THE UNIVERSITY OF MELBOURNE

COMP90087

Module 5: Trust, justice, and accountability

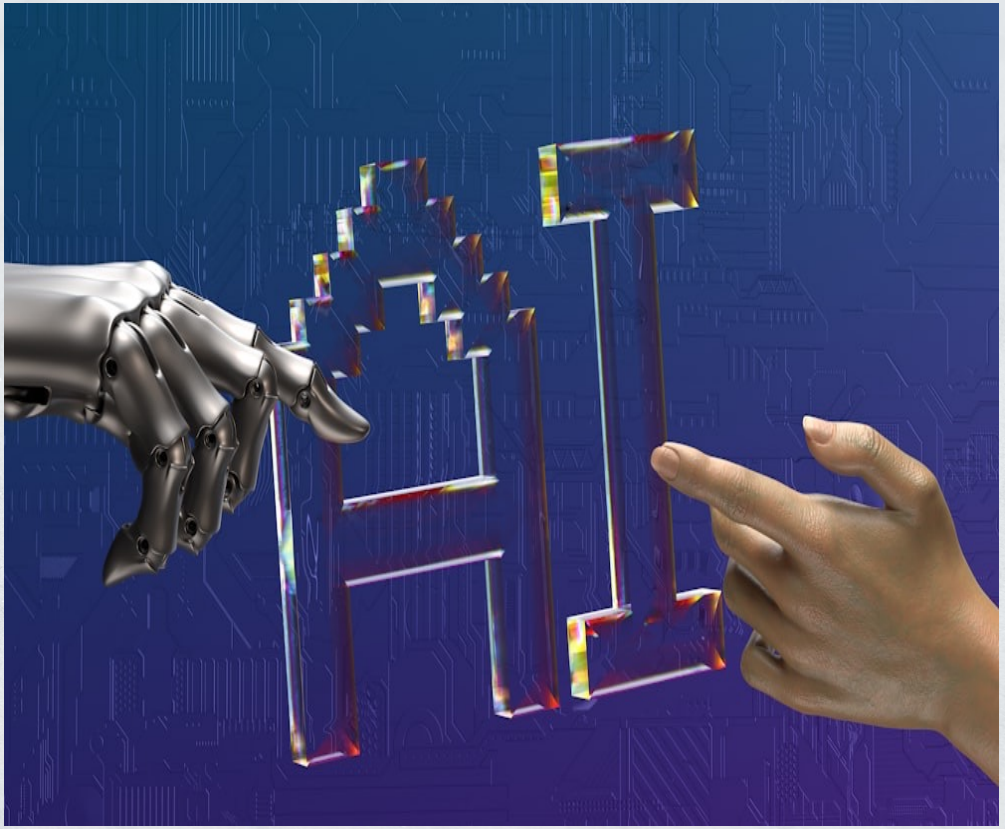
Simon Coghlan

Senior Lecturer
CIS 2024
Unimelb

Thanks to Prof Tim Miller for some material

Learning outcomes today

- Define trust and trustworthiness for AI
- Understand effects of use, misuse, abuse, and disuse of machines when trust is not well calibrated
- Consider issues of fairness and justice
- Understand the nature of accountability for AI



Images Wiki/Unsplash

What is trust and why is it important?

Images Wiki/Unsplash

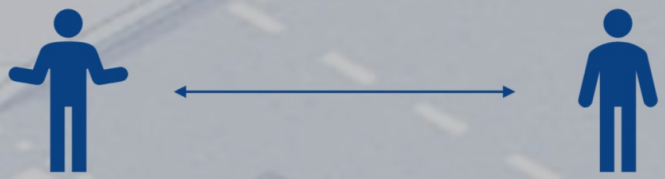
Human-AI trust (Jacovi et al.)

GOALS OF TRUST

Trust between people	Human-machine trust
PREDICTABILITY	PREDICTABILITY
ENABLE COLLABORATION	ENABLE "COLLABORATION"

TRUST IS **NOT** THE END GOAL

TRUST: THE VIEW FROM SOCIOLOGY



Interpersonal trust = humans trusting humans

A trusts B if:

1. A believes that B will act in A's best interests; and
2. A accepts vulnerability to B's actions;

so that A can:

3. Anticipate the impact of B's actions, enabling collaboration

HUMAN-MACHINE TRUST

Human-machine trust = one-way interpersonal trust of machine

H trusts M if:

1. H believes that M will act in H's best interests; and
2. H accepts vulnerability to M's actions;

so that A can:

3. Anticipate the impact of M's decisions, enabling collaboration

DISTRUST vs LACK OF TRUST

H distrusts M if:

1. H believes that M will act *against* H's best interests.

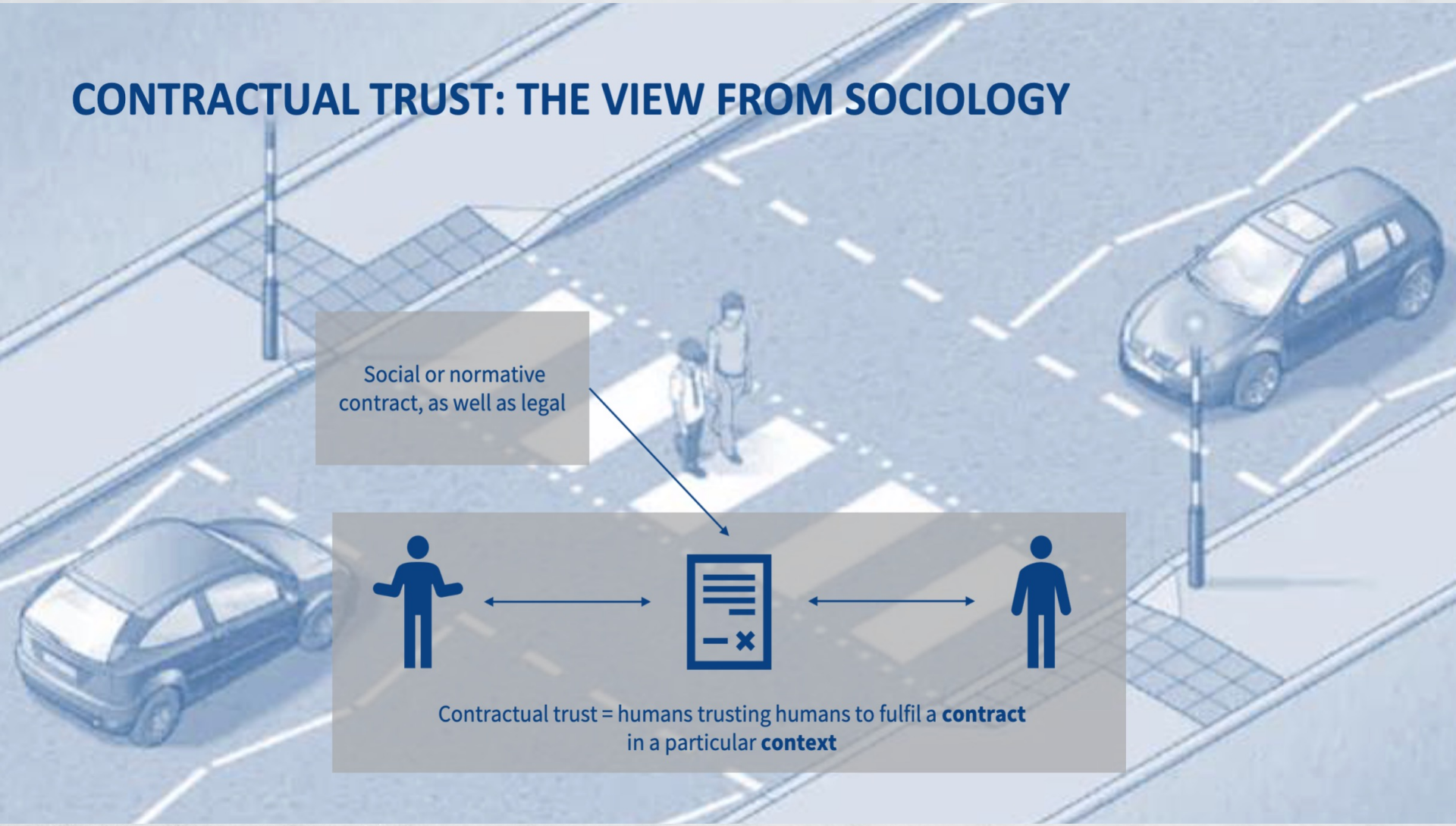
Lack of trust = absence of trust:

