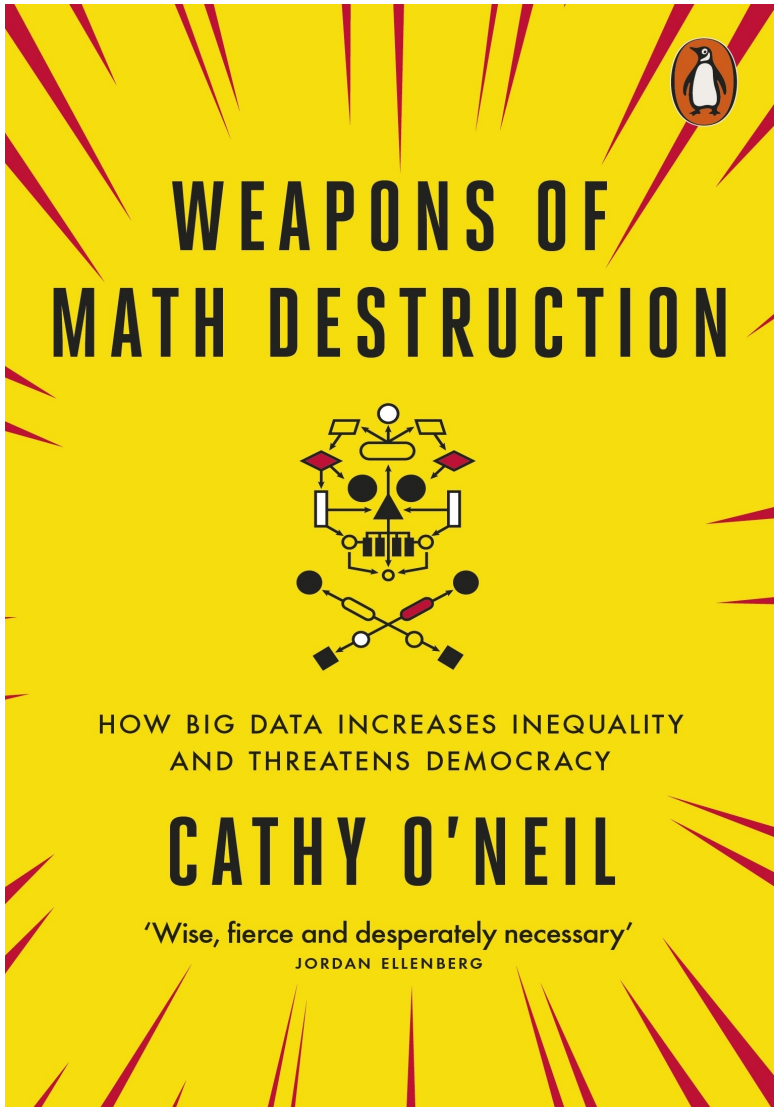
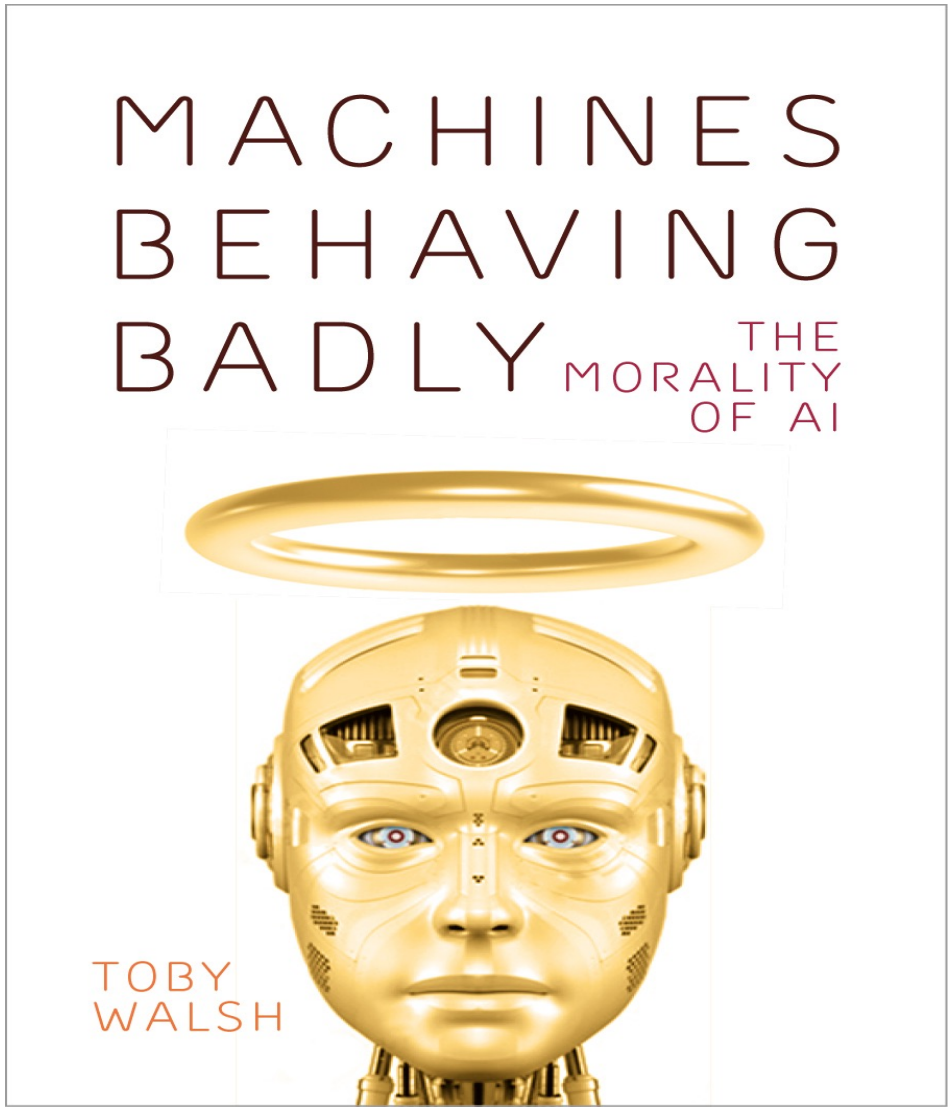
UNSW
THE UNIVERSITY OF NEW SOUTH WALES

