Advances in Health Sciences Education (2023) 28:659–664
https://doi.org/10.1007/s10459-023-10257-4

EDITORIAL

Artificial scholarship: LLMs in health professions education research

Rachel H. Ellaway[1] · **Martin Tolsgaard**[2]

Published online: 19 June 2023

Abstract
This editorial examines the implications of artificial intelligence (AI), specifically large language models (LLMs) such as ChatGPT, on the authorship and authority of academic papers, and the potential ethical concerns and challenges in health professions education (HPE).

Who wrote this editorial? There are two of us named as authors, but how can you tell that this paper wasn't written by a some kind of artificial intelligence (AI)? Or perhaps an AI was an undisclosed third author, what then? Or what if this paper has been edited by multiple AIs between the time we wrote it and it arrived in front of you? What authority does this work have if you cannot tell humans from machines, or you cannot tell when an academic paper is a hybrid of human and AI generated material? Who is accountable for such work? What are the implications for scholarship? These are the starting points for this editorial.

Concerns

Setting aside the substantial research and development into AI, artificial intelligence has long been a cultural meme, almost always portrayed in terms of a threat (such as HAL, The Matrix, various Terminators, and Ultron, to name but a few). Not surprisingly then, now that we are beginning to encounter very real artificial intelligences, opinions are divided as to whether they are useful technologies or existential threats. But what do we mean by 'AI'? Currently, much of the conversation refers to ChatGPT, which is currently the best known of a family of tools known as large language models (LLMs); others include Google

✉ Rachel H. Ellaway
rellaway@gmail.com

[1] Department of Community Health Sciences and Office of Health and Medical Education Scholarship, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[2] University of Copenhagen, Copenhagen Academy for Medical Education and Simulation (CAMES), Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark

Bard and Microsoft Bing. These LLMs can process complex textual inputs and provide textual responses that can be hard to distinguish from those generated by humans. Such is the power and capacity of these emerging LLMs that Springer Nature (our publisher) has banned LLMs from being listed as authors on articles it publishes:

> *"First, no LLM tool will be accepted as a credited author on a research paper. That is because any attribution of authorship carries with it the expectation of accountability for the work, and AIs as machines cannot have such responsibility. Second, researchers using LLM tools should document this use in the methods or acknowledgements sections."* (Anon 2023)

Indeed, we already seem to be in an arms race between those using LLMs to generate academic papers and publishers developing LLM detection tools to identify and presumably stop non-human authorship. Why is it important to distinguish between AI-augmented scholars and those who are not using AI? Assuming that the existing LLMs are truly generative (which admittedly may be a stretch since they often reproduce what has already been written, albeit in new configurations), is the science produced with their help necessarily inferior for this reason alone? Similar concerns were voiced in previous eras (for example, when introducing the radio, television, or the Internet (Rosen et al., 1987). Put another way, if an LLM can produce a paper that is better written, less biased, more concise, and more accessible than a human can then is that not a superior product? Of course, this separates the 'how?' of writing from the 'what?', an LLM might be able to produce a well-written paper but there is still the question as to what is being written about and where the information and ideas come from. It is a fundamental tenet of the ICMJE's recommendations on academic authorship that all named authors on a paper have made:

> *"Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND drafting the work or revising it critically for important intellectual content; AND final approval of the version to be published; AND agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved."* (ICMJE nd)

This would require an LLM to have co-designed a study, been intimately part of its execution, been intimately involved in writing the paper, approved it, and stood accountable for it to be eligible to be an author. Contributing to the writing alone is not sufficient. AHSE fully supports the ICMJE recommendations and these firmly centre on the capabilities of human authors. Please also see this issue's Questions and Queries article on 'Who Should Be an Author on This Paper?' (Kuper et al. THIS ISSUE).

These are not the only issues being raised in the many position papers and perspectives being published including those in our own Springer Nature imprint (see https://www.nature.com/search?q=chatgpt&order=relevance) For instance, Seghier noted equity and cultural concerns of a technology dominated by a single language (English) and culture (US):

> *"French and Arabic are among the world's most commonly spoken languages and they have a widespread presence on the Internet. However, the richness of ChatGPT's response and the intelligibility of its writing in both languages were notably inferior to those in English."* (Seghier 2023)