1. H does NOT believe that M will act in H's best interests;

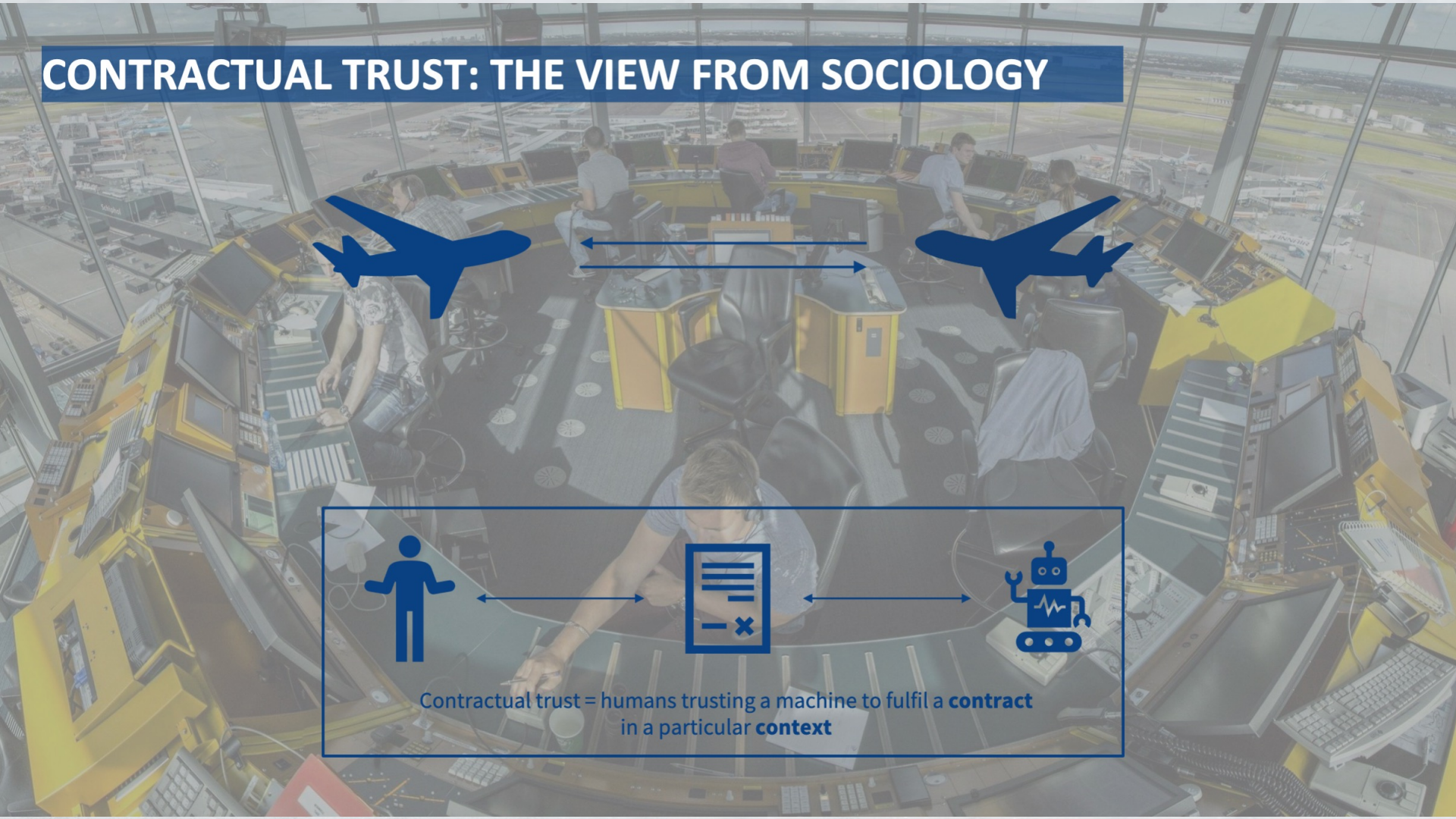
   OR

2. H does NOT accept vulnerability to M's actions;

CONTRACTUAL TRUST: THE VIEW FROM SOCIOLOGY



CONTRACTUAL TRUST: THE VIEW FROM SOCIOLOGY



Contractual trust = humans trusting a machine to fulfil a **contract** in a particular **context**

