Topic 6: Practical Approaches and Ethical Frameworks

DEAKIN BUSINESS SCHOOL

Core Learning Objectives

UNDERSTAND THREE MAIN ETHICAL CHALLENGES OF AI

ARTICULATE EXISTING ETHICAL FRAMEWORKS

ARTICULATE POSSIBLE COURSES OF ACTION WHEN THERE ARE ETHICALLY SENSITIVE ISSUES

6.1 Introduction

THERE HAVE BEEN MANY CASES OF AI BIAS AND DISCRIMINATION BEING REPORTED

LET'S HAVE A LOOK AT EXAMPLES IN SECTION 6.1 ON THE UNIT SITE

6.2 The Ethical Challenges of AI

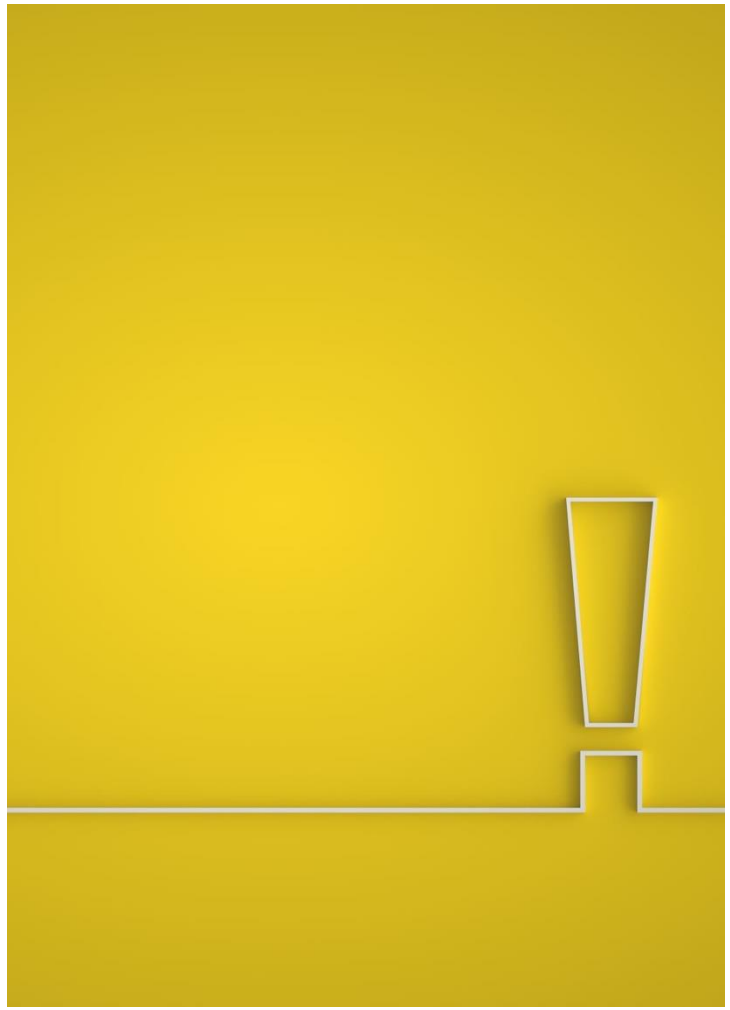
The three major challenges that will be looked at more closely in this topic are

- biased AI
- explainability – black-box algorithms
- privacy concerns

Not all ethical risks that are identified fit precisely in each category

Biggest risks of AI are not the technologies themselves, but the way they are used and the outcomes or results of this use

- The obvious risk for the use of facial recognition technology for surveillance is that it has effects on a person's privacy
- With the potential to cause anxiety, alter peoples' behaviour, erode autonomy
- It creates an imbalance of power



Questions for checking ethical issues that may arise…

Whether people are being treated respectfully?

Whether some product design is manipulative or merely giving reasonable incentives?

Whether some decisions would be cruel or culturally insensitive?

Whether a decision places a burden on people that is too great to reasonably expect of them?

