THE UNIVERSITY OF MELBOURNE

Module 3

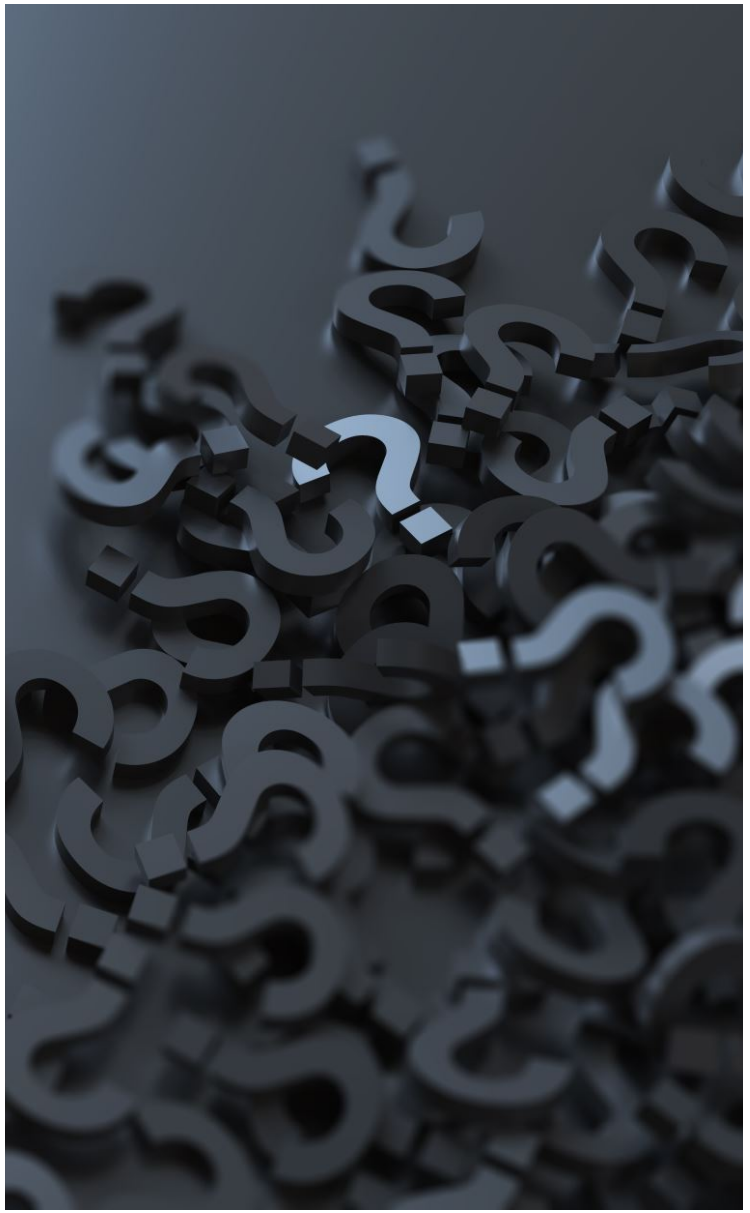
Overview

Simon Coghlan
simon.coghlan@unimelb.edu.au

Learning Outcomes Module 3

At the end of this module, you should be able to:

- Explain why ethics and moral philosophy are important to AI
- Describe key ethical frameworks like utilitarianism, virtue ethics, care ethics, principlism
- Apply ethical frameworks to case studies involving AI



Outline

- Introduction to ethics
- Overview of lecture reading: Thilo Hagendorff: *"The Ethics of AI Ethics: An Evaluation of Guidelines"*
- Why do we need moral philosophy?
- Ethical frameworks – toolkit
- Case study: using the ethical frameworks
- Tutorial: apply the frameworks to a case study for yourselves

Can we trust AI? What good or bad might it bring?

