Peer Response 1

Gavin’s post raises several important points regarding the use of AI writers, highlighting both their benefits and associated risks. One of the most compelling issues he addresses is the potential exposure of sensitive data when using AI tools. Building on this, it is important to consider the ethical implications of data sharing through AI systems. While individuals may often control how their personal data is used, in the context of AI-assisted administrative work, data is frequently shared on their behalf, removing their ability to decide how it is handled (Gomstyn and Jonker, 2024). This raises significant questions about whether AI users should be trained or sensitised to responsibly use these tools, particularly in scenarios where sensitive information is involved (Torkzadehmahani et al., 2020). Should AI writers even be permitted in departments managing highly confidential data? Gavin, do you think training users or sensitising them to these risks would be a feasible and effective way to address this issue?

Another crucial point Gavin raises is the risk of overreliance on AI, potentially leading to a decline in essential cognitive skills such as critical thinking, analysis, and creativity. This issue extends beyond the individual, with far-reaching societal implications (Gerlich, 2025). Gavin's concern about a population increasingly dependent on AI for decision-making struggling to adapt to novel challenges requiring deep human insight is supported by existing literature. The concept of Polanyi's Paradox highlights that much of human knowledge is tacit and not easily codified, posing challenges for AI systems to replicate human adaptability and common-sense reasoning (Muñoz, Mosey & Binks, 2015). Moreover, uncritical acceptance of AI-generated outputs could exacerbate

misinformation and stifle intellectual development (Monteith et al., 2023). Gavin's discussion underscores the importance of balancing AI use, promoting collaboration rather than dependency. His reflections highlight the need for responsible AI integration to mitigate its potential risks to individuals and society.

References:

Gerlich, R. (2025). *The impact of AI reliance on critical thinking: A cognitive offloading perspective*. MDPI. Available at: https://www.mdpi.com/2075-4698/15/1/6 (Accessed 21 January 2025)

Gomstyn, A. and Jonker, A. (2024). *Exploring privacy issues in the age of AI*. IBM Insights. Available at: https://www.ibm.com/think/insights/ai-privacy (Accessed 21 Jan. 2025)

Monteith, S., Glenn, T., Geddes, J.R., Whybrow, P.C., Achtyes, E. and Bauer, M. (2023). *Artificial intelligence and increasing misinformation*. The British Journal of Psychiatry. Available at: https://doi.org/10.1192/bjp.2023.62 (Accessed 21 January 2025)

Muñoz, C., Mosey, S. and Binks, M. (2015). *The tacit mystery: reconciling different approaches to tacit knowledge*. Knowledge Management Research & Practice, 13(3), pp.289–298. Available at: https://doi.org/10.1057/kmrp.2013.50 (Accessed 21 January 2025)

Torkzadehmahani, R., Nasirigerdeh, R., Blumenthal, D.B., Kacprowski, T., List, M., Matschinske, J., Späth, J., Wenke, N.K., Bihari, B., Frisch, T., Hartebrodt, A., Hausschild, A.C., Heider, D., Holzinger, A., Hötzendorfer, W., Kastelitz, M., Mayer, R., Nogales, C., Pustozerova, A., Röttger, R., Schmidt, H.H.H.W., Schwalber, A., Tschohl, C., Wohner, A. and Baumbach, J. (2020). *Privacy-preserving Artificial Intelligence Techniques in Biomedicine*. arXiv preprint. Available at: https://arxiv.org/abs/2007.11621 (Accessed 21 Jan. 2025)

Peer Response 2

As Guilherme aptly noted, while Large Language Models (LLMs) and Artificial Intelligence (AI) are trained on vast amounts of data, they are inherently at risk of replicating and even reinforcing biases related to race, gender, religion, or sexual orientation (Hutson, 2021). A striking example of this issue is the AI recruiting tool developed by Amazon in 2014.

The system was trained primarily on the company's internal data—resumes submitted over the previous decade. Given that the tech industry has historically been male dominated, the model was inadvertently trained on skewed data. This led to a significant bias: the AI disproportionately favoured male candidates, even rejecting resumes that included the word "woman". These biases resulted in discriminatory hiring practices (Dastin, 2018)

The implications of such bias extend beyond employment. If similar biases were to infiltrate critical areas like healthcare, the consequences could be far more severe, potentially jeopardizing lives rather than just career opportunities (Panch, Patel and McCourt, 2023).

The crucial question, therefore, is how such situations could have been prevented. Rigorous testing with gender-specific prompts during the development phase could have identified these issues early on (Sant et al., 2024). Additionally, it should have been

ensured that the model was trained on a more diverse and gender-balanced dataset, particularly since Amazon's internal data was known to predominantly reflect male applicants. Bias mitigation strategies, such as penalizing biased outputs, could also have been implemented to minimize the risk of discriminatory outcomes (Leavy et al., 2020).

In summary, while Large Language Models and Artificial Intelligence hold immense potential to improve lives, their ethical foundation must prioritize fairness and equity. Although this discussion did not specifically focus on AI writers, the solutions to prevent bias remain the same. This commitment to inclusivity should extend not only through their design phase but also during deployment and real-world application. Ultimately, AI should aim to dismantle systemic barriers and create opportunities that are accessible to all members of society. (Floridi & Cowls, 2019).

References:

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Available online at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. (Accessed 23 December 2024).

Floridi, L. and Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). Available online at: https://doi.org/10.1162/99608f92.8cd550d1 (Accessed 23 December 2024)

Hutson, M. (2021) 'Robo-writers: the rise and risks of language-generating AI', Nature, 591(7848), pp. 22–25. Available at: https://doi.org/10.1038/d41586-021-00530-0

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender Bias in Machine Learning Data Sets. arXiv preprint arXiv:2005.06898. Available online at: https://arxiv.org/abs/2005.06898. (Accessed 23 December 2024)

Panch, T., Patel, P. and McCourt, T. (2023). Bias in artificial intelligence: Implications for health equity. PLOS Digital Health, [online] 2(8), p.e0000651. Available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000651 (Accessed 23 December 2024).

Sant, A., Escolano, C., Mash, A., De Luca Fornaciari, F., & Melero, M. (2024). The Power of Prompts: Evaluating and Mitigating Gender Bias in MT with LLMs. Available online at: https://doi.org/10.48550/arXiv.2407.18786 (Accessed 23 December 2024)