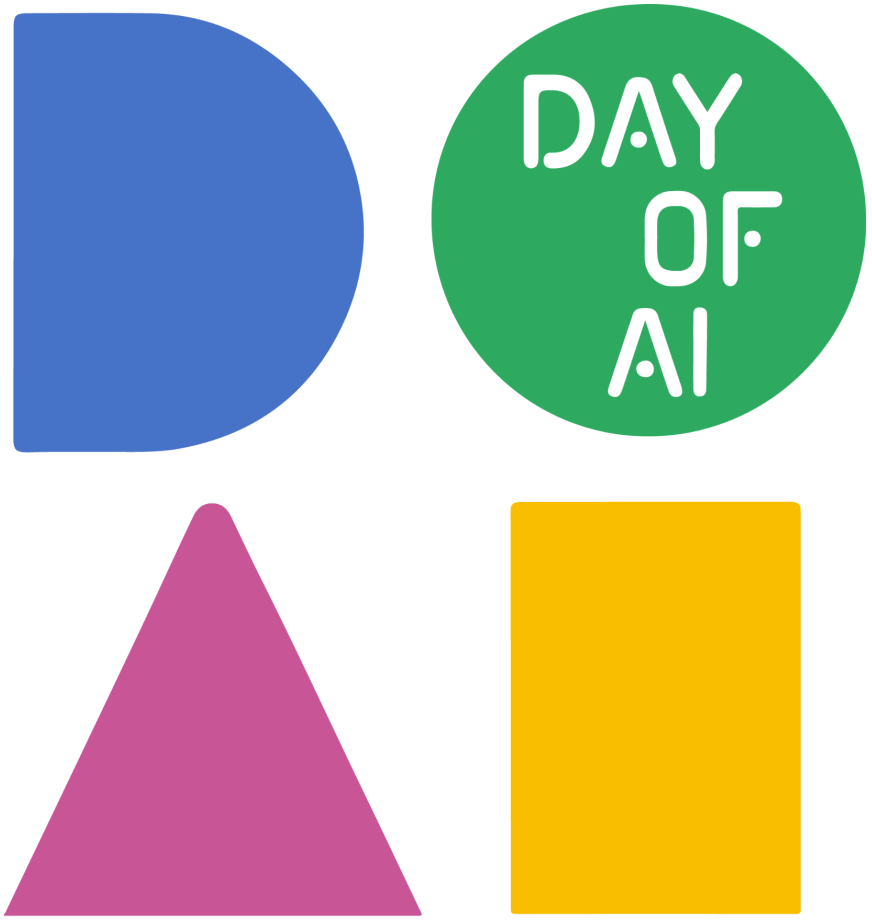

AUSTRALIA

Lesson 4: Ethics and Misinformation in AI

2024 SENIOR PROGRAM

LESSON PLAN

Lesson 4: Ethics and Misinformation in AI | 60 mins

Lesson Summary

In this lesson, students reflect on the ethical considerations surrounding AI usage. They explore societal impacts, analyse AI-generated content in mainstream media, considering biases, plagiarism and responsible AI use. Through discussions and case studies, they examine the perspectives of various stakeholders in the context of journalism, explore the limits of AI, and build an understanding of the ethical implications of AI, fostering critical thinking skills and ethical awareness.

Objectives

- Students will explore **Societal Impact**, the fifth **Big Idea of AI**.
- Students will build an awareness of AI-generated content in mainstream media.
- Students will consider the impact of data bias in generative AI and content derived from it.
- Students will reflect on the responsible use of AI by applying ethical concepts like 'harm', 'benefit', 'misrepresentation' and the 'perspective' of multiple stakeholders.
- Students will learn about the limits of AI, particularly of Large Language Models (LLMs) like ChatGPT.

