Who—or What—Told You? Source Memory for AI- vs. Human-Authored Content

Luise Metzger
University of Mannheim
Mannheim, Germany
luise.metzger@uni-mannheim.de

Neele Bühler
University of Mannheim
Mannheim, Germany
neele.buehler@outlook.de

Sebastian Geiger
University of Mannheim
Mannheim, Germany
sebastian.geiger@students.uni-mannheim.de

Edgar Erdfelder
University of Mannheim
Mannheim, Germany
erdfelder@uni-mannheim.de

Abstract

With the increased availability of LLMs, internet users can increasingly encounter informational content generated by AI. In an initial online study ($N = 101$), we investigate whether people spontaneously categorize content as being AI- or human-authored using the 'Who Said What?' paradigm. Participant response data are analyzed with a multinomial processing tree model to accurately estimate the latent category salience parameter while controlling for confounded cognitive processes. We find poor recall of the AI- vs. human-generated categories and derive steps for a follow-up study.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

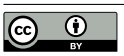
Keywords

human factors, human AI interaction ,LLMs, source monitoring, memory, trust

ACM Reference Format:
Luise Metzger, Neele Bühler, Sebastian Geiger, and Edgar Erdfelder. 2026. Who—or What—Told You? Source Memory for AI- vs. Human-Authored Content. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3772363.3799350

1 Introduction

In recent years, the landscape of online information has changed significantly: Apart from 'traditional' human-curated sources of information—such as forums, online encyclopedias, or news outlets—users now also encounter AI-generated information, e.g., as chatbot responses or automated search summaries. In a recent US-American sample, 64% of respondents indicated that they still use search engines for informational searches while 17% consult AI-chatbots, with the latter percentage reaching 24% among the Gen Z respondents subgroup [2].



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799350

In this research project, we are interested in how internet users handle this change, particularly, whether they spontaneously categorize information they encounter online as AI- vs. human-authored content. To investigate this question, we conducted an initial empirical online study and adapt methods from cognitive psychology, particularly the fields of social categorization and source monitoring.

2 Related Work

With the increasing prevalence of content by generative AI models which effectively emulate human language, several studies have investigated how people perceive and evaluate this content: In general, people are fallible in telling apart human- from AI-authored texts [3–5, 15, 19]. When content is perceived as having been generated by, or assisted by AI, however, this can lead to reduced preferences, lower perceived merit, less trust, and more adverse decisions across several domains [1, 6, 9, 15, 16]. In other words, when prompted—as is usually the case in these studies—people seem to use information about AI vs. human authorship as a relevant cue to adjust their evaluations. In everyday situations, however, they are not usually charged with scrutinizing some piece of content to deduct the nature of its author, asked to explicitly rank directly comparable human- and AI-authored works against each other, or otherwise explicitly pointed to this dichotomy which can then inform further judgements and decisions. In this research, we thus take one step back to ask:

Do people spontaneously categorize informational content they encounter online as AI- or human-authored, and, if so, how strong is this categorization? In other words, we are interested in how *salient* the AI- vs. human-authored categories are. Drawing upon a well-established method from social-cognitive psychology, we operationalize category salience using the 'Who Said What?' source monitoring paradigm: This paradigm was originally introduced by Taylor et al. [18] and has been successfully applied to establish whether people spontaneously categorize sources along specific attributes, e.g., gender or age. In the 'Who Said What?' paradigm, people are presented with a series of statements and multiple sources—for the purpose of our study, online informational websites—which each belong to one of two categories— in our case, websites with AI-authored content (e.g. LLM-based chatbots like ChatGPT) or with human-authored content (e.g. online encyclopedias like Wikipedia). If this categorization is salient to the observer, it will be encoded

in memory in addition to the specific source/website. Therefore, if people are then asked which website a specific statement originated from, even if they do not remember the specific website, they may still recall its category. Subsequently, having recalled the category and trying to reconstruct the specific source website, they will then only guess from within the subset of websites that belong to the same category. For example, if they encountered the statement 'Volleyball was invented in Massachusetts' on ChatGPT.com, they might not remember that it came from ChatGPT, but still know that it was AI-authored. Thus, they would only consider AI-authored websites as a response: They might still correctly guess ChatGPT by chance, or they will respond with another AI-authored website, e.g., Gemini, but the additional information they encoded and recalled, i.e., AI-authorship, would prevent them from guessing out-of-category, so they would not respond with, e.g., Wikipedia or another human-authored website. Consequently, the more salient the AI- vs. human-authored category distinction is, the more within-category guessing errors would be made in comparison to out-of-category guessing errors.

