THE UNIVERSITY OF
MELBOURNE

Trust, machines, and digital ethics

Tim Miller

School of Computing and Information Systems
Centre for AI & Digital Ethics
The University of Melbourne
tmiller@unimelb.edu.au
@tmiller_unimelb

Learning outcomes

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

1. Define *trust* and *trustworthiness* with respect to artificial intelligence.
2. Discuss the effects of use, misuse, abuse, and disuse of machines when trust is not properly calibrated.
3. Discuss the relationship between trust and ethics in artificial intelligence
4. Apply the presented trust model to digital applications to assess trustworthiness at a high level.

Related reading

Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. Alon Jacovi, Ana Marasovic, Tim Miller, and Yoav Goldberg. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2021)*, 2021. https://arxiv.org/abs/2010.07487

Humans and automation: Use, misuse, disuse, abuse. Raja Parasuraman and Victor Riley. *Human factors*, *39*(2), 230-253, 1997. https://stuff.mit.edu/afs/athena.mit.edu/course/16/16.459/OldFiles/www/parasuraman.pdf

Outline

1. Why trust in machines is important
2. Trust and contractual trust
3. Trustworthiness and its relation to trust
4. Warranted and unwarranted trust
5. Intrinsic and extrinsic trust
6. Impact of incorrectly warranted/unwarranted trust
7. Trust and ethics in AI

THE UNIVERSITY OF MELBOURNE

Why trust machines?

Goals of trust

The sociological view of *interpersonal trust* (trust between two people):

- By obtaining trust in someone, we make life more *predictable*, which enables collaboration between people.

The human-machine view:

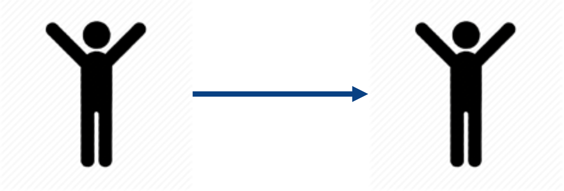
- By obtaining trust in a machine, we make it easier to anticipate the machine's decisions (predictability), which enables human-machine collaboration.

The end goal is NOT trust. Trust is a mechanism to help enable predictability and collaboration

THE UNIVERSITY OF
MELBOURNE

What is trust?

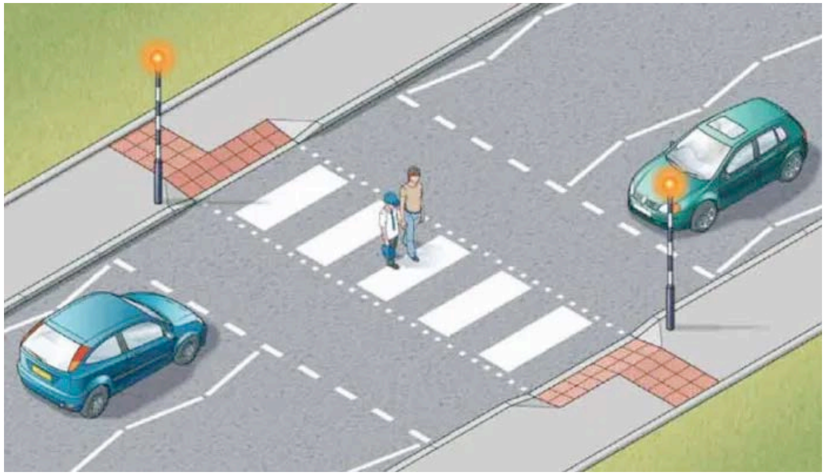
Trust: The view from sociology



Interpersonal trust = humans trusting humans

Person A *trusts* person B *if*:

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

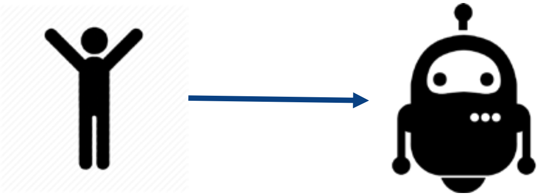


so that A can:

- anticipate the impact of B's actions,

therefore making social life more predictable, enabling collaboration.

Human-AI trust



Human-AI trust = humans trusting AI

H (*human*) trusts M (*machine*) if…

- H believes that M will act in H's best interests;
- H accepts vulnerability to M's actions;

So that H can...

- anticipate the impact of M’s decisions on H

therefore making the interaction more predictable, enabling collaboration.

Belief

Risk

Goal

Distrust and lack of trust

Distrust:

- H believes that M will NOT act in H’s best interest.

A lack of trust is an absence of trust:

- H does not believe M will act in H’s best interest; or
- H does not accept vulnerability to M’s actions.

