THE UNIVERSITY OF MELBOURNE

Philosophy & Ethics

Module 3: Utilitarianism and Deontology

Simon Coghlan

Simon Coghlan

myself

Learning outcomes

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

- Explain what ethics (western philosophy) is and understand some basic features of ethical thinking
- Describe the ethical theories of utilitarianism and deontology
- Begin to apply these ethical theories to a case study involving AI

Simon Coghlan

Power of AI

- Potential for **great good**
- Potential for **great harm**

Digital ethics team

AI ethics

What are the ethics of using AI to:

- **Determine if someone goes to jail?**
- **Help in allocation of police?**
- **Write your essays for you?**
- **Make medical diagnoses?**
- **Exceed human intelligence?**

Ethics and religion

- Buddhism, Montheism (Ch, Islam, Jud), Confucianism, Hinduism, Jainism, Shintoism, African, Dreamtime etc.
- *Source* of moral authority
- Christians: loving God and each other
- Buddhists: universal compassion, nirvana
- Confucianism: respect for parents and ancestors
- Could the source (also) be:
  - Our own/society's attitudes?
  - Reason?

Principles in AI ethics

Fairness

Safety

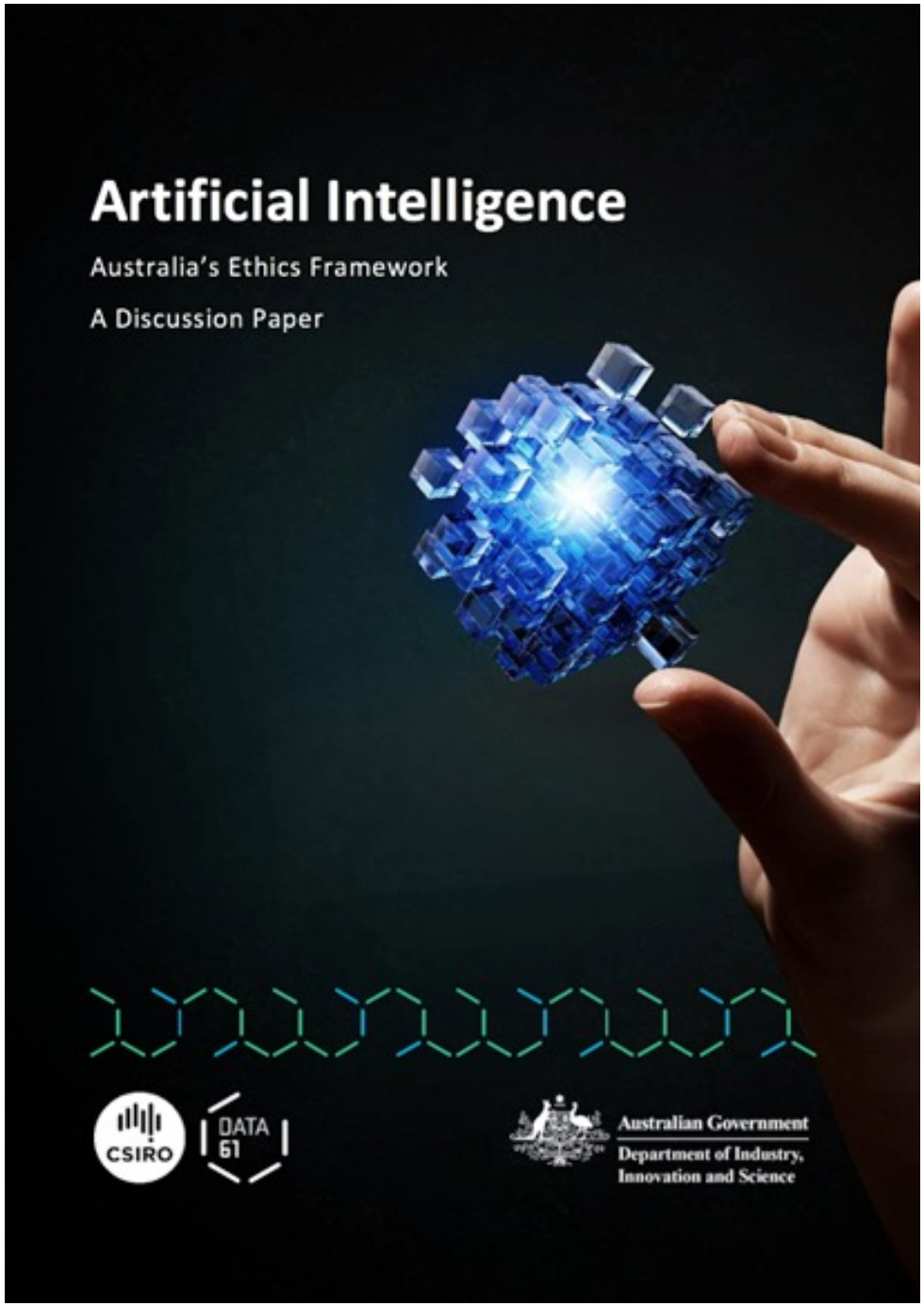
Accountability

Transparency

Benefit

Explainability

Privacy



Simon Coghlan

Ethical theory

Explain *why* we those principles are good or not and how to apply them

Attempt to provide *systematic* and *universal* account that can answer our *practical* questions

Does not rely on religious authority

What is ethics?

- **Socrates:** *How should one live?*
- **Ludwig Wittgenstein:** "*Ethics is the enquiry into what is valuable, or into what is really important, or into the meaning of life, or into what makes life worth living, or into the right way of living.*

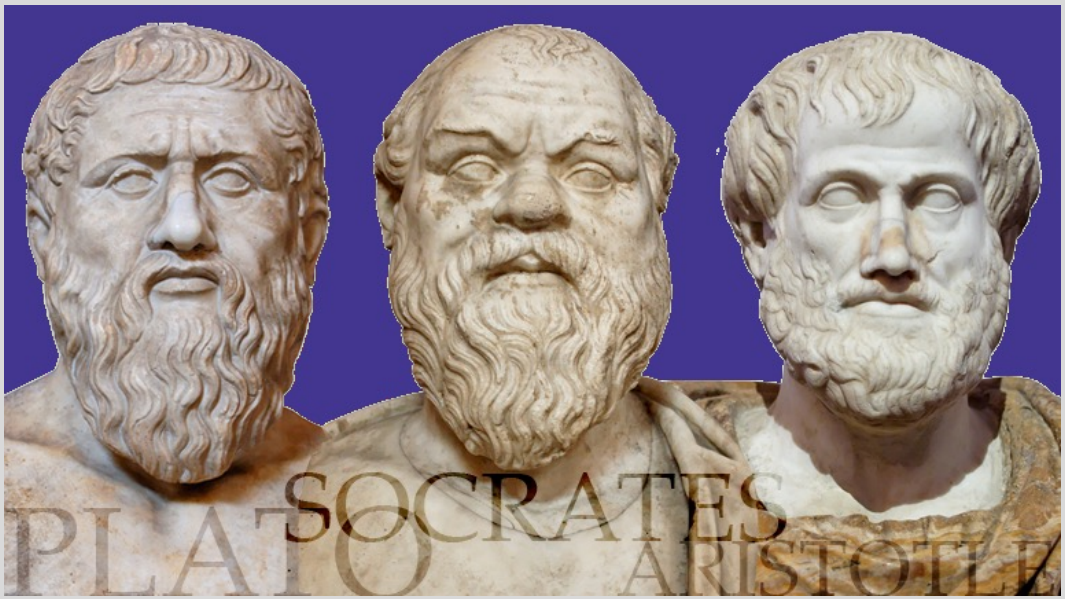
normative how should they do

Ethics about values, morals, good and bad, right and wrong

1. How should I act? What standards of behaviour should I adopt?
2. What sort of person should I be?
3. What sort of professional computer scientist should I be?

Normative versus descriptive (e.g. science)

Ancient and applied ethics - west



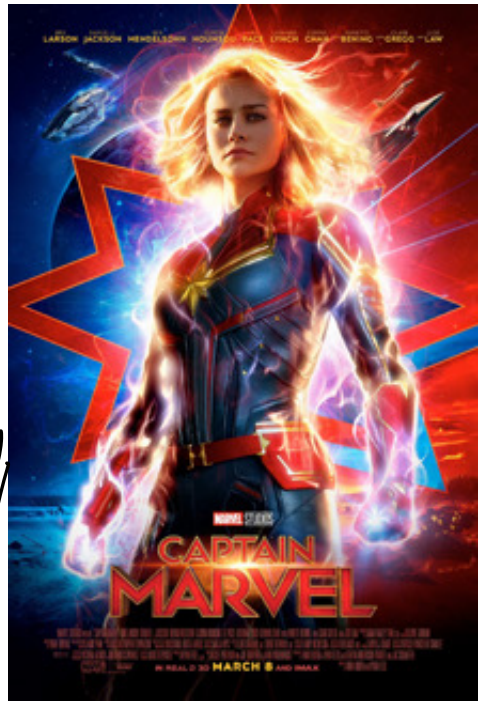
Ancient Greece: Ethics as rational activity

Modern applied ethics – e.g.:

- Is it acceptable to end the lives of suffering people?
- When is OK or not OK to go to war?
- Is it ever justified for a doctor to lie to a patient?
- Can we favor our family over strangers in desperate need?
- Is it right to eat sentient animals?
- Do we have obligations to unborn generations?