Potential benefits

Is this 'rise of the machines' to be feared and resisted? Well, not neccessarily, clearly LLMs can play useful roles. For instance, given that the medical education literature is dominated by contributions from English-speaking countries and those competent in English as a second language, language already creates a barrier that puts non-native English speakers at a disadvantage. In this context, LLMs could be used to translate and correct manuscripts in ways that could reduce language barriers, thereby allowing scholarly work from non-native English-speaking countries to be considered on a more equal footing. Of course, there are also many native English speakers whose writing could benefit from this kind of copy-editing. Indeed, we have colleagues who are already using LLMs to proof their written work prior to submission. If you think this is not appropriate, then ask yourself how different this is from using Grammarly or a spell checker. Answering these kinds of questions comes down to how much is being changed or 'improved' by the LLM. For example, in the AI community, several high-profile conferences have prohibited the use of LLMs in generating work submitted to them, but they allow its use in refining human-generated content. (Vincent, 2023)

Writing support is not the only potential role for LLMs in health professions education research. LLMs could be used to gather information, for instance for a literature review, or conduct a rapid synthesis of a body of information. What about getting LLMs to do the work of librarians in support of systematic reviews? This might be faster and cheaper than engaging humans to do this work but the impacts of removing more and more humans from research processes may be unpredictable at best. There are ethical lines that will need to be drawn (and not crossed), but quite where these lines are and how and when they shift according to context remains unclear.

Ethics and Integrity in Academic Work

It is important to note that those opposing the use of AI in academic writing are not seeking to eschew all technological supports. Many scholars use tools such as Grammarly to help them in their writing, as well as using reference managers, databases, and a cornucopia of other Web services and resources. Indeed, this editorial was written on the authors' laptops using Microsoft Word, we exchanged drafts between Calgary and Copenhagen via email, and we submitted and processed this manuscript using Springer's tools and technologies. If these are acceptable augmentations, then how do we draw a meaningful boundary between accepted and non-accepted AI-augmentation? How much technology use or dependence can there be (or will we accept) without losing professional integrity and accountability?

After all, as much as these technologies could be misused, they can also be very useful, for instance in generating ideas, getting feedback on drafts of text, revising language, providing structure, or completing repetitive tasks. Moreover, although not yet advisable, AI might eventually be used to screen articles submitted to journals such as ours and even to substitute for reviewers when human reviewers cannot be secured. It might even be the case that some journals might someday be run almost entirely by AIs. Processing times would be cut drastically if this were the case, but how reliable and fair would this be? After all, Krügel et al. (2023) raised moral concerns that LLMs have no morality and yet can be very influential of humans who do (presumably) have this capability.

As another example, researchers at the computer lab at Berkeley, University of California, demonstrated shortly after the release of ChatGPT that the AI model exhibited bias against non-male and non-white scientists when tasked with creating a script to identify competent scientists (Alba, 2022). It is axiomatic that all such technologies will reflect their creators' biases and potentially that in the materials they train on such that their impartiality is hollowed out by structural bias. It is also concerning that most LLMs are not fully understood by their makers as they use adaptive deep reinforcement learning rather than explicit content algorithms (the latter having proved too onerous and underpowered to compete with ADRL).

These problems notwithstanding, the rush to employ these emerging technologies is clear. Even as we were preparing this editorial, we have been bombarded by advertisements for courses teaching scholars to use LLMs in writing and research. As an example, an online course from Steel and Fariborzi (2023) offered to teach scholars to use LLMs to automate literature reviews (conducting searches, assessing the relevance and quality of papers, and analyzing and synthesizing the results of searches), to be a part of authoring papers (by synthesizing the literature, drafting papers, preparing them for submission, and responding to reviewers and editors), and grant writing by using LLMs to "synthesize complex research goals into compelling and digestible content for potential national and other funding bodies". Ready or not, the genie is out of the bottle.

What happens if LLMs are tasked with gathering and synthesizing data and writing papers on their findings? What happens if they can do this at a much greater rate and, stylistically at least, of a much higher quality than human researchers or authors can manage? What happens if LLMs on the journals' side are reviewing and accepting and rejecting these LLM-produced papers? What happens if LLMs are editing and revising papers, for other LLMs to publish, so that yet other LLMs consume them? If all of the tasks of academic publishing can be accomplished without human input, then what? While this may seem rather improbable, it is by no means impossible.

However, rather than descending into dystopian anxiety over the academic version of Skynet, perhaps we should ask what it is humans can and do add that machines cannot or should not seek to emulate. We would suggest for instance that human wisdom, creativity, and judgment might be simulated but not emulated, that ethically minds that have both human emotional strengths and fragilities are required to do the work we do, and that accountability to each other is a nonnegotiable part of academic activity. Since humans and AI excel in different areas, the question may not be whether AI will replace us, but rather, how can we collaborate with AI to achieve results that were not possible for either AI or humans alone.