To clearly isolate the latent cognitive process of interest, i.e., category memory as an operationalization of category salience, we follow an extension of the 'Who Said What?' paradigm introduced by Klauer and Wegener [8]. This extension uses multinomial processing tree (MPT) modeling (see [17] for a recent overview) to disentangle the cognitive processes underlying participants' response frequencies, i.e, processes of item discrimination, source discrimination, category discrimination and guessing. Figure 1 illustrates how model parameters D (probability that an old statement is detected as old or a new statement as new), c_{AI} and c_{human} (conditional probability that a specific website is remembered as a statement's source, for AI- and human-authored sources, respectively), d_{AI} and d_{human} (conditional probability that a website's category, AI- or human-authored, is remembered), b (conditional probability that a statement is guessed to be old), and a (conditional probability that a statement is guessed to stem from a website with AI-authored content) are related to responses in the 'Who Said What?' paradigm. As can be seen in the figure, this disentanglement is necessary because different cognitive processes can lead to the same response: Consider the example above, in which a participant is asked on which website they encountered the statement 'Volleyball was invented in Massachusetts' which was originally presented on ChatGPT.com, and they respond with 'Gemini', thus committing a within-category guessing error. This response could have been given because the participant correctly remembered that they encountered the statement before, but did not remember that it came from ChatGPT. However, they did remember that the website the statement came from was AI-authored, and thus guessed from the subset of AI-authored websites, which led to the response 'Gemini'. This would correspons to the third branch from the top in the first tree of Figure 1. However, the same response could have also been given because the participant did not remember that they had seen the statement before, but guessed that it was old, and then guessed Gemini as an AI-authored website, which would correspond to the third branch from the bottom in the first tree of Figure 1. Note that in both cases, the response category is the same, but category memory, reflected by the d_{AI} parameter, is only involved in the first one. For our research question, we thus focus on the d parameters derived from fitting the MPT model as pure, uncontaminated measures of category memory.

3 Method

To assess whether people spontaneously categorize online content as being AI- vs. human-authored, we conducted an initial online study. $N = 101$ participants with German as a first language aged 18–50 ($M = 31$, $SD = 7.15$) were recruited via the Prolific platform [14]. Out of these, 87 resided in Germany, 8 in Austria, and 6 in Switzerland. 53 identified their gender as female, 43 as male, 1 as non-binary, and 4 preferred not to say. For a graphical summary of the study procedure, which followed the adapted version of the 'Who Said What?' paradigm [8, 18], see Figure 2. After they had consented to participate, participants were told to imagine preparing for a pub quiz by researching information on different websites online as a cover story. They were then provided with short descriptions of six websites. Unbeknownst to the participants, we had selected these websites based on a pilot study, so that three of them were assumed to feature AI-authored and the other three to feature human-authored content. Having read the website descriptions, participants then entered the main part of the experiment: In an encoding phase, they were then shown 84 trivia statements. Each statement was presented in the context of a simplified screenshots of one of six websites allocated in a randomized manner for each participant. Trivia statements were German-language and adapted from truth effect items taken from Nadarevic and Erdfelder [11] and Nadarevic et al. [12]. They were selected to be as ambiguous in truth value as possible, so that most participants could be expected to be uncertain about whether a statement was true or false. After a filler task, participants then completed a previously unannounced source monitoring test. In line with the Klauer and Wegener [8] adaptation to allow for MPT-analysis, this test included the 84 statements from the encoding phase plus 42 new statements. For each statement, participants were asked whether it was new, or, if they assumed to have seen it before, which website it had appeared on. We further assessed individual attitude measures as well as additional questions about how the websites and statements were perceived as well as demographic questions about the participants.

4 Results

All MPT models were fitted using multiTree [10] with 500 bootstrap samples. Response frequencies were aggregated across participants.

Maximum-likelihood (ML) estimates for all MPT parameters are depicted in Table 1. The main parameters of interest are the category discrimination parameters, i.e., the respective probabilities to recall either the AI-authored category, d_{AI}, or the human-authored category, d_{human}. Here, both aspects of category discrimination are small: d_{AI} is not significantly different from 0 as demonstrated by a G^2 model comparison test, $\Delta G^2(1) = 0.00, p = .59$. While d_{Human} is significantly larger than 0, $\Delta G^2(1) = 6.03, p < .05$, its ML estimate is descriptively small as well with $d_{human} = 0.12$.

