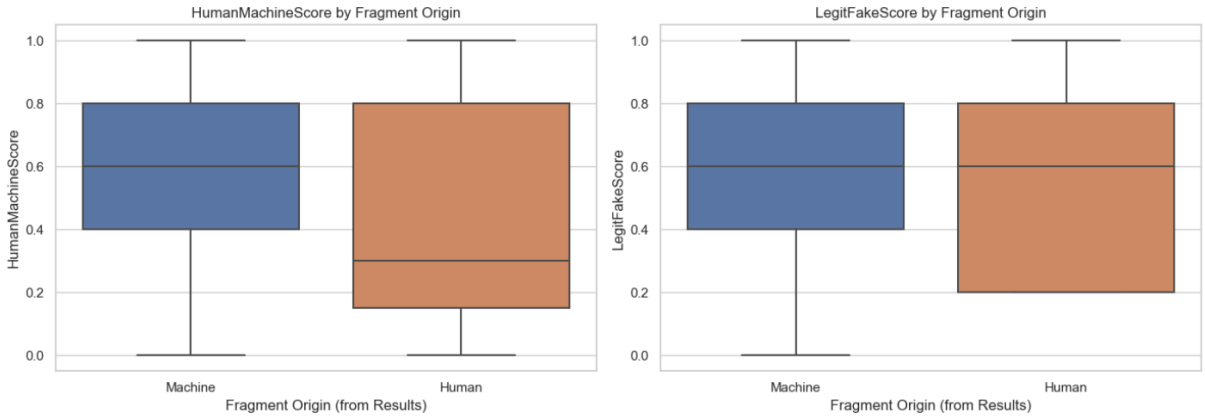

Figure 2: Comparison of HumanMachineScore and LegitFakeScore for fragments by actual source (Human vs. AI). Boxplots show median and interquartile ranges. Independent t-tests found no significant mean differences between human-written and AI-generated groups for either metric, highlighting the challenge of distinguishing foundation model outputs.

(2) **Cognitive Security Education:** Digital literacy programs must pivot from "checking sources" (increasingly impossible with AI-generated content) to "checking logic" (causal consistency). Users must be trained to identify the *absence* of causal reasoning [3] in hallucinations—a skill that targets the fluency trap directly.
(3) **Adversarial User Training:** Platforms should deploy adversarial training for users—periodically presenting synthetic content with verification prompts—to keep Source Monitoring heuristics calibrated to evolving model capabilities. Inspired by Google Jigsaw's EU election pre-bunking campaigns [23], such interventions transform passive consumption into active cognitive exercise.

Limitations and Future Work: The current dataset, while multilingual, is skewed toward educated European participants. We are actively expanding data collection with more diverse populations [anonymous].

From Association to Intervention: Future work must move beyond observing correlations to performing interventions. We propose using [GenTool] to generate text with specific "causal flaws" (e.g., impossible physics, logical contradictions, temporal inconsistencies). By measuring whether users can detect these causal errors better than stylistic errors, we can directly test the "Level 2" causal reasoning hypothesis of Chi et al. [3]. This would reveal whether training users on causal consistency—rather than surface fluency—offers a more robust defense against AI-generated misinformation.

Closing the Trustworthy Cycle: We envision a continuous improvement loop grounded in the framework of Han et al. [6]:

(1) **Detection:** Human feedback via [EvalTool] identifies detection failures and susceptibility patterns.
(2) **Causal Discovery:** Analysis identifies the textual features (M) and user characteristics (C) driving these failures [16].
(3) **Model Improvement:** Discovered features inform "safety classifiers" or RLHF fine-tuning to reduce deceptive mimicry.
(4) **Re-evaluation:** Updated models are re-evaluated by humans, restarting the cycle.

This *Trustworthy Cycle*—integrating human perception, causal inference, and model refinement—provides a framework for developing trustworthy web ecosystems. Future work will incorporate additional foundation models (including Claude, Gemini, and Llama-3 [19]) and investigate whether detection accuracy varies by content domain (e.g., political vs. scientific misinformation). Our ongoing research program includes detailed analysis of the AI-driven disinformation kill chain [anonymous] and reproduction of key findings from the disinformation literature [anonymous].