Self and others

- Nihilism: no standards (do what you want - no *ought*)
- But: self-centred standards
- Egoism: 'I *ought* to do only does the most good for me'
- Ethics is *about* respecting or caring for others as well as self
- Ethics and human life

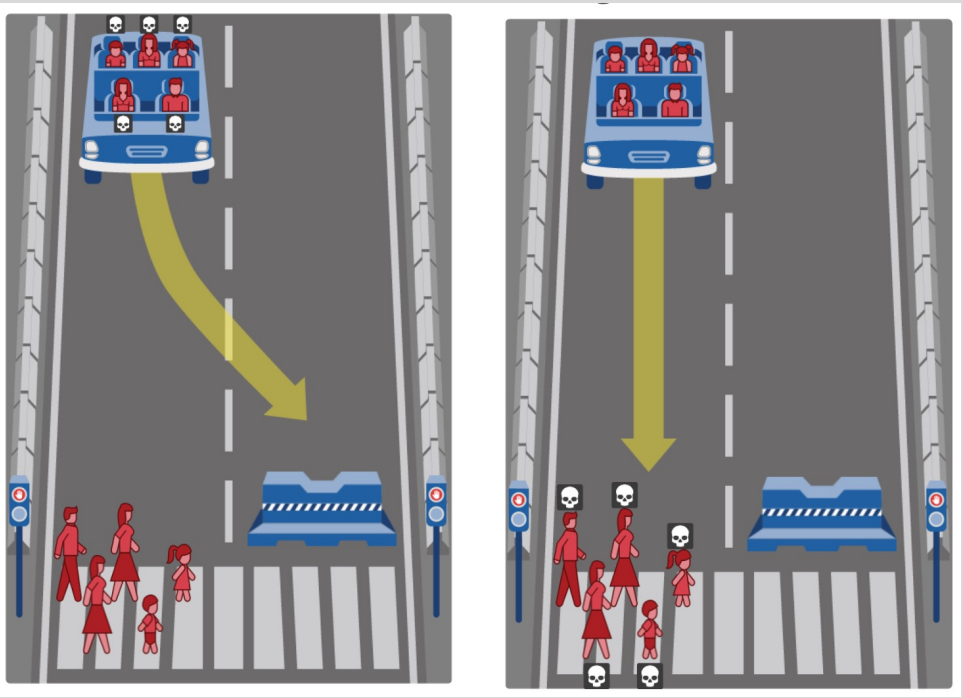


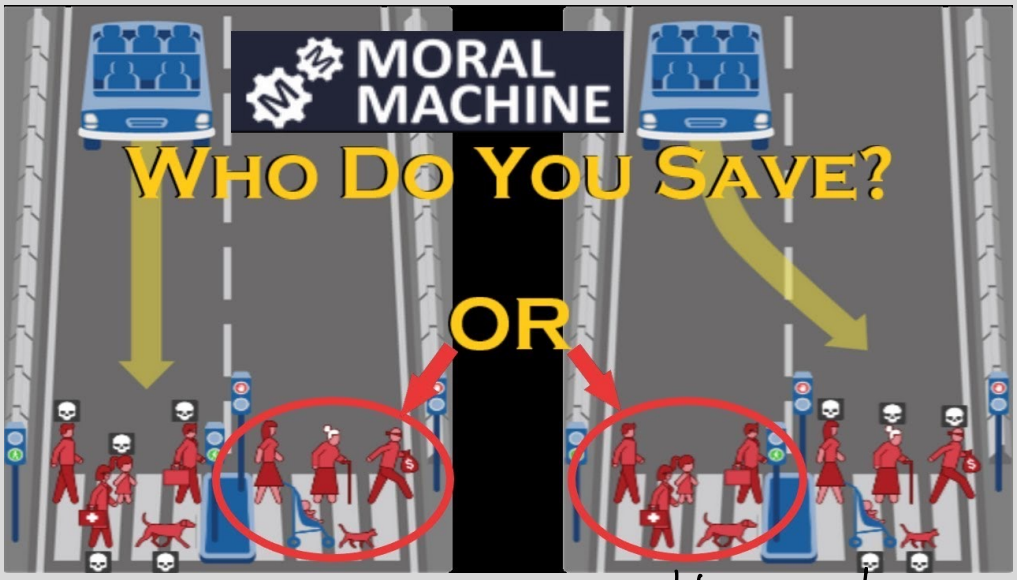
Simon Coghlan

individual ⇒ can disagree within and across cultures

Moral machine

people can disagree within and across cultures

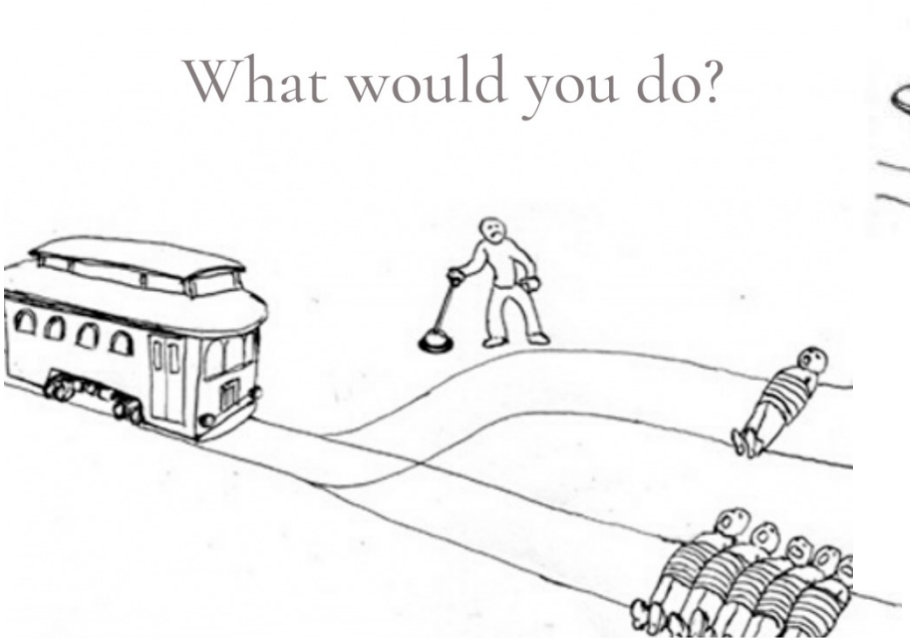


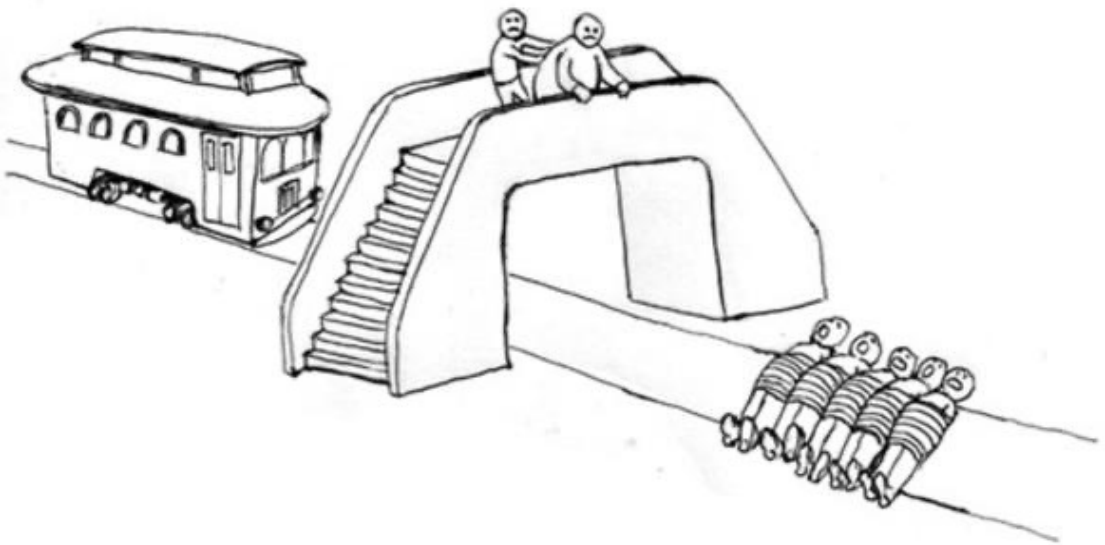


如何 program, based on different cultures ?
(不同文化的价值不同 program 的时候要重点)