Examples of possible benefits

- improve efficiency in our everyday lives
- minimize repetitive human drudgery
- decrease human bias in decision-making
- aid new scientific discoveries
- better address crime, poverty, hunger, disease, find missing people, improve food delivery, humanitarian crises, help environment

Examples of possible harms

- increase discrimination against vulnerable/oppressed groups
- amplify and reinforce existing inequalities
- reduce employment
- harm the environment
- Reduce privacy
- increase power asymmetries
- increase risks of warfare

THE UNIVERSITY OF MELBOURNE

Reading

Thilo Hagendorff

"The Ethics of AI Ethics"

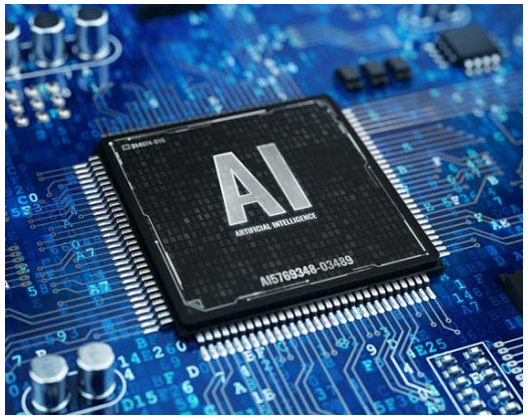
Simon Coghlan

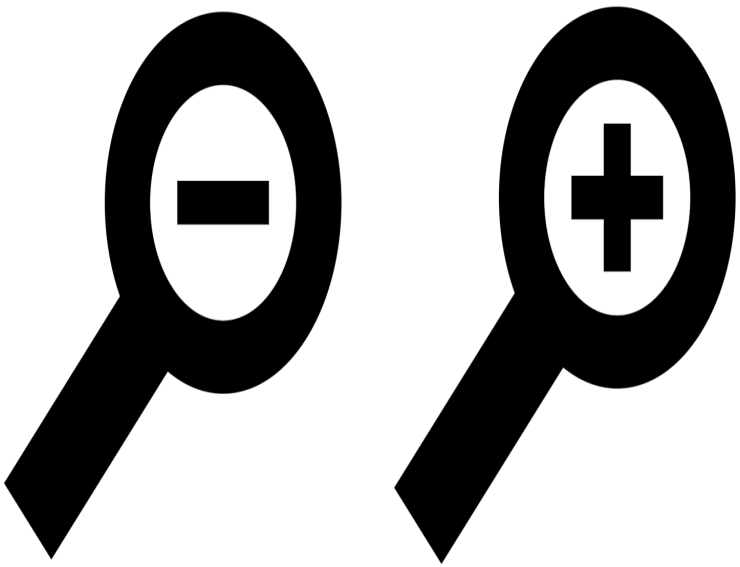
The Ethics of AI Ethics: An Evaluation of Guidelines

Thilo Hagendorff[1]

Received: 1 October 2019 / Accepted: 21 January 2020 / Published online: 1 February 2020

AI





Ethics: what is it good for?

- Is ethics useful?
- Are ethics guidelines useful?
- Or is regulation more important for ensuring good outcomes and holding people accountable?
- Should ethics be taught to all computer scientists?

DeepMind Ethics & Society

22 major AI ethics guidelines

Highly cited principles

- Privacy
- Fairness/justice
- Accountability
- Safety
- Explainability/Interpretability

Less/non cited principles

- Autonomy
- Diversity in AI field
- Whistleblower protection
- Public education about AI risks
- Dual use
- Political abuse
- Social costs
- Public–private partnerships
- Environmental costs
- Existential threats
- Machine/robot ethics

