THE UNIVERSITY OF MELBOURNE

Module 4

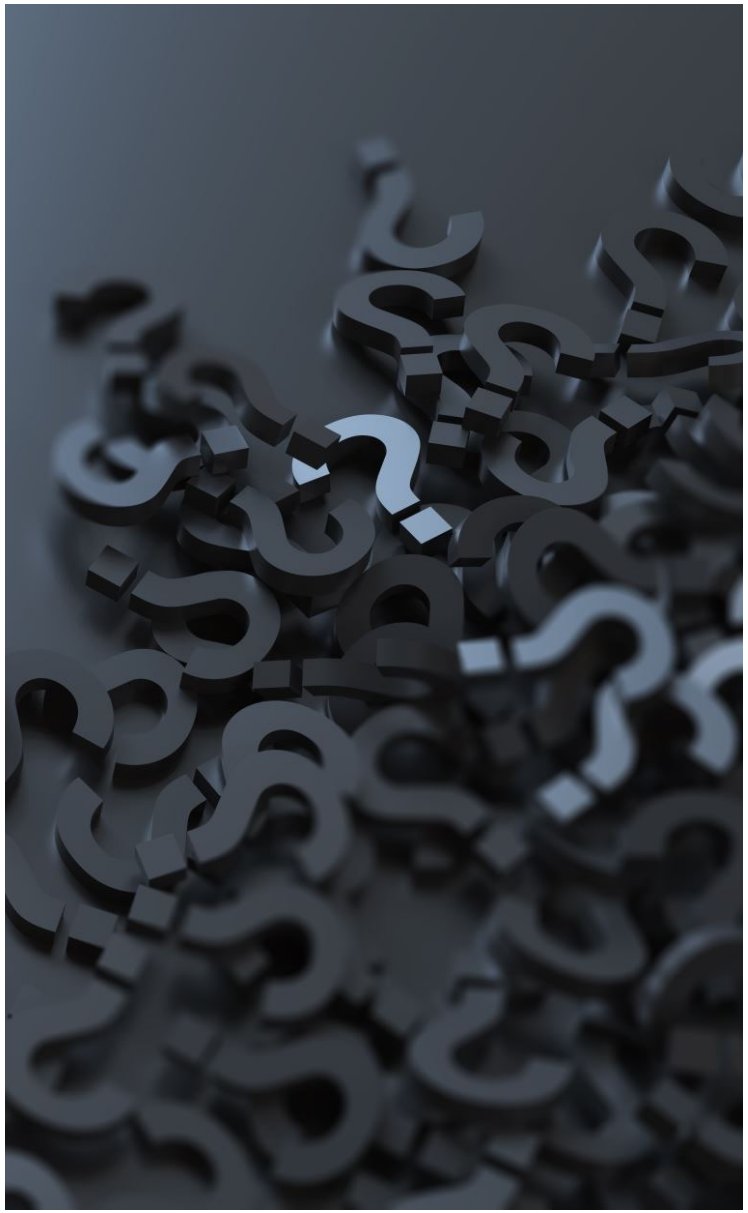
Overview

Simon Coghlan
simon.coghlan@unimelb.edu.au

Learning Outcomes Module 4

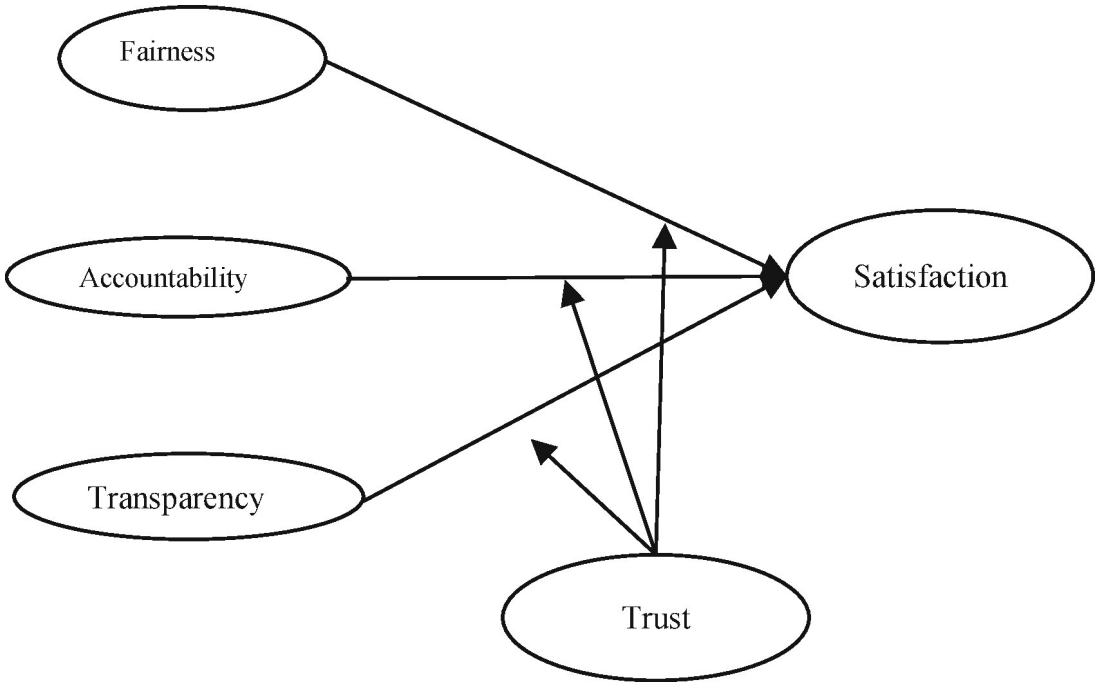
At the end of this module, you should be able to:

- Explain the concepts of fairness and accountability in relation to AI
- Apply the concepts of fairness and accountability to cases involving AI
- Understand elements of the question whether AI can think or understand



Reading 1

- ‘Democratizing Algorithmic Fairness’ (2020) by Pak-Hang Wong
  - Fairness
  - Accountability



- Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277-284.

Reading 2

- ‘Is the Brain’s Mind a Computer Program?’ (1990) by John Searle
  - Can AI think?
  - Can AI be ethical? Should it be held accountable?

THE UNIVERSITY OF MELBOURNE

Module 4

Fairness and Accountability

Simon Coghlan
simon.coghlan@unimelb.edu.au

Justice and Fairness

- Justice-fairness often used interchangeably
- High stakes decisions – AI
  - Bail
  - Sentencing
  - Catching criminals
  - Job applications
  - Welfare
  - Assigning grades
  - Censoring or generating misinformation
  - Diagnosing illness
  - Insurance

Broad definition: ‘Giving each their due’ or ‘what they are owed’.

Treat similar cases similarly; ‘blindness’ to arbitrary differences

Types of justice/fairness

- Equal respect
  - All humans have dignity (Kant – always ends, never merely means)
