THE UNIVERSITY OF MELBOURNE

Module 5

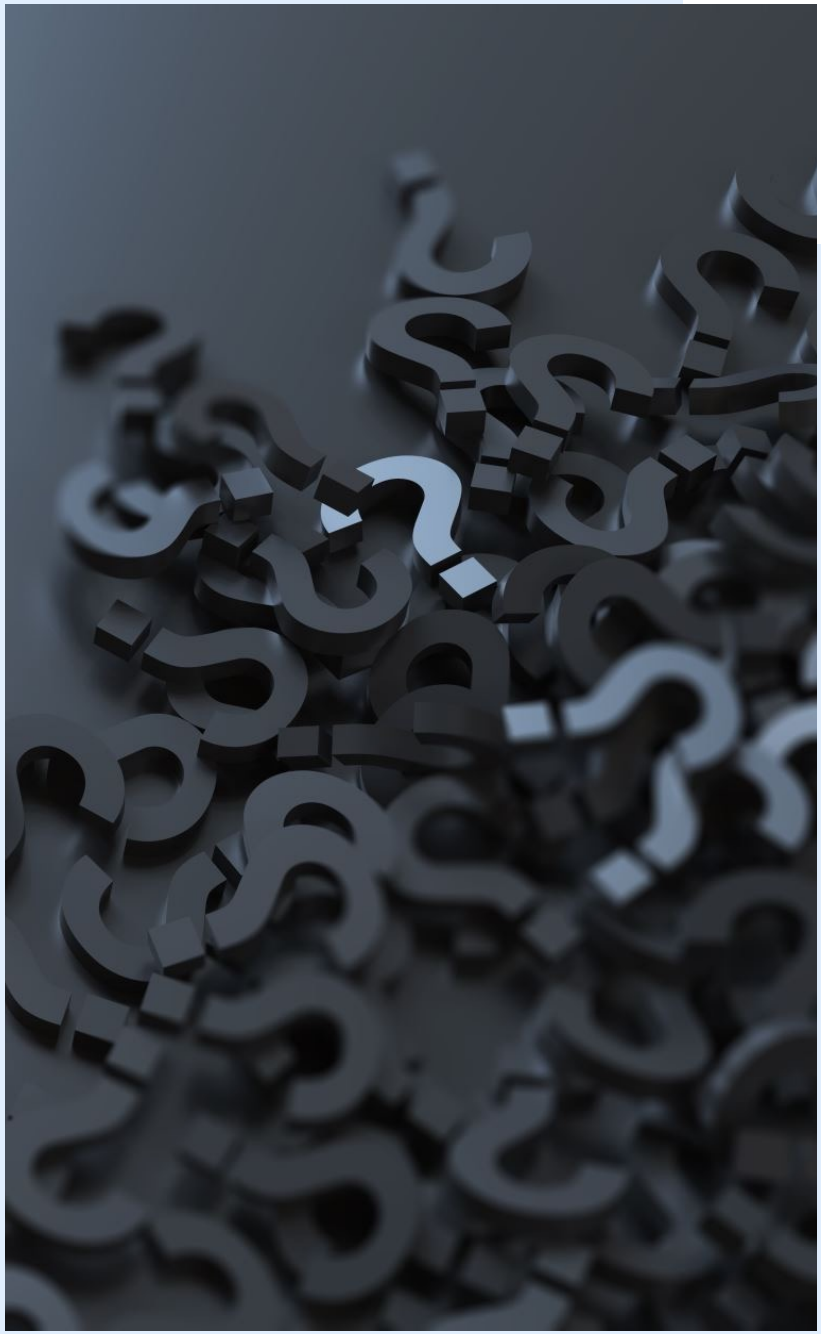
Fairness, Accountability, Principlism

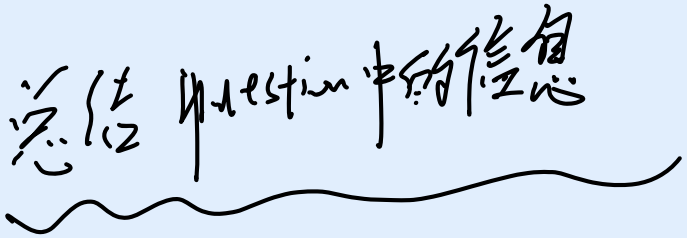
Simon Coghlan
simon.coghlan@unimelb.edu.au