6.3 Biased AI – is it fair?

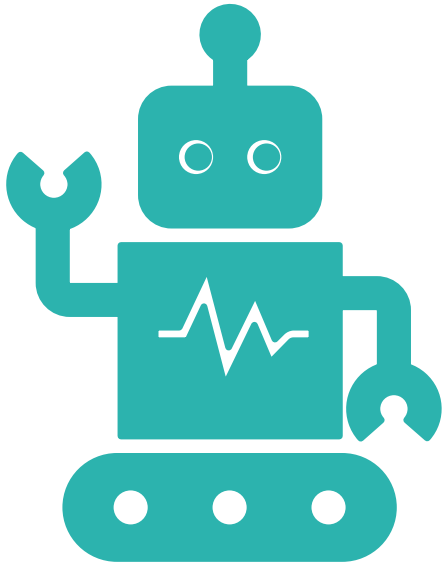
- Any form of preference leading to a judgement or decision
- Some biases can be automatic and unconscious, such as when decisions are based on heuristics, when the brain has created mental shortcuts, that it is hard to detect by the person making those decisions
- Biases can cause problematic situations for humans and equally so for AI/machine learning

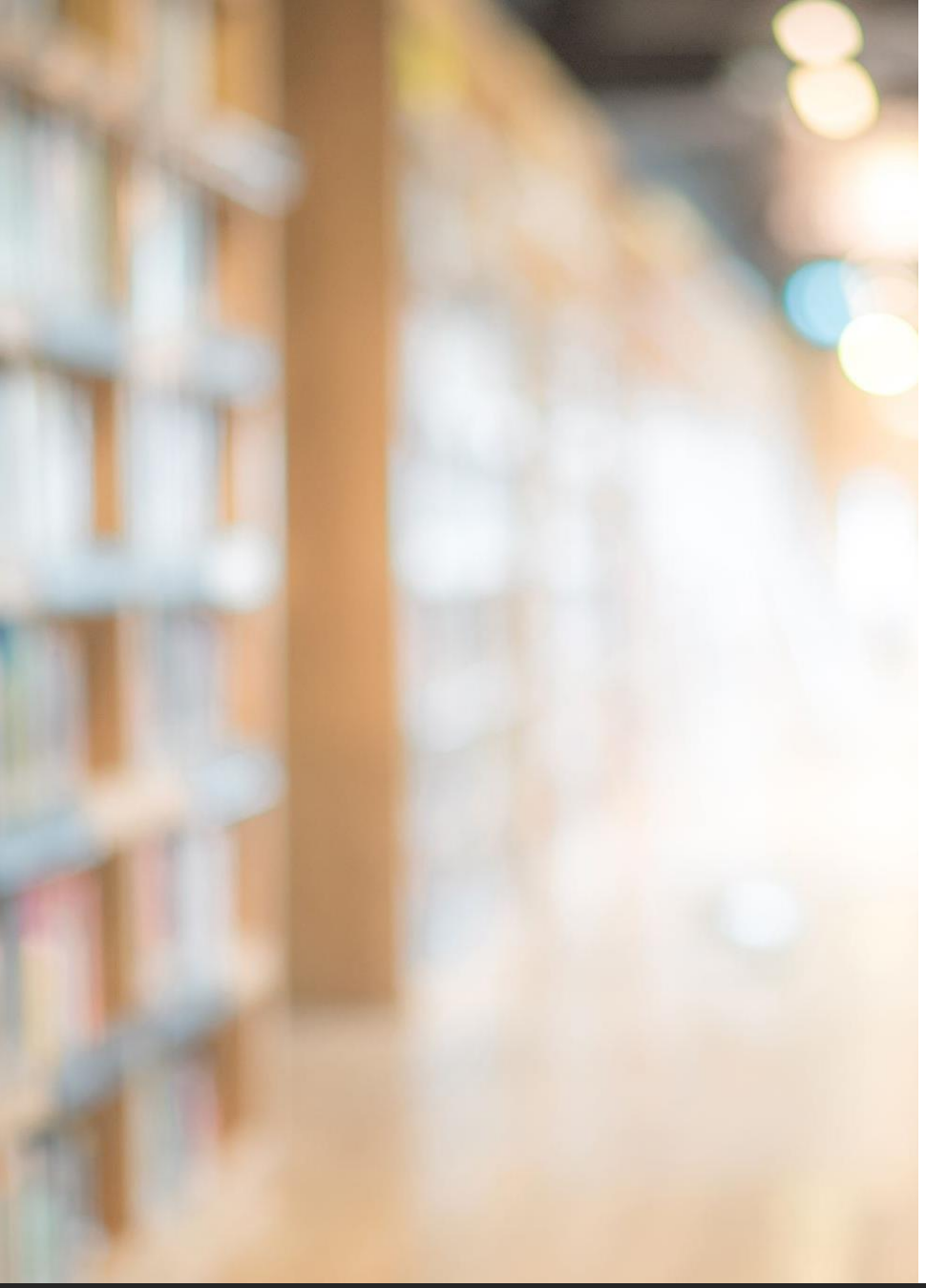
6.3 Biased AI – is it fair: An example

- An organisation was using a resume-reading AI system/tool to select candidates for interviews based on the information in their resumes
- The AI tool was trained by inputting the data of resumes of successful applicants from the previous 10 years in the organisation
- The AI tool was trained to look for patterns that a person would be 'ideal' for an interview

The results from the resume-reading AI system/tool led to interviews of the following people who applied

- 30% of all men
- 20% of all women
- 10% of all Black man
- 5% of all Black women





Where does the potential bias or discrimination in AI come from?

The resume AI tool and other real-life examples have shown that AI biases comes from human prejudices as machine learning originates from the training data

Biases in AI systems or tools can arise from

- Cognitive bias – this is an umbrella term used to describe the tendency when people use their personal experiences and/or emotions to affect their judgements and decisions
- AI's objective/function; biased benchmark; coarse-grained model
- data set used for training includes biases
- Lack of complete data/sample size – if the data set only contains part of the picture it may be biased
- Proxy bias

See the 'racist soap dispenser' example (see section 6.3 - Unit site)

How to mitigate bias in AI and machine learning algorithm?

1. Be aware of contexts in which AI can help correct for bias and those in which there is high risk for AI to exacerbate bias
2. Establish processes and practices to test for and mitigate bias in AI Systems
3. Engage in fact-based conversations about potential biases in human decisions
4. Fully explore how humans and machines can best work together
5. Invest more in bias research, make more data available for research , and adopt a multidisciplinary approach
6. Invest more in diversifying the AI field