5 Discussion

While prior research explicitly asked people to attempt to differentiate between and evaluate AI- vs. human-authored content, we wanted to find out if this distinction is salient enough for internet

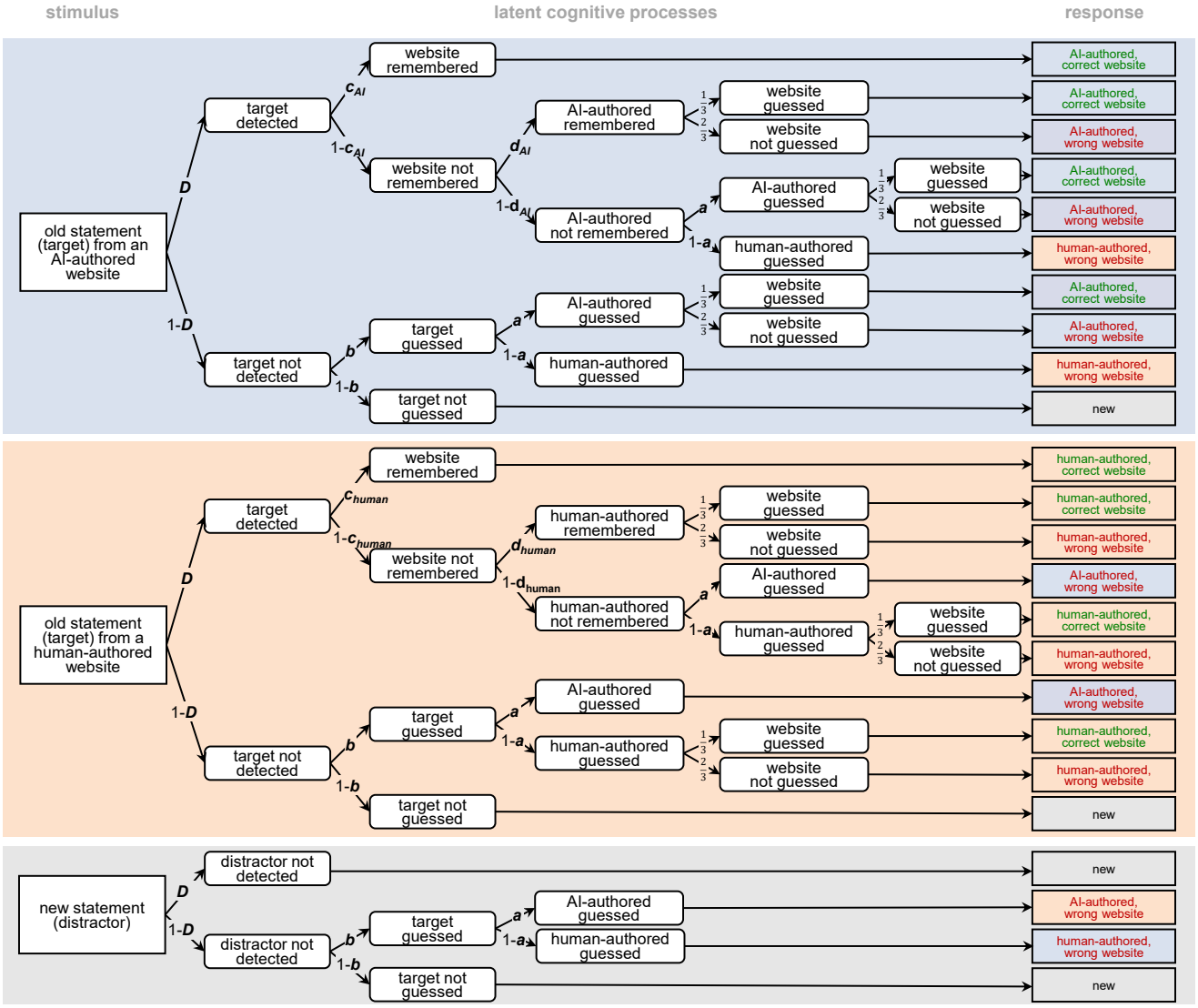


Figure 1: Multinomial processing tree model for the 'Who Said What?' paradigm with AI- vs. human-authored sources. Each branch indicates a possible cognitive process sequence with parameters representing conditional probabilities of process outcomes.

users to apply it even when unprompted. For this, we adapted the 'Who Said What?' paradigm to assess category salience: In an online study, we showed participants several informational statements attributed to websites with either AI- or human-authored content. Our results indicate that participants rarely retrieved the AI- vs. human-authored categorization to a high extent, with the relevant d parameter estimate being equivalent to zero in case of AI-authored, and descriptively low in case of human-authored websites. Despite the changes in the online information landscape in which AI- and human-authored content now both occur frequently, people have not adapted to this by paying attention to the distinction, or do not deem it sufficiently relevant.