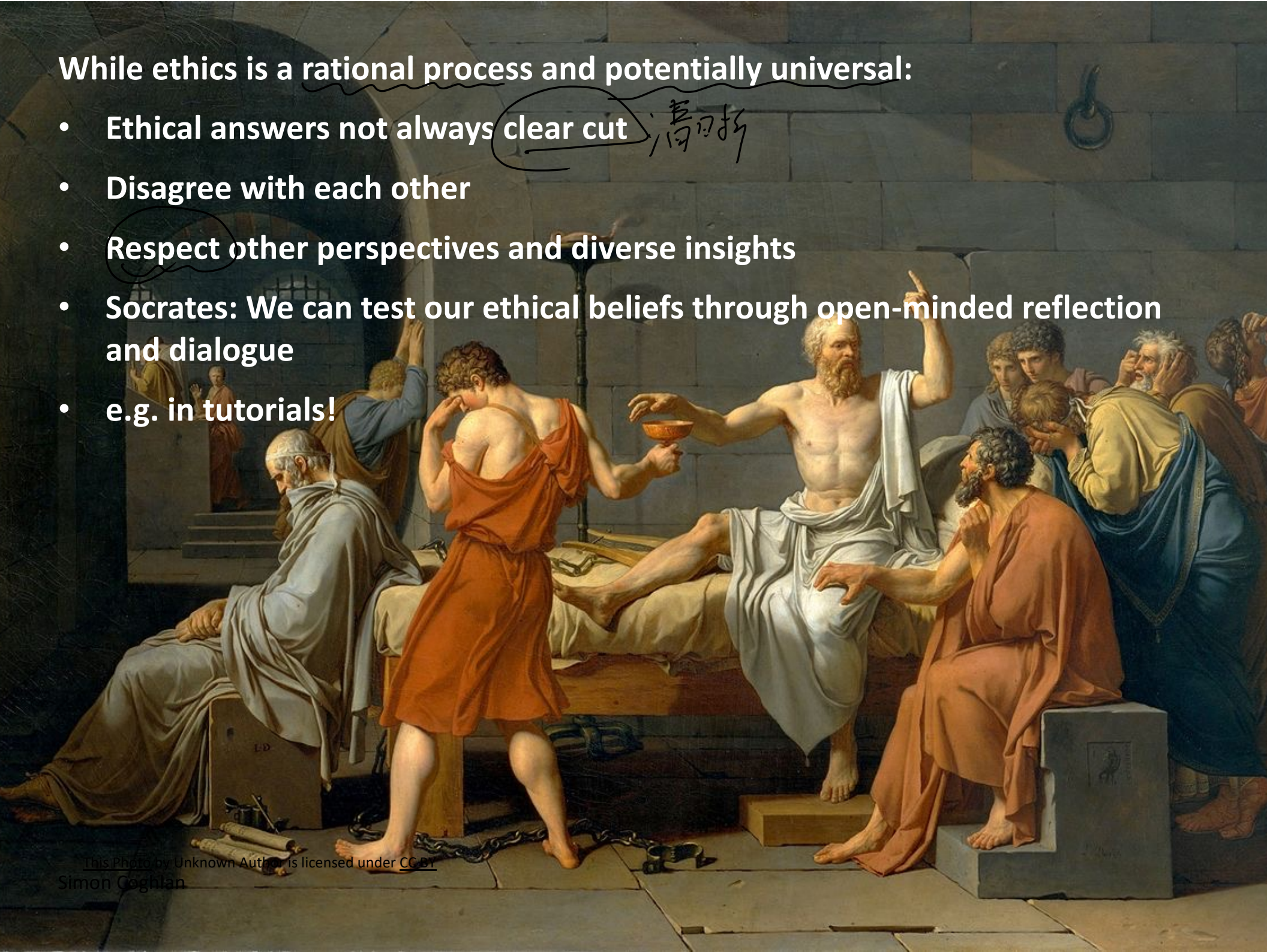
- V. old person vs. child
- Man vs. pregnant woman
- Famous cancer scientist vs. homeless person
- Intelligent animal vs. serial killer

Trolley problem: What would you do and why?





Simon Coghlan



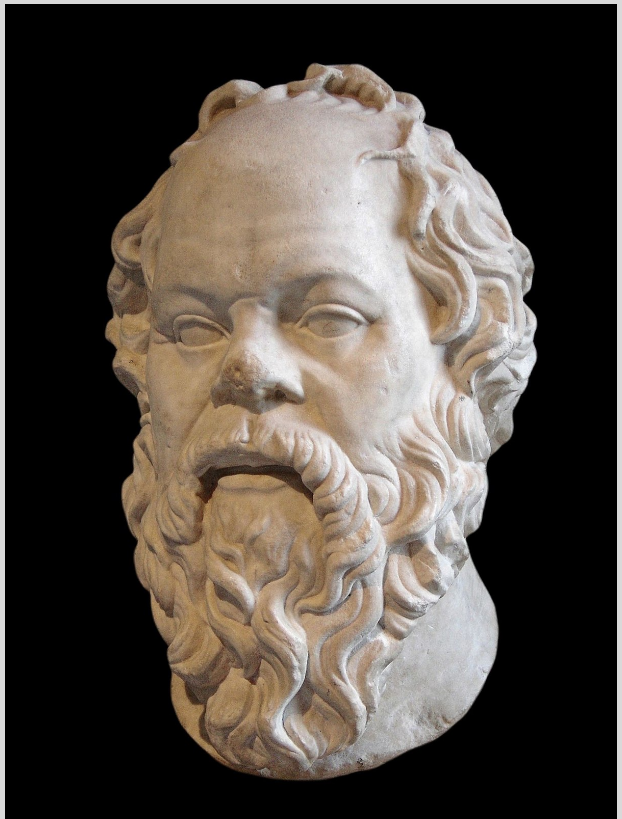
While ethics is a rational process and potentially universal:

- **Ethical answers not always clear cut**
- **Disagree with each other**
- **Respect other perspectives and diverse insights**
- **Socrates: We can test our ethical beliefs through open-minded reflection and dialogue**
- **e.g. in tutorials!**

Why should I be ethical?

- Why not do what I want? Might is right?

- Society expects it
- Successful, co-operative team player
- To gain respect
- Inner peace, avoid guilt
- Just because it is right
- Socrates: "Better to suffer evil than to do it"



Socrates

Simon Coghlan

Ethical theory

Utilitarianism – consequence based

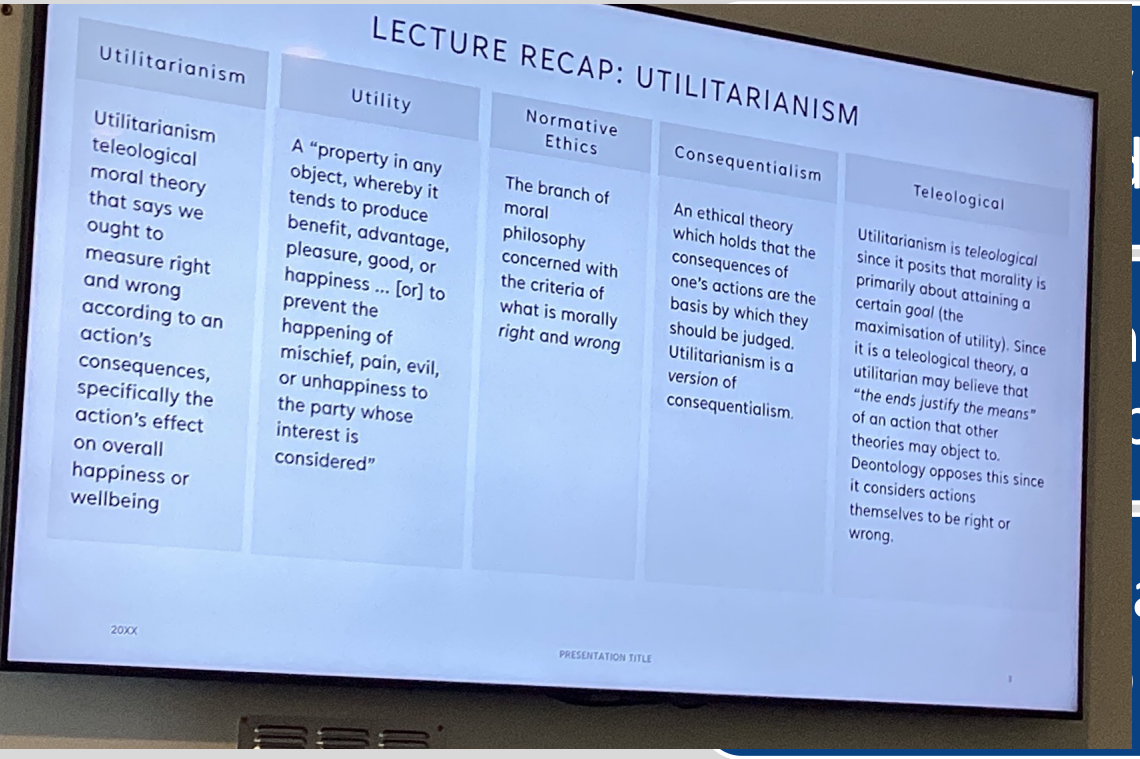
– rule

ics –

based

are –

based



Simon Coghlan

Ethical theories – how to use them

- Why should I care about transparent AI, private AI, explainable AI, etc. ??
- Deeper explanation/justification of what to do, what is right/wrong
- Theories are controversial
- Some philosophers are utilitarians, some deontologists, etc.
- Each theory can be defended and criticized
- But all are well-thought out, complex, taken seriously