CONTRACTS FOR ARTIFICIAL INTELLIGENCE

European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	·Foster fundamental human rights ·Support users' agency ·Enable human oversight	Fairness checklists All N/A	See "Diversity, non-discrimination, fairness" User-centered explanations [62] Explanations in recommender systems [42]
Technical robustness and safety	·Resilience to attack and security ·Fallback plan and general safety ·A high level of accuracy ·Reliability ·Reproducibility	Factsheets (security) N/A Model cards (metrics) Factsheets (concept drift) Reproducibility checklists	Adversarial attacks and defenses [21] N/A N/A Contrast sets [17], behavioral testing [61] "Show your work" [14]
Privacy and data governance	·Ensure privacy and data protection ·Ensure quality and integrity of data ·Establish data access protocols	Datasheets/statements Datasheets/statements Datasheets/statements	Removal of protected attributes [60] Detecting data artifacts [24] N/A
Transparency	·High-standard documentation ·Technical explainability ·Adaptable user-centered explainability ·Make AI systems identifiable as non-human	All Factsheets (explainability) Factsheets (explainability) N/A	N/A Saliency maps [65], self-attention patterns [41], influence functions [39], probing [16] Counterfactual [22], contrastive [54], free-text [28,51], by-example [39], concept-level [20] explanations N/A
Diversity, non-discrimination, fairness	·Avoid unfair bias ·Encourage accessibility and universal design ·Solicit regular feedback from stakeholders	Fairness checklists N/A Fairness checklists	Debiasing using data manipulation [70] N/A N/A
Societal and environmental well-being	·Encourage sustainable and eco-friendly AI ·Assess the impact on individuals ·Assess the impact on society and democracy	Reproducibility checklists Fairness checklists Fairness checklists	Analayzing individual neurons [10] Bias exposure [69] Explanations designed for applications such as fact checking [3] or fake news detection [48]
Accountability	·Auditability of algorithms/data/design ·Minimize and report negative impacts ·Acknowledge and evaluate trade-offs ·Ensure redress	Factsheets (lineage) Fairness checklists N/A Fairness checklists	N/A N/A Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38] N/A

Source: Table 1 from Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. Alon Jacovi, Ana Marasovic, Tim Miller, and Goldberg. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2021)*, 2021.

TRUST AND TRUSTWORTHINESS

TRUSTWORTHY MACHINES

A machine is trustworthy if:

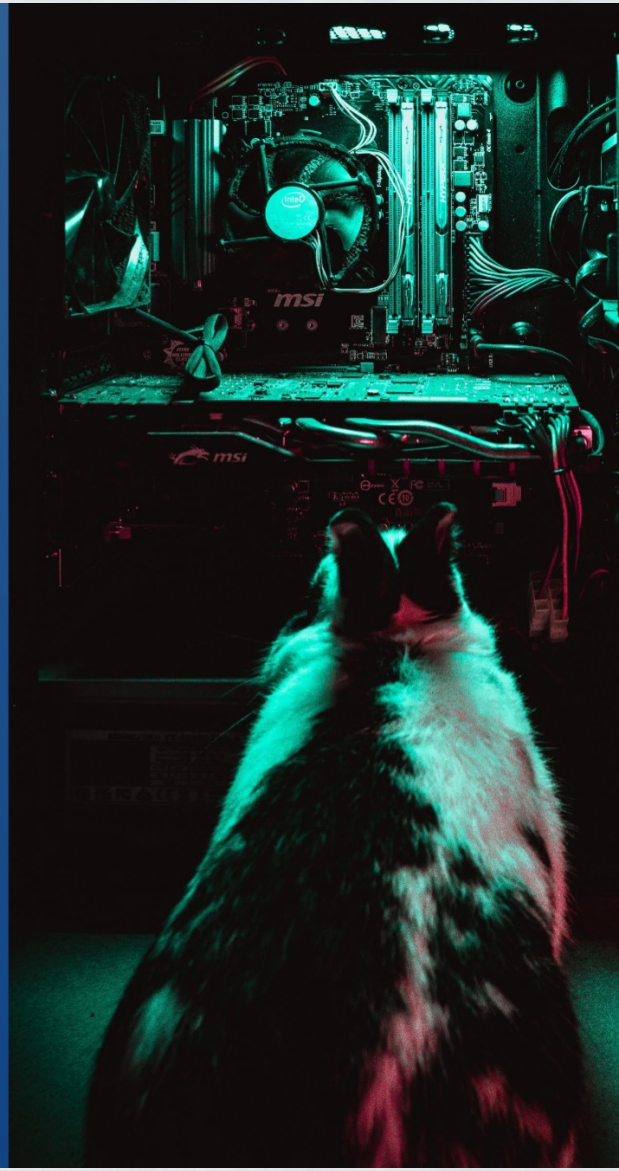
1. It can fulfill its set of contracts

Trust does not imply trustworthiness

Trustworthiness does not imply trust

WARRANTED AND UNWARRANTED TRUST

	Trusted	Distrusted
Trustworthy	Warranted trust	Unwarranted distrust
Not Trustworthy	Unwarranted trust	Warranted distrust



USE, MISUSE, DISUSE, AND ABUSE: UNWARRANTED TRUST AND DISTRUST

FACTORS THAT DETERMINE USE OF AUTOMATION

PARASURAMAN AND RILEY (1997)

MENTAL WORKLOAD

COGNITIVE OVERLOAD

TRUST (!)

MISUSE OF AUTOMATION

Definition: Using automation when it not should be used

Cause: Unwarranted trust

- Over-reliance on automation
- Decision biases and automation biases
- Machine monitoring errors

Impacts: Complacency

DISUSE OF AUTOMATION

Definition: Not using automation when it should be used

Cause: Unwarranted distrust

- Human monitoring errors (high false alarm rate)
- Machine monitoring errors
- Human bias

ABUSE OF AUTOMATION

Definition: Deploying automation when it should not be

Cause: Unwarranted distrust (from designer)

- Distrust in human operators
- Automation bias
- Arrogance

Impacts: Mismatch in human-automation interface

EXAMPLE: THERAC-25

Therac-25 A software-controlled radiation therapy machine

Outcome: Six patients with fatal radiation overdoses

Causes: Software errors from

- Misuse Unwarranted trust from radiographers
- Disuse Hardware interlocks removed
- Abuse Minimal input from radiographers

POWER, TRUST AND MACHINES

WHAT IS POWER?