Activity

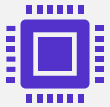
Case of AI tool COMPAS, developed by a private company to predict whether a criminal would reoffend, by assigning individuals with a risk score which are used to inform judges for the sentencing of these people.

(See section 6.3 – unit site)

6.4 Explainability AI - Opening the black box of algorithm

Processes and methods to allow humans to understand and trust the outputs generated by the AI algorithms

To make the mechanisms of AI algorithms transparent and the outputs understandable to humans



If an AI system's output does not match a user's expectations and the mechanisms are not understandable, it can lead to a loss of trust

Explainability helps to confirm predictions, improve, and advance models and to gain new insights into the initial concerns

Explanations are ethically important because

They demonstrate respect for the person being explained to

They enable the person being explained to to make informed decisions and change their behaviour, giving them control over future decisions

They allow for assessment of the model's rules that govern how inputs are turned into outputs against ethical, reputational, regulatory, and legal standards

Classification of Explainable AI techniques: Transparent versus Black box models

Transparent models benefit from their fundamental design which allows to provide explanations based on it whereas black-box models do not provide such insights

	Transparent models	Black-box models
Definition	Models that can fully and understandably explain how an algorithm operates and given an input, can tell what the output will be and why	Models that create internal structures that determine outputs but are opaque to external parties. Even the programmers cannot tell why a particular output was produced
Agnostic	Possible	Possible
Model specific	Not required	Possible
Example Algorithms	Linear regression model rule-based systems decision trees	Feed-forward neural networks, convolutional neural networks recurrent neural networks generative adversarial networks

(Hamm et al., 2021, p. 3)

Activity

Test your knowledge: After reading the table above comparing Transparent and Black-box models can you match the following example Algorithms

See section 6.4 – Unit site

How to reduce the issues of black box algorithms?