Learning Outcomes Module 4

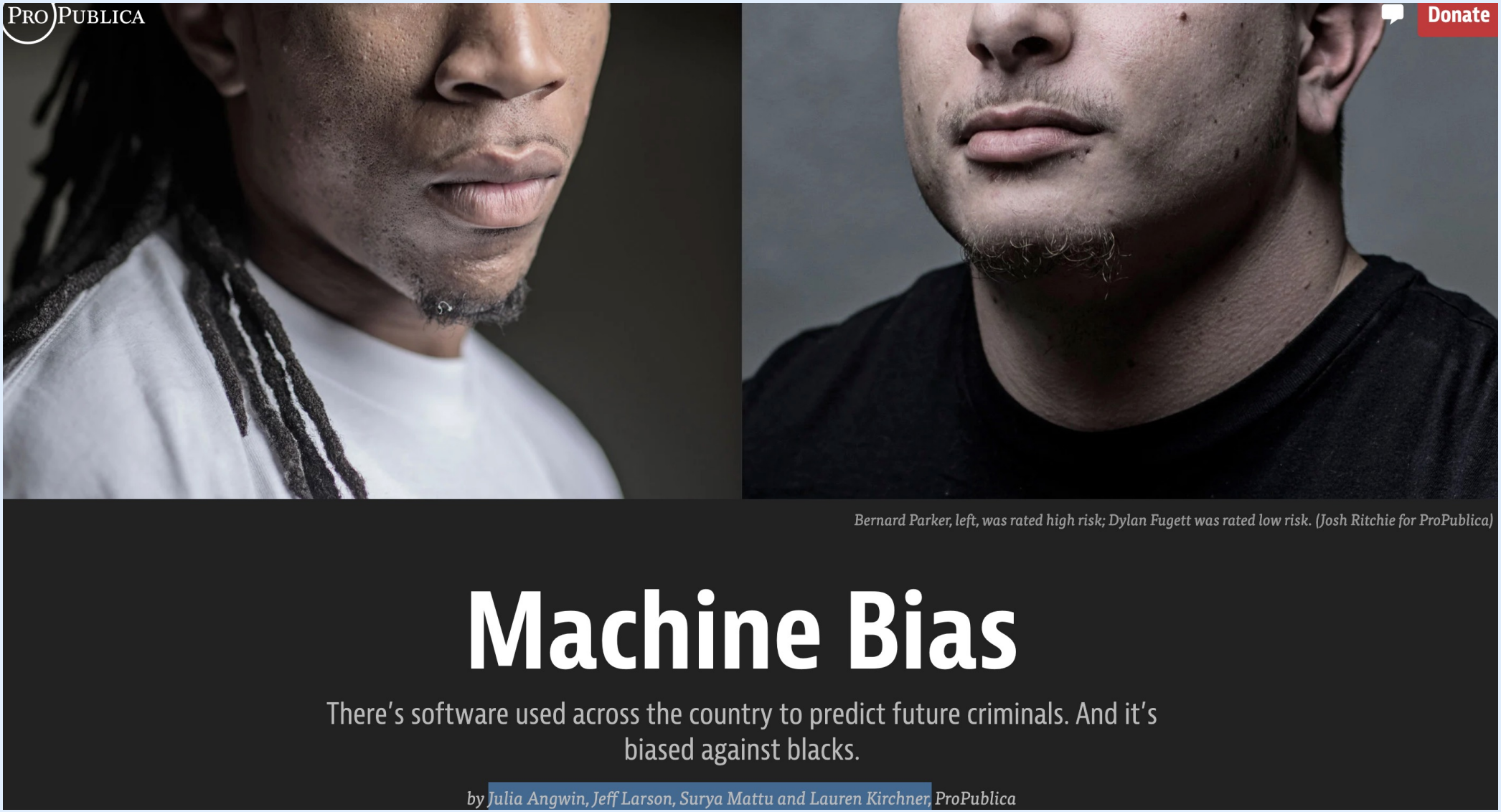
At the end of this module, you should be able to:

- Explain framework of Principlism in AI ethics
- Explain the concepts of fairness and accountability in relation to AI
- Intelligently apply the concepts of fairness and accountability to cases involving AI





What questions do you have on anything so far in the subject??