  - Essentially equal regardless of race, religion, class, sex, gender, sexual orientation, etc.

- Distributive justice
  - Resources, opportunities
  - Need, merit, contracts etc.
  - E.g. Should everyone be given the same career opportunities regardless of talent? Income? Should we help people suffer bad luck? Should we favour people who are morally responsible rather than selfish?
  - Positive/reverse discrimination: more resources/opportunities to disadvantaged and historically oppressed groups

Types cont'

- Procedural justice
  - Fair procedure or process to allocate benefits or harms
  - E.g. only 10 spaces at uni and 15 equally qualified candidates
  - Random, queue
  - Social psychologists: many people care more about being treated fairly by institutions than about actual outcomes
  - Pure procedural justice: no question of need or merit etc. E.g. I have one extra lollipop and two friends – who gets it?

- Retributive justice
  - Impose penalty due to wrongdoing
  - Based on actual guilt and fair procedure (trial)

- Reparative justice
  - Remediation for unfair treatment

Accountability

- Assuming accountability (responsibility)
  - E.g. holding myself to fair procedures and distribution
- Being held accountable (responsible)
  - By external pressures and mechanisms
  - E.g. law, profession, codes of practice, colleagues, elections

- *Who* is accountable? E.g. researchers, engineers, organizations (private and public), deployers, authorities
- *What* mechanisms are the right and fair ones?