Utilitarianism

Deontology

Virtue ethics

Ethics of care

Ethical theories – cont'

- You don't need to embrace one or any of them (though you can!)
- Can adopt them to defend or criticise an ethical perspective (e.g. "research into 'Human-level AI' is bad")
- To do that – need to understand theories and their strengths/weaknesses
- Mistakes:
  - (1) 'Each theory will only give ONE answer about what to do for any issue' – No, it depends how the theory is *applied*
  - (2) 'Different theories will always give different answers about what to do for any issue' – Sometimes, but not necessarily; they may agree about what to do, but give different *reasons*

Utilitarianism

Religious ideas dominated pre-Enlightenment

Christianity strict prohibitions and looked to happiness in the *next* world

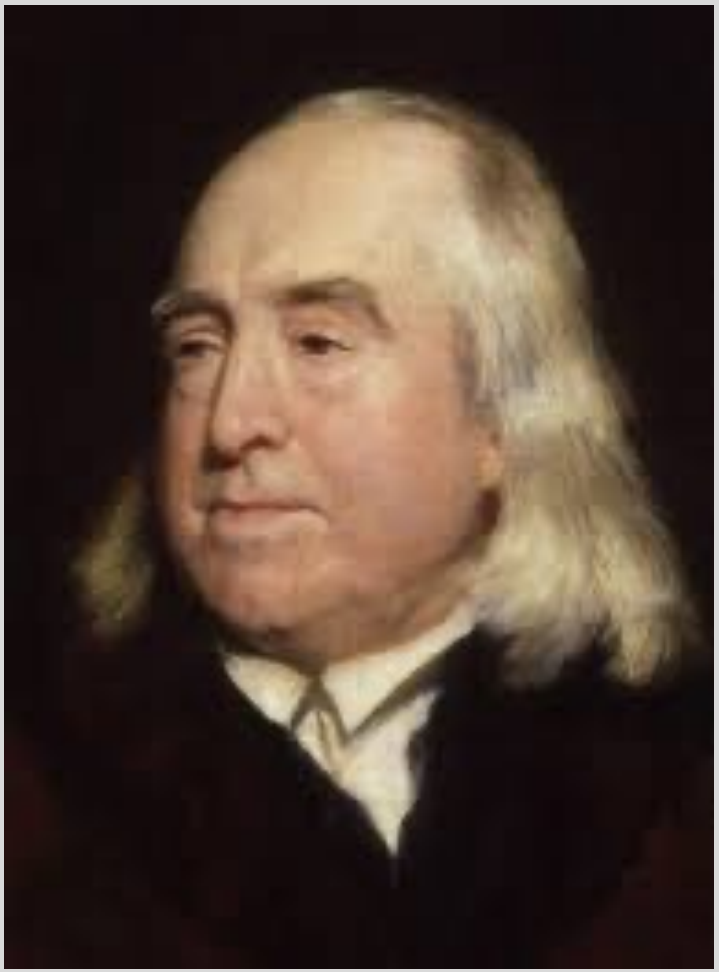
Revolutionary

Against abstract rules "written in the heavens"

Progressive – social change e.g. more liberty

Not opposed to pleasure

Partial return to Greek philosophers (Socrates, Aristotle, Plato) – reason



Bentham (1748—1832)

Utilitarianism

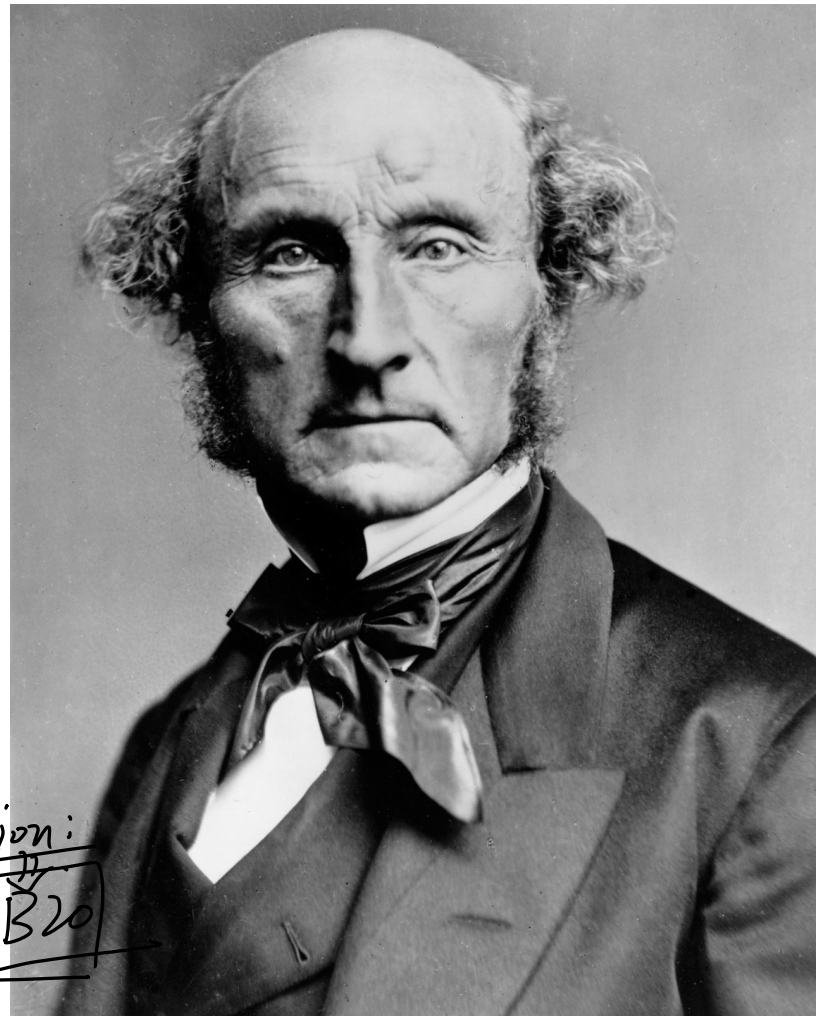
Consequentialism: consequences alone determine action's rightness

What consequences?

Greatest-Happiness Principle: "actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness" (John Stuart Mill 1806-1873)

"Greatest happiness of the greatest number of people"

Principle of utility: Right = maximise total net happiness/wellbeing

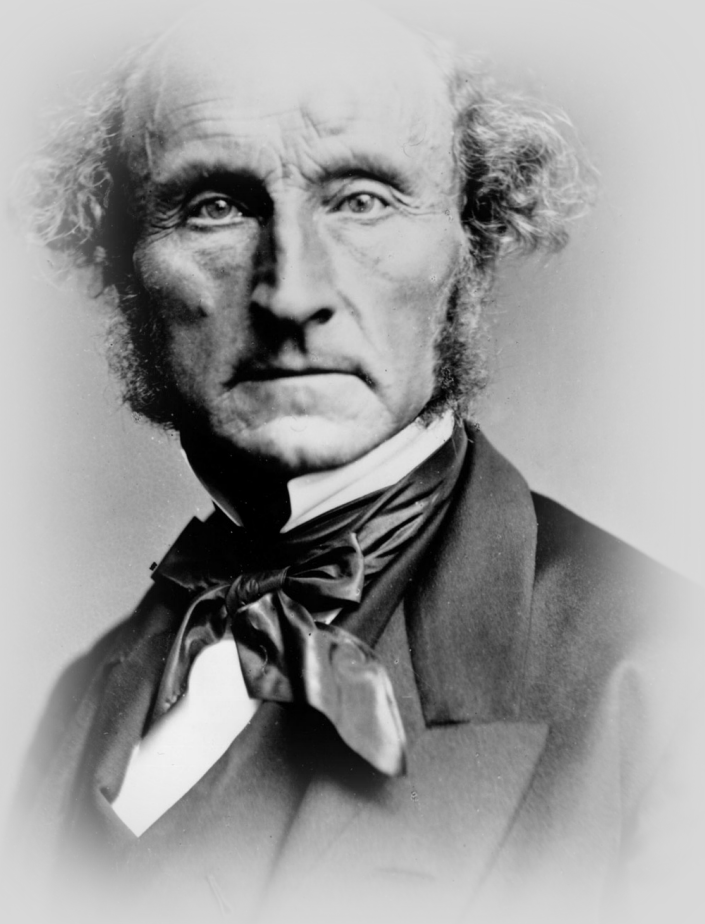


John Stuart Mill (1806-1873)