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

Correctional Offender Management Profiling for Alternative Sanctions

Northpointe (now *equivant* – tagline: “Software for justice”)

Recidivism algorithm

- Risk score for reoffending (‘recidivating’) after initial arrest
- Violent crimes, nonviolent crimes
- Blacks and whites involved
- Guides officers in determining bail
- May reduce rates of detention (allow bail)

4 principle in ethics AI medical

① Non-maleficence - do no harm
predict and minimize harm
② Beneficence - do good.
Anticipate good outcomes
③ Respect autonomy

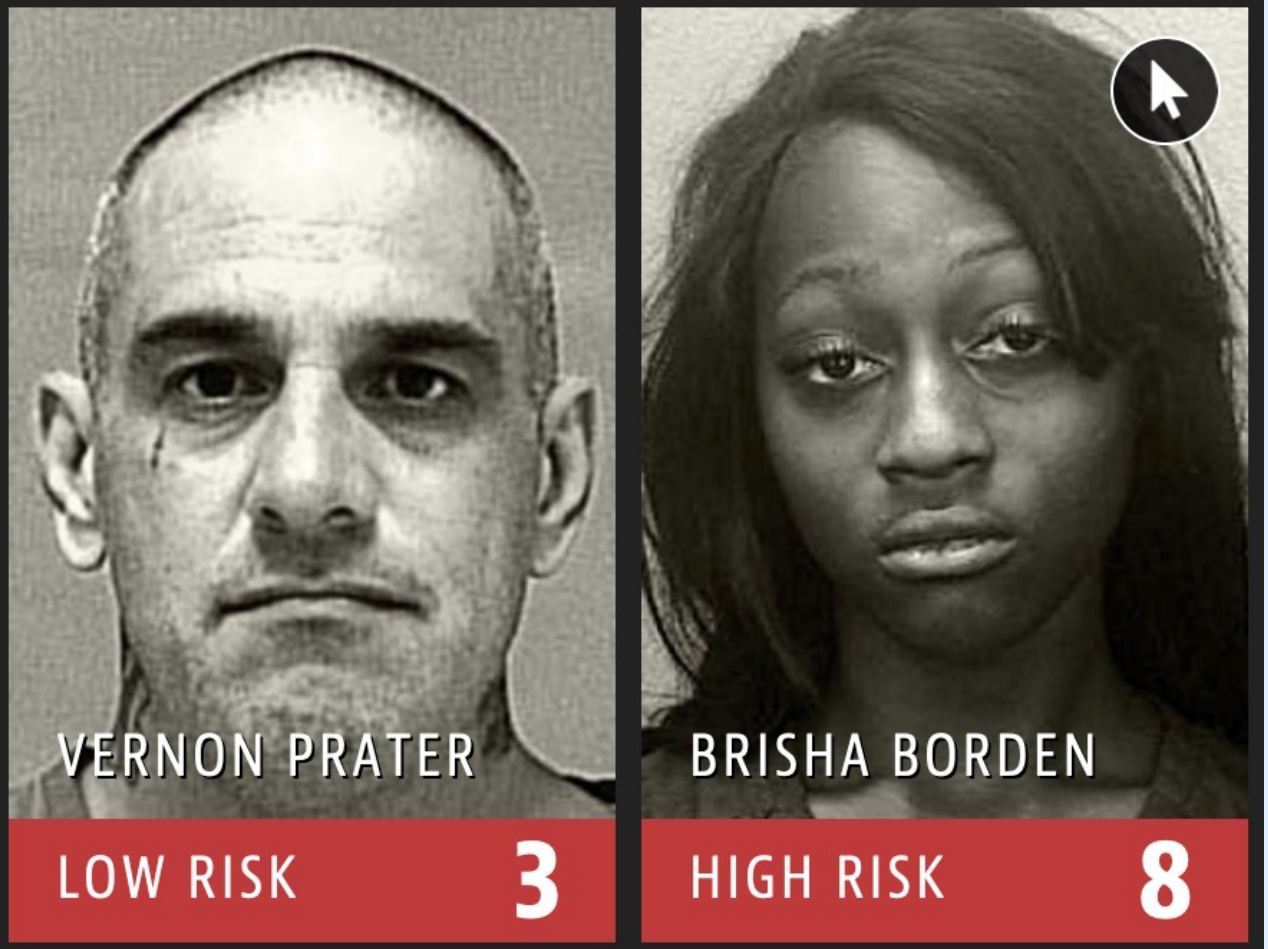
COMPAS

- Factors may include: current charges, pending charges, prior arrest history, previous pretrial failure, residential stability, employment status, community ties, substance abuse, age
- NOT race - 'protected attribute' like e.g. gender
- Actual algorithms are trade secrets

Julia Angwin et al

- Used actual re-arrest rates to determine actual offending post-COMPAS
- Biased against blacks
- Black nonoffenders higher risk scores than white nonoffenders
- Disparate impact (different impact on 2 groups) ethnic