THE UNIVERSITY OF MELBOURNE

Reading (part 1):

Fairness

Democratizing Algorithmic Fairness

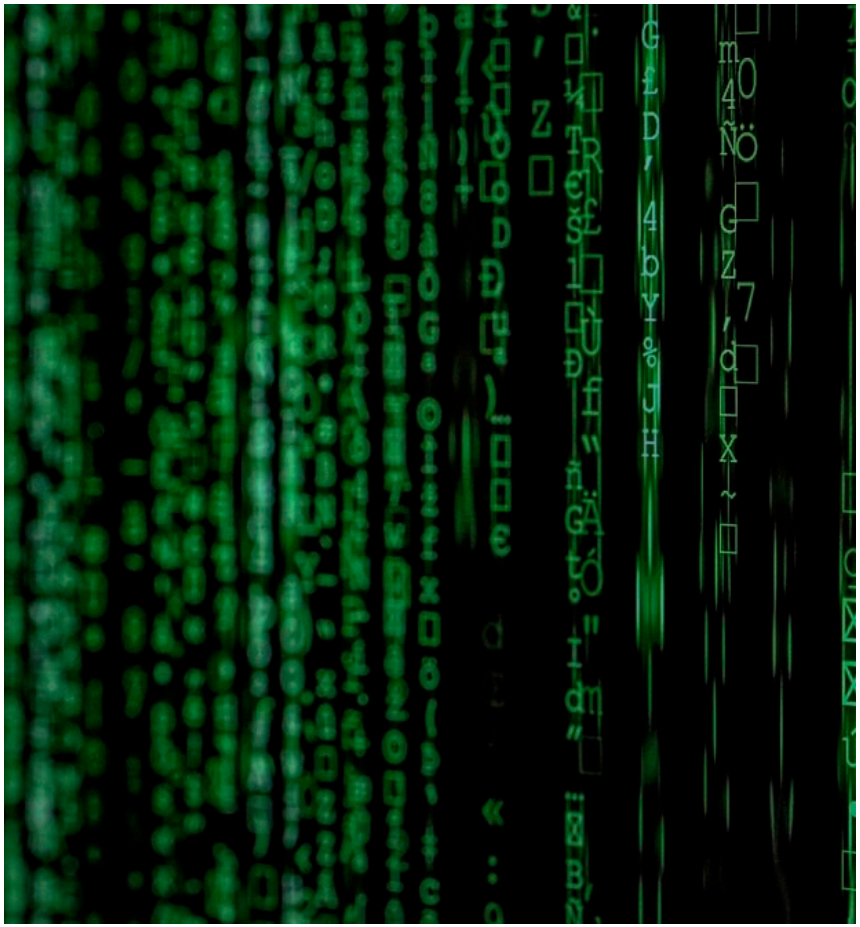
By Pak-Hang Wong

Algorithmic Fairness

- ML algorithms can have embedded bias – unfair
  - E.g. Discriminate against groups unfairly e.g. race, gender
  - Either explicitly or by prox

- *Technical* solutions to minimize unfairness
  - E.g. change inputs
  - E.g. improve processing of dataset
  - E.g. change weighting of false –ves vs. +ves

- Mathematical measures e.g. (Berk et al 2018)
  1. overall accuracy equality
  2. statistical parity
  3. conditional procedure accuracy equality
  4. conditional use accuracy equality
  5. treatment equality
  6. total fairness (1-5 achieved)



But…

- Ethical vs. technical problem: Whether mathematical fairness really is fair depends on the *standard* of fairness adopted. And this standard is disputed.
  - Hence: ethical question and debate
- Also…Cannot always have perfectly fair algorithms due to:
- The Impossibility Theorem: "mathematically impossible for an algorithm to simultaneously satisfy different popular fairness measures"
  - E.g. group parity that unfairly punishes group X who broke the law less often
  - →same treatment and disparate impact
- The Inherent Tradeoff: between fairness and performance
  - E.g. Increased group fairness→decreased accuracy of recidivism prediction for bail
  - Decrease false positives (defendants falsely scored as high risk) but increase false negatives (miss some high risk defendants) (social cost)

ProPublica vs. Northpointe

- COMPAS recidivism algorithm
  - Risk score for reoffending: blacks and whites
  - Determines bail, cf. potentially biased humans
  - May reduce rates of incaceration
- **ProPublica**
  - AI racially biased against blacks
  - *Disparate impact*: black nonoffenders given higher risk scores than white nonoffenders

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

- **Northpointe**
  - *Not disparate treatment*
  - No use of arbitrary differences (race) in algorithm
  - Reoffending rate for blacks and whites equal at each COMPAS scale