Resulting from the unprompting nature of our experiment and our waiver of explicit learning instructions to address spontaneous processes of categorization in source monitoring, it is important to note a limitation: While participants evidently paid attention to the presented statements—as evidenced by the good memory performance in this regard, i.e., the high value of the item discrimination parameter D— not only was category discrimination d poor, but also the memory for the specific websites as reflected in low values of the c parameters of source discrimination. Accordingly, the rather poor category discrimination might also have resulted from people not having paid attention to sources in the first place.

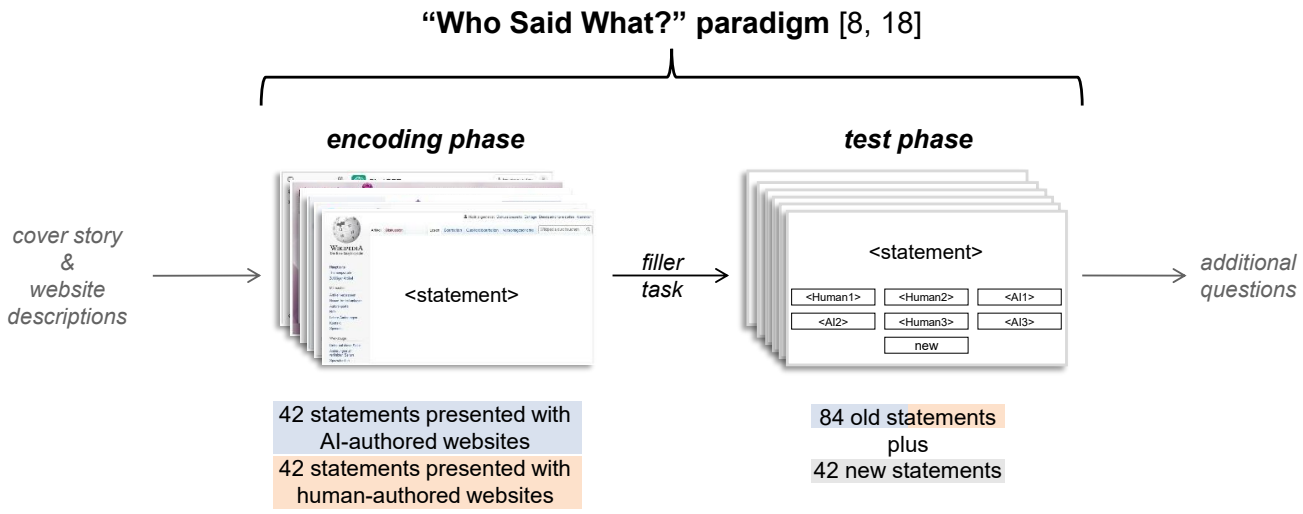


Figure 2: Procedure of the Online Experiment.

parameter		estimate (SE)
label	**meaning**	
D	item discrimination	0.79(0.01)
c_{AI}	source discrimination for AI-authored websites	0.01(0.01)
c_{human}	source discrimination for human-authored websites	0.05(0.01)
d_{AI}	category discrimination for AI-authored websites	0.00(0.06)
d_{human}	category discrimination for human-authored websites	0.12(0.05)
b	guessing old	0.45(0.01)
a	guessing AI-authored	0.51(0.02)

Table 1: MPT Results: Maximum-likelihood estimates and standard errors on the probability scale

Nevertheless, especially since the lines between human- and AI-generated content blur and the discussion around ethical issues like hallucinations or accountability continues [7], this is an interesting finding: While there is evidence that people do use information about AI- vs. human authorship as a cue for how they evaluate content, they apparently do not categorize AI- vs. human-authored sources spontaneously in memory, thus forgoing potentially important insights. The question remains what the limit is for where these categories become more salient, e.g., if this can be influenced by user interface design decisions such as including disclaimers about the nature of a source, or establishing explicit labels for AI- vs. human-authored content. We plan to further investigate this issue in a follow-up in which we extend the current research to a US sample. In this next step, we will increase source salience to test whether category salience emerges under this facilitated condition. Further, should category salience emerge, it will then be interesting to see which factors influence this both on the front of UI design, as well as on on the individual level. Again drawing from human factors and social-cognitive psychology research alike, we assume trust to be an important factor. Trust is known to be relevant for proper adoption behavior, and has previously been linked to source monitoring in human-human interaction contexts, finding that lower levels of trust relate to stricter source monitoring [13].