Two Petty Theft Arrests

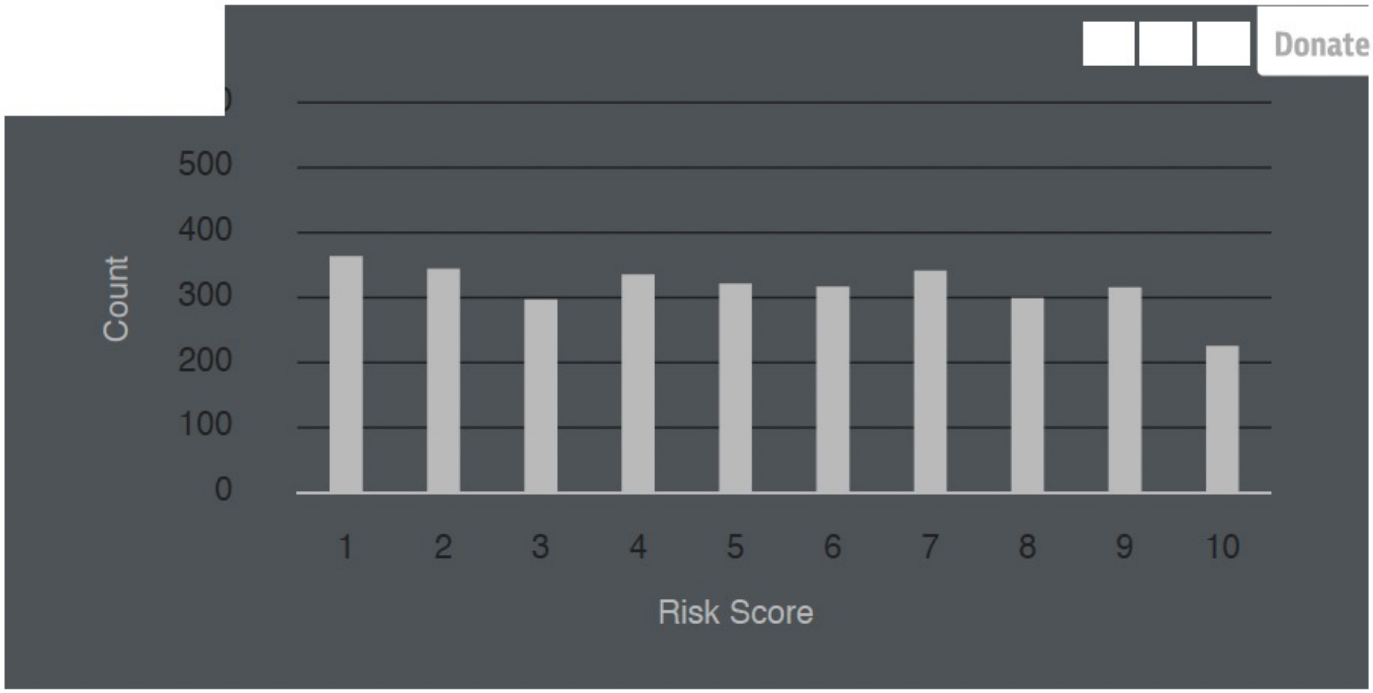


Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

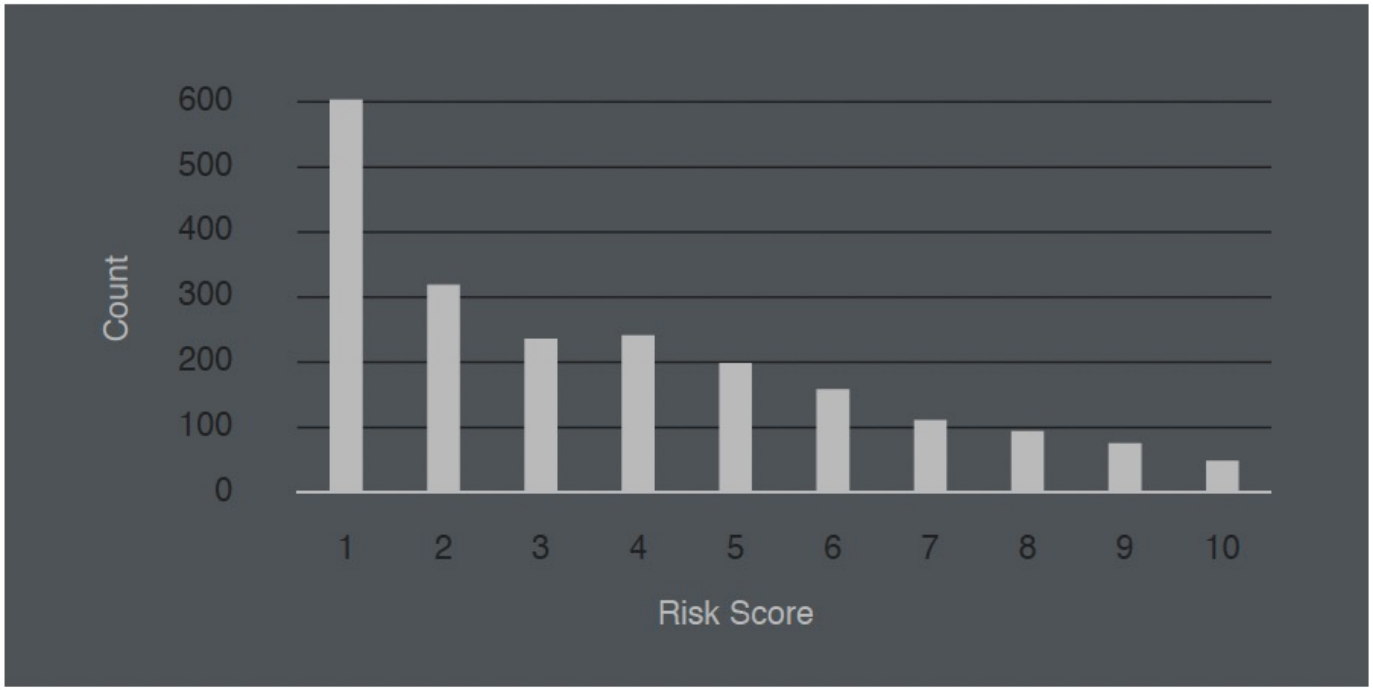
Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

"I'm surprised [my risk score] is so low. I spent five years in state prison in Massachusetts." (Josh Ritchie for ProPublica)



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. ***(Source: ProPublica analysis of data from Broward County, Fla.)***

ProPublica

	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%

- Equal accuracy (61%) for both whites and blacks – true
- But – the wrong predictions (39%) went wrong in different ways

Replies

- Not disparate treatment - COMPAS fair

- No explicit use of arbitrary differences (race)
- Yes, can be proxies for race (e.g. suburb)…but…
- Same accuracy for each race
- Reoffending rate for blacks and whites equal at each COMPAS scale ('calibration')
- E.g. At risk factor 3, blacks and whites had same rate of actual reoffending

sacrifice accuracy if change

only a tool — final choice made by human

Other replies

- Better than biased humans
- Historically: inconsistent, gut instinct
- One study (Dressel et al): COMPAS more accurate than individuals lacking criminal justice expertise and slightly less accurate than groups of individuals
- Decreased accuracy: more crime?
- Problem identified by ProPublica is less bad than e.g. poor ‘calibration’ (could violate discrimination laws)

- How do we decide what is fair?

Question

Is algorithmic (un)fairness purely a technical problem?

Who else apart from ‘AI experts' might be required to help solve it?

Principlism

- Distill theories into handy principles for AI ethics
- Midlevel principles: b/w theory and detailed rules
- Theory can *guide* these broad principles
- Derived from medical ethics principles

