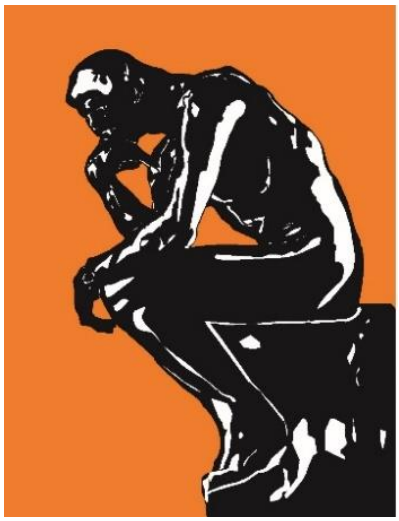
Artificial intelligence and retracted science

Minh-Hoang Nguyen [1]*, Quan-Hoang Vuong [1,2]

[1] Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam

[2] A.I. for Social Data Lab (AISDL), Hanoi, Vietnam

* Correspondence: hoang.nguyenminh@phenikaa-uni.edu.vn



August 19, 2024

[*Original working draft v2 / Un-peer-reviewed*]

> "[...] The report is still completely honest, trustworthy, and ethical, even though the data are fabricated and measurements are falsified."
>
> —In "GHG Emissions"; *The Kingfisher Story Collection* (2022)

Technology firms are now purchasing access to research papers from academic publishers to train their artificial intelligence (AI) models. It came to light that United Kingdom publisher Taylor & Francis signed a $10 million agreement with Microsoft last month, providing the company with access to the scientific contents of nearly 3000 academic journals to enhance its AI capabilities. Meanwhile, in June 2024, an investor report revealed that Wiley made $23 million by permitting an undisclosed company to use its content for training generative AI models [1].

Using scientific content to train AI can come with multiple benefits. Journal articles are generally precise, reliable, and well-structured due to the scientific publication standards and peer-review process, so they are high-quality data for improving the trustworthiness of the AI's generated outcomes. Moreover, the content in journal articles is written with highly specialized knowledge and terminology, which can help enhance the AI's capability of understanding and processing issues across a wide range of topics. The inherent logical reasoning of scientific content can also improve the AI's ability to analyze information, make logical deductions, and draw conclusions.

Journal articles are generally considered reliable because of the rigorous peer review system. However, the evaluation process is constrained by several limitations, with one of the most significant being subjectivity. Since editors and reviewers are human, they are inevitably influenced by personal biases, prejudices, and the limits of their own expertise. Moreover, many scientific conducts, like result manipulation, data fabrication, and falsification, cannot always be detected during the peer-review process. Some publishing systems are also manipulated to have unqualified studies published (e.g., through exploiting Special Issue publishing, author-suggested reviewers mechanism, and creating fake reviewer accounts). As a result, a number of unreliable and invalid studies have been published.

Retraction is often viewed as a crucial self-correction mechanism within science, allowing the academic community to identify and flag seriously flawed research that has been published. In recent years, there has been a significant surge in the number of retracted scientific articles, particularly following the COVID-19 pandemic. Even studies authored by Nobel laureates and published in highly prestigious journals such as *Nature*, *Science*, *PNAS*, *The New England Journal of Medicine*, and *The Lancet* have been found to contain serious flaws and subsequently retracted. Since the pandemic ended, the annual number of retractions has not decreased; instead, they continued to reach a new peak in 2023, with over 10,000 post-review papers retracted.

If retracted articles are used to train AI, it could lead to the spread of misinformation. Given the power of AI, this misinformation could be disseminated on a large scale and in a very short time, which might misinform and significantly increase the entropy (uncertainty) within society [2].

In reality, the likelihood that unreliable and invalid articles are used for training AI models is high. Both publishers that recently signed agreements with technology companies to grant access to their scientific content have faced a significant number of retractions in recent years. Specifically, the number of retractions for research articles in Taylor & Francis Group

journals sharply increased from 2017 to 2022, with over 350 papers retracted in 2022 alone. For Wiley, the situation is even more severe. Its subsidiary, Hindawi, had to retract over 8,000 articles in 2023 due to concerns about the compromised integrity of the peer review process and systematic manipulation of the publication and review procedures.

When a published paper is retracted, it is not removed from the literature but rather marked as retracted, signaling that its results and conclusions are no longer reliable and should not be cited or reused. Retractions are typically accompanied by a notice from the editors or authors explaining the reasons for the retraction. These transparent retraction notices act as historical records, preventing future studies from replicating or building upon flawed results and helping to uphold the integrity of the scientific record [3]. Therefore, AI developers can use legitimate information about these retractions to exclude seriously flawed articles from the training process.

Nevertheless, once a paper has been used as training data for a model, it cannot be removed from the model's knowledge base after the training is complete. Then, if the scientific content used to train AI is later discovered to be seriously distorted (e.g., data manipulation, fraudulent data, unsupported conclusions, questionable data validity, non-replicability, or data errors), what can be done to address the problem? Are there any mechanisms or principles that AI developers and managers can apply in this situation?

If so, how much human effort would be required to counteract the chaos caused by misinformation stemming from AI trained on retracted scientific studies?

As AI is increasingly applied in critical areas such as education and decision-making, training AI on seriously flawed scientific findings could lead to significant systematic consequences. Therefore, collaboration between the scientific community and AI developers to address these issues is crucial.

References

[1] Gibney E. (2024). Has your paper been used to train an AI model? Almost certainly. *Nature*. https://www.nature.com/articles/d41586-024-02599-9

[2] Vuong QH, Nguyen MH. (2024). Further on informational quanta, interactions, and entropy under the granular view of value formation. https://philpapers.org/rec/VUOARN

[3] Vuong QH. (2020). Reform retractions to make them more transparent. *Nature*, 582, 149. https://www.nature.com/articles/d41586-020-01694-x