THE UNIVERSITY OF MELBOURNE

Model Number: 1

Accuracy

Over all: 68.4%

White American: 71.3%

African American: 65.5%

Disparity

Disparity in Accuracy: 5.8%

Disparity in FPR: 6.1%

Disparity in FNR: 14.9%

False Positive Rate

Over all: 8.0%

→ White American: 5.2%

→ African American: 11.3%

False Negative Rate

Over all: 74.6%

→ White American: 83.2%

→ African American: 68.3%

Ethical frameworks and fairness/justice

- **Utilitarianism:** Fair/just = maximizes net wellbeing, even if some individual's must be made worse off than others. Everyone's similar interests are still considered equal
- **Kant's deontology:** Recognise human dignity and respect autonomy. Treat autonomous agents always as ends, never merely as means
- **Virtue ethics:** Consider what a fair and just person would do
- **Ethics of care:** Consider special relationships and roles and responsibilities that flow from them. Consider impacts on the most vulnerable
- All these frameworks recognize basic human equality
- They may have different/similar views on algorithms that:
  - Reinforce existing disadvantage e.g. increasing policing for some groups
  - Overlook past oppression e.g. that affect algorithmic prediction that results in biases against disadvantaged groups or don't positively discriminate to help disadvantaged groups

THE UNIVERSITY OF MELBOURNE

Reading (part 2):

Accountability

Democratizing Algorithmic Fairness

By Pak-Hang Wong

Pak-Hang Wong

- AI may necessarily create winners and losers – harms and benefits for different people.
- Determining what is fair in high stakes AI is not purely a technical task, but an ethical one
- But: how to ensure this determination is itself fair?
- Recall: no consensus about what is fair + perfect algorithmic fairness is impossible
- Wong: procedural justice
- What mechanism is fairest for holding AI designers and owners accountable?
- E.g. panel of AI ethics experts?
- *Political* mechanism

Accountability for reasonableness (AFR)

- AFR developed from health ethics
- Wong: “ensure decisions are morally and politically acceptable to those affected by algorithms through inclusion and accommodation of their views and voices”
- AFR assumes no totally final and ‘right’ answer: answers emerge through open, democratic, good-faith dialogue and reason-giving involving stakeholders
- Not just developers and researchers determining what is fair AI
- Four conditions

Four conditions for AFR

1. **Publicity condition:** Decisions about algorithmic fairness and their rationales must be publicly accessible, transparent, and understandable to non-technical people.
2. **Full Acceptability condition:** Provide a reasonable explanation of the chosen fairness parameters i.e. give evidence, principles, reasons that fair-minded persons could accept – for all affected stakeholders, especially the vulnerable.
3. **Revision and appeals condition:** Have ongoing (not one-off) mechanisms for challenge and dispute resolution and revision of policies.
4. **Regulative condition:** Have ongoing public regulation of process to ensure conditions (1)–(3) are met.

Northpointe's COMPAS – recidivism prediction AI

1. **Publicity condition:** Explain clearly what measures of fairness will be used to predict re-offending
2. **Full Acceptability condition:** Justify why the chosen parameters and impacts are relevant. Could those impacted accept these reasons? E.g. allowing 'disparate impact' on historically disadvantaged black people while 'avoiding disparate treatment'? Using education level or being the victim of crime - that negatively affects certain racial groups more?
3. **Revision and appeals condition:** Mechanism for those impacted to contest those reasons e.g. vulnerable groups, representatives of broader society
4. **Regulative condition:** Media put spotlight on COMPAS, but no stronger regulation or enforcement

Northpointe did not follow the AFR approach and were not held accountable by public regulation

A final point: should AI be used at all for this purpose?