Forces influencing the march of AI

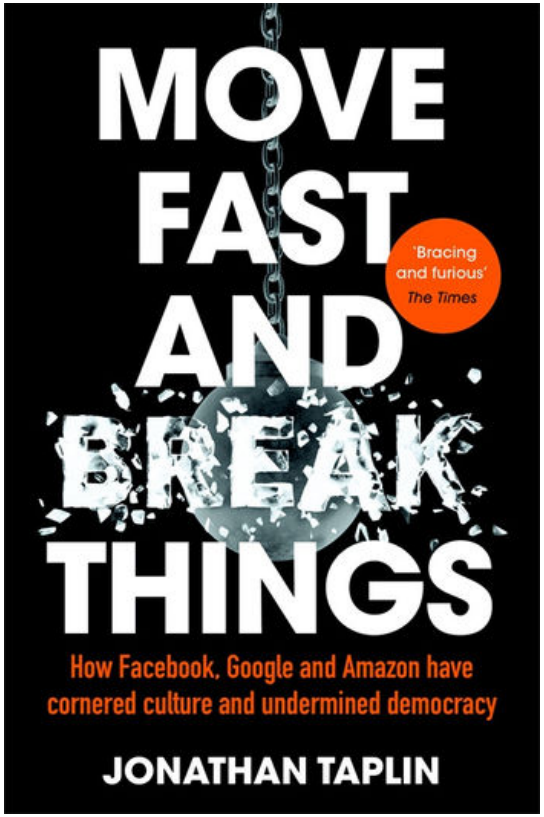
- Highly cited: more amenable to **technical solutions**
- Cf. less 'algorithmic' ones based in wider social context

- Reflect male domination of field (including the guideline authors)?
- 80% of the professors at world leading universities are male
- perhaps more feminist modes of thought required?

- Corporate money e.g. fund AI research
- Corporate power e.g. Facebook
- Military backing
- State e.g. surveillance
- = capacity for lot of harm as well as good

Cutthroat competition

- AI superpowers: USA, EU, China ----> AI arms race
- *"Move fast and break things"* cf. **AI4People** or **AI4Social good**
- Competitors seen as enemies
- Economic logic
- **BUT:** Ethics does not favour competition at all costs
- Rather ethics generally favours: collaboration, cooperation, trust
- But! Ethics "plays the role of a bicycle brake on an intercontinental airplane"



However

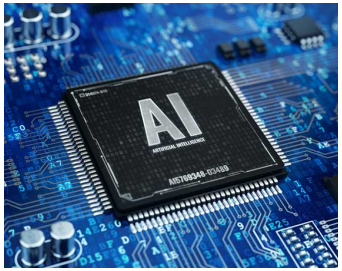
- Individual ethical values can have an effect
- Google working on computer vision for Project Maven (AI military drones)
- Microsoft working on data science and and AI for ICE (Immigration and Customs Enforcement) which was separating children from parents

- Individuals with technical skills can have impact
  e.g. Joy Buolamwini & Timnit Gebru's work on racially biased facial recognition

Broad ethical frameworks

- Deontology – rule based
- Utilitarianism – consequence based
- Virtue ethics – character based
- Ethics of care – relationship based
- Explore in module

- AI practitioners need to consider the short and long term moral implications of their work

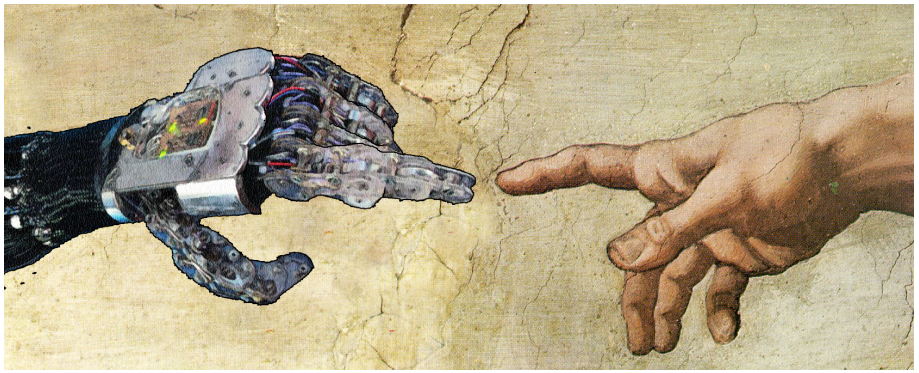


THE UNIVERSITY OF MELBOURNE

Brief Introduction to ethics

Simon Coghlan

AI and ethics



- **Philosophy** – ultimate, ‘big’ questions
- **Philosophy of Mind**
- **Moral Philosophy**

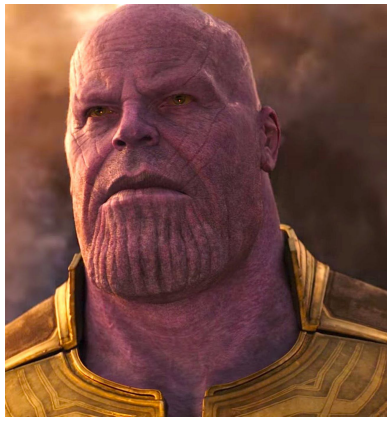
- **Digital ethics**
  - Computer ethics
  - Robot ethics
  - Machine ethics
  - Data ethics
  - AI ethics

What is ethics?

