Peer Response

by Saleh Almarzooqi - Thursday, 9 October 2025, 1:44 AM

I appreciate your consideration of the ethical issues of the DALL·E and ChatGPT models. I liked in particular that you have bridged their creative potential and the threat of misinformation and deepfakes. Now, as Bansal, Nawal, Chamola, and Herencsar (2024) mention, generative systems are capable of creating convincing but completely generated pictures, text, and video on a large scale. The implications of this for the general trust and integrity of information are severe, especially when elections are taking place or when the whole country is facing a crisis. Your proposal of watermarking and provenance-tracking is a powerful one, as it may provide journalists, platforms, and ordinary users with a practical mechanism to check the sources of the content.

By the same token, I liked the part quoted in the intellectual property to which you directed me. This is simply copyright, but there are other matters of moral rights related to just and appropriate remuneration for the many creative individuals who put models together. To achieve this objective, Radke (2024) has suggested training opt-out registries, licensing programs, or collective bargaining agreements to ensure that the artists, writers, and photographers are informed of the proposed training use and have an opportunity to decide if they wish to be trained on the subject under consideration. This can be seen as a supplement to the transparency measures you are talking about.

On algorithmic bias and accountability, you are absolutely correct; there's a need for audit, but it's complicated. We discuss one potential solution, using regular bias audits in conjunction with a corresponding set of pre-deployment model cards and impact analyses, both for developers and legal experts to learn how the system was trained and tested (Mökander, Jakob, et al. 2024). This is a good definition for accountability, and enabling people to take it even further

In short, your post suggests that good governance, as we know it today, is multi-factorial: all factors need to be incorporated: technical security, legal/statutory framework, and the culture of an organisation.

References:

Bansal, G., Nawal, A., Chamola, V. and Herencsar, N., 2024. Revolutionizing visuals: the role of generative AI in modern image generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(11), pp.1-22.

Mökander, J., Schuett, J., Kirk, H.R. and Floridi, L., 2024. Auditing large language models: a three-layered approach. *AI and Ethics*, *4*(4), pp.1085-1115.

Radke, A., 2024. *Collective Societies, Artists' Associations, and Other Desperate Measures to Prevent the Tidal Wave: The Role of Collective Bargaining Mechanisms in Modern Canadian Copyright Contracts* (Master's thesis, McGill University (Canada)).

Maximum rating: -

Permalink Show parent Reply

Re: Peer Response

by Saleh Almarzooqi - Thursday, 9 October 2025, 1:42 AM

I would like to thank you for presenting such a succinct yet detailed list of risks and potential security measures against deep learning systems. I found it very helpful that your post systematically linked the ethical issues to practical action, instead of discussing them abstractly.

Your focus on data consent and provenance, in particular, I find especially appropriate. There are no clear records of the origin of training data, so that bias, intellectual property, or privacy risks cannot be evaluated. To build on your argument, according to Brajovic, Danilo, et al. (2023), there may be the introduction of so-called model cards and data cards as standard documentation to make end-users and regulators understand how a model was constructed and tested.

Embedding watermarks and tags is also an important solution to address the misinformation that you suggested. A recent study by Cao (2025) indicates that strong watermarking, along with open provenance specifications, can assist websites and news reporters in locating AI-generated material before it goes viral.

And lastly, I like your observation of the keeping of human beings in a circle. Besides review processes, organisations might introduce obvious red lines in areas where AI cannot be employed without direct supervision, such as in medical or legal decision-making (Jago, Robert, et al. 2021). This would ease the accountability and trust, and still enjoy the benefits of automation.

Generally speaking, your post emphasises that the development of AI responsibly requires a combination of technical, legal, and organisational efforts instead of a single solution.

References:

Brajovic, D., Renner, N., Goebels, V.P., Wagner, P., Fresz, B., Biller, M., Klaeb, M., Kutz, J., Neuhuettler, J. and Huber, M.F., 2023. Model reporting for certifiable AI: A proposal for merging EU regulation into AI development. *arXiv preprint arXiv:2307.11525*.

Cao, L., 2025. A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content. *arXiv preprint arXiv:2504.02898*.

Jago, R., van der Gaag, A., Stathis, K., Petej, I., Lertvittayakumjorn, P., Krishnamurthy, Y., Gao, Y., Silva, J.C., Webster, M., Gallagher, A., and Austin, Z., 2021. Use of artificial intelligence in regulatory decision-making. *Journal of Nursing Regulation*, *12*(3), pp.11-19.

Maximum rating: -

Permalink Show parent Reply