Collaborative Discussion 3: Deep Learning

Discussion Topic: 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?

Learning Outcomes

- The knowledge and skills required to develop, deploy and evaluate the tools and techniques of intelligent systems to solve real-world problems.

Initial Post

by Anastasia Rizzo - Tuesday, 4 July 2023, 1:28 AM

Number of replies: 2

For generative AI, transparency is one of the primary ethical concerns. Transparency refers to the ability to understand and explain how an AI system reaches its decisions or generates its output. Deep Learning models, such as Dall-E (openai, 2023) and ChatGPT (OpenAI, 2023), are often considered black boxes due to the complexity of these models, making it challenging to understand how they generate specific outputs or trace the decision-making process.

The European Parliament is currently reviewing The EU AI Act (European Parliament, 2023), which will regulate the development of artificial intelligence systems and protect citizens from their potential risks. The EU AI Act represents the first comprehensive set of regulations specifically designed for AI, addressing various aspects related to its deployment and usage (Browne, 2023). The implementation of this law is expected no earlier than 2025.

Of particular focus is generative AI, which enables the generation of fresh content based on user inputs.

In accordance with the Artificial Intelligence Act (European Parliament, 2023), Amendment 399, Amendment, Article 28(b), Obligations of the provider of a foundation model, paragraph 4:

"Providers of foundation models used in AI systems specifically intended to generate, with varying levels of autonomy, content such as complex text, images, audio, or video ('generative AI') and providers who specialise a foundation model into a generative AI system, shall, in addition: a) comply with the transparency obligations outlined in Article 52(1); b) train, and where applicable, design and develop the foundation model in such a way as to ensure adequate safeguards against the generation of content in breach of Union law in line with the generally-acknowledged state of the art, and without prejudice to fundamental rights, including the freedom of expression; c) without prejudice to Union or national or Union legislation on copyright, document and make publicly available a sufficiently detailed summary of the use of training data protected under copyright law."

The citation above explains that generative AI systems, such as ChatGPT, derived from such models, would need to adhere to transparency mandates by clearly indicating that the content is generated by AI. These systems should also assist in differentiating deep-fake images from genuine ones, thus preventing potential deception. Moreover, stringent measures should be in place to prevent the generation of unlawful content. Additionally, comprehensive summaries of the copyrighted data employed during their training must be made accessible to the public (European Parliament, 2023).

Transparency is crucial because it enables users and developers to understand why and how AI systems produce certain outputs. The lack of transparency can lead to a loss of control and accountability, making it difficult to address potential biases, errors, or unethical outputs generated by the model.

This lack of transparency raises several ethical concerns:

Accountability: It is crucial to understand how AI systems generate outputs to ensure accountability for biassed, inaccurate, or harmful results. The lack of transparency hinders holding algorithms responsible for their actions.

Bias and Discrimination: If Deep Learning models are trained on biassed or discriminatory data, they may perpetuate those biases. The lack of transparency complicates identifying and mitigating such biases, leading to unfair outcomes.

Trust and Informed Consent: Transparency is vital for establishing trust between users and AI systems. Users need to comprehend how their data is used and decisions are made. Without transparency, trust and informed consent become challenging to obtain.

Manipulation and Misinformation: Deep Learning models have the potential to create convincing content, raising concerns about their misuse for manipulation and spreading misinformation. Transparency aids in detecting and addressing such manipulations.

References:

Browne, R., 2023. EU lawmakers pass landmark artificial intelligence regulation. Available from: https://www.cnbc.com/2023/06/14/eu-lawmakers-pass-landmark-artificial-intelligence-regulation.html [Accessed 04 July 2023].

European Parliament, 2023. Artificial Intelligence Act. Available from: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html [Accessed 04 July 2023].

European Parliament, 2023. MEPs ready to negotiate first-ever rules for safe and transparent AI. Available from: https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai [Accessed 04 July 2023].

openai, 2023. DALL·E 2 Preview - Risks and Limitations. Available from: https://github.com/openai/dalle-2-preview/blob/main/system-card.md [Accessed 04 July 2023].

OpenAI, 2023. GPT-4 Technical Report, arXiv:2303.08774

Peer Response:

by Vasilisa Lukashevich - Sunday, 9 July 2023, 8:47 AM

Hello, Anastasia, and thank you for the post!

I agree with you regarding the issue of transparency in the field of artificial intelligence and the ethical concerns it raises. Furthermore, I would like to emphasise that the lack of explainability is apparent not only in complex generative AI tools that employ deep

learning and pre-trained transformers, but also in relatively straightforward machine learning tasks.

As you pointed out, one reason for this lack of transparency is the opaque nature of the decision-making process in machine learning, often referred to as the "black-box problem". Even the developers of the ML models may struggle to comprehend the underlying mechanisms driving the decision-making (Shook et al., 2017).

A second reason lies in the inherent conflict between transparency and privacy. Strobel (2019) highlighted the issue of data privacy in training models. However, I suppose, that it is important to acknowledge that application manufacturers, as business entities, also have proprietary secrets to protect their intellectual property. This further hinders our understanding of the rationales behind AI decisions.

I fully support your perspective that AI-generated content should be publicly displayed with a clear indication that it is generated by artificial intelligence. I suggest that the relevant information should be displayed over the picture or video, similar to warnings on cigarette packs. Otherwise, a significant portion of the audience may mistakenly believe that the Pope wears a white down jacket and that Donald Trump has been arrested (Cartter, 2023; Devlin & Cheetham, 2023). This misleading affects people's trust in the media and in each other, causing problems in society.

References

Cartter, E. (Mar 28, 2023) The Pope Francis Puffer Photo Was Real in Our Hearts. GQ. Available from: https://www.gq.com/story/pope-puffer-jacket-midjourney-ai-meme [Accessed 9 Jul 2023]

Devlin, K. & Cheetham, J. (Mar 24, 2023) Fake Trump arrest photos: How to spot an AI-generated image. BBC News. Available from: https://www.bbc.com/news/world-us-canada-65069316 [Accessed 9 Jul 2023]

Shook, J., Smith, R. & Antonio, A. (2017) Transparency and fairness in machine learning applications. Tex. A&M J. Prop. L., 4, p.443.

Strobel, M. (2019) Aspects of transparency in machine learning. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (pp. 2449-2451).