Simon Coghlan

Utility

- *Teleological* theory: ends (good) determines the right
- Utility: value to individuals
- Interests: Harms and benefits
- Psychological, physical, emotional, social, economic, spiritual
- Benefits (goods) = positive utility
- Harms (bads) = negative utility
- Maximise utility = increase benefits, decrease harms



Hedonism: Bentham vs Mill

Utility: *intrinsically* good and bad for us (not just instrumentally good e.g. money, exercise)

Hedonism: "Two masters": pleasure & pain

Bentham: Pain/suffering is bad

Intensity, duration

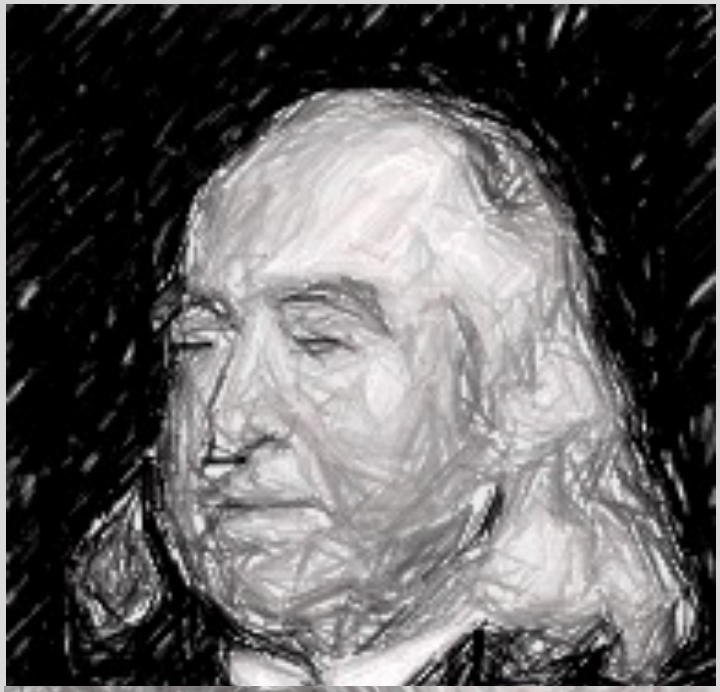
All pleasures are equal

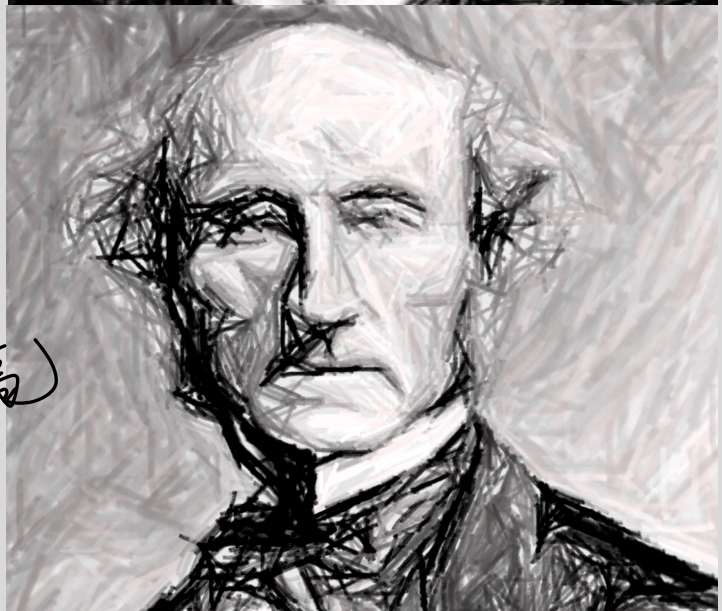
"Pushpin as good as poetry"