Principlism: 4 principles +1

1. Non-maleficence – do no harm

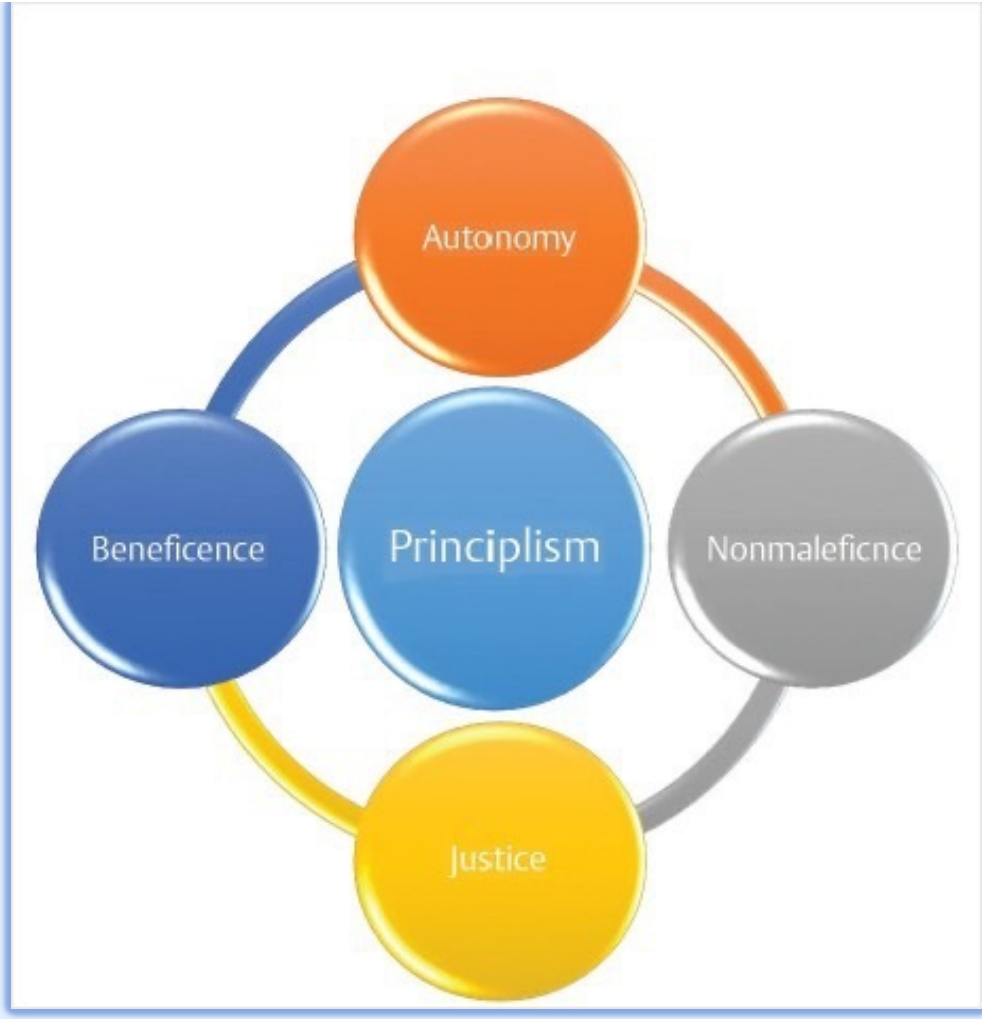
- Predict harm, avoid causing harm, minimize harm, short and long term

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

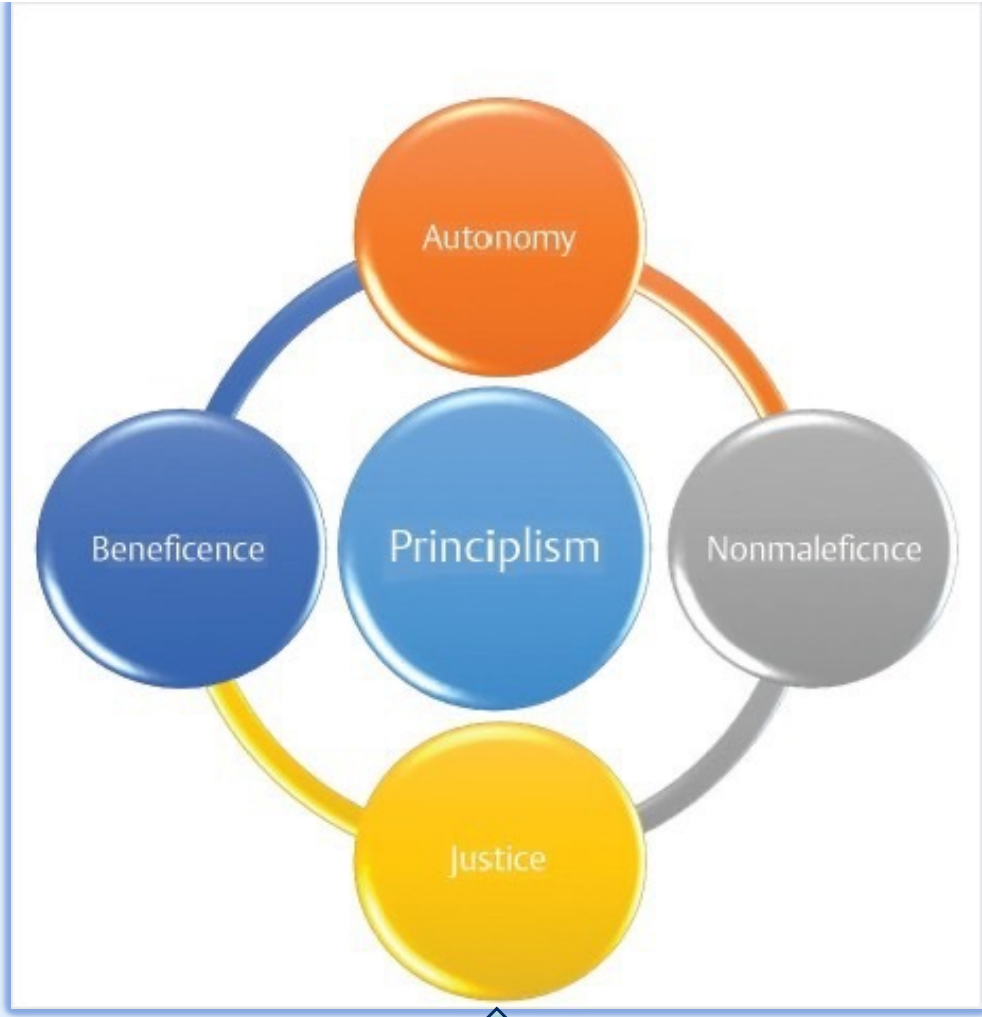
- Distribute benefits and harms fairly, fair processes, don't unfairly discriminate