10b. Deep Learning and Ethics

Dr Sonit Singh
School of Computer Science and Engineering
Faculty of Engineering
The University of New South Wales, Sydney, Australia
sonit.singh@unsw.edu.au

Increasing concerns on the rise of AI





Source:

Potential issues

- Potential harms arising from the design and use of AI systems:
  - algorithmic bias
  - lack of explainability
  - data privacy violations
  - militarization
  - fraud
  - environmental concerns

Value Alignment

- When we design AI systems, we wish to ensure that their "values" (objectives) are aligned with those of humanity, called value alignment problem.
- Challenging problem due to:
  - it is difficult to define our values completely and correctly.
  - it is hard to encode these values as objectives of an AI model.
  - it is hard to ensure that the model learns to carry out these objectives

Value Alignment

- In ML/DL, the loss function is a proxy for our true objectives.

- For example:
  - consider training an RL agent to play chess. If the agent is rewarded for capturing pieces, this may result in many drawn games rather than the desired behavior (to win the game).

Bias and fairness

- In AI, bias is when illegitimate factors impact an output.
- For example,
  - gender is irrelevant to job performance, so it is illegitimate to use gender as a basis for hiring a candidate.
  - similarly, race is irrelevant to criminality, so it is illegitimate to use race as a feature for recidivism prediction.

Bias and fairness

Data collection	Pre-processing	Training	Post-processing
• Identify lack of examples or variates and collect	• Modify labels • Modify input data • Modify input/ output pairs	• Adversarial training • Regularize for fairness • Constrain to be fair	• Change thresholds • Trade-off accuracy for fairness

Figure 21.2 Bias mitigation. Methods have been proposed to compensate for bias at all stages of the training pipeline, from data collection to post-processing of already trained models. See Barocas et al. (2023) and Mehrabi et al. (2022).

Artificial moral agency

- Many decision spaces do not include actions that carry moral weight. For example, choosing the next chess move has no obvious moral consequence.
- However, some actions can carry moral weight. For example:
  - decision-making in autonomous vehicles
  - lethal autonomous weapons systems
  - professional service robots for childcare, elderly care, and healthcare.

Artificial moral agency

- These concerns leads to artificial moral agency.
- An artificial moral agent is an autonomous AI system capable of making moral judgements.
- The field of machine ethics seeks approaches to creating artificial moral agents.

Transparency and opacity

- A complex computational system is *transparent* if all of the details of its operation are known.
- A system is *explainable* if humans can understand how it makes decision.
- Due to absence of transparency and explainability – hard to ensure value alignment.
- GPT-4 is not transparent at all.