- Pause: what do *you* think ethics is?
- Ethics involves a BIG question: *How should we live?* (Socrates) – meaning of life
- Ethics is about values, morals, good and bad behaviour, right and wrong

1. How should I act? (What standards of behaviour should I adopt?)
2. What sort of person should I be?

Egoism



- ‘I ought to do only what benefits me’
- Ethics is about respecting and looking out for others

Relativism

- No *universal* right and wrong
- **Relativism**: Right/wrong just *means* what a given culture believes is right/wrong
- E.g. human rights
- Problem: Does right/wrong mean those things?
- People can disagree with their own or others' cultural beliefs
- Give reasons to each other; reasons can be *strong or weak*
- So: the relativist doesn't show that there are not universal rights and wrongs

Suttee

However…

While ethics is a rational process and there possibly *can* be things that are universally good and bad, and right and wrong answers to ethical questions :

- Ethical answers not always clear cut
- No textbook with the ethical answers
- We are free to disagree with each other
- We can respect each others' perspectives
- We can respect cultural diversity and insight
- Socrates: We can better understand ethics and and test our ethical beliefs through open-minded reflection and dialogue with others
- e.g. in tutorials!

Why should I be ethical?

- Society expects it
- To be a successful team player
- To gain others' respect
- To avoid guilt and maintain integrity
- To live a fulfilling life
- Because I care about others
- Just because it is the right thing to do
- Because: "the unexamined life is not worth living for human beings"

Socrates

THE UNIVERSITY OF MELBOURNE

Ethics frameworks

Simon Coghlan

Ethics frameworks

1. Utilitarianism – consequence based
2. Deontology – rule based
3. Virtue ethics – character based
4. Ethics of care – relationship based

5. Principlism – simplified integration of 1-4

Deeper values *underlying* our decisions, including for AI

Images herein: Wiki commons and unsplash

Utilitarianism

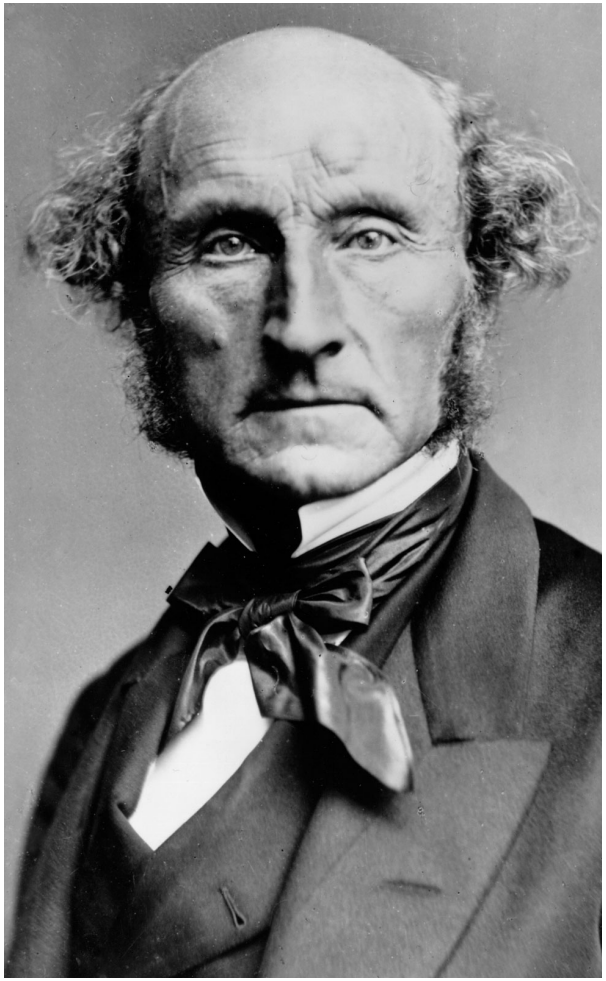
- Consequentialism: consequences alone determine action's rightness

