Summary Post

As technologies of Deep Learning, especially generative models like DALL·E and ChatGPT, raise new ethical challenges, my previous post articulated impacts pertaining to the concerns of authenticity, IP, bias, and the environment. Ending up with highly realistic content that challenges the line between human and machine, these systems contribute to trust erosion, deep fakes, and the misinformation crisis (Vincent, 2022). Moreover, unregulated and large datasets used to train these systems invoke ethical concerns of ownership and fairness (Bender et al., 2021). I considered how biased datasets can strengthen harmful stereotypes (Buolamwini and Gebru, 2018) and the training of large models impacts highly the environment (Strubell, Ganesh and McCallum, 2019).

Feedback from peers helped broaden these aspects. Rayyan Alnaqbi highlighted the erosion of human agency and creativity when generative AI performs human tasks, which, according to her, could lead to the devaluation of artistic and intellectual work (Zhou and Zafar, 2023). She also mentioned the opacity and lack of explainability of generative models, which quasi shifts responsibility from humans, making it problematic when used in the law and healthcare, domain of high stakes (Doshi-Velez and Kim, 2017). She even noted the global AI impacts which calls the need for ethically inclusive and diverse frameworks (Crawford, 2021).

By relating the unit content to the said perspectives, one can conclude that the fundamentals of responsible AI development are ethical governance, engagement of the public, transparency with datasets, explainable AI, and AI design with consideration of energy use. These are actions that correspond with the values of fairness, accountability, and sustainability. Thus, ethical AI has to do with positive value as it is also aligned with the advancement of technology that facilitates human imagination, trust, and social value to people in the world as the core of community integration, unlike the common belief that ethical AI is just in the prevention of harm.

References

Bender, E.M. *et al.* (2021) ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (Accessed: 7 October 2025).

Buolamwini, J. and Gebru, T. (2018) ‘Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification’. Available at: https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed: 7 October 2025).

Crawford, K. (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. Available at: https://yalebooks.yale.edu/book/9780300209570/atlas-of-ai/ (Accessed: 13 October 2025).

Doshi-Velez, F. and Kim, B. (2017) ‘Towards a rigorous science of interpretable machine learning’, *arXiv preprint*. Available at: https://arxiv.org/abs/1702.08608 (Accessed: 13 October 2025).

Strubell, E., Ganesh, A. and McCallum, A. (2019) ‘Energy and Policy Considerations for Deep Learning in NLP’, *arXiv preprint*. Available at: https://arxiv.org/abs/1906.02243 (Accessed: 7 October 2025).

Vincent, J. (2022) ‘The ethical challenges of AI art’, *The Verge*, 20 September. Available at: **https://www.theverge.com/2022/9/20/23360380/ai-art-ethical-challenges-generative-models** (Accessed: 7 October 2025).

Zhou, L. and Zafar, M.B. (2023) ‘Human creativity and the rise of generative AI: A double-edged sword?’, *AI & Society*. Available at: **https://doi.org/10.1007/s00146-023-01586-1** (Accessed: 13 October 2025).