Mill: higher and lower pleasures

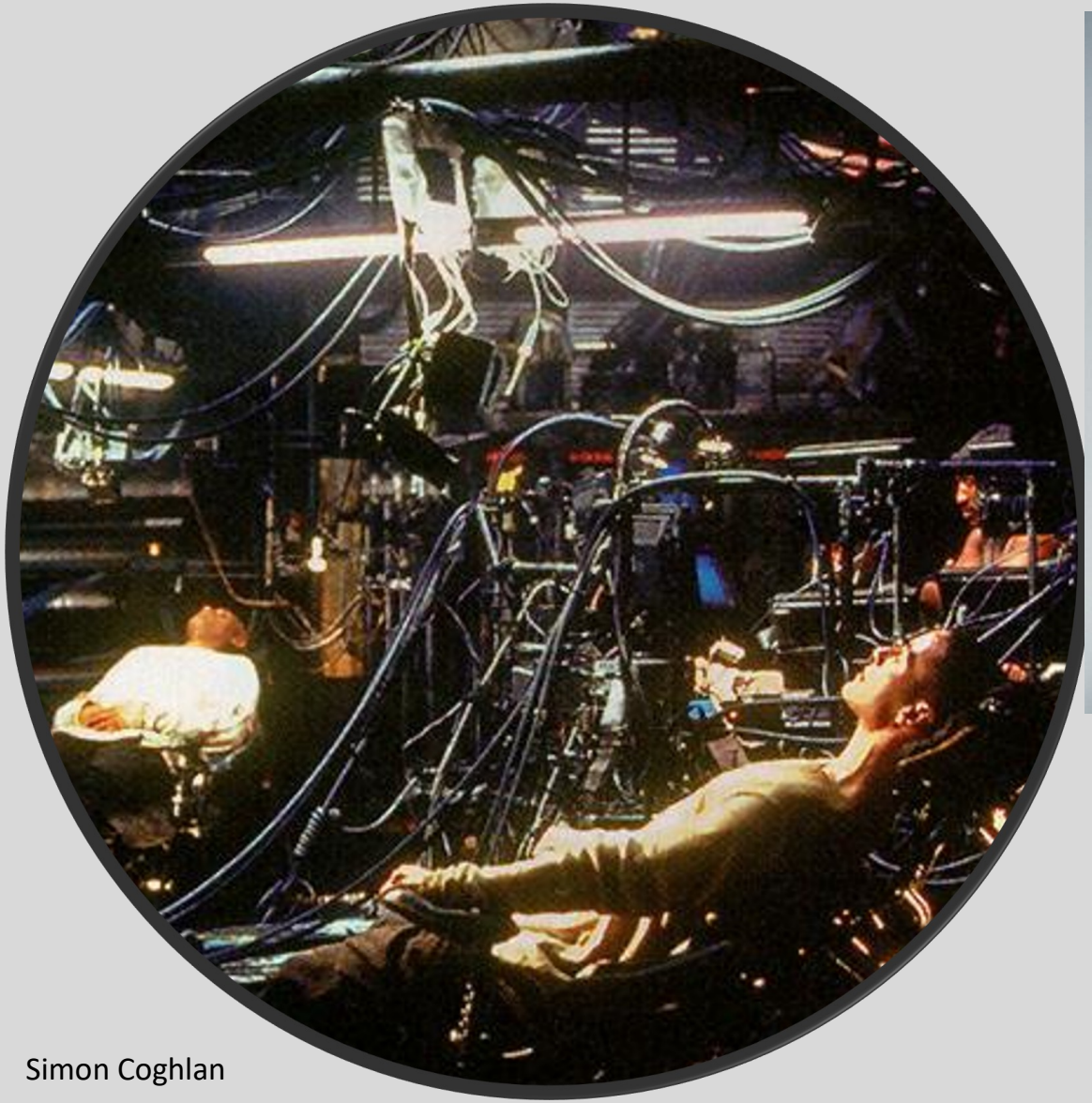
Poetry better than pushpin

Socrates unsafisfied>satisfied fool

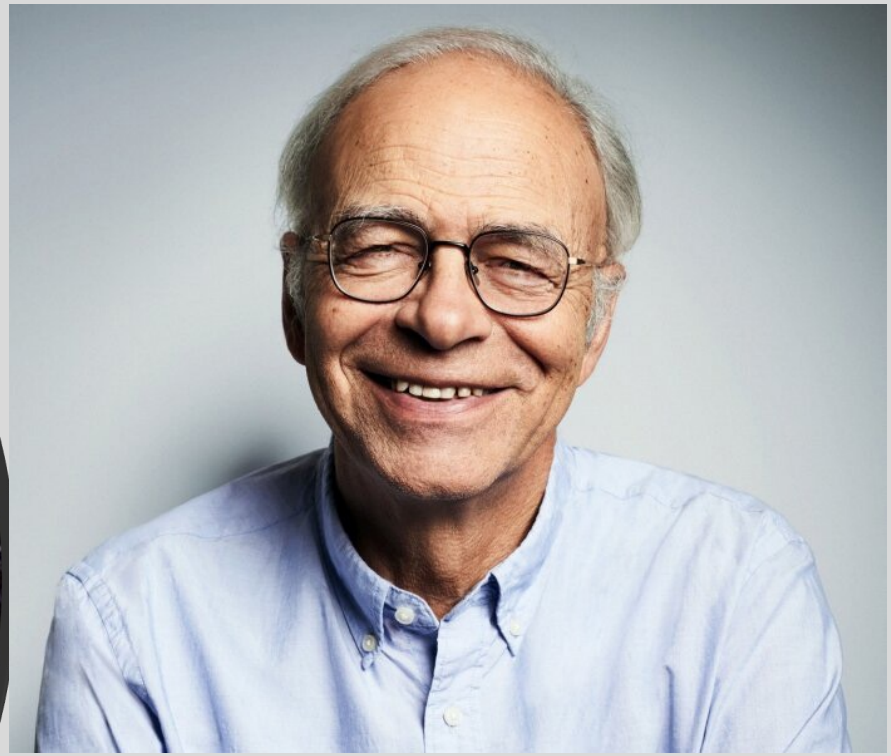




Preference utilitarianism



Simon Coghlan



Peter Singer (1946-)

Intrinsic good = satisfied preferences
Intrinsic bad = thwarted preferences
Happiness = overall balance
Preference calculus

The right

Utilitarianism

Best overall state of affairs
Net (not gross) pleasure/pref
All pleasure/pain matters equally
In that sense: all individuals are equal (including you)
Consider ALL the consequences
Good and bad
Near and far
Magnitude
Probability
Calculate best outcome
Hedonic calculus/preference calculus

Satisfy most interests

Utilitarianism

Utilitarianism

Utilitarianism

Normative ethic
The branch of moral philosophy concerned with the criteria of what is morally right and wrong

Simon Coghlan

Strengths

Surely consequences matter

Surely happiness/wellbeing matter

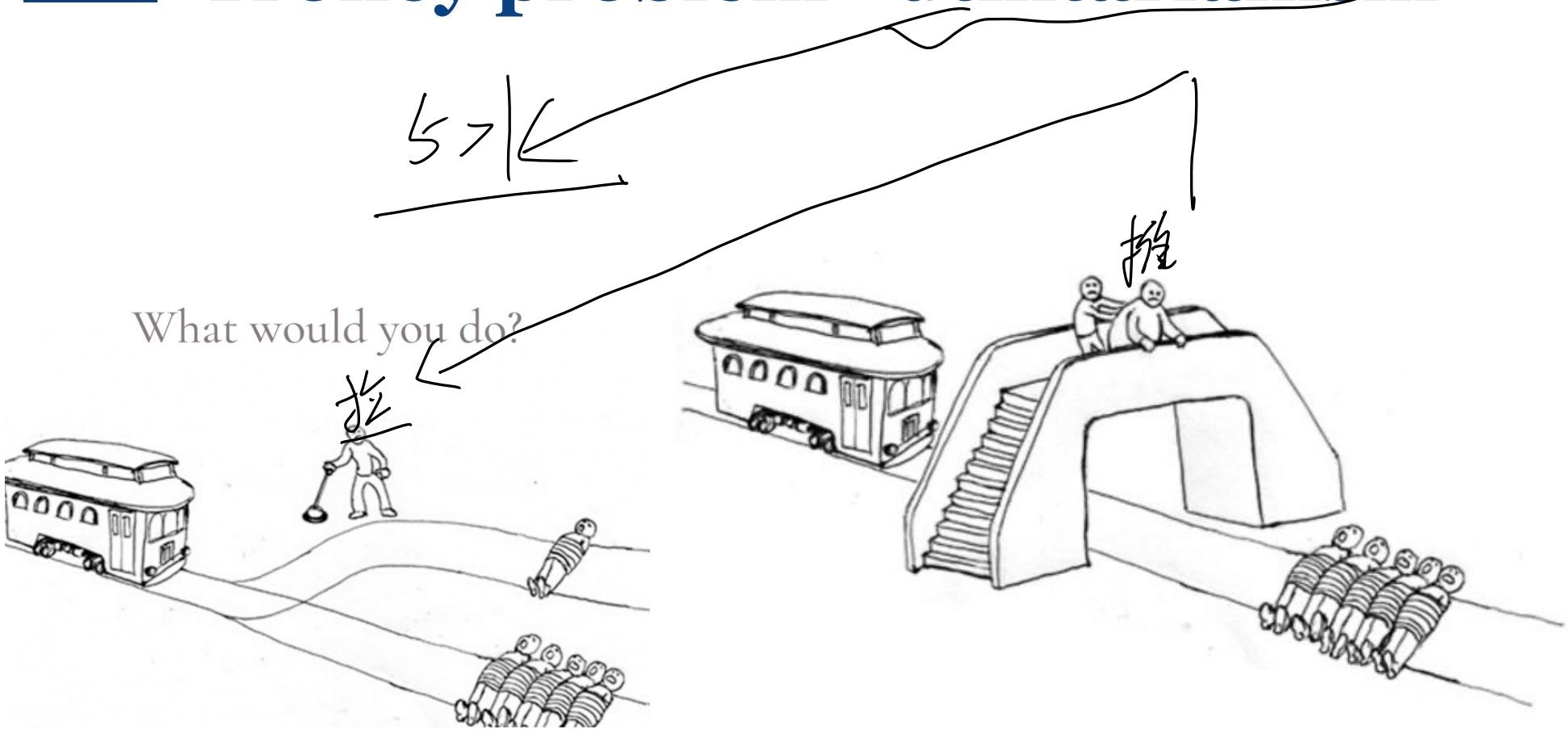
If they matter, isn't more happiness better than less?

Simple, clear decision procedure: Principle of Utility

Rational (cf. accepting authority)

Equality: all count for 1: no class, race, gender, intelligence, etc. favouritism

Trolley problem - utilitarianism



Simon Coghlan

Simon Coghlan

use to justify whether action themselves can be moral

Deontology

attack utilitirism

- Non-teleological (ends)
- Not purely consequentialist
- Deontic = duty, obligation
- Rules or principles
- We learn rules and principles from childhood (be fair).
- These capture best what morality is about
- Can refine and alter rules via reflection and argument

V: the end justifies the means

D: the end doesn't justify the means

Simon Coghlan

Deontology

Examples:

Keep promises

Don't steal

Be honest, don't deceive

Be just and fair

Repay kindnesses

Avoid doing harm

Look after your friends and family

Don't deliberately harm/kill the innocent

Simon Coghlan

WD Ross

1. Fidelity. Keep promises and be honest and truthful.

2. Reparation. Make amends when we have wronged someone.

3. Gratitude. Be grateful when others benefit us and try to return favor.

4. Non-injury (or non-maleficence). Refrain from harming others physically or psychologically.

5. Beneficence. Be kind and try to improve others' health, wisdom, security, happiness, and well-being.

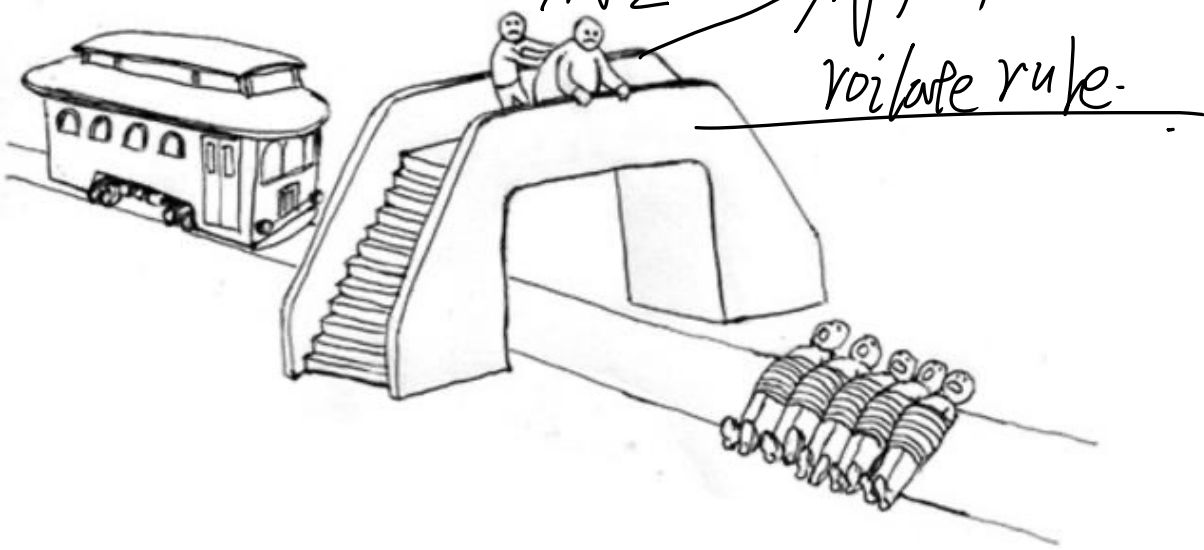
6. Self-improvement. Strive to improve our own health, wisdom, security, happiness, and well-being.