- *"Greatest-Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness"* (John Stuart Mill)

- Consequences: happiness or wellbeing
- Interests: Harms and benefits can be psychological, physical, emotional, social, economic, spiritual

- Right action: Best overall state of affairs
- **Maximises net wellbeing or interests**

- *AI:* Consider ALL the consequences
- Their MAGNITUDE
- And PROBABILITY



Deontology

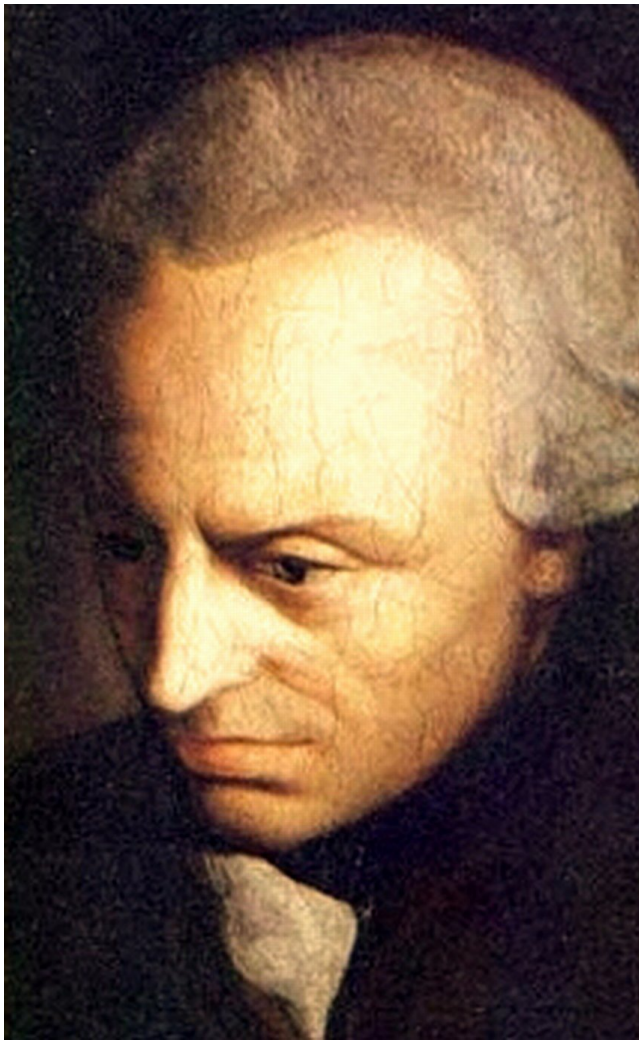
- Rule-based
- E.g. Keep promises, tell truth, act fairly, show gratitude, don't harm, do good, improve yourself

- Right is *not necessarily* what produces best state of affairs

- **Categorical Imperative:** *'Always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end'* (Immanuel Kant)

