AI in the News: The Next Step in Journalism or the Source of its Demise?

Over the past few years, Artificial Intelligence has made its way into countless areas of society, many of which involve the average individual. Widespread adoption of such an impactful technology has not been seen since the birth of the smartphone and can even be compared to the rise of the internet itself; both of which have undoubtedly had an influence on how information is dispersed. Journalism is highly dependent on the available methods of news reporting, with organizations being forced to adapt with new technology or lose their audience. Just as the printing press once eliminated any hand-written papers, which then was replaced by the radio, then television, the internet, mobile computing, and now AI; many believe that Artificial Intelligence is the next step in the cycle [1]. AI is already in use by journalists today, with tools for article summarization, audio transcription, and comment moderation being some of the most popular [2]. While there is nothing inherently wrong with these tools, they make a connection between AI and journalism which may be exploited in the future, specifically on the end of the consumer.

The main factor driving each of the above-listed news innovations is convenience. Consumers will gravitate towards cheap and easy-to-obtain sources of news, pushing publications to produce ‘convenient’ news to meet the demand [3]. For many, interaction with AI will become inevitable somewhere in the news cycle, whether it be through tools provided by the publisher, article summarization requests in the user’s own model, or even direct queries to chatbots to “tell me what important happened today” which exclude the publisher entirely. The latter is the most dangerous, as many processes behind an AI model’s ‘black box’ directly violate the core principles of journalism. For instance, valid sources should be traceable back to an origin point; however, information given by an AI model that was learned through training is near impossible to trace, and even when feasible, the original sources are often unreliable due to being part of an unfiltered training dataset [4]. Similarly, AI models can suffer from hallucination, which is the generation of false or misleading information as a result of lapses in training or training datasets [5].

These issues are just a few among many sources of inaccuracy that are present in Large Language Models (LLMs), which are the primary type of model used in most text-based AI tools. Both hallucination and the lack of data traceability in LLMs will be further explored in this article, along with an investigation into some of the specific AI tools used by news publishers and their perception by the public. Additionally, some possible solutions to the problem will be discussed, each centering around minimizing the impacts of AI in journalism rather than outlawing it entirely. Lastly, different ethical frameworks will be applied to help determine situations where the use of AI may be morally justified or unjustified.

Following OpenAI's release of ChatGPT to the public in December 2022, many news outlets and publications began integrating it (and other tools powered by it) into their daily workflow. Several of these tools do genuinely have a positive impact in the journalists day-to-day work; for instance, automatic transcription of audio-based material and podcast generation from written material makes the journalist's content more accessible to a border audience. AI can also help automate repetitive and time consuming tasks, allowing journalists to devote more time to important work and engage with the story in a much deeper manner than they would be able to otherwise [1]. Many journalists use AI for research, data analysis, and writing as well, which while not always necessarily an ill-advised decision, these tasks require a higher standard of accuracy and reliability from the AI outputs. Current models cannot guarantee the level of precision needed to comfortably implement them into these tasks.

The incorporation of Artificial Intelligence into the news cycle has presented many problems regarding the validity and ethicality of the content being published. The use of AI algorithms in the selection and prioritization of news stories raises critical concerns due to bias and transparency. The inner-workings of these algorithms are often shrouded, allowing biases present from their training data to be perpetuated into the content they produce and interact with [1, p. 7]. These biases can also lead to echo-chambers, which occur when a user is only presented with content that aligns with their existing beliefs [1, p. 7]. AI-generated content can even be used by bad actors in a deceptive manner to impersonate another individual or source (commonly known as deepfakes). Even in a perfect-world scenario where unbiased training data is used, algorithms properly distribute content, and tools are used in good-faith only, two unavoidable issues remain that plague the credibility of AI journalism: hallucination and a lack of traceability.

Hallucination Invalidates the Credibility of AI-Generated Output

AI hallucination is when a model generates information that is factually incorrect, misleading, fabricated, or nonsensical. These false outputs are often difficult to detect due to the degree of confidence they are presented with and their appearance next to factually correct information [6]. Hallucinations can materialize in various different ways, ranging from minor inaccuracies to entirely fake citations or sources [6]. The phenomena of hallucination is not necessarily an issue in itself as much as it is the result of numerous other compounded issues. Some of these include limitations and overfitting to training data, overgeneralizations based on patterns, a lack of real-world grounding, and epistemic and aleatoric uncertainties [7].