Open Science: Both platforms are available as open-source projects. [EvalTool] [anonymous] is accessible at [anonymized] (code: [anonymized]), and [GenTool] [anonymous] at [anonymized]. We invite researchers to contribute to this ongoing investigation of human-AI interaction in the context of misinformation.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 610–623. doi:10.1145/3442188.3445922
[2] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. 2021. Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation (Computational Propaganda Research Project).
[3] Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2025. Unveiling Causal Reasoning in Large Language Models: Reality or Mirage? *arXiv preprint arXiv:2506.21215* (2025).
[4] Elizabeth Clark, Tal August, Sofia Serber, 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 (ACL)*. 7282–7296. doi:10.18653/v1/2021.acl-long.565
[5] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12463–12492.
[6] Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo, and Feng Liu. 2025. Trustworthy Machine Learning: From Data to Models. *Foundations and Trends® in Privacy and Security* 7, 2-3 (2025), 74–246. doi:10.1561/3300000043

Figure 3: Pair-plot visualization of participant judging behavior clusters. Each point represents a participant, colored by cluster. The plots show relationships between HumanMachineScore, LegitFakeScore, TimeToAnswer, PoliticalView, and FNewsExperience. Distinct clusters indicate different judgment behavior patterns that may be causally linked to participant characteristics.

[7] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, et al. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[8] Yue Huang and Lichao Sun. 2024. FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models. arXiv:2310.05046 [cs.CL]

[9] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content. arXiv:2309.02524 [cs.HC] https://arxiv.org/abs/2309.02524

[10] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press. doi:10.1017/CBO9781139025751

[11] 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 (ACL)*. 2808–2819. doi:10.18653/v1/2020.acl-main.164

[12] Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. 1993. Source monitoring. *Psychological bulletin* 114, 1 (1993), 3.

[13] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux.

[14] Stephan Lewandowsky, John Cook, Ullrich K. H. Ecker, Dolores Albarracín, Michelle A. Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn J. Newman, Gordon Pennycook, Ethan Porter, et al. 2020. The Debunking Handbook 2020. *Databrary* (2020). doi:10.17910/b7.1182

[15] Hu Li, Haoang Chi, Mingyue Liu, and Wei Yang. 2024. Look within, why LLMs hallucinate: A causal perspective. *arXiv preprint arXiv:2407.10153* (2024).

[16] Feng Liu, Jiangchao Yao, Bo Han, and Tongliang Liu. 2025. Causal Discovery with Language Models as Imperfect Experts. *arXiv preprint arXiv:2501.00063* (2025).

[17] Rakoen Maertens, Friedrich M Götz, Hudson F Golino, Jon Roozenbeek, Claudia R Schneider, Yara Kyrychenko, John R Kerr, Stefan Stieger, William P McClanahan,

Figure 4: Judgment Scores vs. Response Order. HumanMachineScore and LegitFakeScore are plotted over the sequence of trials for each participant (with smoothing). The HumanMachineScore tends to decrease initially (learning effect), while the LegitFakeScore drops later in the session (possible fatigue or strategy shift toward increased skepticism).

Karly Drabot, et al. 2024. The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods* 56, 3 (2024), 1863–1899.

[18] David M Markowitz and Jeffrey T Hancock. 2024. Generative AI are more truth-biased than humans: A replication and extension of core truth-default theory principles. *Journal of Language and Social Psychology* 43, 2 (2024), 261–267.

[19] Meta AI. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[20] Judea Pearl. 2009. *Causality.* Cambridge university press.

[21] Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50. doi:10.1016/j.cognition.2018.06.011

[22] Jon Roozenbeek and Sander van der Linden. 2024. Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist* 29, 3 (2024), 189–205.

[23] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Stephan Lewandowsky, et al. 2024. Prebunking interventions based on the psychological theory of inoculation can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review* (2024). Google Jigsaw collaboration on EU election pre-bunking campaigns.

[24] 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.

[25] Sander Van Der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine* 28, 3 (2022), 460–467.

[26] Yaqing Wang. 2018. *yaqingwang/EANN-KDD18.* https://github.com/yaqingwang/EANN-KDD18 original-date: 2018-11-26T16:32:54Z.

[27] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of Risks Posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022), 214–229. doi:10.1145/3531146.3533088

[28] Jeongwon Yang and Yu Tian. 2021. "Others are more vulnerable to fake news than I Am": Third-person effect of COVID-19 fake news on social media users. 125 (2021), 106950. doi:10.1016/j.chb.2021.106950

[29] Jiangchao Yao, Feng Liu, and Bo Han. 2025. Trustworthy AI under Imperfect Web Data. In *Companion Proceedings of the ACM Web Conference (WWW).* 65–68.

[30] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems* 32 (2019).