7. Justice. Be fair and distribute benefits and burdens equably and evenly.

https://iep.utm.edu/ross-wd/

Simon Coghlan

Trolley problem - deontology



Simon Coghlan

D attacks U

For some deontologists: Consequences can matter – e.g. generosity requires calculating benefit. *But* – more to ethics than calculating consequences!

Maximising ethic too demanding – give up much of own wellbeing for strangers. Singer - give up high percentage of income.

Evil-doing – pushing the large man

Not as helpful a decision-making procedure as U think – difficult to impossible calculation

Fairness – although each person's similar interests counts equally, maximizing wellbeing can cause apparent injustice

Angry mob example

Angry Mob

Murder in town

Townsfolk want justice

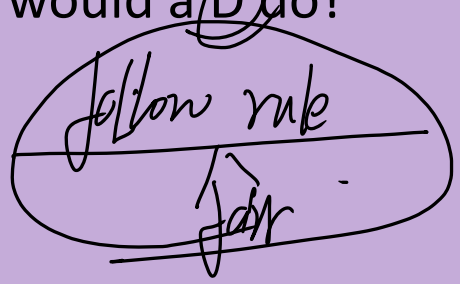
Captured a suspect – a homeless loner

As sheriff, you know suspect is innocent

But if man not hanged – rampage!

What would a U do?

What would a D do?



Deontology: act is good or not.

Simon Coghlan

Prima facie vs. absolute rules

Prima facie rules: 'on the face of it' (WD Ross)

- Rule applies presumptively
- Rules can conflict: need judgement to resolve (e.g. break promise to save a life)
- Some rules win out, others are overridden

Absolute rules: unconditional

- Don't have exceptions
- Don't yield to other rules
- Greatest protagonist: German philosopher Immanuel Kant (1724-1804)

Kant's ethics

A special *kind* of deontology

Absolute duties

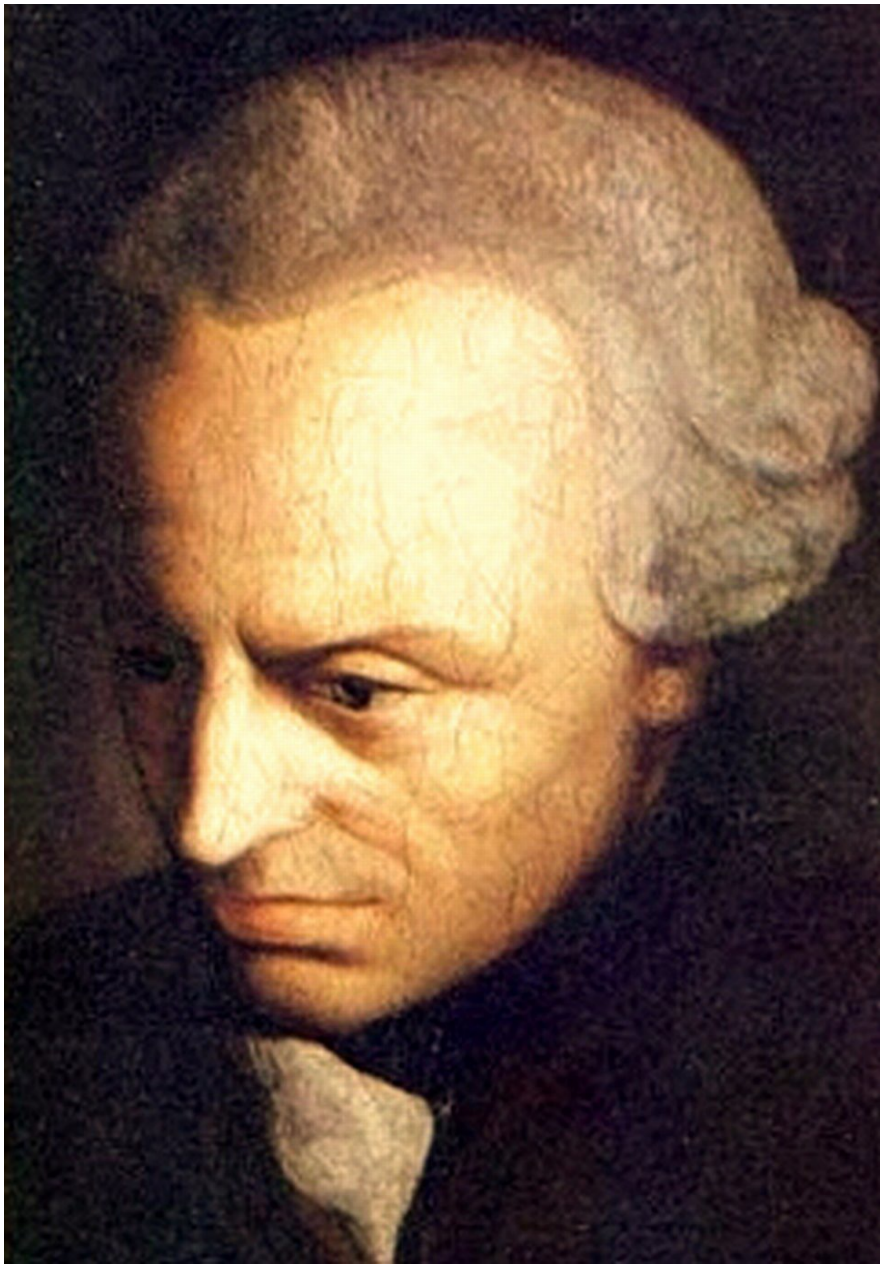
Despised "the serpent-windings" of U

Aiming to produce good consequences ≠ right

Right = a "good will"; acting for right reasons; acting for *duty's* sake

Morality is rational

But rationality is opposed to consequentialism



Immanuel Kant (1724-1804)

Simon Coghlan

Categorical imperative - first

- Actions must be *universalisable*
- We act on rules in ethics
- But moral rules don't just apply to *you*
- Can't make exceptions for ourselves
- ***"Act only according to that maxim (rule) whereby you can at the same time will that it should become a universal law"***
- E.g.: Rule: it's OK for me to lie
- This means: it's OK for anyone to lie
- But if everyone lies when it suits them: truth collapses > lying becomes impossible
- Hence: lying is irrational
- Same for promise-breaking