POWER

The ability to control our circumstances

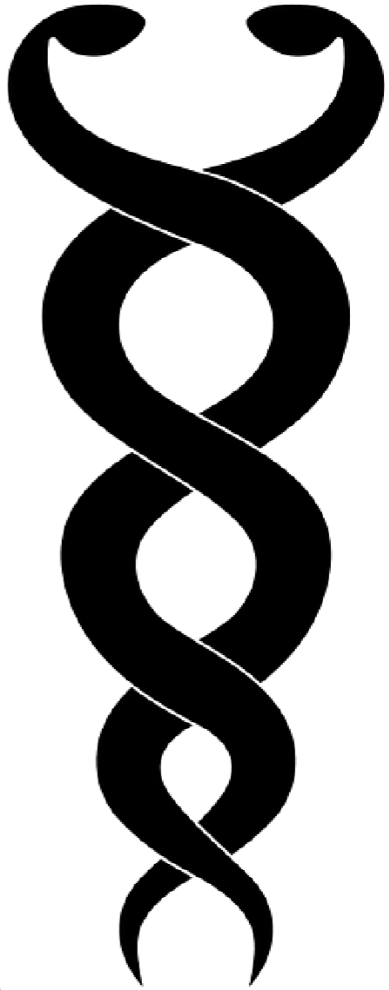
POWER TO DO …

POWER OVER …

POWER, TRUST, AND ETHICS

Trust ≠ Ethics ≠ Power

But! They are closely related and cannot be separated.



USER TRUST

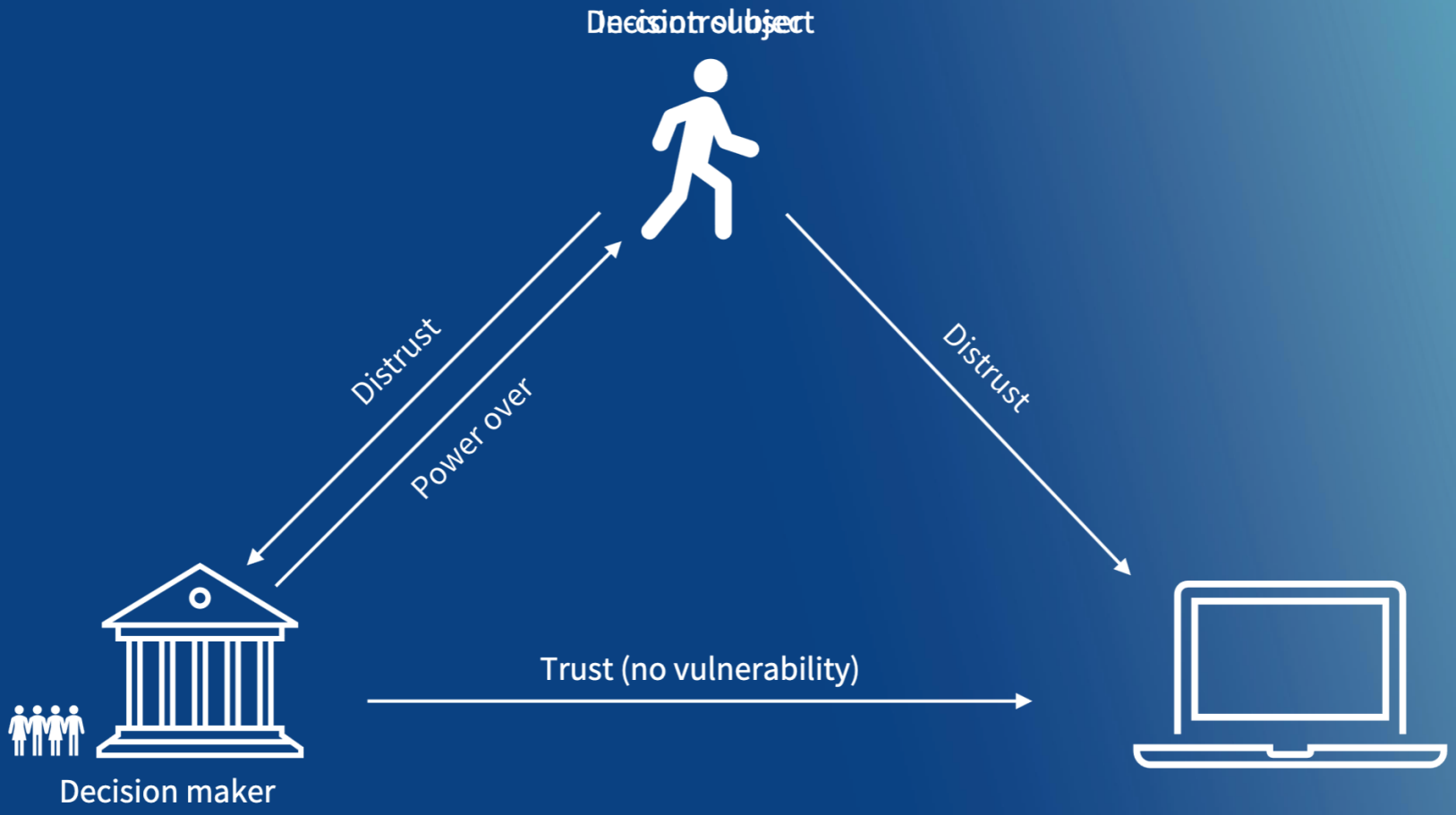
In-control user

Trust

In-control user

Distrust

USER TRUST


Decision subject
Distrust
Power over
Distrust
Trust (no vulnerability)
Decision maker

POWER, TRUST, AND MACHINES: SUMMARY

TRUST AND POWER

Belief in acting 'in my interests'
Accepting vulnerability
Anticipating impact of decisions