THE UNIVERSITY OF MELBOURNE

Module 4

AI and Minds

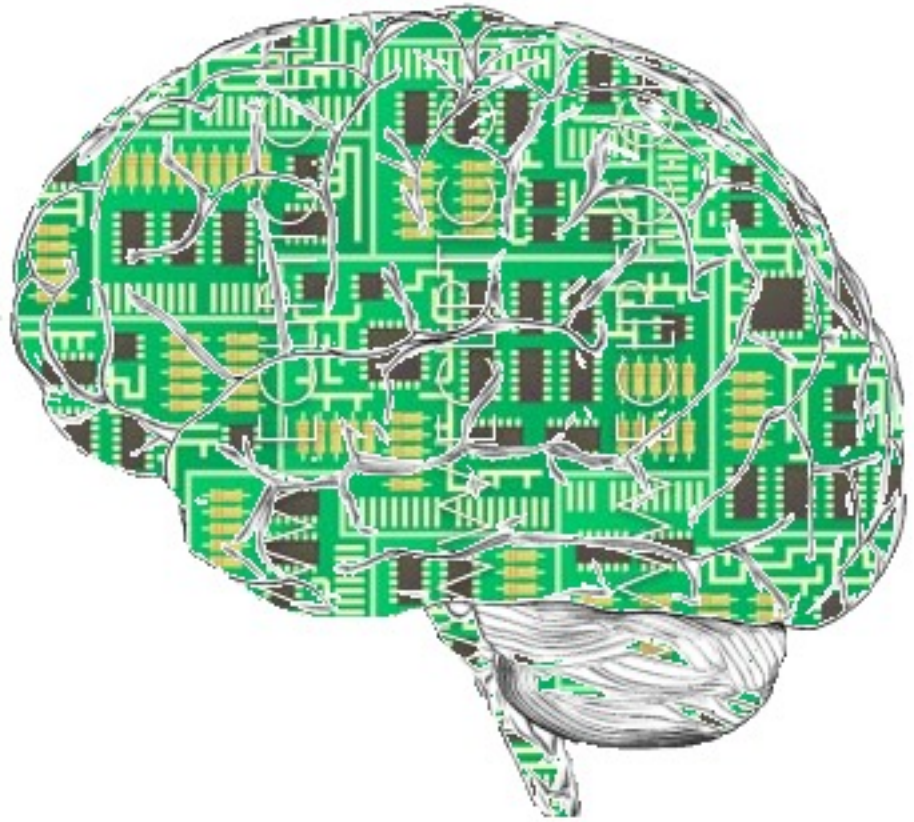
Simon Coghlan
simon.coghlan@unimelb.edu.au

Philosophy of Mind

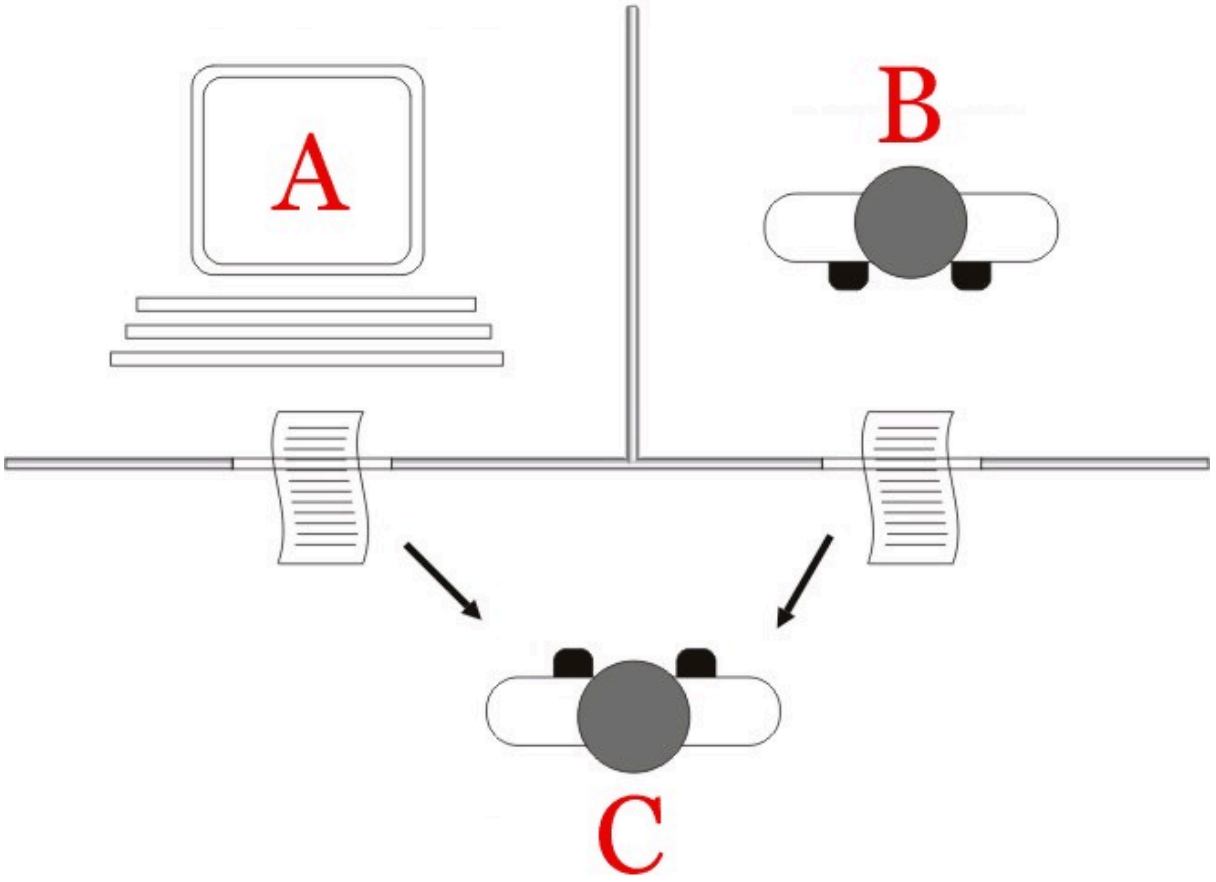
- What is the nature of a mind?
- Is it the same as matter?
- Is the mind the brain?