Categorical imperative - second

Second part of Moral Law

Connected to the first

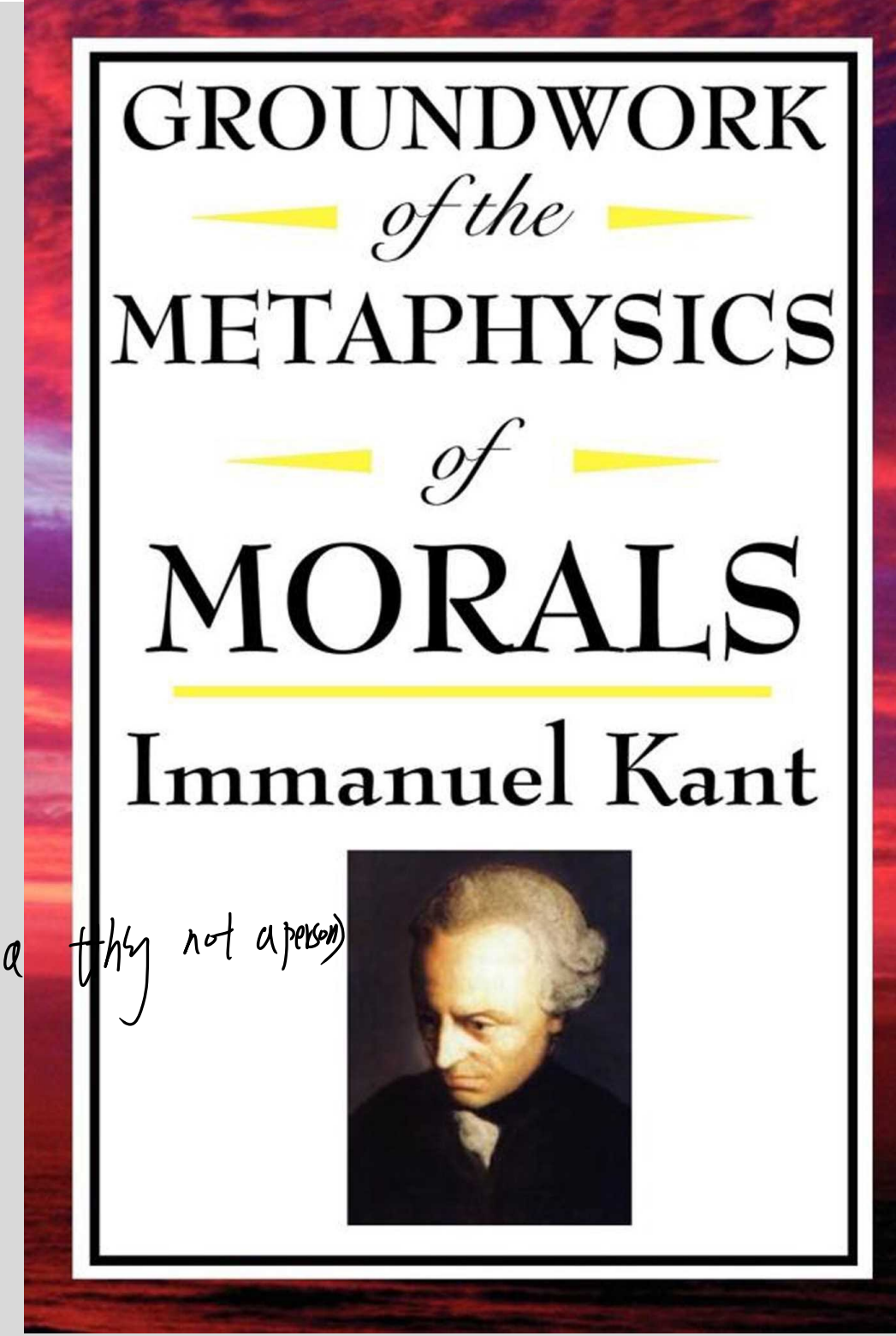
All rational beings can grasp and follow the moral law

They have *autonomy*

"Act in such a way that you treat humanity, whether in your own person or in the person of any other, ***never merely as a means to an end, but always at the same time as an end"***

Ends and merely means

- All autonomous beings are ends-in-themselves
- Sublime and equal dignity (infinite worth)
- Never treat them as *merely* means
- Hire a plumber – use as means
- But *mere* means and not ends-in-themselves: e.g., deceiving, breaking promises, manipulating
- Never allowed
- 'I feel used' — (you treated me like a thing not a person)
- E.g. murderer asking if you have hidden his intended victim – may you lie?
- *Autonomy* must be respected



Modern notion of autonomy

Autonomy is the ability to think for ourselves, plan our lives, act on our values

To respect autonomy, need people's *informed consent* (e.g. collecting private information about them)

We should aim to:

- Respect the autonomy of others
- Try to understand people's values and beliefs and get their consent
- Respect control over personal information
- Remember both powerful and weak have autonomy
- Be honest with people, including when things go wrong (deceiving people disrespects their autonomy)

U and rules

- U: consequences matter more than rules
- But: rules matter *if* they affect consequences!
- E.g. Social rule against punishing innocent – good U rule?
- *Some* rules, laws, basic rights are important (e.g. don't kill, don't torture etc.)
- *But*: must be changed if not best consequences!
- U: 'Morality made for people, not people for morality'

Example:
AI headbands
Morally justified or not?

Simon Coghlan

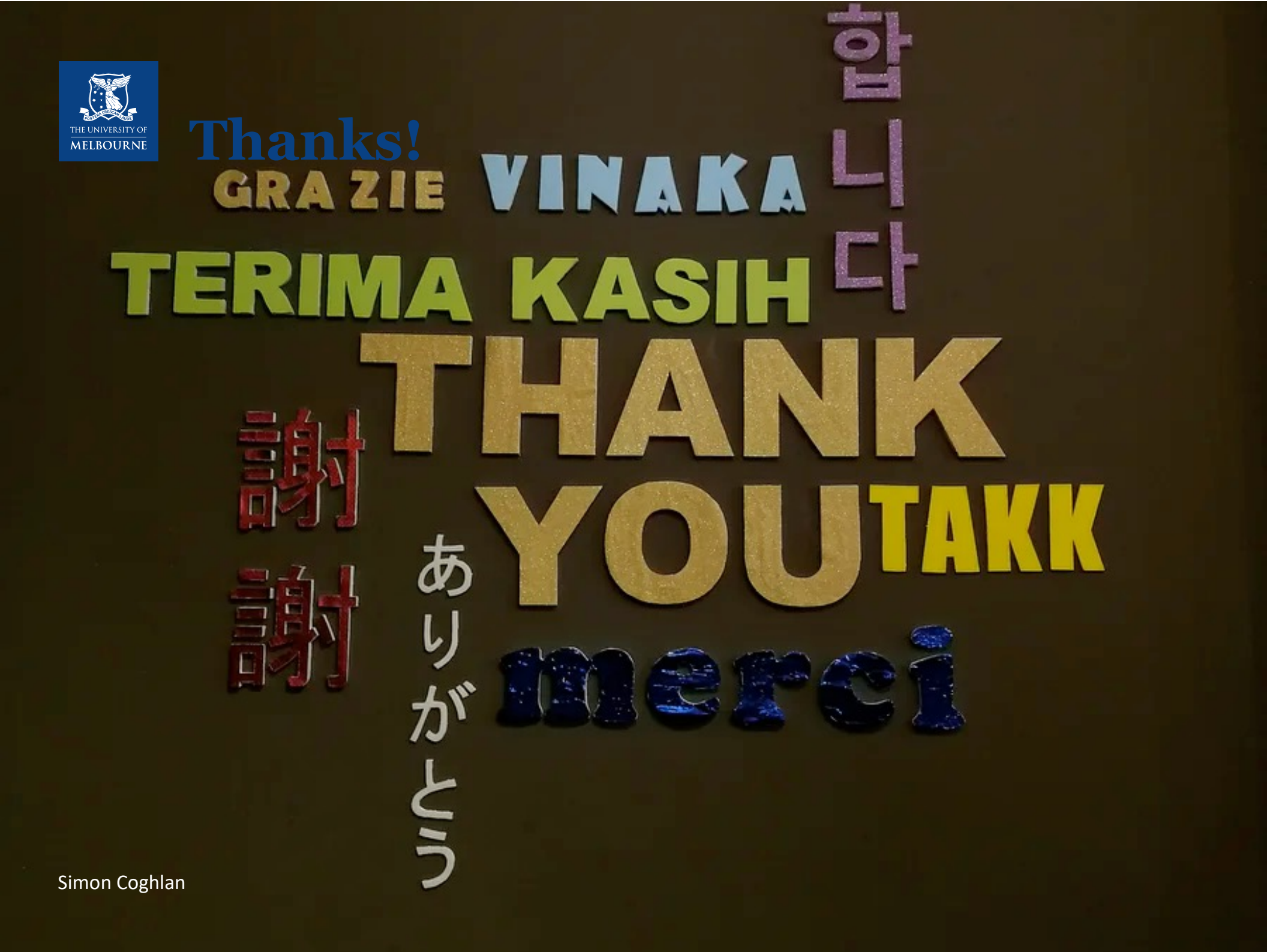
Wall Street Journal video (https://www.wsj.com/articles/chinas-efforts-to-lead-the-way-in-ai-start-in-its-classrooms-11571958181)

AI headbands example

- Study in selected classrooms to collect data, train AI, and improve the headbands
- Uses electroencephalography (EEG) sensors to measure brain signals and AI algorithm to translate signals into real-time focus levels
- Colours displayed on band
- Also designed to help students focus through neurofeedback
- Results of classrooms with and without compared
- Data from students kept on company server and sold to other companies
- Compulsory and students and parents not told about details of the study
- **What might U and Kant say about this?**

Summary

- Nature of ethics
- Religion
- Egoism, relativism
- Moral reason
- Utilitarianism
- Deontology and Kant
- Begin to apply to AI



Simon Coghlan