CORE139:

LARGE LANGUAGE MODELS

ETHICS AND LLMs

NEWS

- https://towardsdatascience.com/the-name-that-broke-chatgpt-who-is-david-mayer-f03f0dc74877
- https://sora.com/
- https://openai.com/index/video-generation-models-as-world-simulators/

ETHIC AND AI

REFLECTION AND VISION OF THE FUTURE IN LITERATURE

- 1818 - **Frankenstein** - Mary Shelley (explores the ethics of creating intelligent beings and the responsibility of creators)
- 1872 - **Erewhon** - Samuel Butler (speculates about the development of machines with consciousness and their possible domination)
- 1920 - **R.U.R.** - Karel Čapek (introduces the word "robot" and describes the rebellion of artificial beings against humanity)
- 1950 - **I, robot** - Isaac Asimov (formulates the three laws of robotics and explores the relationship between humans and robots)

(by Vít Šebestík)

ETHIC CHALLENGES AND LLMs

models have become an integral part of applications ranging from customer service automation to content generation. Here are the key ethical considerations for their use:

- Accountability and transparency
- Bias and discrimination
- Privacy and security
- Misinformation and manipulation
- Impact on employment

ACCOUNTABILITY AND TRANSPARENCY

Transparency in how LLMs operate is crucial for building trust among users. Stakeholders must understand how these models make decisions and generate outputs. This includes providing clear explanations for model behavior and ensuring accountability for the consequences of their use. Developing ethical frameworks that outline the responsibilities of AI developers can help foster accountability within the industry.

BIAS AND DISCRIMINATION

Large language models can inadvertently reproduce or amplify existing biases in society:

Identifying Bias: It is essential to analyze training data and model outputs for the presence of stereotypes and discriminatory views.

Impact mitigation: organisations should implement strategies to minimise bias and ensure fair treatment of all users.

PRIVACY AND SECURITY

The use of LLM may have privacy implications:

Data protection: it is necessary to ensure that models are not trained on sensitive personal data without users' consent.

Information security: It is important to protect data from misuse or unauthorized access.

MISINFORMATION AND MANIPULATION

LLM can be misused to create misinformation:

Preventing misinformation: organizations should have clear policies against using AI to spread false information.

User education: it is important to educate the public on how to recognize AI-generated misinformation.

IMPACT ON EMPLOYMENT

Automation through LLM can impact the labour market:

Transition to new roles: It is essential to support workers in transitioning to new roles created by technological advances.

Training and retraining: Organisations should invest in training and retraining programmes for employees.

PAIR TASK

ACTIVITY

- **1: Content Generation:** A language model generates fake news. What are the risks, and who should bear responsibility?
- **2: Bias in Models:** The model favors a specific culture or religion. How would you address this issue?
- **3: Use in Education/Research:** Students/Researches use the model to write essays/papers. Is this ethical?
- **4: Data Privacy:** The model is trained on users' sensitive data. How can privacy be ensured?

DATA

WHAT TO KNOW ABOUT DATA?

- **Data provenance:** refers to the documentation of the origin of data, including its source, context, and how it has been transformed over time. Understanding data provenance is vital for ensuring ethical data use, as it provides insights into the quality and reliability of the data used to train LLMs.
- **Data lineage:** refers to the documentation of the origin of data, including its source, context, and how it has been transformed over time. Understanding data provenance is vital for ensuring ethical data use, as it provides insights into the quality and reliability of the data used to train LLMs.

DISCUSSION

WHAT ARE YOU AFRAID ABOUT?

AI REGULATIONS

WORLDWIDE

The global approach to AI regulation reflects a diverse range of strategies aimed at balancing innovation with ethical considerations and public safety. As countries continue to develop their regulatory frameworks, ongoing collaboration and dialogue will be essential in establishing effective governance that can adapt to the rapidly changing landscape of AI technology.

WORLDWIDE

The global approach to AI regulation reflects a diverse range of strategies aimed at balancing innovation with ethical considerations and public safety. As countries continue to develop their regulatory frameworks, ongoing collaboration and dialogue will be essential in establishing effective governance that can adapt to the rapidly changing landscape of AI technology.

EUROPEAN UNION

- **AI Act**: The EU is pioneering comprehensive AI regulation with the AI Act, which establishes a legal framework addressing the risks associated with AI. It categorizes AI applications into four risk levels, sets requirements for high-risk applications, and mandates conformity assessments before deployment. The Act aims to ensure safety and protect fundamental rights while promoting innovation within the EU. It is expected to be fully applicable **by June 2026** after a two-year grace period.
- **AI Liability Directive**: This proposed directive aims to clarify civil liability related to damages caused by AI systems, addressing gaps in existing product liability laws.

UNITED STATES OF AMERICA

- **Artificial Intelligence Environmental Impacts Act of 2024**: Introduced in February 2024, this legislation directs the National Institute of Standards and Technology (NIST) to develop standards for measuring AI's environmental impacts, promoting transparency in reporting these effects.
- **State-Level Initiatives**: Various states are exploring their own regulations, leading to a patchwork of laws across the country. This includes guidelines on data privacy and ethical AI use in sectors such as healthcare and finance.

UNITED KINGDOM

- **AI Security Guidelines**: Developed by the UK National Cyber Security Centre, these guidelines focus on enhancing cybersecurity measures for AI systems. They cover secure design practices and have been endorsed by multiple countries..
- **ICO Consultation**: The Information Commissioner’s Office (ICO) is consulting on how existing data protection laws apply to generative AI models, emphasizing the need for transparency in data usage.

ASIA

- **China**: In March 2024, China published a draft **Artificial Intelligence Law**, which aims to establish strict governance over AI technologies. This law emphasizes safety and ethical considerations in AI development.
- **Singapore**: The Cyber Security Agency launched guidelines to secure AI systems across their lifecycle, focusing on best practices for mitigating cybersecurity risks..
- **Hong Kong**: The Hong Kong Monetary Authority issued guiding principles for using generative AI in customer-facing applications, emphasizing governance, fairness, and transparency.

MIDDLE EAST

- **Qatar Central Bank Guidelines**: These guidelines mandate licensed entities to develop an AI strategy that includes accountability measures and risk management protocol.

UNESCO IDEA

UNITED NATIONS, UNITED IDEAS

AI regulation should not be viewed from the perspective of national interests, but from the perspective of humanity as a whole.


AI

QUESTIONS