Commentary on Porsdam Mann and colleagues, “AUTOGEN: A Personalized Large Language Model for Academic Enhancement – Ethics and Proof of Principle”

Alexandre Erler
alexandre.erler@philosophy.oxon.org

This is the final, pre-publication draft of an Open Peer Commentary published in the American Journal of Bioethics*, 23:10 (2023), 94-96, DOI:* https://doi.org/10.1080/15265161.2023.2250291. *Please cite the published version.*

Publish with AUTOGEN or Perish? Some Pitfalls to Avoid in the Pursuit of Academic Enhancement Via Personalized Large Language Models

The potential of using personalized Large Language Models (LLMs) or “generative AI” (GenAI) to enhance productivity in academic research, as highlighted by Porsdam Mann and colleagues (Porsdam Mann et al., 2023), is of relevance not only to bioethics but virtually all academic fields. Here, I wish to elaborate on some of the authors’ remarks regarding the pitfalls that might result from the increasing use of GenAI for academic enhancement combined with the “publish or perish” imperative prevalent in academia today.

First, the authors mention the risk that perverse incentives might arise from such a combination. These incentives might lead researchers to produce ever more papers of barely publishable quality to boost their CVs and career prospects, resulting in a race to the bottom. How can such a phenomenon be discouraged? The solution does not seem to lie in an official prohibition on the use of GenAI for academic research, even one limited to the more “substantive” or original parts of a text. Indeed, any such prohibition is likely to prove very difficult to enforce, given the apparent inability of existing tools to reliably detect AI-generated text, especially given the possibility of slightly modifying or paraphrasing the AI’s output (Weber-Wulff et al., 2023). Furthermore, even novel and valuable content can conceivably be produced via a judicious mix of contributions from both humans and AI. A more constructive approach might simply lie in upholding proper standards of quality in research: that is, standards researchers can only meet by substantially contributing to the end result, and not by re-using a LLM’s response to prompts more or less *as is*. This will be the responsibility of editors and reviewers at respected journals and book publishers, but also of academics tasked with making hiring and promotion decisions, who will need to take special care not to unduly privilege quantity over quality when assessing a candidate’s output.

Secondly, if the potential of GenAI to enhance productivity does get realized, the question arises of how authors, journal editors and reviewers will deal with the consequent growth in article submissions and in the volume of published literature. Whether or not finding reviewers for papers submitted to academic journals has generally become more challenging is currently a disputed issue (compare for instance Flaherty, 2022, with Zupanc, 2023), yet there is little doubt that GenAI could make such a challenge very real. Unless the efficiency of the reviewing process can somehow be enhanced in proportion to the rise in submitted articles, journals are likely to face a growing backlog of submissions. A natural suggestion here is to pursue greater efficiency by integrating AI into the reviewing process itself. As it happens, this is already being done, to assist with the initial screening of submissions (Checco et al., 2021). In the future, the role of AI can be expanded further: for instance, reviewers will be able to use GenAI to write reviews more efficiently. While this is certainly a plausible development, it remains to be seen whether it will suffice to successfully cope with the growth in submissions, especially if the quality of peer review is not to suffer as a result. (We may suspect that it will if reviewers use AI to avoid having to read the articles they are reviewing in full, so as to complete more assignments within the same amount of time.)

Similar remarks apply to the challenge of keeping up with an ever more voluminous published literature. Here again, GenAI might be part of the solution, for example by providing researchers with reviews and summaries of the relevant readings. The question is how to accomplish this without hurting the quality of the final output. Already today, a researcher could seek to accelerate their literature review by only reading abstracts, rather than full articles, yet they would thereby be at risk of overlooking or misconstruing many important aspects of the writings that formed the basis of their discussion. Reliance on GenAI could present similar risks. Unlike, say, a postdoctoral researcher tasked with conducting a literature review for a co-authored paper, it is not clear that GenAI could, at least for the foreseeable future, properly rectify errors or omissions in the sections written by other collaborators who may not be as familiar with the current literature.