- Human dignity
- Respect autonomy – moral choosers

- Strong or absolute duties eg never knowingly jail an innocent person



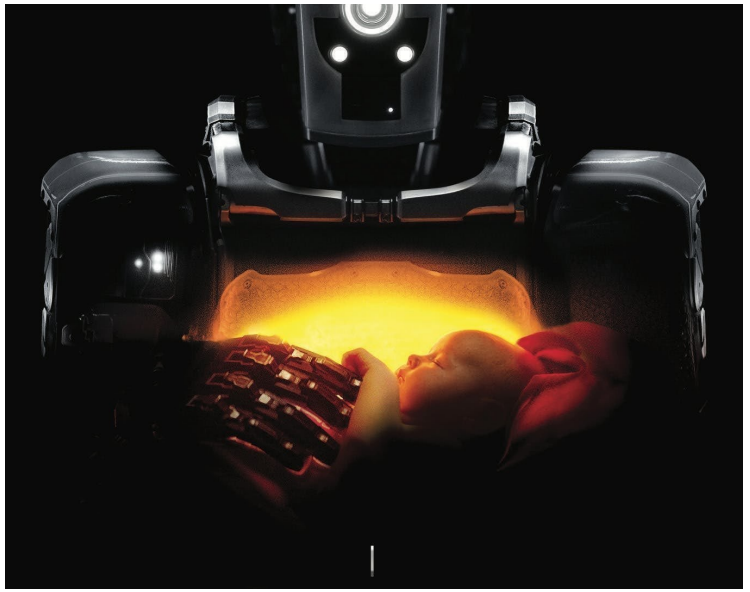
Virtue ethics

- Focus on character: dispositions involving emotion and action
- **Virtues:** compassion, courage, justice, benevolence, honesty, kindness, gratitude, self-control, charity, loyalty, trust
- **Vice:** intolerance, malice, reckless, selfish, blindly loyal, gullibility
- Right = how a virtuous person would behave
- The *right amount* of feeling and the *right* action for a given situation
- E.g. Timnit Gebru/Joy Boulamwini: Stood up to Big Tech - called for fairness, transparency, diversity in AI development
- →Justice/fairness, honesty, courage

Bloomberg. https://www.bloomberg.com/features/2018-bloomberg-50/

Ethics of care

- Feminist ethics
- Cf. rationalism, abstract principles, severe impartiality, lack of emotion
- Relational and interpersonal – emotions and actions
- Caring, attending, loving, taking responsibility
- Recognising vulnerability, powerlessness



Principlism: 4 principles +1

Derived from medical ethics

1. Non-maleficence – do no harm

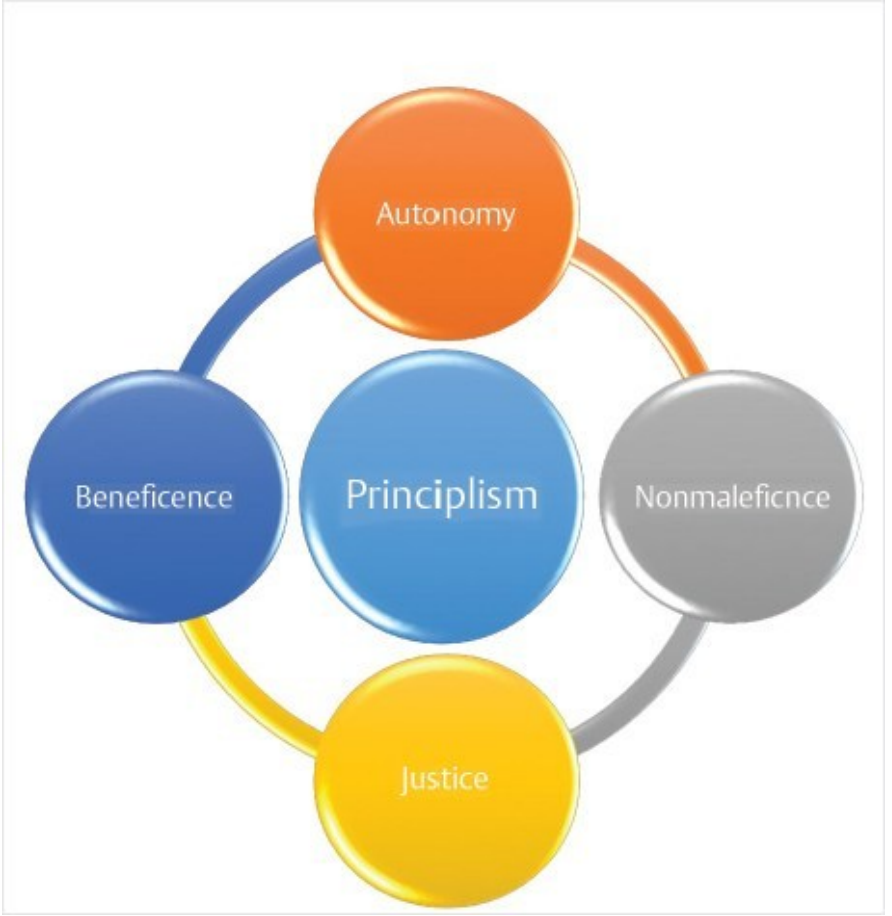
- Predict harm, avoid causing harm, minimize harm, short and long term