Large Language Models are usually trained on massive datasets scraped from the internet, which while being the most convenient way of obtaining the large amounts of information needed, it does not allow for the data to be properly filtered and cleaned. This is what causes the previously-mentioned biases which can be reproduced in the model's outputs. Training datasets, regardless of their size or quality, can be incomplete or suffer from gaps; when the model is prompted with a query that it has insufficient or no knowledge of, it may 'fill in the gaps' by generating plausible-sounding but entirely fabricated information [6], [8]. Models can also develop an over-reliance on language priors, where they learn to predict what sounds plausible based on patterns in the training data. This is a form of overfitting, which can occur when a model learns the training data too well to the point where it gets confused by new unseen data [7]. The following example displays a case of hallucination caused by learned patterns from training data:

> If an AI-powered journalism tool was tasked with writing an article about a mayor's speech to the public about the opening of a new park, it might generate something along the lines of "*Mayor John Doe's decision to approve funding for the park was met with thunderous applause from the attendees*," even when there was no mention of clapping or unanimous approval. This is due to applause being commonly associated with positive announcements, leading the model to also assume that the decision was unanimously embraced by the public.

While most modern LLMs are intelligent enough to avoid simple overgeneralizations such as the example above, it still demonstrates the idea that AI does not just report the provided facts, it reproduces common patterns and details from its training set that it expects to see in the given type of story. This is quite problematic in the context of journalism, where reporting information exactly as it is from the source is of utmost importance.

Another key factor linked to hallucination in LLMs is their lack of real-world grounding, which is the idea that the models do not possess a true, real-world understanding of the concepts that they process. Humans anchor knowledge to logic, sensory experiences, and external reality; in contrast, AI models generate text based solely on probability [5]. While this statistical process of predicting the next word is effective in creating plausible-sounding text, it is disconnected from validation against the real world [6]. These uncertainties fall under two general categories: epistemic and aleatoric. Epistemic uncertainty is related to the model, and reflects the unpredictability in the model's weights, parameters, and architecture; aleatoric uncertainty refers to the randomness within the data itself [7]. The main difference between the two is that epistemic uncertainty is preventable with better model design and a more exhaustive data-collection process, whereas aleatoric uncertainty is largely inevitable regardless of the quality or quantity of data collected [7]. A higher level of epistemic or aleatoric uncertainty typically corresponds to a higher chance of hallucination. These uncertainties and lack of real-world understanding can cause AI models to struggle with comprehending context and nuance, both of which are vital in news reporting. One particular instance of this occurred when Microsoft implemented AI-powered editors for its MSN website in 2020; the AI, lacking the subtle understanding of a human editor, incorrectly associated an image with a story about the popular music group Little Mix, resulting in a racially insensitive blunder [1]. This example highlights the failure of AI models to understand real-world context, especially in relation to socially and culturally sensitive topics.

LLMs Lack the Ability to Trace Data Back to its Source

Hallucination is not the only fundamental issue with Large Language Models; there exists a flaw in the design of modern AI architecture that inhibits its use as a credible tool in practices of journalism. Quite similarly to the previously discussed aleatoric uncertainty, this flaw, known as the black box problem, is unavoidable due to the use of massive neural networks that require vast amounts of training data. The black-box problem refers to the difficulty in understanding how advanced AI systems arrive at their decisions due to the opacity in their internal workings. This lack of transparency and interpretability creates challenges in tracing information in the AI's responses back to an exact training source [9]. A misconception that is commonly associated with the black-box problem is the idea that the model architecture itself is hidden or difficult to understand. This is largely untrue; AI scientists and engineers thoroughly research and test the exact ordering of each layer to ensure that the model performs at its absolute best. In fact, this level of attention-to-detail and focus on even the smallest elements when designing an AI system is what gives some models a slight edge over others from competing companies [9]. Even with a satisfactory understanding of the network architecture, it is nearly impossible for even the model's creators to follow the reasoning behind a specific decision back through the entire neural network back to an origin [9], [10]. Many of the issues that lead to hallucination also can be attributed to the black box problem, such as the use of uncleaned and undisclosed datasets, which can result in the model learning patterns from the huge corpus of data rather than storing specific sources that can be referenced later [11]. Additionally, because AI models operate on mathematical probabilities to make decisions, what might seem like the 'obvious path' for the network to take may differ from the path that it actually selects [9], [11].