Thirdly and finally, Porsdam Mann and colleagues make an important point when they mention the concern that GenAI might have a "homogenizing effect" on the writing style of users – although this risk may have greater applicability to "general" than to personalized LLMs, and also to writings in some fields (such as the Arts and Humanities) than others (say, the hard sciences, in which clarity and accuracy may matter more than having a unique "voice").[1] That said, the risk that these tools might homogenize users' *thinking* might be even more significant. This risk stems from the combination of two factors. The first one is an environment in which a few Big Tech companies dominate the GenAI market, and provide tools that harbor hidden (and potentially overlapping) biases, which could have been inherited either from a company's ideology, or simply from the data the tool was trained on. As described by Porsdam Mann and colleagues, a researcher might use a tool like ChatGPT to suggest ideas for a journal article: for instance, salient ethical issues related to a contemporary biomedical or other technological development. Because of the company's values, or of dominant trends in the scientific literature on which the tool was trained (resulting in a form of "majority bias"; Nam et al., 2023), or both, the AI might tend to highlight a specific set of issues or avenues of thought as promising, and might present this same set to large numbers of researchers in response to their (similar) prompts.[2]

The second factor relevant to the risk of homogenization is the aforesaid pressure to publish, which can limit the opportunity for researchers to let their own ideas mature. If this pressure intensifies in the future as a result of the productivity enhancements fostered by GenAI, the incentive to turn to GenAI itself for "instant inspiration" on what to write about will grow in turn, in a self-reinforcing feedback loop. This would mean that LLMs could increasingly shape the kind of topics, and perhaps even views and arguments that academics will consider

[1] This may in turn imply a differential impact on highly interdisciplinary fields like Bioethics, which encompasses a wide variety of different approaches.

[2] I assume that even a fine-tuned LLM would need to have been trained mostly on writings other than the researcher's own, if it is to prove useful for idea generation. Also, it is not clear that personalized LLMs can help avoid the homogenization problem with regards to thought as they can in relation to writing styles. Faithfully replicating a user's reasoning and creative abilities seems significantly more challenging than merely replicating their way of writing. Achieving the former may entail crossing the key threshold of artificial general intelligence or AGI (McLean et al., 2023).

in their work, and those they will not. Even though different researchers can certainly develop similar suggestions from GenAI in different ways, and although intellectual bubbles clearly predate the advent of such tools, the aforesaid combination of factors nevertheless suggests that a degree of thought homogenization as a result of the use of GenAI for academic enhancement is a distinct possibility.

While I believe that such pitfalls deserve our attention, I do not mean to imply that they are an unavoidable consequence of the reliance on GenAI in academic research. Ultimately, it all depends on how exactly we use these new tools. It is certainly conceivable that GenAI's responses could be used not as a pre-set framework, but rather as stepping stones towards genuinely novel ideas that owed as much, or even more, to a researcher's own creativity, and to their discussions with others. GenAI could even be deliberately prompted to suggest topics or lines of thought that had been neglected in the literature, or that challenge current orthodoxy in promising ways. This would simply require an effort to avoid taking the quickest and easiest route to publication. Furthermore, even when homogenization of thought does occur, it need not always be a bad thing: suppose for instance that the AI's recommendations caused many researchers to focus their attention on an issue of paramount importance for humanity. Homogenization becomes especially problematic when it causes significant topics or ideas to get overlooked, and when these could have been identified if researchers had taken the time to ponder the issues and let their own ideas incubate, rather than letting GenAI set their research agenda for them.

In conclusion, the issues involved are complex and nuanced, and they do not justify renouncing the use of GenAI in academic research. Rather, they highlight the need to develop a clear set of guidelines for the responsible use of GenAI to boost academic productivity, a goal that can be achieved by continuing the discussion that Porsdman Mann and colleagues have initiated.
--

References:

1. Checco, A., Bracciale, L., Loreti, P., Pinfield, S. & Bianchi, G. 2021. AI-Assisted Peer Review. *Humanities & Social Sciences Communications,* 8.
2. Flaherty, C. 2022. The Peer-Review Crisis. *Inside Higher Ed* [Online]. Available: https://www.insidehighered.com/news/2022/06/13/peer-review-crisis-creates-problems-journals-and-scholars [Accessed 06/08/2023].
3. Mclean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A. & Salmon, P. M. 2023. The Risks Associated with Artificial General Intelligence: A Systematic Review. *Journal of Experimental & Theoretical Artificial Intelligence,* 35**,** 649-663.
4. Nam, J., Mo, S., Lee, J. & Shin, J. 2023. Breaking the Spurious Causality of Conditional Generation via Fairness Intervention with Corrective Sampling. *arXiv*.
5. Porsdam Mann, S., Earp, B. D., Møller, N., Vynn, S. & Savulescu, J. 2023. AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle. *American Journal of Bioethics*.
6. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. & Wadding, L. 2023. Testing of Detection Tools for AI-Generated Text. *arXiv*.
7. Zupanc, G. K. H. 2023. "It Is Becoming Increasingly Difficult to Find Reviewers" – Myths and Facts About Peer Review. *J Comp Physiol A*. https://doi.org/10.1007/s00359-023-01642-w.