- What is thinking and understanding?
- What is consciousness?

- Can AI have a mind, think, understand, be conscious?



Turing Test



Thinking like a human = Giving outputs like a human

John Searle

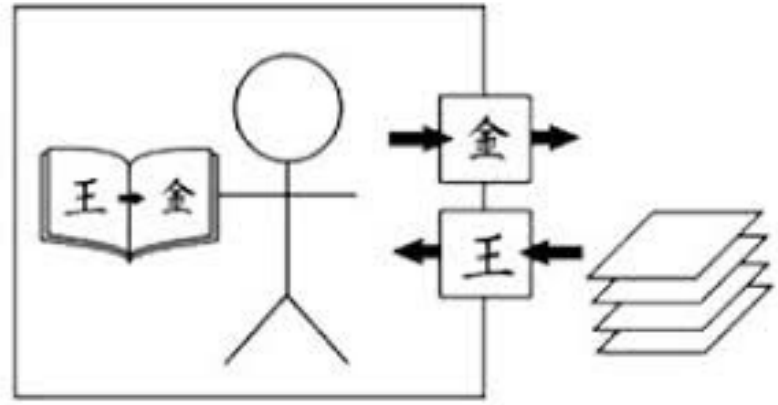
- The brain and human mind is not a computer
- AI that performs computations can never have a mind
- i.e. think, understand, be conscious *in the way humans can and are*
- Turing: thinking is a kind of computation and is independent of the matter it occurs in
- i.e. mind could be instantiated in any material: microchips, vacuum tubes, beer cans
- Strong AI: AI can have a mind
- Weak AI: simulates human thinking in a specific area e.g. chess, medical diagnosis

Chinese Room

Chinese Room cont'

- Rule book = computer program
- Writers of rule book = programmers
- Person = computer
- Inputs(question)→computer→outputs (answers)
- Pass Turing Test
- But – has the person understood?
- No! Just manipulated symbols
- Symbols have no meaning to the person – in the relevant sense, they are not thinking at all

Chinese Room cont'

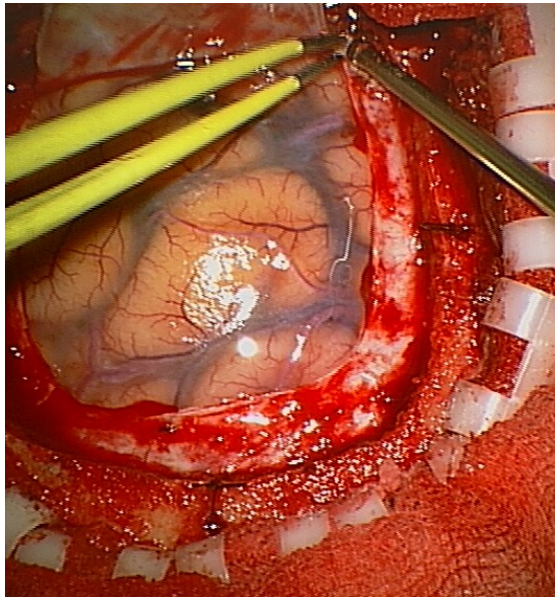


- Computation: Syntax vs. semantics
- E.g. the person/computer has no idea what the symbols are *about*
- Even if minds perform computations, that is insufficient for thinking, understanding etc.
- Would not help to make the system more complex (e.g. parallel processing, neural networks)…
- …or increase its Turing Test performance and so make it seem more 'human' etc. (e.g. making it appear quick-witted , funny, knowldgable etc.)
- AI *simulates* but doesn't *duplicate* minds

