Intelligent Agents – group discussion 3

Discussion 3: Deep Learning

The advent of new technologies supported by Deep Learning models mean that it is now possible to generate 'new' content, for example, Dall-E AI to generate images or ChatGPT to create prose.

Do you think that these new technologies offer any ethical issues that should be considered, and if not, why not?

My initial post:

Deep Learning technologies, such as DALL-E for image generation and ChatGPT for text creation, have disrupted digital content creation, enabling the fast and automated generation of new, creative outputs.But these developments also raise significant ethical concerns that warrant careful consideration and discussion.

One major issue is the potential for misuse, including the creation of misleading or harmful content. For instance, AI-generated images can be manipulated to produce deepfakes, posing threats to privacy and trust (Chesney & Citron, 2019).Deepfakes have been used in the spread of propaganda, election tampering and incriminating campaigns against politically exposed individuals. Additionally, the lack of transparency in how these models generate content can lead to challenges in accountability,intellectual property rights and particularly when AI-generated outputs propagate misinformation or reflect biases present in training data (Binns, 2018).

What'a more is the discussion about intellectual property rights, as AI-generated content challenges traditional notions of authorship and ownership (McCormack et al., 2019). As we integrate these technologies into various sectors, it is crucial to develop ethical frameworks and guidelines that address these challenges, ensuring that AI serves society responsibly while promoting innovation.

In conclusion, while Deep Learning technologies present exciting opportunities, their ethical implications must be rigorously examined to foster trust and accountability in AI-generated content.

References:

- Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency.
- Chesney, R., & Citron, D. K. (2019). "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." California Law Review, 107(6), 1753-1821.
- McCormack, J., Gifford, T., & Hutchings, P. (2019). "Creativity and AI: Is It Possible to Own AI-Generated Content?" Computer Law & Security Review, 35(4), 111-120.

Initial post Christopher Butterworth:

The depth of deep learning lies in the multiple hidden layers between the input and output layers of its neural networks. That, and the ability to engineer their own features, is what makes deep learning networks stand out from simpler machine learning models (Brundage et al., 2018).

The ethical dimension of artificial intelligence, and in particular generative AI, such as ChatGPT™ (OpenAI, 2024) and other systems that can output "original" material such as text, images and video, is coming under scrutiny as people begin to worry where the rise of this technology will leave them. There are concerns about fake news being disseminated by AI systems which have ingested opinionated posts on social media, without adequate guarantees of objectivity. This is not deliberate bias but the result of unchecked biased inputs echoing around various deep learning systems. There is also the problem of copyright when information such as writing, images and music are output without acknowledgement of their original sources, which can get lost among the many layers of these systems (Brundage et al, 2018).

Countermeasures are being developed, themselves using AI methods of detecting bias in order to clean up data sources. This process has been called *algorithmic hygiene* (Lee, 2019) and it is currently focused on the areas of recruitment, facial recognition and online advertising. It is top be hoped that deep learning can be given an ethical framework.

References

Brundage, M. et al. (2018) The Malicious use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Available from: https://arxiv.org/pdf/1802.07228 [Accessed 18 October 2024].

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press. Available from: https://www.deeplearningbook.org/ [Accessed 18 October 2024].

Lee, N.T., Resnick, P. & Barton, G. (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Available at: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

OpenAI. (2024) *ChatGPT* (October 2024 version). Available at: https://chat.openai.com [Accessed 18 October 2024].

My answer:

Hello Christopher,

You do raise critical and valid points about the ethical challenges posed by deep learning technologies, especially in generative AI like ChatGPT. The reliance on extensive hidden layers in neural networks allows these systems to engineer features autonomously, which can inadvertently perpetuate biases present in their training data. As you mentioned, the issue of fake news and misinformation is particularly topical and important, given that these models can amplify biased perspectives without adequate checks on objectivity (Brundage et al., 2018).

Additionally, the copyright dilemma highlights the need for accountability in AI outputs, as original sources may become obscured amidst layers of data processing. The concept of "algorithmic hygiene" is promising, aiming to address bias and ensure data integrity (Lee et al., 2019). Establishing a robust ethical framework for deep learning is essential, not only to mitigate these concerns but also to foster trust in AI technologies as they become increasingly integrated into our lives. Addressing these challenges proactively will shape the responsible development of AI.