In journalism, credibility is built on verifiable sourcing, rigor, and accountability [11]. The lack of traceability within the AI black box directly undermines this core value. LLMs excel at making genuine-sounding claims, but often lack the ability to back up their claims with evidence sourced from a credible point of origin [12]. Additionally, many AI journalism tools are fine-tuned to have higher confidence and fluency to mimic human journalists, which only increases the difficulty of being detected by readers [6]. One common consequence of indiscernible AI content is a feedback loop with other artificial content. For instance, when a journalist attempts to verify the authenticity of AI-generated information, they might find other

AI-generated content that backs up the same falsehood [10]. The black box can also lead to gaps in accountability whenever misinformation occurs; if an AI model makes a harmful decision, it is often unclear who should hold responsiblity—the developer, the journalist using the tool, or the news publication [11]. The continued use of AI in journalism breaks the principle of being able to trace claims back to credible sources, and threatens the legitimacy and trustworthiness of the entire journalism industry.

While there are currently no methods of eliminating the black box problem entirely, some solutions have been proposed with the goal of improving data traceability. These fall into two general categories: technical safeguards, which focus on making the algorithms more explainable, and organizational governance, which focuses on the aspect of human accountability. Technical safeguards center around the idea of Explainable AI (XAI) and aim to provide clear reasoning behind model decisions [6]. Techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are used to interpret how many specific data features contributed to an AI's decision or prediction [9]. Another critical approach to grounding AI outputs in verifiable fact is Retrieval Augmented Generation (RAG). RAG integrates external sources, such as uploaded documents, databases, or even real-time internet access into the LLM, which decreases the model's reliance on its training data [6]. This technique has been shown to reduce hallucinations, provide better data traceability, and significantly increase overall accuracy in domain-specific applications [6]. Technical fixes can be supplemented with organizational and governance frameworks to better ensure ethical use. Human-in-the-loop oversight is a core ethical safeguard that certifies that editorial decisions, especially those with widespread societal impact, remain firmly under the control of qualified individuals [11]. Journalists must always retain the authority to reject, revise, or recontextualize any AI-generated content. Proactive disclosure policies, such as the European Union's AI Act, impose requirements for content transparency and explicit labeling of AI-generated information [11]. This also allows for accountability to be more effectively placed on the correct individual. Within the context of journalism, readers tend to view AI-written news more favorably when they perceive more human-like features, suggesting that readers will gravitate towards content that better conceals its AI authorship [8]. News organizations should not fall into this trap, and always prioritize transparency of where an article came from and who wrote the article; it only takes one breach of trust for a publication to permanently damage their reputation.

Ethical Analysis of Incorporating AI into Journalism

The use of Artificial Intelligence in the news cycle presents significant ethical dilemmas due to its violation of the key journalistic principle of perceived credibility. The integration of AI is primarily viewed as unethical; however, when looking through the lens of different ethical frameworks (such as Social Contract Theory, Virtue Ethics, and Utilitarianism), the reasoning for its immorality may shift or even disappear entirely.

Social Contract Theory states that morality consists of a set of rules that govern how people treat each other; rational people will agree to accept them on the condition that others will follow them as well for the mutual benefit of everyone [13]. In a democratic society, journalism operates under an implicit contract to provide accurate and truthful information in exchange for public discourse. AI breaches this contract with its violation of accuracy through hallucination, its violation of transparency through model opacity, and its violation of accountability through the lack of data traceability [11], [12]. The condition that there must be mutual benefit is not met either, as many consumers of news will suffer as a result of the increased quantity of misinformation.

Virtue Ethics centers around the intent of an individual, and posits that an action is right if a virtuous person, acting in character, would do under the same circumstances [13]. The irresponsible deployment of AI into the field of journalism demonstrates a failure to uphold moral virtues. Additionally, the foreknowledge that these models can generate misleading or even entirely false information violates the virtue of honesty. Over-reliance on AI risks the deterioration of journalist's skills by shifting their role away from critical investigators to mere model validators [11]. The prioritization of convenience over commitment to truth undermines the foundational moral virtues required of a professional news organization.

