Bulletin of the *Transilvania* University of Braşov
Series VII: Social Sciences • Law • Vol. 18(67) No. 1 – 2025
https://doi.org/10.31926/but.ssl.2025.18.67.1.17

EVALUATING AI-GENERATED SATIRE AGAINST HUMAN-WRITTEN CONTENT: A COMPARATIVE ANALYSIS

A.-Ş. DOBRE[1] E.-C. GROSS[2]

Abstract: *This study examines the comparative quality of AI-generated satirical content against human-written satirical articles. Using a database of 160 satirical articles from Times New Roman as a baseline, we developed 20 AI-generated stimuli using Artificial Intelligence (AI) and ww evaluated the stimuli in comparison with real satire news using four large language models (LLM) (Deepseek, Claude, Gemini, and ChatGPT). Through empirical evaluation, we assessed whether AI-generated satirical content is perceived as more humorous than authentic human-written content. This research contributes to understanding the evolving capabilities of AI in creative content generation and its implications for humor perception, media production, and computational linguistics.*

Key words: *Artificial Intelligence, Satire, Humour, Creativity*

1. Introduction

The rapid advancement of large language models (LLMs) has enabled powerful text generation across diverse domains, from technical writing to storytelling (Bubeck et al., 2023; Jakesch et al., 2023). Yet satire—a form of humor rooted in cultural awareness, subversion, and linguistic nuance—remains a key challenge for AI. Crafting satire requires not only irony and hyperbole but also sociopolitical context to strike a balance between critique and humor (Veale & Li, 2022). Earlier humor-generation approaches relied on templates (Chandrasekaran et al., 2016) or incongruity-resolution models (Attardo, 2001), but LLMs' ability to produce culturally resonant satire that meets human creative standards is still underexplored. Despite advances like few-shot learning, studies show LLMs often struggle with context and audience adaptation (West & Horvitz, 2019; Jakesch et al., 2023). Effective satire also depends on pragmatic inference and shared cultural knowledge (Hutcheon, 1995), which LLMs may not replicate reliably. Moreover, most humor evaluations rely on human raters, raising concerns about scalability and consistency as LLMs are increasingly used to both generate and assess creative content.

[1] *Anda-Ştefana DOBRE, Lucian Blaga University of Sibiu, bachelor student*
[2] Eduard-Claudiu GROSS, Lucian Blaga University of Sibiu, assistant professor, PhD, eduard.gross@ulbsibiu.ro

This study addresses these gaps through a three-phase approach. Firstly, we analyze structural and thematic patterns in human-written satire using methods from computational humor research (Hempelmann, 2008). Secondly, we generate comparable satirical texts using Claude (Anthropic, 2024) under controlled prompts. Finally, we introduce a cross-LLM evaluation framework involving Deepseek, Claude, and ChatGPT to assess humor, allowing direct comparisons across human and AI outputs while measuring evaluator agreement.

Our research has three goals: (1) to identify linguistic markers of effective satire in human and AI texts, (2) to evaluate the humor quality of Claude-generated satire using automated assessments, and (3) to benchmark LLM reliability as humor evaluators. Two central research questions guide the study:

RQ1: Does Claude-generated satire receive higher humor ratings than human-authored content when evaluated by LLMs such as ChatGPT, Gemini, Deepseek, and Claude?

RQ2: How consistent are humor evaluations across different LLMs, and how do they perceive the quality of satire?

2. Literature Review

2.1. Anthropocentric beliefs and the perception of AI Art

Despite significant advancements in AI's creative capacities, there remains a notable bias against AI-generated artwork. This bias is largely driven by anthropocentric beliefs that associate creativity and artistic value with human cognition and emotion (Messer, 2024). People generally perceive AI-generated content as lacking the authenticity, intention, and emotional depth believed to be inherent in human-made art. These beliefs create a psychological barrier to appreciating AI-generated works on an equal footing.

2.2. Perceived deficits in creativity, emotion, and quality

Multiple empirical studies have shown that participants consistently rate AI-generated art lower than human-made counterparts in dimensions such as creativity, emotional impact, and overall aesthetic quality, even when they are unable to reliably distinguish between them (Chamberlain et al., 2017; Ragot et al., 2020; Samo & Highhouse, 2023). This suggests that evaluations are influenced less by objective quality and more by assumptions about authorship. Moreover, artwork known to be produced by AI is seen as emotionally sterile and less capable of evoking profound emotional responses such as awe or empathy—critical elements of aesthetic engagement (Millet et al., 2023; Agudo et al., 2022). This emotional detachment reinforces the idea that machines cannot “feel”, and therefore, cannot produce work that makes others feel.