Responses

Biased against non-biological systems (cf. 'wet slimy stuff')

- Searle: must be something about neurobiological systems (brains) that is vital for causing mind states
- But: Searle is open to non-biological minds that are not merely computational



The whole system is the mind (person+book+basket of symbols)

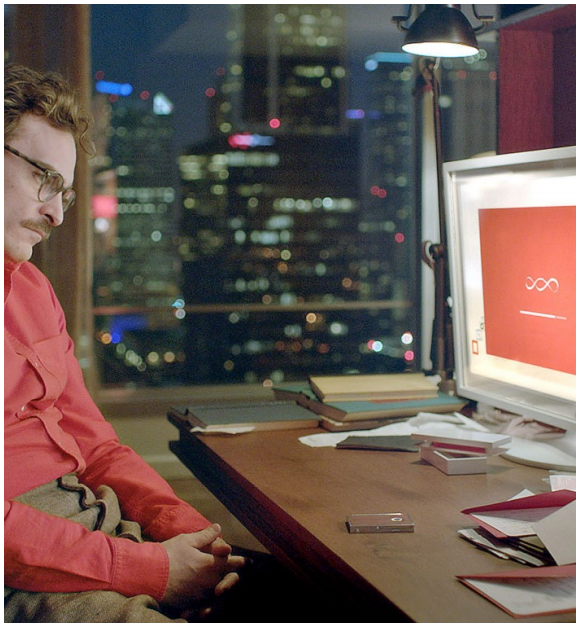
- Searle: person memorises the rules etc.

- **Robot Reply**

  Computers are isolated from the world – but robots are causally connected to it: AI that can sense and respond in physical space and to objects may be different to Searle's CR e.g. could play chess *knowingly*

- **Turing Test revisited**

  While it seems counterintuitive that person in CR understands, it seems counterintuitive to deny that a fully conversant future AI (e.g. movie Her) understands



Sum up

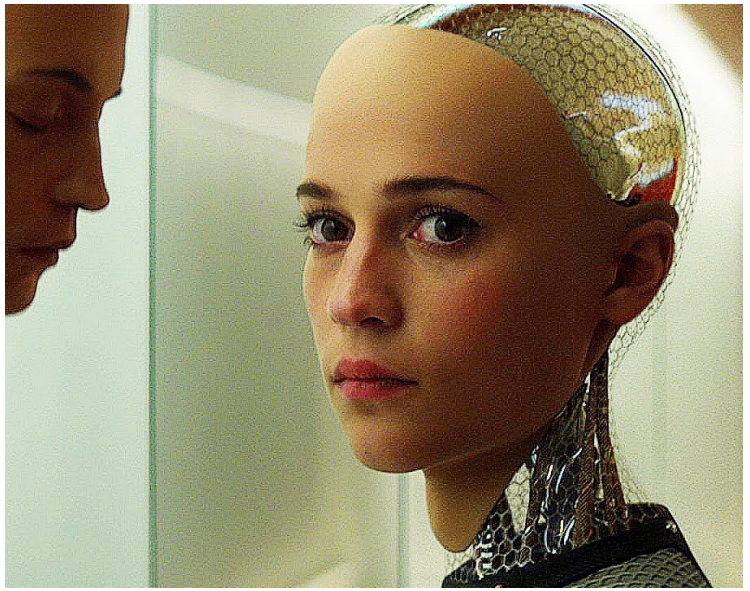
So: can AI be a mind, think, feel etc.?

-- Is strong AI possible?

Can AI have *ethical* agency? Be morally responsible for its actions? Presumably it can't, if it cannot think etc.

Complex questions in philosophy of mind

Understand: the nature and difficulty of the questions



THE UNIVERSITY OF MELBOURNE

Take homes
Module 4

Take homes

Different types of fairness/justice

People disagree about what is fair use of AI

Impossibility Theorem and Inherent Tradeoffs: make ‘perfect’ fairness unachievable for AI

Accountability frameworks for procedurally fair outcomes e.g. reasoned, transparent, democratic discussion

Is AI *itself* accountable? Can AI understand? Turing Test vs. Chinese Room