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) GRK 2277 "Statistical Modeling in Psychology". The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

[1] Joy Buchanan and William Hickman. 2024. Do People Trust Humans More than ChatGPT? *Journal of Behavioral and Experimental Economics* 112 (Oct. 2024), 102239. doi:10.1016/j.socec.2024.102239

[2] Claneo. 2025. State of Search 2025: Insights into American Online Search Behavior. https://www.claneo.com/en/state-of-search-us/.

[3] Alexandra Fiedler and Jörg Döpke. 2025. Do Humans Identify AI-generated Text Better than Machines? Evidence Based on Excerpts from German Theses. *International Review of Economics Education* 49 (June 2025), 100321. doi:10.1016/j.iree.2025.100321

[4] Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan D. Keller, Olaf Köller, and Jens Möller. 2024. Do Teachers Spot AI? Evaluating the Detectability of AI-generated Texts among Student Essays. *Computers and Education: Artificial Intelligence* 6 (June 2024), 100209. doi:10.1016/j.caeai.2024.100209

[5] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing Scientific

Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers. *npj Digital Medicine* 6, 1 (April 2023), 75. doi:10.1038/s41746-023-00819-6

[6] Jakub Harasta, Tereza Novotná, and Jaromir Savelka. 2024. It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human. doi:10.48550/ARXIV.2407.06798

[7] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. 2025. Navigating LLM Ethics: Advancements, Challenges, and Future Directions. *AI and Ethics* 5, 6 (Dec. 2025), 5795–5819. doi:10.1007/s43681-025-00814-5

[8] Karl Christoph Klauer and Ingo Wegener. 1998. Unraveling Social Categorization in the "Who Said What?" Paradigm. *Journal of Personality and Social Psychology* 75, 5 (Nov. 1998), 1155–1178. doi:10.1037/0022-3514.75.5.1155

[9] Edward Lee and Andrew Moshirnia. 2024. The AI Penalty: Is There a Bias against AI-Generated Works? *Michigan State Law Review* 2024, 3 (2024), 641–734.

[10] Morten Moshagen. 2010. multiTree: A Computer Program for the Analysis of Multinomial Processing Tree Models. *Behavior Research Methods* 42, 1 (2010), 42–54. doi:10.3758/BRM.42.1.42

[11] Lena Nadarevic and Edgar Erdfelder. 2025. On the Relationship between Recognition Judgments and Truth Judgments: Memory States Moderate the Recognition-Based Truth Effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 51, 11 (Nov. 2025), 1778–1795. doi:10.1037/xlm0001460

[12] Lena Nadarevic, Martin Schnuerch, and Marlena J. Stegemann. 2021. Judging Fast and Slow: The Truth Effect Does Not Increase under Time-Pressure Conditions. *Judgment and Decision Making* 16, 5 (Sept. 2021), 1234–1266. doi:10.1017/S193029750000841X

[13] Michelle M. Pena, J. Zoe Klemfuss, Elizabeth F. Loftus, and Amelia Mindthoff. 2017. The Effects of Exposure to Differing Amounts of Misinformation and Source Credibility Perception on Source Monitoring and Memory Accuracy. *Psychology of Consciousness: Theory, Research, and Practice* 4, 4 (Dec. 2017), 337–347. doi:10.1037/cns0000137

[14] Prolific. 2025. *Prolific*. London, United Kingdom. https://www.prolific.com/

[15] Kristina Radivojevic, Matthew Chou, Karla Badillo-Urquiola, and Paul Brenner. 2024. Human Perception of LLM-generated Text Content in Social Media Environments. doi:10.48550/ARXIV.2409.06653

[16] Manav Raj, Justin M. Berg, and Rob Seamans. 2026. The Artificial Intelligence Disclosure Penalty: Humans Persistently Devalue AI-generated Creative Writing. *Journal of Experimental Psychology: General* (Jan. 2026). doi:10.1037/xge0001889 Advance online publication.

[17] Oliver Schmidt, Edgar Erdfelder, and Daniel W. Heck. 2025. How to Develop, Test, and Extend Multinomial Processing Tree Models: A Tutorial. *Psychological Methods* 30, 4 (Aug. 2025), 720–743. doi:10.1037/met0000561

[18] Shelley E. Taylor, Susan T. Fiske, Nancy L. Etcoff, and Audrey J. Ruderman. 1978. Categorical and Contextual Bases of Person Memory and Stereotyping. *Journal of Personality and Social Psychology* 36, 7 (July 1978), 778–793. doi:10.1037/0022-3514.36.7.778

[19] Tal Waltzer, Celeste Pilegard, and Gail D. Heyman. 2024. Can You Spot the Bot? Identifying AI-generated Writing in College Essays. *International Journal for Educational Integrity* 20, 1 (July 2024), 11. doi:10.1007/s40979-024-00158-3