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 one 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.



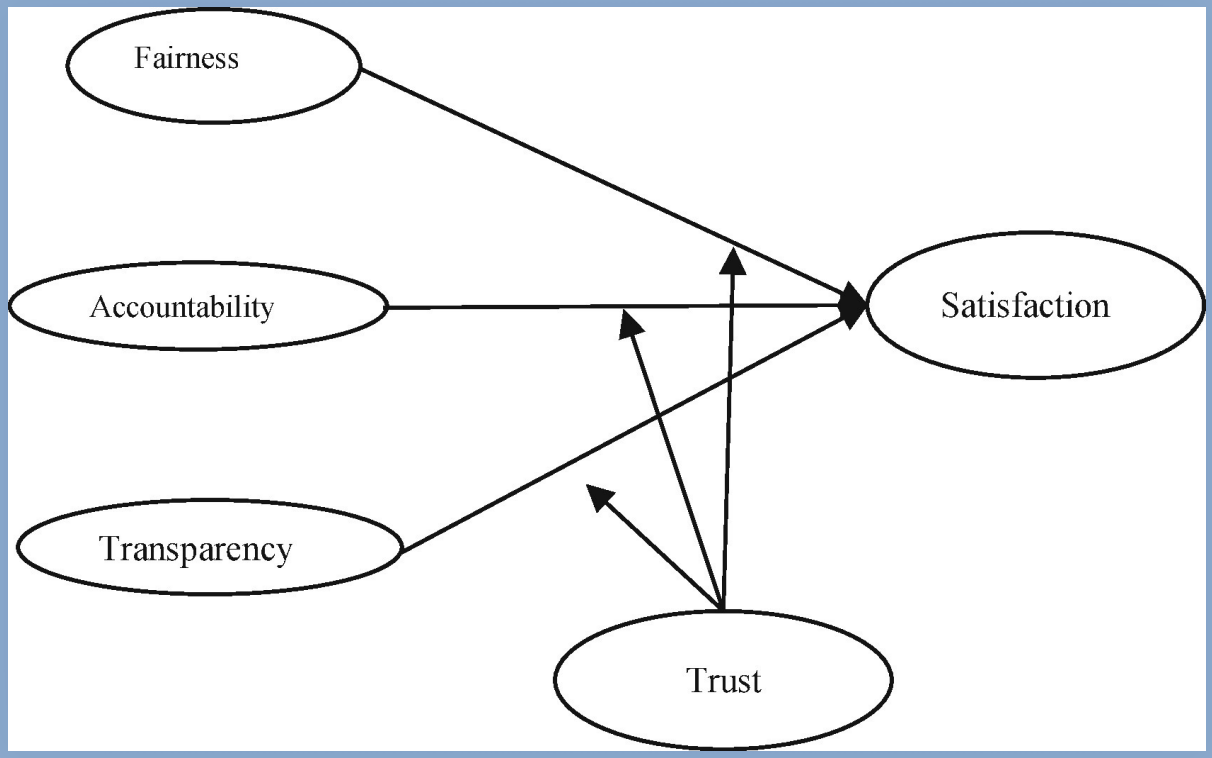
no hierarchy => balance according to situation

(what harm could do? (city)

whether it's fair
whether it's affect autonomy?

Reading 2

- '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.