Contractual trust

Warranted and unwarranted trust and distrust

Use, misuse,disuse, and abuse of technology

Power is ability to control circumstances

KEY TAKEAWAYS

Be explicit about which contracts holds for your systems/applications

Ethically, trust is only desirable if it is warranted

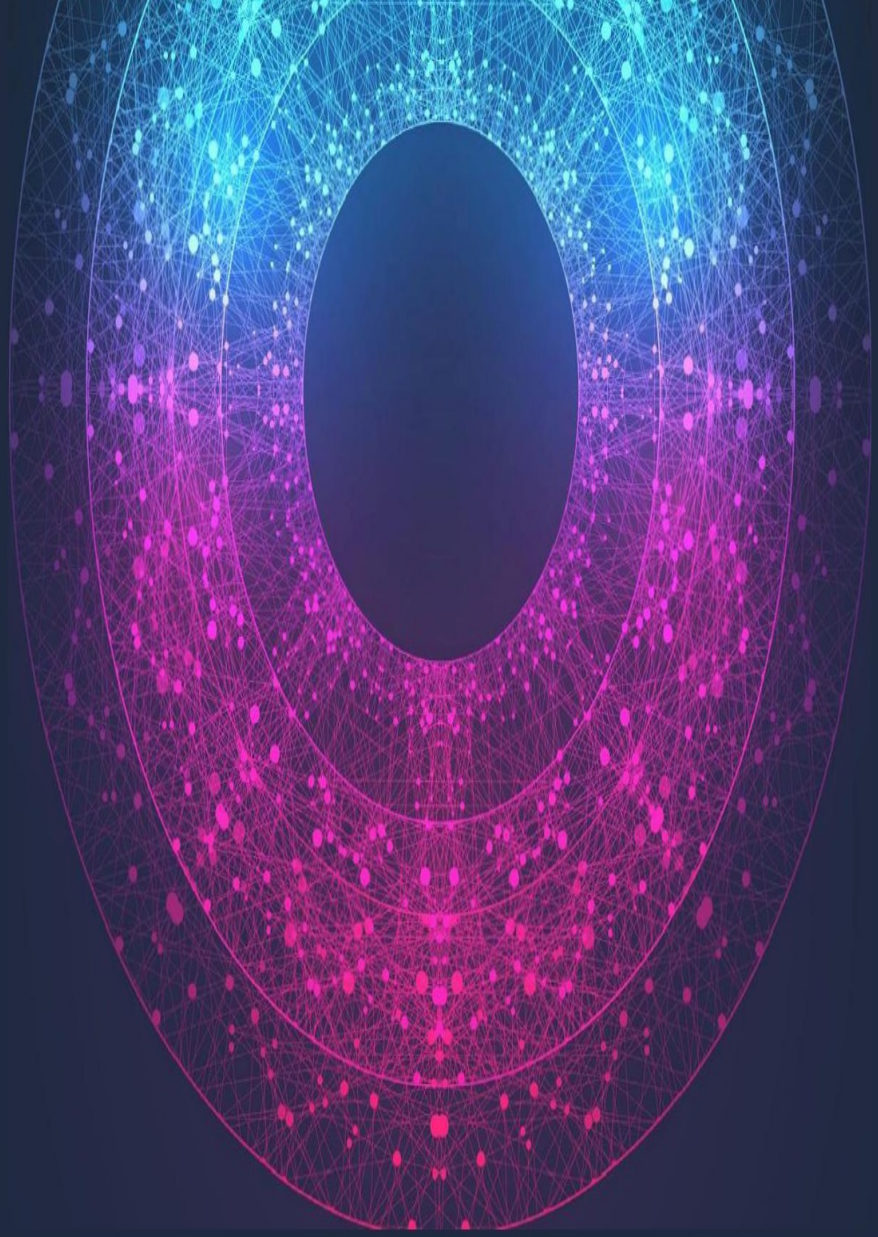
Distrust is desirable if it is warranted

Incorrectly calibrated trust leads to real problems

Ethical issues emerge from (real or perceived) power imbalances between groups with different interests

TRUST, POWER & JUSTICE:

AN AI EXAMPLE



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

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”

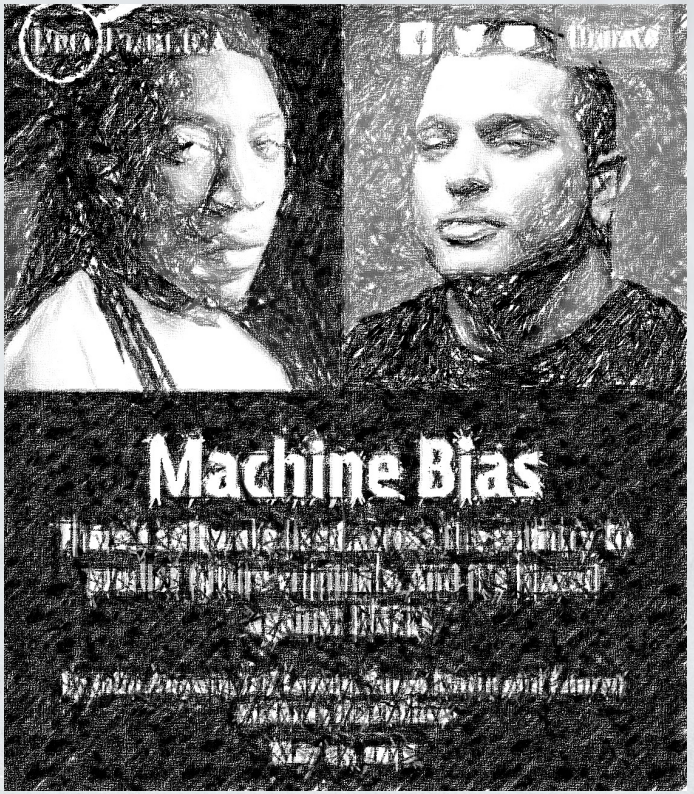
Recidivism algorithm

- Risk score for reoffending (‘recidivating’) after initial arrest
- Guides officers in determining bail
- May reduce rates of detention

Data used

Factors may include:

…current charges, arrest history, residential stability, employment status, community ties, substance abuse, age