Organisations must consider the importance of explainability for each specific use case and determine the balance against other goals, such as accuracy

Using the following guidelines from European General Data Protection Regulation (GDPR) (see next slides)

According to GDPR, Machine explainability is necessary when…

showing respect is ethically required

people need to understand how to improve results

people need to know how to approach the AI that makes a decision

the outputs are unexpected

justification of action is needed from an ethical, regulatory, or legal perspective)

Appropriate explanations should be…

true

easy, efficient, and effective for its intended use

understandable for the intended audience

explanation should also include justification of decisions, actions, processes, etc.

6.5 Privacy

Refers to the extent to which people have knowledge and control over their data without undue pressure

Privacy issues have received a fair amount of media coverage in recent years, with citizens, consumers and employees demanding how governments should regulate issues regarding privacy

There have been some countries that have implemented regulations (e.g. Australia, UK, France)

Regulatory compliance, data integrity and security, and ethics are the major aspects to privacy

Ethical Treatment of Privacy

Transparency

- Users, customers, or citizens should be able to know what information is being collected about them, what's being done with it, what decisions they contribute to, who it's being shared with or sold to
- If organisations don't advise customers about those, or worse, they don't know either, that's an ethical problem

Data control

- Giving the capacity for users, customers, or citizen to be able to correct, edit, or delete information about themselves to opt out of being treated in a certain way
- By doing so, the organisation shows their respect to the stakeholders' privacy and provides the stakeholders an opportunity to fix incorrect information the organisation has about them

Ethical Treatment of Privacy (ctn.)

Opt out by default

- Without users' consent, many organisations currently collect data by default via the registration process and while a user is active on the company's site
- This means that users are automatically "opted-in" for the data collection about them by the organisations, raising another ethical issue of privacy

Full services offered

- Organisations may also increase or decrease the services they provide based on what data a user provides
- The ethical issue of privacy will occur depending on how essential the services are.
- If the services are essential and an organisation only provides the services with the condition that you provide them with an access to your data that you are not comfortable with, then this raises an ethical issue of privacy

	Level 1	Level 2	Level 3	Level 4	Level 5
	Blindfolded and handcuffed	Handcuffed	Pressured	Slightly curtailed	Grateful
Transparency		✓	✓	✓	✓
Data control			✓	✓	✓
Opt out by default				✓	✓
Full services	✓	✓			✓

Activity: Five ethical levels of privacy

(see Section 6.5 – unit site)

6.7. AI Ethical Frameworks

AI regulation with EU AI Act

One of the very first attempts to regulate AI is the AI Act developed and proposed by the EU to provide developers, deployers, and users with clear requirements and obligations regarding specific uses of AI

In April 2021, the EU Commission proposed the first-ever legal framework on AI addressing the risk of AI and positions EU countries to play a leading role globally

In December 2023, the EU Act has been approved by the European Parliament and the Council of the European Union; it came into force on August 1, 2024 and will be fully effective from August 2, 2026

UNESCO's Recommendation on AI Ethics

- The very first global standard-setting instrument on the subject to not only protect but also promote human rights and human dignity
- Serves as an ethical guiding compass and a global normative bedrock allowing to build strong respect for the rule of law in the digital world
- Aims to provide a basis to make AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm
- Focus on 10 aspects/principles: Proportionality and Do No Harm
- Safety and security, Fairness and non-discrimination, Sustainability, Right to Privacy and Data Protection
- Human oversight and determination, transparency and explainability
- Responsibility and accountability, Awareness and literacy, Multi-stakeholder and adaptive governance and collaboration

Australia's AI Ethics Framework

To guide businesses and governments to responsibly design, develop and implement AI

Aims to make Australia a global leader in responsible and inclusive AI and to obtain the immense potential of AI by making sure it is safe, secure and reliable to use

(Let have a look at the details in section 6.7 on the unit site)

We will discuss this topic further in this week's seminars.

THANK YOU!