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

Justice/fairness

- Broad definition: 'Giving each their due' or 'what they are owed'.
- Treat similar cases similarly; 'blind' to arbitrary differences

- Equal regardless of race, religion, class, sex, gender, sexual orientation, etc.
- 'Equal treatment/consideration/respect'

  - U – all similar interests are counted equal in calculus (maximize happiness)
  - D – duty of justice
  - Kant – Dignity equal; 2nd C.I.: always ends, never merely means
  - always ends, never merely means
  - VE – justice: praiseworthy trait
  - CE – justice in care

Distributive justice

- Resources, opportunities (分 pie)
- Need, merit, contracts etc.
  - Should everyone be given the same career opportunities regardless of talent?
  - Income? reg lardless how hard they work?
  - Should we help people suffer bad luck?
  - Should we favour people who are morally responsible rather than selfish?

- Positive/reverse discrimination
  - Affirmative action
  - More resources/opportunities/advantages to disadvantaged and historically oppressed groups

Procedural justice/fairness

- Fair procedure or process 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/merit etc. E.g. one lollipop and two best friends – who gets it?

Other types

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

Kant – retributive justice warranted irrespective of consequeses (不管后果 put in to prision)

- Reparative justice
  - Remediation for unfair treatment

V: oppose 'mere' retribution but deterrance can be justified (even if unjust)

Accountability

- *Assuming* accountability (responsibility)
  - E.g. holding myself to fair procedures
- *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?

Algorithmic Fairness

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

- *Technical* solutions to minimize unfairness
  - E.g. change inputs to reduce bia
  - 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)

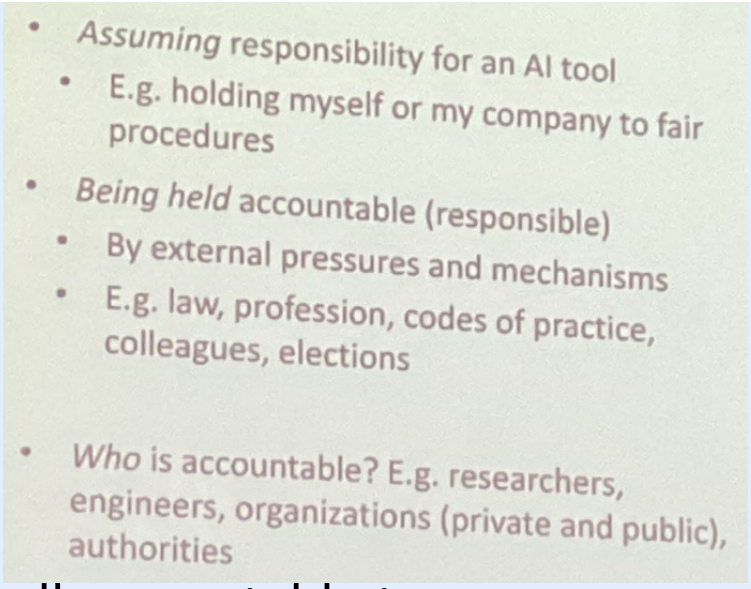
Ethical frameworks and fairness/justice

- **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 (absolute)
- **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

Pak-Hang Wong (reading)

- 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. ongoing respectful reason-giving and reason-responding
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?

Other accountability mechanisms?

- Audits
- Committees
- 'Turing' stamps - suitable body
- Open-source software – 'crowd'
- Others?

Rigorous testing

Ethics review

Model cards describe AI models, data used, weakness: bias etc)

AI codes of practice

hard law

FAIRNESS & ACCOUNTABILITY

Fairness	Fairness applicable when:	Accountability	Accountability for Reasonableness:
Fairness (sometimes 'justice') means: • giving each person their due (what they deserve), • treating similar cases similarly (being blind to arbitrary differences), • treating people equally regardless of race, religion, class, sex, gender, sexual orientation, etc. • equal consideration and respect for different beliefs and characters.	• allocating social resources (distributive justice), • implementing procedures that allocate benefits or harms (procedural justice), • determining penalties for wrongdoing (retributive justice), • providing remediation/compensation for unfair treatment (reparative justice)	The state of being *accountable*. Accountability for AI systems requires answering for their failures and shortcomings, being responsible for blame, and providing an account for why the system exists in the way it does. Accountability often involves legal liability.	• According to AFR, a decision making process must satisfy 4 conditions in order to be *legitimate and fair*: • Publicity Condition • Relevance (or Full Acceptability) Condition • Revision and Appeals Condition • Regulative Condition