What Julia Angwin et al found

Used actual re-arrest rates
to determine actual offending post-COMPAS

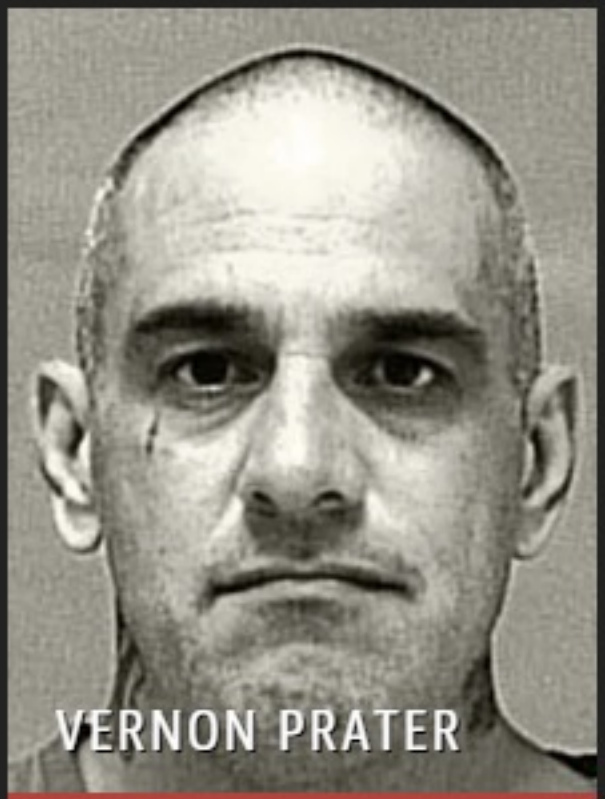
Biased against blacks

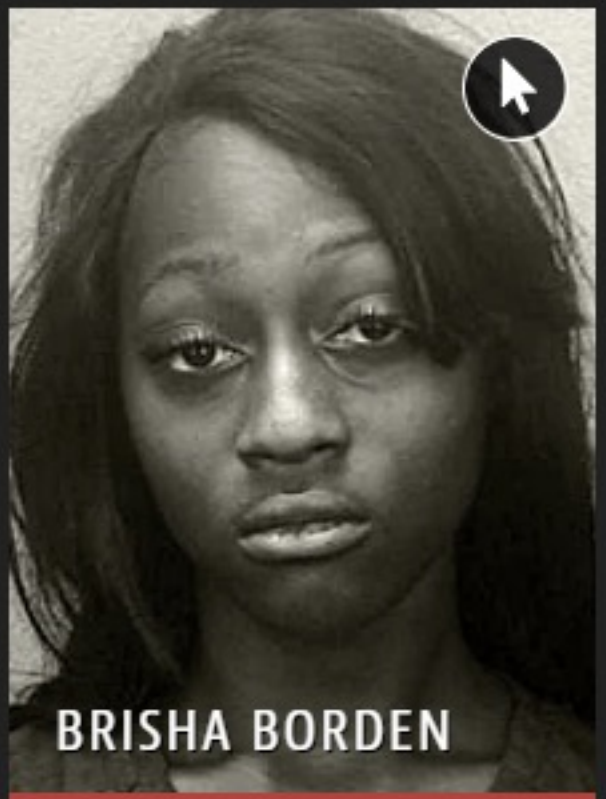
More black nonoffenders given higher risk scores than white nonoffenders

>Disparate impact

Images Wiki/Unsplash

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.

Disparate impact

Prediction Fails Differently for Black Defendants		
	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

What AI vendor said

COMPAS *is* fair:

- Excluded race as a factor
- Overall same accuracy for each race
- Differences in risk scores due to underlying differences in recidivism rates for blacks and whites, not inherent bias in COMPAS

So, what is fairness or justice?

Is algorithmic fairness purely a technical problem?

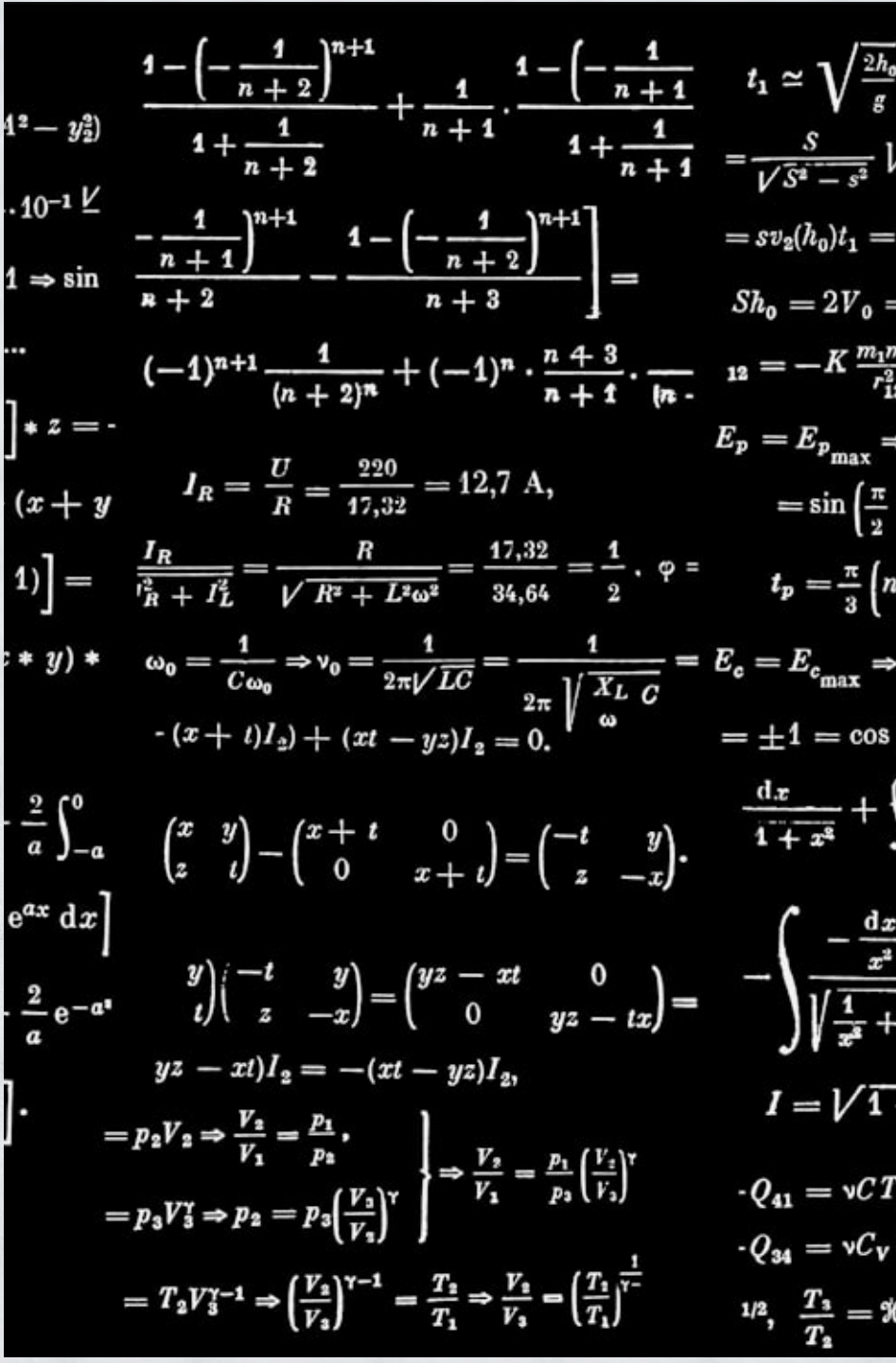
Can we solve it mathematically?

Not entirely: depends what we judge is ‘fair’

Remember our moral theories?

U – Fair = maximises utility

D/VE – No! Fair might be e.g. giving everyone the same opportunity, not worsening disadvantage etc.



Justice or fairness

- Broad definition: ‘Giving each their due’ or ‘what they are owed’
- Treat similar cases similarly; different cases differently
- Watch out for potentially arbitrary or irrelevant factors e.g. gender
- COMPAS: Should avoid disparate impact?

Images Wiki/Unsplash

Also….

Mathematical measures of fairness (Berk et al 2018)

- overall accuracy equality
- statistical parity
- conditional procedure accuracy equality
- conditional use accuracy equality
- treatment equality
- Etc.

