AI and Ethics

About this unit

- Authors: Christina Nick and Natasha McKeever
- 2018/19 Unit Tutor: Andrew Kirton

This is the first ethics unit of XJCO2611, Artificial Intelligence.

In this unit we will look at the ethical issues surrounding artificial intelligence.

Activity Period: 27th April - 1st May

You should aim to have completed your reading and preparation for this unit, and made your initial contributions to the group discussions by the end of **Wednesday 29th April**.

You can print the whole unit out, or read it online.

Introduction

Artificial intelligence or AI is what we call intelligence that is exhibited by machines and systems. There are two different kinds of intelligence that we could be talking about:

1. Weak AI: exhibit machine learning with regard to a specific domain (e.g. chess playing)
2. Strong AI: exhibit general intelligence and self-awareness (like a human)

We will look into the ethical issues surrounding weak and strong AI in turn.

Exercise 1.1: Examples of weak AI

Can you think of any examples of Weak AI?

Write down your answer in your **personal learning blog**.

Weak AI (Part 1)

What is currently decided by AI in our personal lives? Here are some examples:

- Job applications
- Stock market
- Insurance premiums
- Results of internet searches
- Online dating
- University places

There are some clear benefits to using weak AI for everyday processes. It can process information in regard to a specific domain a lot quicker than any human could. This means that using weak AI is an efficient use of resources. We might also think that weak AI, in comparison to humans, are more impartial and neutral. They will not have any of the prejudices or biases that all humans inevitably have. The results of a decision made by weak AI may therefore seem fairer. But is this really the case?

Video

Watch the following talk until 7:11

[Joy Buolamwini: How I'm fighting bias in algorithms]

Exercise 1.2: Questions for the video

Please answer the following questions when watching the video:

1. What are the problems of "algorithmic bias"?
2. Why was the facial recognition algorithm biased?
3. Why are algorithms called "weapons of math destruction" in a book by Cathy O'Neill?

Write down your answer in your **personal learning blog**.

Weak AI (Part 2)

Biases in algorithms, just like biases in humans, can lead us to exclude, target, or disadvantage certain groups of people based on arbitrary factors (e.g. race, gender, disability). The problem with algorithmic bias is that it can do so at a greater speed across the globe without us realizing.

Weak AI is based on machine learning. In the case of facial recognition, for example, the computer is presented with a training set of faces. The problem is that if the training set is not diverse, then the computer will not be able to recognize faces that look different from what it has been taught.

Algorithms are widespread and have an increasing impact on society as they are used in more and more domains to make decisions that affect our lives. At the same time, they remain relatively mysterious to most people. It is not transparent how the algorithms that make decisions about our lives work. Even coders will sometimes struggle to understand why exactly a weak AI is targeting a certain group of people (e.g. when the "algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution" – Bostrom and Yudkowsky). These two qualities can make the biases inherent in algorithms very destructive. They make it difficult to reconstruct the decision-making process and it is therefore harder to intervene when weak AI reinforces human biases or prejudices.

The question then is, what can we do to make machine learning and weak AI less biased and more impartial?

Example: Association for Computing Machinery

7 Principles for Algorithmic Transparency and Accountability:

1. *Awareness*: Owners, users, and designers have to be aware of potential biases and the harm that could results from them
2. *Access and redress*: Groups that are adversely affected should have channels to question these algorithms
3. *Accountability*: Institutions should be accountable for the decisions of their algorithms
4. *Explanation*: Those using algorithms should do their best to reconstruct the decision-making process
5. *Data provenance*: The training sets used to teach AI should be assessed for potential biases
6. *Auditability*: All decisions and algorithms should be recorded so that they can be investigated if any harm is suspected
7. *Validation and Testing*: Algorithms should be routinely tested for potential biases

Source: US Association for Computing Machinery

Strong AI (Part 1)

So far we have talked about AI that is highly skilled – often more skilled than humans – at a specific task. What they are missing, however, is the ability to act outside of this domain. For example, while a weak AI could beat any human at chess, unlike a human it is unable to teach itself a different game or to read or to exhibit any other skill. When we talk about strong, as opposed to weak, AI we talk about the idea that at some point in the future there may be a machine that could show intelligence across all domains like humans do.