Implications for HPE research and publishing

The potential impact of LLMs and other AIs on research, practically, ethically, and morally is large, and this has created a reappraisal of much of academic processes and standards. For instance, Van Dis et al. (2023) asked a set of essential questions for researchers and publishers including: which research tasks should or should not be outsourced, what steps in an AI-assisted research process require human verification, and how should LLMs be incorporated into the education and training of researchers?

So far, the response from journals has been cautious and critical. As an example, the editors of Academic Medicine recently (DeVilbiss & Roberts, 2023) stated their position in terms of four broad principles: that authors must be accountable for their work (noting that LLMs do not meet this standard); that any use of LLMs in developing or writing papers submitted to the journal must be disclosed; that any use of LLMs in the research process must be clearly described in any paper reporting on that research; and, as these technologies are changing so quickly, that policies will need to adapt over time to track these changes. These are somewhat generic concerns, albeit ones that we support at Advances.

There are bigger issues to consider. For instance, what implications do LLMs have for our use of theory, the conceptualization of assessment practices, learning and performance research etc.? In a field (assuming that HPE is a field) where many of the most cited and downloaded publications in the major journals consist of non-empirical work, commentaries, letters, and reflective pieces, will the use of LLMs further water down the few original research contributions that remain published? Will the scientific currency and integrity of our work be diminished if anyone can produce an LLM-generated review within minutes?

There are also the interfaces between HPE scholarship and HPE to be considered, as well as our interfaces with healthcare as a whole. AIs are already finding their way into medicine (Li et al., 2019) and into medical education (Tolsgaard et al., 2020; Katznelson & Gerke, 2021). Indeed, we are seeing a great many speculative opinion pieces being submitted to AHSE and other journals about the importance, opportunities, risks, and other consequences of AIs. As a journal this is an interesting issue, but we look for substantial theoretical and philosophical or empirical work, not speculation. In many ways, we predict a new wave of atheoretical and non-empirical work similar to that we saw during the COVID pandemic but now focused on the hopes, expectations, and dangers of LLMs and AI technology.

The problems we have noted are not trivial, indeed they may prove transformative and even disruptive. We encourage our scientific community therefore to think deeply about how concepts, theories, and practices are being shaped by LLMs and the implications thereof. We also ask that all work submitted to Advances meets the emerging disclosure standards for the use of LLMs, and that any work submitted that takes the use of LLMs as its subject is thoughtful and critical of the many issues involved and their implications for advances in health sciences education, both in theory and in practice.

References

Alba, D. (2022). OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. Bloomberg, December 8, 2022. https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results accessed May 5 2023.

Anon. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, *613*(7945), 612.

DeVilbiss, M. B., & Roberts, L. W. (2023). Artificial Intelligence Tools in Scholarly Publishing: Guidance for Academic Medicine Authors. *Academic Medicine*. https://doi.org/10.1097/ACM.0000000000005261.

ICMJE (Accessed May 5 2023). (nd.) Defining the Role of Authors and Contributors. International Committee of Medical Journal Editors. https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Katznelson, G., & Gerke, S. (2021). The need for health AI ethics in medical school education. *Adv in Health Sci Educ*, *26*, 1447–1458.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*. 13;4569.

Kuper, A., O'Sullivan, P., & Cleland, J. (2023). *Who should be an author on this paper?'*. THIS ISSUE – PLEASE COMPLETE).

Li, D., Kulasegaram, K., & Hodges, B. D. (2019). Why we needn't fear the Machines: Opportunities for Medicine in a machine Learning World. *Academic Medicine*, *94*(5), 623–625.

Rosen, L. D., Sears, D. C., & Weil, M. M. (1987). Computerphobia. Behavior Research Methods. *Instruments & Computers*, *19*(2), 167–179.

Seghier, M. L. (2023). ChatGPT: Not all languages are equal. *Nature*, *615*(7951), 216.

Steel, P., & Fariborzi, H. (2023). Using ChatGPT for Academic Publications and Grants. Online course offering from the Institute for Statistical and Data Science Pty. (Australia): https://instats.org/structured-course/chatgpt-for-academic-paper-writing2515 accessed 25th May 2023.

Tolsgaard, M. G., Boscardin, C. K., Park, Y. S., Cuddy, M. M., & Sebok-Syer, S. S. (2020). The role of data science and machine learning in Health Professions Education: Practical applications, theoretical contributions, and epistemic beliefs. *Adv in Health Sci Educ*, *25*, 1057–1086.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226.

Vincent, J. (2023). Top AI conference bans use of ChatGPT and AI language tools to write academic papers. The Verge, Jan 5, 2023: https://www.theverge.com/2023/1/5/23540291/chatgpt-ai-writing-tool-banned-writing-academic-icml-paper accessed 27 April 2023.