Explainability and interpretability

- Certain regulations such as EU GDPR suggests all data subjects should have the right to “obtain an explanation of the decision reached”
- Deep neural networks have multiple layers and have billions of parameters, making them less explainable in terms of how they work.
- The sub-field of Explainable AI
- Local explanations such as Grad-CAM, LIME, SHAP, etc.

Source:

Militarisation and political interference

- Governments have a vested interest in funding AI research in the name of national security and state building.
- This risks an arm race between nation-states, which carries with it “high rates of investment, a lack of transparency, mutual suspicion and fear, and a perceived intent to deploy first”

Fraud

- Generative AI can be used to deceive people into thinking they are interacting with a legitimate entity or generate fake documents that mislead or deceive people.
- AI could increase the sophistication of cyber-attacks, such as by generating more convincing phishing emails.
- ChatGPT have been used to write software and emails that could be used for espionage, ransomware, and other malware.

Data privacy

- Modern DL methods rely on huge crowd-sourced datasets, which may contain sensitive or private information.
- Studies showed that even when sensitive information is removed, auxiliary knowledge and redundant encodings can be used to trace individuals.
- Need for privacy-first design to ensure security of individuals.

Intellectual Property

- Many AI models are trained on copyrighted material. Consequently, these models' deployment can pose legal and ethical risks and run afoul of IP rights.
- Can the output of a ML/DL model (e.g., art, music, code, text) be copyrighted or patented?
- Is it morally acceptable or legal to fine-tune a model on a particular artist's work to reproduce that artist's style?

Automation bias and moral deskilling

- As society relies more on AI systems, there is an increased risk of automation bias (i.e., expectations that the model outputs are correct because they are ”objective”).
- Off-loading cognitive skills like memory into technology may cause a decrease in our capacity to remember things.

Environmental impact

- Training deep neural networks requires a significant computational power.
- For example, Training a Transformer model with 213 million parameters emitted around 284 tonnes of CO2.

Employment and society

- Technology innovation has a history of job displacement.
- McKinsey Global Institute suggests that up to 30% of the global workforce could have their jobs displaced due to AI.
- In summary:

Human + AI > Human

Democratise AI

- To avoid concentration of power, there is a big push towards democratizing AI.

- Make DL technologies more widely available and easier to use via open-source and open science initiatives.

- Reduce barriers to entry and increase access to AI while cutting down costs, ensuring model accuracy, and increasing participation and inclusion.

Summary

- Ethical AI is a collective action problem.
- When designing/developing or using AI systems, consider the moral and ethical implications of their work.

References

[1] Ho et al., Denoising Diffusion Probabilistic Models. NeurIPS 2020.

https://arxiv.org/pdf/2006.11239.pdf?ref=assemblyai.com

[2] Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. 2022.

https://arxiv.org/pdf/2204.06125.pdf

[3] Desai and Johnson. VirTex: Learning Visual Representations from Textual Annotations

https://arxiv.org/pdf/2006.06666.pdf

[4] Zhang et al., Contrastive Learning of Medical Visual Representations from Paired Images and Text. 2022.

https://arxiv.org/pdf/2010.00747.pdf

[5] Radford et al., Learning Transferable Visual Models From Natural Language Supervision. 2021.

https://arxiv.org/pdf/2103.00020.pdf

[6] Vaswani et al., Attention is All You Need. NeurIPS 2017.

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] Dosovitskiy et al., An Image is Worth 16 x 16 words: Transformers from image recognition at scale. ICLR 2021.

https://openreview.net/pdf?id=YicbFdNTTy

[8] Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022.

https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf

[9] Ronneberger et al., U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015.

https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

[10] Jay Alammar. The Illustrated Stable Diffusion

https://jalammar.github.io/illustrated-stable-diffusion/

Questions?

myExperience

Tell us about your experience. Shape the future of education at UNSW.

Visit Moodle to complete the myExperience survey

UNSW SYDNEY

Tell us about your experience. Shape the future of education at UNSW.

Visit Moodle to complete the myExperience survey

Please be mindful of the **UNSW Student Code of Conduct** as you provide feedback. At UNSW we aim to provide a respectful community and ask you to be careful to avoid any language that is sexist, racist or likely to be hurtful. You should feel confident that you can provide both positive and negative feedback, but please be considerate in how you communicate.

UNSW SYDNEY

Please complete myExperience survey

https://myexperience.unsw.edu.au/unsw/