There is a whole new set of moral questions that we are faced with when we consider the possibility of such strong AI. In particular, we will look at the following two questions:

1. How should strong AI treat us?
2. How should we treat strong AI?

Exercise 1.3: How should strong Ai treat us?

Could and should we design strong AI in a way that includes ethical principles? If so, which principles should we give them?

Please post your answer in the **group discussion forum**. Remember to ensure that you offer reasons and arguments for the views that you express. Once you have posted your initial contribution please read through and comment on the contributions made by others.

Strong AI (Part 2)

Here are some complications to bear in mind when thinking about giving strong AI ethical principles:

- There is reasonable disagreement about how we should make ethical decisions (e.g. should we always follow our duties and obligations? What if that would result in very bad consequences?), so it is unclear what principles we should use when programming.
- The nature of ethical principles makes them difficult to codify through an algorithm (e.g. what should be done when two ethical principles clash? Can they be put into a neat hierarchy from most to least important?)
- The environment in which we act is very complex and it is impossible for an AI to analyse all of the potentially infinite amount of options available; it will therefore be difficult to take everything into consideration that may matter ethically.
- To develop general artificial intelligence the system will have to learn and evolve independently, but this will require it having a certain amount of freedom which cannot be controlled by humans.

Optional Further Reading

Miles Brundage: Limitations and Risks of Machine Ethics

So far we have considered how strong AI should treat us. What we will now be exploring is how we should treat strong AI; in particular, we should be concerned with whether such AI could have rights just like humans do. There are several ways in which we might want to show that someone or something can have rights.

The first one would be to show that they are a moral agent who is capable of conscious moral decisions and whom we can therefore hold responsible for their actions. If an entity satisfies these conditions it will also be worthy of having certain rights. This is the approach that, for example, was taken by early feminists to vindicate the rights of women.

The second approach is to figure out whether someone or something is a moral patient. A moral patient is any entity that is capable of suffering, the idea being that anything that can experience pain should have their pain taken into account in our moral decision-making. Any entity that satisfies this condition will also be worthy of having certain rights. This is the approach often taken, for example, by those wanting to vindicate the rights of animals.

The final approach also starts by considering who or what could be a moral patient, but it uses a different criterion to suffering. On this view anything that exists as a coherent body of information could be worthy of rights that protect it from being destroyed or corrupted in any way. This is a very recent approach that some philosophers have taken to vindicate the rights of machines and technology.

Optional Further Reading

David J. Gunkel: A Vindication of the Rights of Machines

Exercise 1.4: How should we treat strong AI?

If an AI with genuine general intelligence was to exist in the future, should it have the same rights as humans? If so, why? If not, why not?

Please post your answer in the **group discussion forum**. Remember to ensure that you offer reasons and arguments for the views that you express. Once you have posted your initial contribution please read through and comment on the contributions made by others.

Summary

There are two different kinds of artificial intelligence: weak AI have already surpassed the capabilities of humans in performing domain-specific tasks and strong AI which are currently not in existence but which might show human-like general intelligence in the future. Each of these kinds of AI pose a unique set of ethical problems. Weak AI is already impacting our everyday lives and we should be particularly concerned with its transparency and the ability to hold people accountable for the outcomes of using such systems. When we consider strong AI we have to ask ourselves how ethical principles ought to be programmed into them and what their moral status would be.

End of unit

This is the end of the unit. You can revisit any of the sections of the unit using the document menu on the right-hand side, or you can print the whole unit out for your records.

Please make sure that you have completed all of the exercises and have actively contributed to the group discussions.

Previous Next