2. Beneficence – do good

- Anticipate good outcomes, short and long term

3. Respect autonomy – respect people's values, choices, life plans

- Understand what others' value, don't override their choices, be honest,



Principlism

4. Justice – fairness

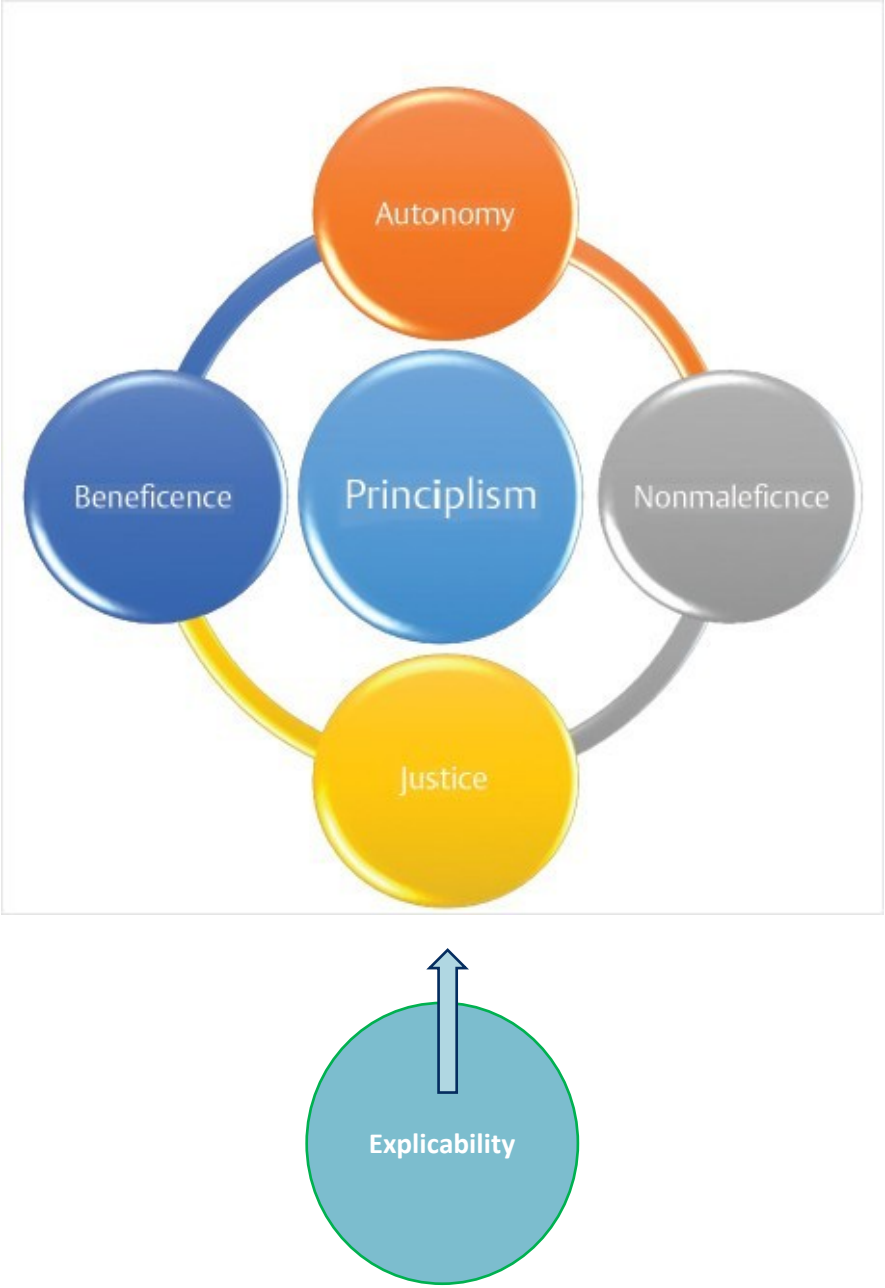
- Distribute benefits and harms fairly, fair processes, don't unfairly discriminate