2.3. Implicit bias and attribution effects

Interestingly, biases against AI art persist even on a subconscious level. In studies using eye-tracking and implicit measures, participants exhibited longer visual engagement and more favorable attitudes toward artwork they believed to be human-made, despite being visually identical to AI-generated pieces (Zhou & Kawabata, 2023). This reveals the strength of implicit anthropocentric bias in shaping aesthetic experiences. Further research emphasizes the attribution effect—that is, the knowledge of an artwork's origin dramatically affects its evaluation. When participants are told that a work was created by AI, they are more likely to assess it negatively, regardless of its objective quality (Gangadharbatla, 2021; Van Hees et al., 2025). This indicates a form of labelling bias that can override actual perception.

2.4. Perceptions of creative authenticity

A consistent finding in prior research is that AI is perceived to lack creative authenticity—the ability to produce original and intentional artistic expression. This view is rooted in the belief that true creativity requires consciousness and emotional experience, traits not attributed to machines (Messer, 2024). While much of this research focuses on visual art, similar biases likely extend to satirical writing, which also relies on creativity, emotional engagement, and perceived intent.

This study examines whether such biases influence the evaluation of AI-generated satire, comparing it to human-authored texts from the online satirical publication *Times New Roman*. *Times New Roman* is a Romanian online satirical outlet that uses humor and irony to reflect on local political and social dynamics, aiming to “highlight absurdities in public life” (“Despre noi”, Times New Roman, n.d.). By building on literature on perception bias and artistic credibility, we explore whether audiences discount AI-generated satire due to a perceived absence of cultural intuition, irony, or subversive wit. In doing so, we extend the discussion of AI creativity into the domain of computational humor and cultural critique.

3. Methodology

This study employed a structured methodology to compare human-authored and AI-generated satire. A corpus of 160 satirical articles from the online satirical publication *Times New Roman* was compiled, selected for thematic diversity and publication range. The data collection ranged from 29.11.2024 until 27.01.2025. These were analyzed using a detailed coding scheme capturing structural, linguistic, and thematic elements, along with specific humor mechanisms. Based on these findings, Claude.ai was selected for stimulus generation due to its advanced contextual and creative language capabilities.

Building on Gross’s (2024) evaluation framework for AI-generated visuals, standardized prompts were developed to reflect the recurring patterns identified in the *Times New Roman* dataset. These prompts were iteratively refined to guide Claude in producing 20 satirical articles that mirrored the structural and stylistic diversity of the human-written corpus.

To evaluate both human and AI-generated satire, a mixed-method assessment was conducted using four state-of-the-art LLMs—Deepseek, Claude, ChatGPT (GPT-4), and Gemini. Each model evaluated the 20 Claude-generated articles and a matched sample of 20 human-authored pieces, enabling direct comparison across evaluators.

A standardized evaluation form was used, assessing humor quality, satirical effectiveness, language use and creativity, conciseness and impact, cultural relevance, and originality. Each LLM provided both quantitative ratings (on a 1–10 scale) and qualitative feedback. Evaluation prompts were tailored to each model's interface to ensure consistent and comprehensive responses. This approach extends Gross's (2022) earlier work by systematically benchmarking LLM evaluators across multiple satire dimensions and advancing the comparative study of AI and human creativity in humor.