Curriculum Alignment

This lesson is linked to the following Australian Curriculum (Version 9) content descriptors:

- **Years 7 and 8**
  - AC9M8P01: recognise that complementary events have a combined probability of one; use this relationship to calculate probabilities in applied contexts.
  - AC9M8ST01: investigate techniques for data collection including census, sampling, experiment and observation, and explain the practicalities and implications of obtaining data through these techniques.
  - AC9M8ST02: analyse and report on the distribution of data from primary and secondary sources using random and non-random sampling techniques to select and study samples.
  - AC9TDI8P10: evaluate existing and student solutions against the design criteria, user stories and possible future impact.
  - AC9TDI8P14: investigate and manage the digital footprint existing systems and student solutions collect and assess if the data is essential to their purpose.
- **Years 9 and 10**
  - AC9M10ST01: analyse claims, inferences and conclusions of statistical reports in the media, including ethical considerations and identification of potential sources of bias.
  - AC9M9ST01: analyse reports of surveys in digital media and elsewhere for information on how data was obtained to estimate population means and medians.
  - AC9M9ST02: analyse how different sampling methods can affect the results of surveys and how choice of representation can be used to support a particular point of view.

- AC9S9H04: examine how the values and needs of society influence the focus of scientific research.
- AC9TDI10P10: evaluate existing and student solutions against the design criteria, user stories, possible future impact and opportunities for enterprise.

Vocabulary

- **Ethics,** n. the study of morality and values, specifically what is 'right' and 'wrong'.
- **AI Ethics,** n. the application of ethics to AI research and development.
- **Beneficial,** adj. something that results in a positive effect for a person, or a group of people.
- **Harmful,** adj. something that results in a negative effect for a person, or a group of people.
- **Stakeholder,** n. a person, or a group of people, who has an interest in a particular decision or process i.e. they hold a 'stake' and may be affected by the outcome.
- **Bias,** n. a preference or a prejudice towards someone or something, especially in a way considered to be unfair or unrepresentative.
- **Misinformation,** n. false or inaccurate information, especially that which is deliberately intended to deceive.
- **Misrepresentation,** n. the action of giving a false or misleading account of the nature of something or someone.

Resources

- Presentation Slides for lesson (available on website once logged in)
- Miley Cyrus voice (embedded in the lesson slides)
- Journalism stakeholders handout (optional, for students to reference)

Activity Steps

1. **1 min.** Introduce **Societal Impact** as the fifth 'big idea' of AI.
2. **1 min.** Set up a simple model of ethics using 'harm' and 'benefit'. 'Harm' is something that negatively affects people, 'benefit' is something that positively affects people.
3. **2 mins.** News article scenario. Walk through the example of a journalist or news producer using AI to generate an article to meet a deadline, rather than writing the article themselves. Point out that the harms and benefits in this scenario apply to different people of interest (the journalist/writer, the news outlet, the reader).

    **EXTENSION QUESTION:** What other stakeholders are affected here? What harms and benefits might apply to them?
4. **4 mins.** In pairs, ask students to discuss whether the way you use technology influences whether or not it is harmful or of benefit.

   Feel free to give an example such as using the internet to research things you want to learn more about versus using the internet to illegally download movies.
5. **1 min.** Introduce the concept of perspective taking, and the idea of a **stakeholder.** Return to the plagiarism example. Point out again how the harms and benefits apply differently to different **stakeholders.**
6. **6 mins.** Play the audio clip of 'Miley Cyrus'. Ask students to discuss in pairs or small groups the following questions:

 Is the claim itself true or false that *"The problem with internet quotes is that they can be false."*

 Do you believe that Miley Cyrus made the claim?

 Does it matter whether Miley Cyrus made the claim or not?

 Is the audio clip ethical? Why, or why not?

7. **10 mins.** Run through the questions together with students. Reveal that the audio clip was made in under a minute using an online Deepfake voice generator called Vidnoz, which utilises AI. The following is some guidance on answering the questions:

 **Is the claim itself true or false?**
The claim, as a qualified claim, seems true enough, if it is accepted or 'believed' it may lead to a moderate scepticism about internet quotes, and fact-checking. This is beneficial, not harmful.

 **Did Miley Cyrus make the claim?**
Now that we know it was made by AI, we know that Miley Cyrus did not make the claim. It is a false attribution and the claim leads to a false belief about Miley Cyrus. It is misinformation: so while the claim itself might be true, we must still consider it misinformation since Miley Cyrus herself did not say it.