Initial post by Zhu Zang:

Generative AI models have already brought about significant changes in healthcare. However, it has also created unprecedented problems

AI models require large amounts of data for training, and if this data includes sensitive or personally identifiable information (PII), then this information may be inadvertently exposed through the generated content. For example, the collection and processing of sensitive patient data, along with tasks such as model training, model building, and implementing generative AI systems, present potential security and privacy risks. (Chen & Esmaeilzadeh, 2024)

Second, The issue of bias being exhibited, perpetuated, or even amplified by AI algorithms is an increasing concern within healthcare.(Mittermaier et al., 2023)

Impact of model bias: if the training data is predominantly from a particular group (e.g. white patients), the generated AI model may show better recognition and processing of data from that group, while it may not perform as well for other groups (e.g. people of different races, genders or ages). This is particularly dangerous in the healthcare field, where incorrect or missed diagnoses can lead to serious health consequences, even life-threatening ones.

Algorithmic bias can also lead to an unfair distribution of healthcare resources. For example, if an AI system is used to determine which patients are prioritised for treatment or access to a certain medication, and the system is biased, certain groups may be systematically overlooked for necessary healthcare. Such a situation could further widen health disparities and make it even more challenging for those who are already disadvantaged.

Chen, Y., & Esmaeilzadeh, P. (2024). Generative AI in medical practice: in-depth exploration of privacy and security challenges. J*ournal of Medical Internet Research*, 26, e53008. Available from: https://www.jmir.org/2024/1/e53008/ [Accessed: 16 Oct 2024].

Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digital Medicine*, 6(1), 113. Available from: https://www.nature.com/articles/s41746-023-00858-z [Accessed: 16 Oct 2024]

My answer post:

Hello Zhu,

Your post highlights important concerns regarding generative AI in healthcare, particularly related to data privacy and algorithmic bias. The potential exposure of sensitive patient information during model training poses significant privacy risks, as was also emphasized by Chen and Esmaeilzadeh (2024). Additionally, the issue of bias in AI models, as noted by Mittermaier et al. (2023), can lead to disparities in diagnosis and treatment down the line, which can ultimately affect patient outcomes. If AI systems are trained mainly on data from specific demographics (for example gener, age etc), they may inadequately serve underrepresented groups, exacerbating existing health disparities. Addressing these ethical and operational challenges is vital to ensure equitable and safe healthcare practices.

My summary post:

This discussion about the ethics of deep learning technologies like DALL-E and ChatGPT showcase both the disruptive potential of generative AI in content creation, but also significant ethical challenges that require debate and attention.

One reocurring concern is the potential for misuse, particularly regarding the generation of misleading content, such as deepfakes, which can undermine trust and privacy (Chesney & Citron, 2019). As noted by Christopher Butterworth, the presence of unchecked biases in training data can lead to the dissemination of fake news, further complicating the landscape of information integrity (Brundage et al., 2018).

Furthermore, the problem of copyright and ownership of AI-generated content raises important questions about accountability and legal regulation, as original sources may become obscured in the data processing layers (McCormack et al., 2019). Zhu Zang highlighted in his post the ethical implications of generative AI in healthcare, noting risks associated with the exposure of sensitive patient information and the potential for algorithmic bias to exacerbate health disparities (Chen & Esmaeilzadeh, 2024; Mittermaier et al., 2023).

As generative AI technologies continue to evolve, establishing ethical frameworks to address these challenges is crucial. This involves proactive measures to ensure transparency, accountability, and equity, fostering trust in AI systems as they integrate into various sectors of society. By engaging with these ethical considerations, we can better harness the potential of deep learning while mitigating its risks.

References:

- Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency.
- Brundage, M. et al. (2018). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." Available from: arXiv.
- Chesney, R., & Citron, D. K. (2019). "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." California Law Review, 107(6), 1753-1821.
- Chen, Y., & Esmaeilzadeh, P. (2024). "Generative AI in Medical Practice: In-depth Exploration of Privacy and Security Challenges." Journal of Medical Internet Research, 26, e53008.
- McCormack, J., Gifford, T., & Hutchings, P. (2019). "Creativity and AI: Is It Possible to Own AI-Generated Content?" Computer Law & Security Review, 35(4), 111-120.
- Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). "Bias in AI-based Models for Medical Applications: Challenges and Mitigation Strategies." NPJ Digital Medicine, 6(1), 113.