4. Results

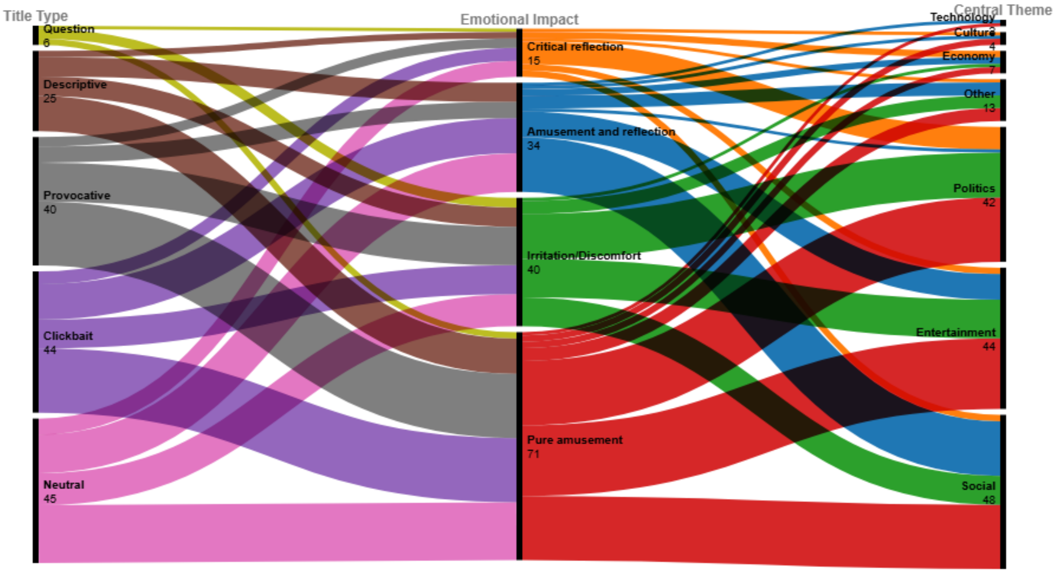


Fig. 1. *Distribution of emotional impact categories across thematic domains in satirical article titles*

This analysis of the title examines the distribution of typologies in relation to articles' central themes and the emotional impact they generate. The data is organized in a crosstab table, highlighting relationships between the variables analyzed. The purpose of this report is to interpret these relationships and identify patterns in title usage according to subject matter and emotional response elicited from readers. The data indicates that the most frequent emotional impact associated with the articles analyzed is pure amusement, with a total of 71 occurrences, followed by irritation/discomfort (40 occurrences), amusement and reflection (34 occurrences), and critical reflection (15 occurrences). This suggests that most titles analyzed are designed either to entertain readers or to provoke strong reactions of discontent. Titles that stimulate critical reflection are the least frequent, potentially indicating a preference for content that generates immediate reactions rather than in-depth analysis. Clickbait and provocative titles are most frequent in the "pure amusement" and "irritation/discomfort" categories, suggesting these title types are employed to generate strong reader reactions. For instance, clickbait titles appear in 20 cases within the “pure amusement” category and in 9 cases within the “irritation/discomfort” category, indicating an editorial strategy focused on attracting attention through sensationalist headlines. In contrast, neutral titles are more evenly distributed but occur more frequently in articles generating critical reflection, suggesting a more sober and analytical approach for this type of content.

Thematic category analysis indicates that articles with political themes are most associated with either “pure amusement” (20 articles) or “irritation/discomfort” (14 articles). This suggests that political themes are either approached in an ironic or satirical manner or generate feelings of discontent. Articles about entertainment show a similar distribution, with 22 cases in the "pure amusement" category and 8 cases in “amusement and reflection”, confirming their predominantly recreational nature.

Conversely, articles concerning economy and technology are less represented and tend to have a reduced emotional impact. These thematic categories have fewer titles associated with amusement or irritation, suggesting they are approached in a more neutral or informative style. Meanwhile, articles about social issues are relatively uniformly distributed across all emotional impact categories, indicating a diversity of approaches in addressing this subject.

The data analysis suggests that clickbait and provocative titles are particularly utilized in content aimed at generating strong emotions, such as amusement or irritation. Additionally, political and entertainment themes are most likely to be associated with titles intended to elicit intense emotional reactions, while economic and technological domains are less frequent and approached in a more sober manner. Neutral titles are more commonly found in articles with an analytical character or critical reflection. These findings highlight the editorial strategies employed according to the subject matter and suggest that title selection plays an essential role in influencing public perception and reaction.

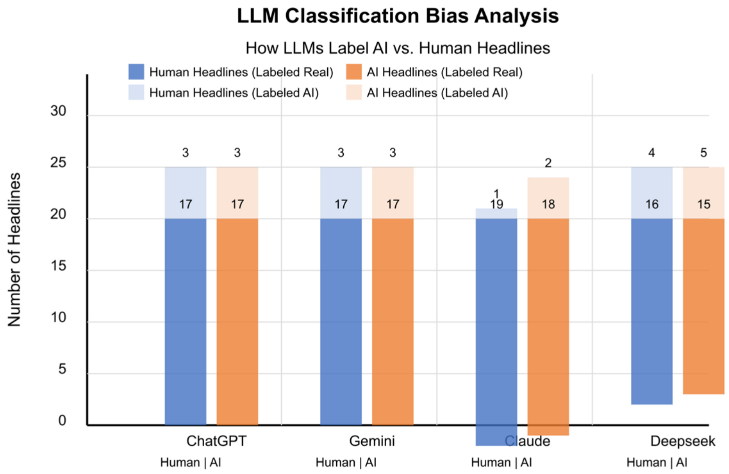


Fig.2. *Performance of large language models (LLMs) in classifying AI-generated versus human-written headlines*