The Impossibility Theorem: mathematically impossible for an algorithm to simultaneously satisfy different popular fairness measures

Images Wiki/Unsplash

The Inherent Trade-off

…between fairness and performance

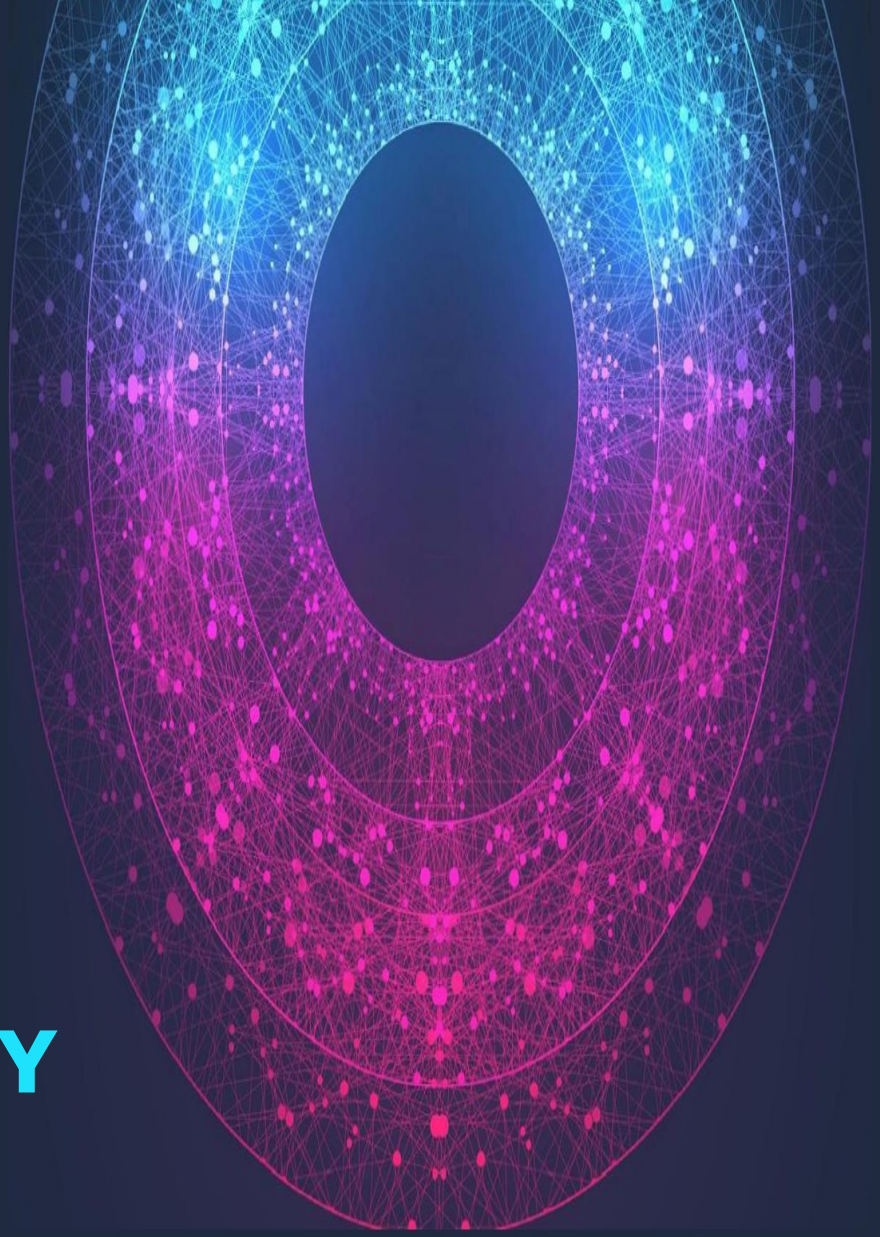
E.g. Increased group fairness → decreased accuracy of recidivism prediction for bail

Decrease F+ (defendants falsely scored as high risk)

--- but increase F- (miss some high-risk defendants)

Images Wiki/Unsplash

ACCOUNTABILITY



Accountability

Being responsible for things going wrong:

- preventing
- addressing (intervening in harms, paying compensation, apologising etc)
- having mechanisms (legal, codes, principles, contracts)
- promoting warranted trust
- recognising relevant power imbalances

Images Wiki/Unsplash

Transparency

- Being open, honest, and forthcoming
- COMPAS algorithm are trade secrets – not open to being scrutinized
- Thus: low trust and unfair (?) use of power

Images Wiki/Unsplash

Fair *procedures*

If harms or unequal impact unavoidable: fair procedures for deciding use of AI

Wong: *"ensure decisions are morally and politically acceptable to those affected by algorithms through inclusion and accommodation of their views and voices"*

No totally final and 'right' answer: answers emerge through open, democratic, good-faith dialogue and reason-giving involving stakeholders

Pak-hang Wong

https://conditiohumana.io/interview-pak-hang-wong/

Accountability for reasonableness (AFR) framework

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: give 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

Wong, Pak-Hang. "Democratizing algorithmic fairness." *Philosophy & Technology* 33 (2020): 225-244

AFR framework

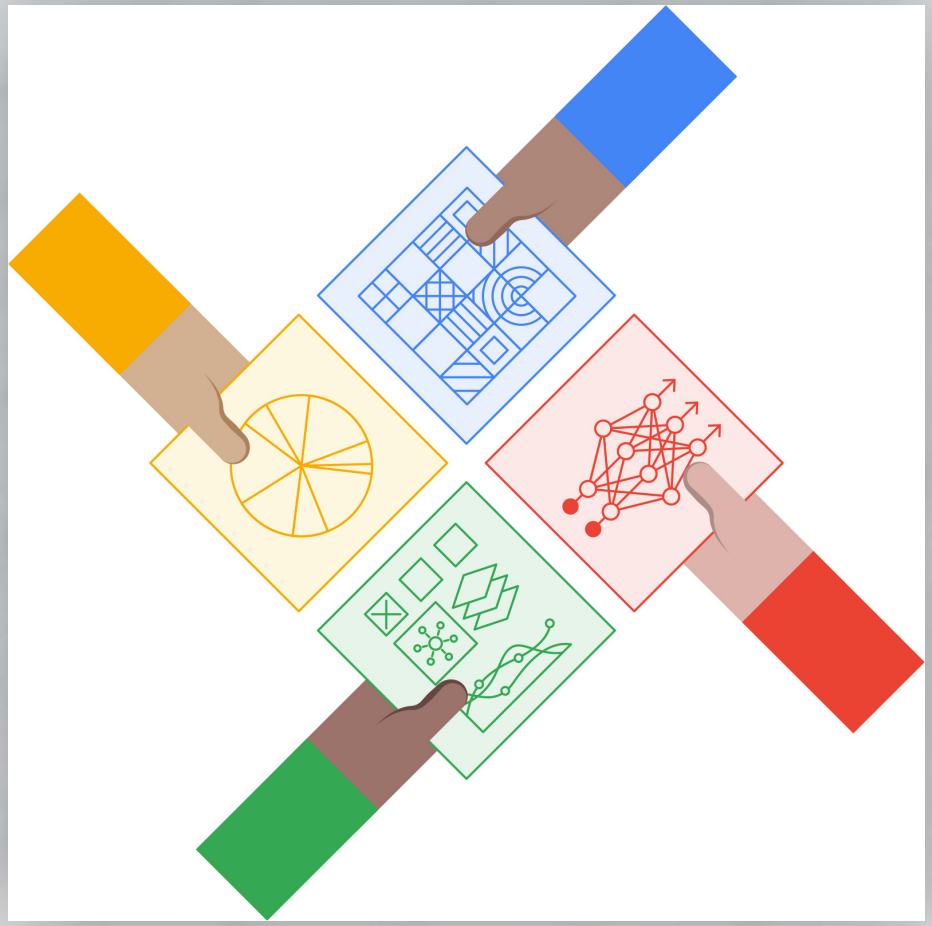
3.Revision and appeals condition: ongoing (not one-off) mechanisms for challenge and dispute resolution and revision of policies