Utilitarianism follows the theory that an action is good if its benefits exceed its harms, and an action is bad if its harms exceed its benefits [13]. Under Rule Utilitarianism, where actions are assessed in their conformity to a general rule, the deployment of AI systems is generally unethical due to the widespread distribution of misinformation and erosion of public trust. However, under Act Utilitarianism, which focuses on examining individual acts, an argument can be made for the use of AI in journalism. AI can automate routine jobs, allowing human journalists to focus on more complex tasks and news organizations to increase coverage volume and speed. Some studies also suggest that readers perceive AI-written news as more objective, which can be seen as an overall benefit [8].

The incorporation of Artificial Intelligence and Large Language Models into the news cycle presents a massive technological shift, forcing journalists and publications to adapt. The fundamental design limitations of these models, particularly the phenomena of hallucination and the black box problem (which makes tracing data back to an original source difficult) invalidate the credibility of the resulting content. When applying ethical frameworks to the scenario, AI's use in journalistic settings is deemed to be generally unethical, as it violates the public contract for accuracy and accountability while undermining the moral virtues of professional journalism. AI offers many utilities and automations that do have genuinely positive impacts; for this reason, an approach of careful management should be taken rather than outright banning its use. Efforts to mitigate hallucination and model opacity such as Retrieval Augmented Generation have seen promising levels of success and should be explored further in the future, especially within the field of journalism [8].

Reference List

[1] P. N. Amponsah and A. M. Atianashie, “Navigating the New Frontier: A Comprehensive Review of Ai in journalism,” *Advances in Journalism and Communication*, vol. 12, no. 01, pp. 1–17, Mar. 2024. doi:10.4236/ajc.2024.121001

[2] T. Ryan-Mosley, “How generative AI is boosting the spread of disinformation and Propaganda,” MIT Technology Review, https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/ (accessed Sep. 14, 2025).

[3] A. Pilolli, “The Future of News: Navigating the changing digital landscape,” Local Media Association + Local Media Foundation, https://localmedia.org/2024/11/the-future-of-news-navigating-the-changing-digital-landscape/ (accessed Sep. 14, 2025).

[4] S. Lee, S. Nah, D. S. Chung, and J. Kim, “Predicting AI news credibility: Communicative or social capital or both?,” *Communication Studies*, vol. 71, no. 3, pp. 428–447, May 2020. doi:10.1080/10510974.2020.1779769

[5] The Science Desk Authors, “Ai hallucinations: Causes, risks, and fixes,” Science News Today, https://www.sciencenewstoday.org/ai-hallucinations-causes-risks-and-fixes#google_vignette (accessed Sep. 14, 2025).

[6] S. Joshi, “Comprehensive Review of AI hallucinations: Impacts and mitigation strategies for financial and business applications,” *International Journal of Computer Applications Technology and Research*, vol. 14, no. 6, pp. 38–50, May 2025. doi:10.7753/ijcatr1406.1003

[7] Y. Xiao and W. Y. Wang, “On hallucination and predictive uncertainty in conditional language generation,” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744, Mar. 2021. doi:10.18653/v1/2021.eacl-main.236

[8] D. C. Lee, J. Jhang, and T. H. Baek, “AI-generated news content: The impact of AI writer identity and perceived AI human-likeness,” *International Journal of Human–Computer Interaction*, pp. 1–13, Mar. 2025. doi:10.1080/10447318.2025.2477739

[9] S. Clark, “The AI Black Box: The hidden risk behind every algorithmic decision,” VKTR.com, https://www.vktr.com/digital-experience/cracking-the-ai-black-box-can-we-ever-truly-understand-ais-decisions/ (accessed Sep. 30, 2025).

[10] R. Thomas, “Vetting sources: Containing the spread of misinformation in AI-generated content,” nDash.com, https://www.ndash.com/blog/vetting-sources-containing-the-spread-of-misinformation-in-ai-generated-content (accessed Sep. 30, 2025).

[11] S. Olufemi Olanipekun and O. Olakoyenikan, “Ethical implications of generative AI in journalism: Balancing innovation, truth, and Public Communication Trust,” *World Journal of Advanced Research and Reviews*, vol. 16, no. 3, pp. 1293–1311, Dec. 2022. doi:10.30574/wjarr.2022.16.3.1159

[12] Q. Liu, “Generative AI and journalism ethics: Controversies over Chatgpt,” *Journal of Information, Technology and Policy*, pp. 1–6, Apr. 2025. doi:10.62836/jitp.2025.346

[13] M. J. Quinn, *Ethics for the Information Age*, 9th ed. Hoboken, New Jersey: Pearson, 2025.