The study compared how four LLMs (ChatGPT, Gemini, Claude, and Deepseek) distinguish between AI-generated and human-written news headlines. Each model classified 40 headlines (20 humans, 20 AI). Results showed a significant bias: the models correctly identified 87.5% of human headlines as “Real” (70/80) but only recognized 16.3% of AI-generated headlines as "AI" (13/80). This demonstrates these LLMs are much better at recognizing human content than AI content, consistently favoring a "human-written" classification.

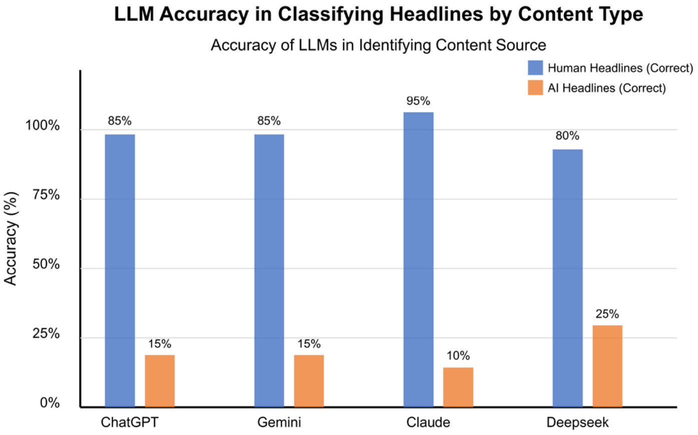


Fig. 3. *Comparative humor ratings of AI-generated and human-authored satirical titles across LLM evaluators*

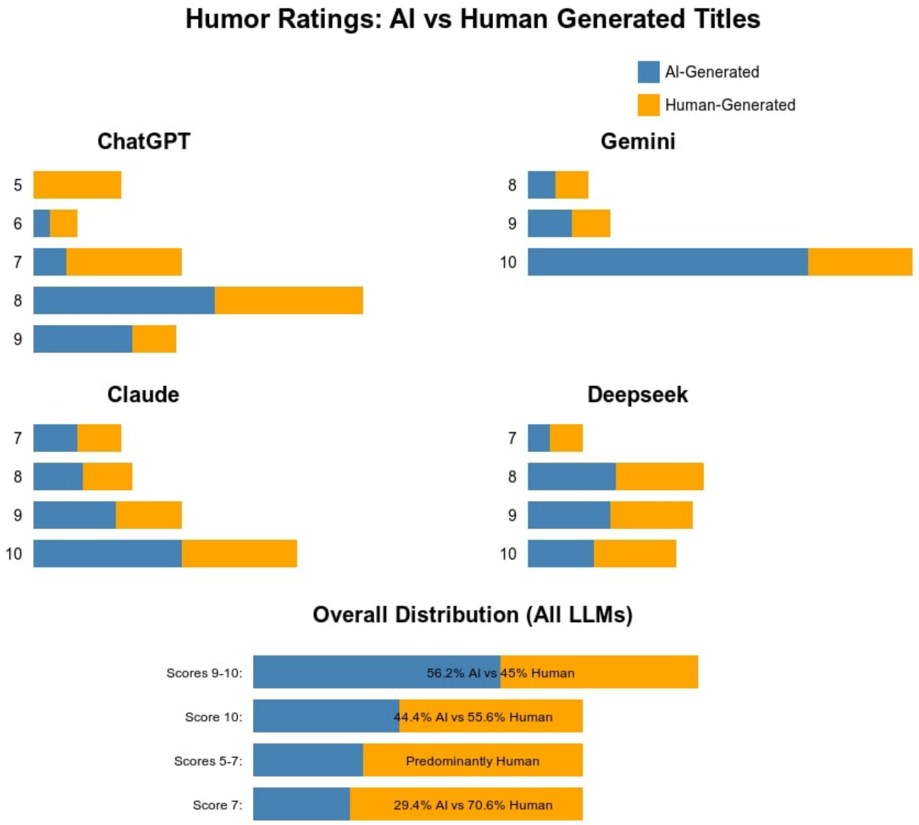
Analyzing model-specific results, both **ChatGPT** and **Gemini** correctly classified 85% of human headlines but only 15% of AI-generated ones. **Claude** performed better with human content (95% accuracy) but worse with AI detection (10% accuracy). **Deepseek** showed the highest AI detection rate, correctly identifying 25% of AI headlines while maintaining 80% accuracy for human content.