 **Does it matter whether Miley Cyrus made the claim or not?**
This answer is not straightforward and you could arrive at different answers depending on your point of view and that of your students.

i. **'No'** would be an acceptable response for the reasons outlined above: the claim is true but the false attribution leads to a misrepresentation of Miley Cyrus.
ii. **'Yes'** would be an acceptable response if you are prepared to accept the recording as a joke that ironically points out how common it is for internet quotes to be false. However, you could argue that the joke could be lost on some people, or that the context in which it's presented could be different. Encourage discussion on this with your students.

 **EXTENSION QUESTION:** Imagine this was a political statement, made to look like a politician said it. What could the potential impact of that be?

✎ **NOTE**

Your students may be interested to know that the audio clip was made in under a minute using an online Deepfake voice generator called Vidnoz, which utilises AI.

8. **5 mins. CASE STUDY:** Researchers at the University of Queensland (UQ) have developed an AI tool which has been trained on input from human experts and trusted datasets to identify misinformation. When the system looks at a new piece of information, it classifies it as potential misinformation or not, which is then confirmed and refined by humans to ensure quality, agency and accountability.
    a. **Here's how it works:**
        i. Humans provide expert inputs, including expert datasets like the RMIT ABC Fact Check dataset, to train the AI. The RMIT ABC Fact Check dataset is information on statements made by Australian politicians and whether those statements are true, false or in-between.
        ii. The AI can then review large amounts of online information and generate labels and explanations as to why a claim may be misleading. These labels and explanations are provided to users to help them make better choices about the information they trust and use.
        iii. Humans are then also involved to check and refine the labels and explanations, as both experts and crowd workers (people who have volunteered to work on these sorts of projects) to help maintain the quality, agency, and accountability of the process.
        iv. Using AI in this way means huge amounts of information can be processed - much more than could be done by humans alone.
9. **4 mins.** Introduce the concept of 'bias', and data/algorithmic bias in AI, giving an example of data bias in identifying heart disease in men and women.
10. **2 mins.** Introduce Large Language Models (LLMs) as an AI model that generates text, and is trained on a huge dataset of text. ChatGPT in particular is an LLM which generates human-like conversational dialogue.

> ✎ **NOTE**
>
> **If you have already completed Lesson 3 with your students, they will already be familiar with LLMs, ChatGPT and how they work. You can spend less time on this example, or spend more time on it to jog their memory.**

11. **4 min.** Ask the class about the kind of bias that might occur in LLMs in a variety of formats, including novels, social media posts, political campaigns and news articles.
    For example, if the LLM is trained only on social media posts from 18-25 year olds, it will be biased towards producing text in a style that appeals to that age range and wouldn't consider the perspective of other age groups.
12. **10 mins.** Introduce three imaginary stakeholders in the journalism industry. In pairs or small groups, students should select one person of interest. For the stakeholder they have chosen, they should discuss:


What sort of things would they **care** about?


What **concerns** might they have about generative AI?


How do they stand to **benefit (positive)** from generative AI?

What might be a potential **harm (negative)** from generative AI?

How could they be affected by potential **bias** in generative AI?

NOTE

This handout transcribes each of the stakeholders, which you can print and share with students in advance if you would like them to have a physical copy to refer to.

NOTE

Your students may be interested to know that these stakeholders were generated by ChatGPT!

13. **8 mins**. Discuss the limits of AI and LLMs. Encourage students to consider other limits of AI, and how they should think about using the technology. These are a few prompts to begin the discussion:

LLMs are just making predictions. Like any AI, LLMs are ultimately just guessing at what you want. Sometimes they're right, sometimes they're not. They don't really understand the things they're predicting.

LLMs are frozen in time. Once trained, they do not learn new information. E.g. ChatGPT version 3.5 was trained in January 2022. It doesn't know anything after that.

LLMs do not always give the same answer, even for the same prompt. The LLM generates a new prediction every single time, and generating two identical responses is extremely unlikely.

LLMs can be biased. They can be trained on data that leads to inaccurate or wrong predictions, just like humans. You should always fact check an answer you get from an LLM.

END OF LESSON PLAN