Group 9

People are skeptical of headlines labelled as AI-generated, even if true or human-made, because they assume full AI automation

Paper: https://academic.oup.com/pnasnexus/article/3/10/pgae403/7795946

Paper Overview

The paper is recent, having been published in 2024. It was written by Sacha Altay and Fabrizio Gilardi. It explores how “AI-generated” labels on news headlines change people’s perception of how accurate the news article is and their willingness to share the article.

Study 1: Each person saw 16 headlines (true/false × human/AI). Participants were randomly placed in labeling environments: no labels, all AI labeled, some AI missing, noisy (some human mislabeled as AI), or false (all false tagged “False”). They rated accuracy or share intent (on a scale of 1-6). The goals of this experiment included: Test if labeling headlines as “AI-generated” lowers trust, compare AI labels to False labels, Measure broader effects of labeling on news trust, Check for spillover effects.

Study 2: People saw 10 headlines. The labels came with definitions of what the AI generated label meant. The goal was to understand how people interpret the term “AI-generated” and to see how adding a definition to the label would affect people's willingness to share + impact their perceived accuracy of the article.

Key findings

Labelling an article's headline as AI generated reduced a participant's perceived accuracy of the article and also reduced their likelihood of sharing the article. It did not matter whether the article headline was AI generated, or human made, and it did not matter whether the article headline was true or false. Regardless of the origin and veracity, the AI-generated label reduced trust in the headline.

The study showed that participants did not necessarily equate “AI-generated” with being false. However, it was observed that when something was labeled as AI-generated, they were more suspicious of it.

- The second study explored why people trusted AI-labeled headlines less. They found that when given explanation, the negative effect decreased.
- People often assume full AI control unless told otherwise.
- Recommendation: Focusing on labeling harmful content rather than just labelling as AI generated.

Key takeaways

In Study 1, the participants were not given a clear explanation of what “AI generated” means, it was left up to the participants' own interpretation. We found it interesting how this was done to reflect how many social media platforms will give content a generic “AI-generated” label, leaving their users to interpret what that means for themselves. We felt that this made the study more reflective of real-world context, where we are left with these ambiguous definitions, which can cause different perceptions from user to user.

If an unlabelled headline is human-written, people might think it is AI-generated due to writing style, like using an em-dash. Could this influence perception more than the label?

Labeling content as “AI-generated” can unintentionally reinforce existing skepticism or negative perceptions toward AI, regardless of the content’s actual quality or origin. We found it particularly interesting that the study explored how participants’ attitudes shifted after learning more about the actual role and limitations of AI in the content creation process, highlighting the importance of transparency and education in shaping public perception.

What We Would Do Differently / Future Research

- Measure actual behaviour (clicks, reading time, comments) instead of only self-reported accuracy and sharing intention.
- Ask about trust in the source (publisher/author), not just the headline.
- Capture emotional reactions (surprise, anger, curiosity) to see how labels affect feelings.
- Include sharing rationale questions (e.g., “I share to inform others” vs. “I avoid uncertain content”).
- Test multiple AI models, not just ChatGPT 3, to see if reactions are different by model.
- Different demographics (age, education, cultural background): makes the study more diverse.
- Study the role of writing style or topic. People might mistake a human headline for AI based on style (like using em-dash)?

Our Reflections & Critical Takeaways

- Bias disclaimer: Acknowledging our bias, our perspective as AIML students means we might be assuming too much about AI education of the public.
- Even between the 4 of us who have similarities (educational background, age, etc.) there was disagreement about how best to handle AI generated content and

it's labelling. We can gather from this how difficult it is to label AI content in such a way that appeals to a way bigger demographic (who are more diverse).

- The study asked participants: "Outside of this study, when a headline is labeled 'AI-generated,' what do you think that means?" They rated several descriptions, and the results showed that most people believe AI does everything. From our perspective, we agree that when a headline says "AI-generated," most people including us assume AI handled most of the work, though we still expect some level of human supervision.
- Why are people less willing to share news? We think because of privacy and judgement fears, which make users cautious about sharing any news, AI labeled or not, because they worry about being judged or spreading fake news.
- We found it interesting and expected that human-written headlines labeled as "AI-generated" experienced the same drop as actual AI-generated headlines. This clearly shows that it's the label, not the content itself, that causes skepticism. If the label alone makes people doubt a headline, regardless of its true origin, then does labeling actually help?
- We agree that the effect likely generalizes beyond headlines to full articles.
- Transparency trade-off: Explaining AI's role helps, but too much detail can overwhelm users and make it more confusing, but oversimplifying might also not be super transparent. We suggest a layered approach (simple summary + optional technical details).
- Since "False" labels have a three-times larger effect than "AI-generated" labels platforms should focus more on harmfulness than origin.
- As the paper mentions, the results of this study might change overtime. For example, if AI becomes common, skepticism might lessen.
- Generic "AI-generated" labels reinforce negative perceptions, even if the content is high-quality.
- Study 2's approach of adding clear definitions showed how crucial label clarity is for public trust.
- The real-world mimicking of ambiguous labels made the experiments feel relevant to current social-media practices.

Broader context

As AI tools become more integrated in journalism, society needs to recognise that "AI-generated" is not a binary label. It consists of a spectrum of uses, including polishing grammar, writing entirely new stories, writing the first draft, only writing the headline, polishing up a final draft to match the editorial style of a certain publisher, etc. We can learn that there is more nuance to what "AI-generated" means; it is not so black and white. Without this explanation, vague labelling can mislead readers and unfairly harm journalists and writers. Which also begs the question, who is responsible for the harm a false positive creates? Does a false positive do more harm than a correct one does

good? Although the intentions behind AI flagging are well-meaning, it may confuse and affect the trust between journalists and readers.

Given that the study’s median age was 40, future research could explore younger people's perception of AI flagging. It is also worth considering how authors and journalists might respond to this research, and the need for them to have a voice in how functional it is to continue this flagging system.

While transparency is essential, the research shows an imbalance in AI’s perception. To move forward, we need to define AI’s role more clearly, ensure labels include context, and consider the potential harm of labelling content.