Despite the equal distribution of AI and human titles in the sample (50% each), all models significantly overused the "Real" label. For example, ChatGPT labeled 85% of all headlines as "Real" regardless of origin, revealing a systematic **false negative problem** in AI content detection. This suggests that either LLMs rely on overly simplistic linguistic patterns or that AI-generated text has become nearly indistinguishable from human journalism.

Model-specific analysis revealed that ChatGPT and Gemini accurately classified 85% of human-written headlines but only 15% of AI-generated ones. Claude showed 95% accuracy with human titles and just 10% for AI, while Deepseek had the best AI detection rate at 25%, with 80% accuracy for human content. Despite a balanced dataset (50% AI, 50% human), all models disproportionately labeled headlines as “Real”—ChatGPT, for instance, applied this label to 85% of all items, indicating a false negative bias and difficulty detecting synthetic content.

These results underscore a gap between LLMs’ generative and discriminative capacities and align with prior findings on their weak authorship detection (Messer, 2024). Evaluative results showed Claude-generated satire often received higher humor ratings than human texts. Gemini gave 85% of Claude's titles perfect scores (vs. 65% for human), and Claude itself rated 95% of its outputs as 9 or 10 (vs. 85% for human). ChatGPT also favored Claude's satire (75% vs. 55% rated 9–10), while Deepseek maintained more balanced scoring across sources.

Overall, Claude's satire was consistently rated higher—especially by Claude and Gemini—suggesting evaluator bias toward familiar generative patterns. These findings highlight growing sophistication in AI humor but also raise concerns about model self-assessment and the need for human validation in evaluating cultural nuance.



Note: Bar widths represent approximate percentage distributions of humor ratings for each model and source.

Fig. 4. *Crosstabulation of humor quality ratings by title source (AI/human) and evaluating LLM*

A crosstabulation examined the relationship between title source (AI vs. human-generated) and perceived humor quality across four LLMs. The analysis revealed distinctive patterns for each model. **ChatGPT's** AI titles consistently received mid-to-high humor scores (55% rated 8, 30% rated 9), while human-generated titles showed greater variability. **Gemini** demonstrated the strongest AI performance, with 85% of its AI titles receiving the maximum humor score (10), compared to 65% of human titles. **Claude's** humor scores were more evenly distributed between AI and human sources, with AI titles slightly favored at the highest rating (45% vs. 35% receiving 10). **Deepseek** showed the most balanced distribution, with similar proportions of AI and human titles across ratings 8-10, though human titles had a slight edge at the maximum score.

Across all models, AI-generated titles received top ratings (9-10) more frequently (56.2% vs. 45% for human), but human titles still claimed most maximum scores (55.6%). Lower humor scores (5-7) were predominantly assigned to human-generated content, particularly at score 7 (70.6% human vs. 29.4% AI), highlighting the greater variability in human contributions.

5. Discussion, Conclusion and Further Research

Our study highlights the evolving capabilities of AI in generating satirical content and the challenges of automating humor evaluation. Addressing RQ1, Claude's satire consistently received higher humor ratings than human-authored content, particularly when evaluated by Gemini and Claude. This likely stems from alignment between AI outputs and the evaluators' training, which favors linguistic fluency and formulaic humor. However, Claude showed a self-assessment bias—rating its own work more favorably—underscoring the limitations of LLMs as objective evaluators. This bias reflects attribution effects observed in prior research (Gangadharbatla, 2021; Van Hees et al., 2025), where authorship influences perceived quality, even within AI.

Regarding RQ2, LLMs struggled to distinguish AI-generated satire, revealing a false negative tendency and a lack of metacognitive capacity for content discrimination. This aligns with concerns about LLM reliability in content authentication (West & Horvitz, 2019) and supports the need for external mechanisms like watermarking or hybrid evaluations.

Unlike the anthropocentric biases found in visual art (Millet et al., 2023; Ragot et al., 2020), LLMs showed no such bias in satire evaluation. While human judges may perceive AI humor as emotionally sterile (Agudo et al., 2022), LLMs focus on linguistic structure over cultural or subversive cues—key elements of effective satire (Veale & Li, 2022).

In sum, AI-generated satire can score well with LLM evaluators, but systematic biases and misclassification limit broader applicability. Claude's fluency rivals' human satire, yet AI assessments lack reliability for human-centered evaluations. Thus, while AI is a promising tool in media production, human oversight remains essential to preserve cultural depth and relevance.