4+1. Explicability – transparency and accountability (Floridi*)

- Complements the 4 principles
- Ensure those potentially impacted have sufficient understanding of the AI and that relevant people are held to account

Principles: need to be balanced against on another; all are 'equal'

*Floridi, Luciano, et al. "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations." *Minds and Machines* 28.4 (2018): 689-707.



THE UNIVERSITY OF MELBOURNE

Applying ethics frameworks

Simon Coghlan

Ethics frameworks

- Utilitarianism
- Deontology
- Virtue ethics
- Ethics of care
- Principlism

One example: AI headbands
Morally justified or not?

Wall Street Journal video (https://www.wsj.com/articles/chinas-efforts-to-lead-the-way-in-ai-start-in-its-classrooms-11571958181)

AI headbands

- BrainCo Focus headbands
- Primary school in Jinhua, east China's Zhejiang province
- Three sensors, two behind the ears and one on the forehead
- *"It uses electroencephalography (EEG) sensors to measure brain signals and uses an AI algorithm to transmit and then translate signals into real-time focus levels"*
- Monitor students' focus
- Colours displayed on band
- Apparently designed to help students focus through neurofeedback
- Data kept on company server
- Caused controversy and parents objected
- Was the tech company justified in making them? The school in deploying them?

EETimes. https://www.eetimes.com/china-a-headband-for-your-thoughts/

Utilitarianism

Type of consequentialism: Bring about the best **consequences**

Formal rule: Maximize net wellbeing (principle of utility)

Everyone's interests count

Magnitude and probability

- <u>Good consequences?</u>
  - Improve concentration
  - Better education outcomes
  - Assist teachers
  - Reassure parents
  - Refine use of AI - science
- <u>Bad consequences?</u>
  - Physical effects
  - Anxiety and stress for students
  - Worse education outcomes
  - Stressed parents
  - Does it actually work?

Wall Street Journal video (https://www.wsj.com/articles/chinas-efforts-to-lead-the-way-in-ai-start-in-its-classrooms-11571958181)

Deontology

- Rule based
- Key rules
- Is it fair to the students?
- Was it honest to the parents?

- **Even if it produced overall good – e.g. by improving AI monitoring technology and making a scientific breakthrough – that would not necessarily make it right**

- Immanuel Kant: Did it respect autonomy and treat people as ends in themselves, not merely as means?

Virtue ethics

- Look to people of exemplary virtue (good character) for guidance
- What would a **teacher** you respect think of this? Would they sign up to it?
- What virtues are relevant here?
- e.g. honesty, compassion, fairness, courage
- Trust: gullibility vs. cynicism

unsplash

Ethics of care

- Relational, contextual, caring responses: nurturing, loving, taking responsibility, recognizing vulnerability and lack of power

- Suppose you are on the development team for the BrainCo Focus headband

- You might ask:
- How should I respond if I were a **parent** of these children?
- Is developing the BrainCo Headband consistent with care, nurture, or love for these children?

unsplash

Principlism (4+1 principles)

- **1. Non-maleficence** and **2. Beneficence**
- **3. Respect for autonomy**
  - Parents
  - Children
  - Privacy – company stored and owned data
  - Informed consent
- **4. Justice**
  - Aim to improve educational outcomes --> but…ultimately at the expense of the children
- **4+1 Explicability**
  - Transparency to parents, children, teachers
  - Accountability of education authorities, schools, tech company,

Take Homes – Intro to ethics

- Ethical answers are not always clear cut
- But there are better and worse ways of doing ethics
- Ethical Frameworks: utilitarianism, deontology, virtue ethics, ethics of care, principlism
- All frameworks may have something to offer
- You don't need to use every single framework when answering questions
- Socrates: we gain wisdom through interrogating our assumptions and listening to and testing our ideas in dialogue with others e.g. tutorials, seminars