4.Regulative condition: ongoing public regulation of process to ensure conditions (1)–(3) are met

Wong, Pak-Hang. "Democratizing algorithmic fairness." *Philosophy & Technology* 33 (2020): 225-244

Supplying information

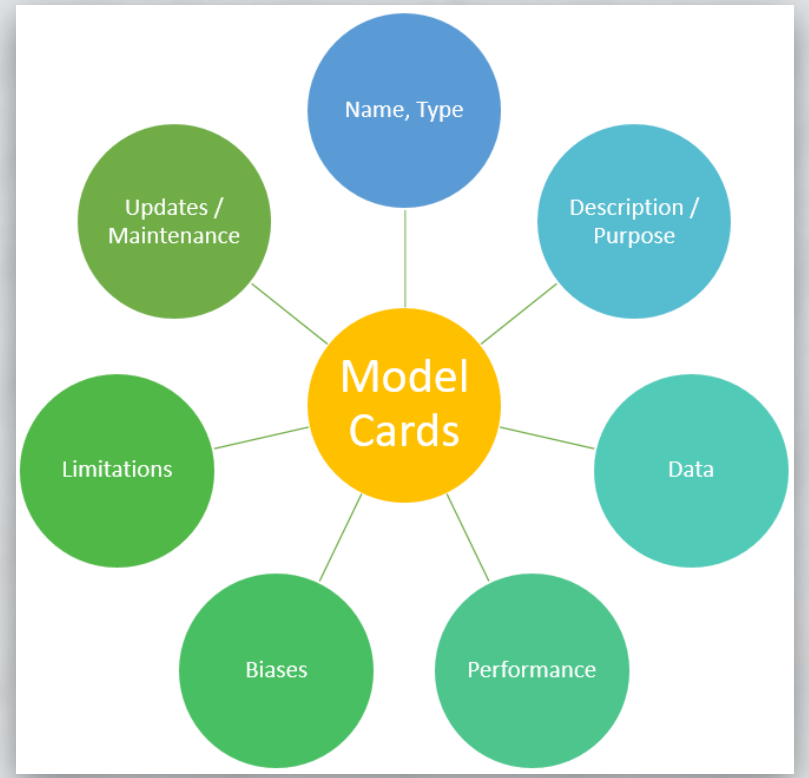
- Reproducibility Checklists
- Fairness Checklists
- Factsheets
- Data Statements
- Datasheets for Datasets
- Model Cards



https://modelcards.withgoogle.com/about

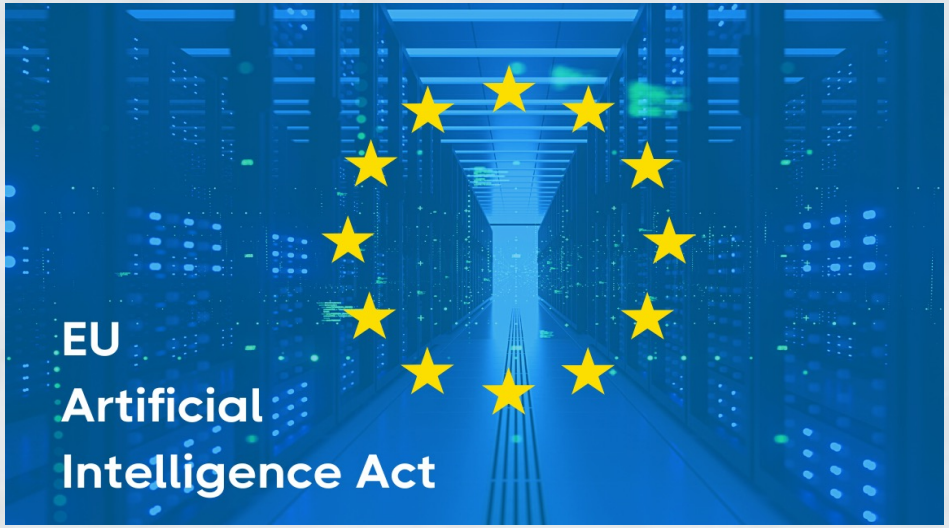
AI model information

- performance
- code
- potential biases
- architecture
- training data
- data pre-processing
- sensitive attributes
- evaluation metrics (e.g. disparate impact)
- potential use cases



https://vitalflux.com/model-cards-example-machine-learning/

Self-regulation versus law



Teaching ethics

Home › Science and Engineering Ethics › Article

What Do We Teach to Engineering Students: Embedded Ethics, Morality, and Politics

Avigail Ferdman & Emanuele Ratti

Provocation

- Can we hold AI *itself* responsible??
- ChatGPT, the hyperintelligent robot, HAL, etc. is to blame!?
- Or is that a cop out?

Images Wiki/Unsplash

Assigning accountability

Autonomous robot mistakes civilian for soldier

- Private ordered to launch it?
- Army commander?
- Software engineer?
- War tech company? CEO?
- Government?

Images Wiki/Unsplash

Moral crumple zone

- Blaming nearest human with limited control over AI
- Scapegoating to protect AI system or vendor

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction *Engaging Science, Technology, and Society*

Images Wiki/Unsplash

Take homes

- Trust, trustworthiness, power affect ethics
- Justice/fairness disagreements cannot be solved purely mathematically (recall COMPAS)
- Accountability may be vital for fair and ethical design/use of AI

Images Wiki/Unsplash