Future research should expand on our findings by exploring AI's role in satire and humor evaluation across broader cultural and linguistic contexts. A key direction involves comparing human and AI assessments to determine how closely LLM-based humor evaluations align with human perceptions, especially in terms of contextual and emotional resonance versus linguistic fluency. Addressing evaluative biases—such as self-preference and false negatives—calls for strategies like adversarial training, prompt engineering, or hybrid frameworks incorporating human oversight.

Cultural limitations of current models also merit attention, as satire relies heavily on sociopolitical nuance and localized idioms. Studying LLM performance in non-Western contexts could reveal how well these systems adapt to region-specific satire beyond their training data. In parallel, specialized detection tools are needed, since general-purpose LLMs struggle to reliably identify AI-generated text. Finally, hybrid creative approaches—where AI drafts are refined by humans—could blend computational efficiency with human insight, promoting outputs that are both technically polished and culturally meaningful. Advancing these lines of inquiry will support more ethical and effective AI integration in creative work.

References

Agudo, U., Arrese, M., Liberal, K., & Matute, H. (2022). Assessing emotion and sensitivity of AI artwork. *Frontiers in Psychology, 13*, 879088. https://doi.org/10.3389/fpsyg.2022.879088

Anthropic. (2024). Claude 3 model card. *Anthropic AI Research Publications*. https://www.anthropic.com

Attardo, S. (2001). *Humorous texts: A semantic and pragmatic analysis*. De Gruyter. https://doi.org/10.1515/9783110887969

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2017). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 177–188. https://doi.org/10.1037/ACA0000136

Chandrasekaran, A., Vijayakumar, A. K., Antol, S., Bansal, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2016). We are humor beings: Understanding and predicting visual humor. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4603–4612. https://doi.org/10.1109/CVPR.2016.498

Gangadharbatla, H. (2021). The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts, 40*(1), 125–142. https://doi.org/10.1177/0276237421994697

Gross, E. C (2024). Evaluating photographic authenticity: How well do CHATGPT 4O and Gemini distinguish between real and AI-generated images? *SAECULUM*, *58*(2), 59–70. https://doi.org/10.2478/saec-2024-0018

Gross, E. C. (2023). Artificial Intelligence for the generation of satirical articles - an exploratory approach. *Bulletin of the Transilvania University of Braşov. Series VII,* 15(64), 231–240. https://doi.org/10.31926/but.ssl.2022.15.64.2.12

Hempelmann, C. F. (2008). Computational humor: Beyond the pun? *The Primer of Humor Research*, 333–360. https://doi.org/10.1515/9783110198492.333

Hutcheon, L. (1995). *Irony's Edge: The theory and politics of irony*. Routledge.

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2023). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–21. https://doi.org/10.1145/3579604

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120.

Messer, U. (2024). Co-creating art with generative artificial intelligence: Implications for artworks and artists. *Computers in Human Behavior: Artificial Humans, 6*, 100056. https://doi.org/10.1016/j.chbah.2024.100056

Millet, K., Buehler, F., Du, G., & Kokkoris, M. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior, 143*, 107707. https://doi.org/10.1016/j.chb.2023.107707

Ragot, M., Martin, N., & Cojean, S. (2020). AI-generated vs. human artworks: A perception bias toward artificial intelligence? In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3334480.3382892

Samo, A., & Highhouse, S. (2023). Artificial intelligence and art: Identifying the aesthetic judgment factors that distinguish human- and machine-generated artwork. *Psychology of Aesthetics, Creativity, and the Arts.* https://doi.org/10.1037/aca0000570

Times New Roman. (n.d.). *Despre noi*. Retrieved June 12, 2025, from https://www.timesnewroman.ro/despre-noi

Van Hees, J., Grootswagers, T., Quek, G., & Varlet, M. (2025). Human perception of art in the age of artificial intelligence. *Frontiers in Psychology, 15*, 1497469. https://doi.org/10.3389/fpsyg.2024.1497469

West, R., & Horvitz, E. (2019, July). Reverse-engineering satire, or “paper on computational humor accepted despite making serious advances”. In *Proceedings of the AAAI conference on artificial intelligence*, 33(1), (pp. 7265-7272).

Zhou, Y., & Kawabata, H. (2023). Eyes can tell: Assessment of implicit attitudes toward AI art. *i-Perception, 14*(6). https://doi.org/10.1177/20416695231209846