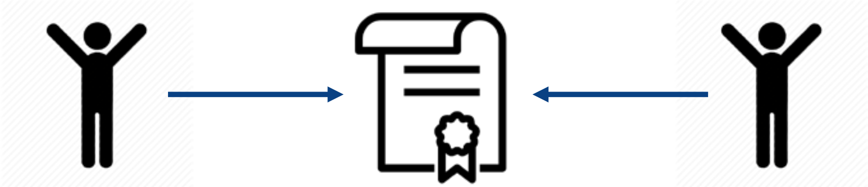
http://www.flickr.com/photos/lwr/

Trust can exist *regardless* of whether the H can anticipate the impact of M’s actions on H

THE UNIVERSITY OF
MELBOURNE

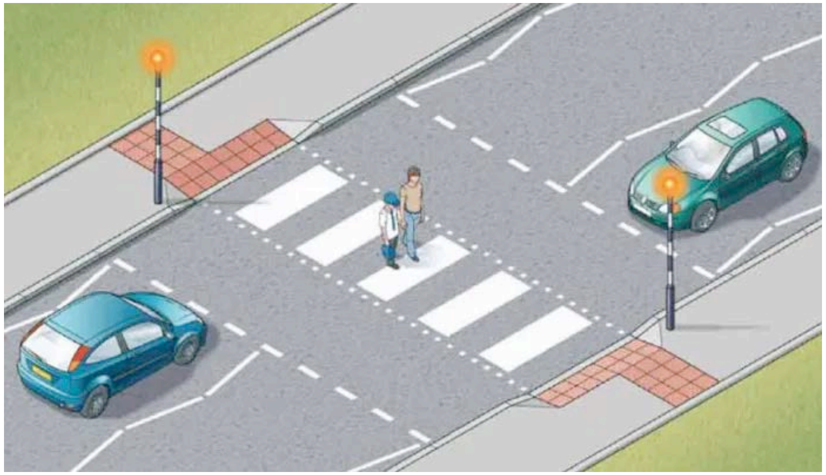
Contractual trust

Contractual trust: The view from sociology

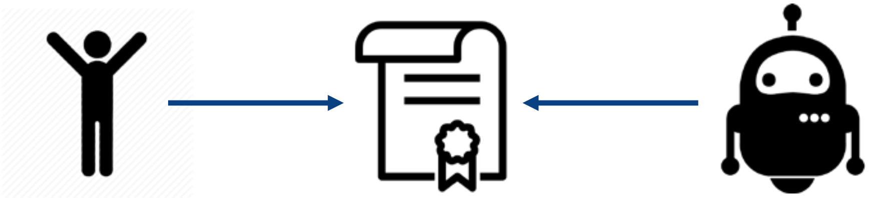


Contractual trust = humans trusting humans to ***fulfill a contract*** *in a particular* ***context***

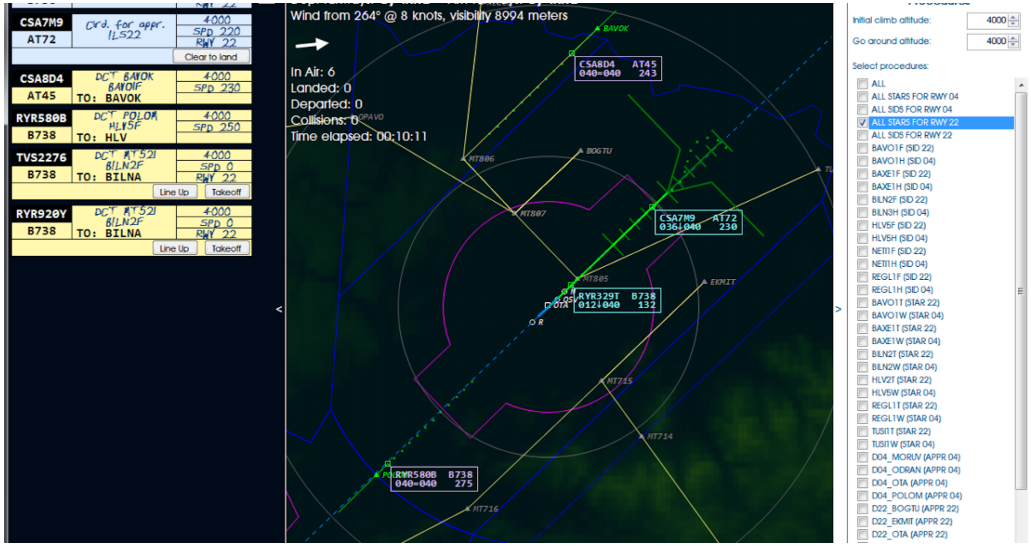
the contract can be social/normative, not just legal



Contractual human-AI trust



Contractual trust = humans trusting an AI model to ***fulfill a contract*** *in a particular* ***context***



CSA7M9 AT72	Clrd. for appr. ILS22	4000 SPD 220 RWY 22
		Clear to land
CSA8D4 AT45	DCT BAVOK BAVO1F TO: BAVOK	4000 SPD 230
RYR580B B738	DCT POLOM HLV5F TO: HLV	4000 SPD 250
TVS2276 B738	DCT MT521 BILN2F TO: BILNA	4000 SPD 0 RWY 22
		Line Up · Takeoff
RYR920Y B738	DCT MT521 BILN2F TO: BILNA	4000 SPD 0 RWY 22
		Line Up · Takeoff

Wind from 264° @ 8 knots, visibility 8994 meters

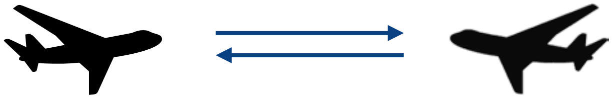
In Air: 6
Landed: 0
Departed: 0
Collisions: 0
Time elapsed: 00:10:11

Initial climb altitude: 4000

Go around altitude: 4000

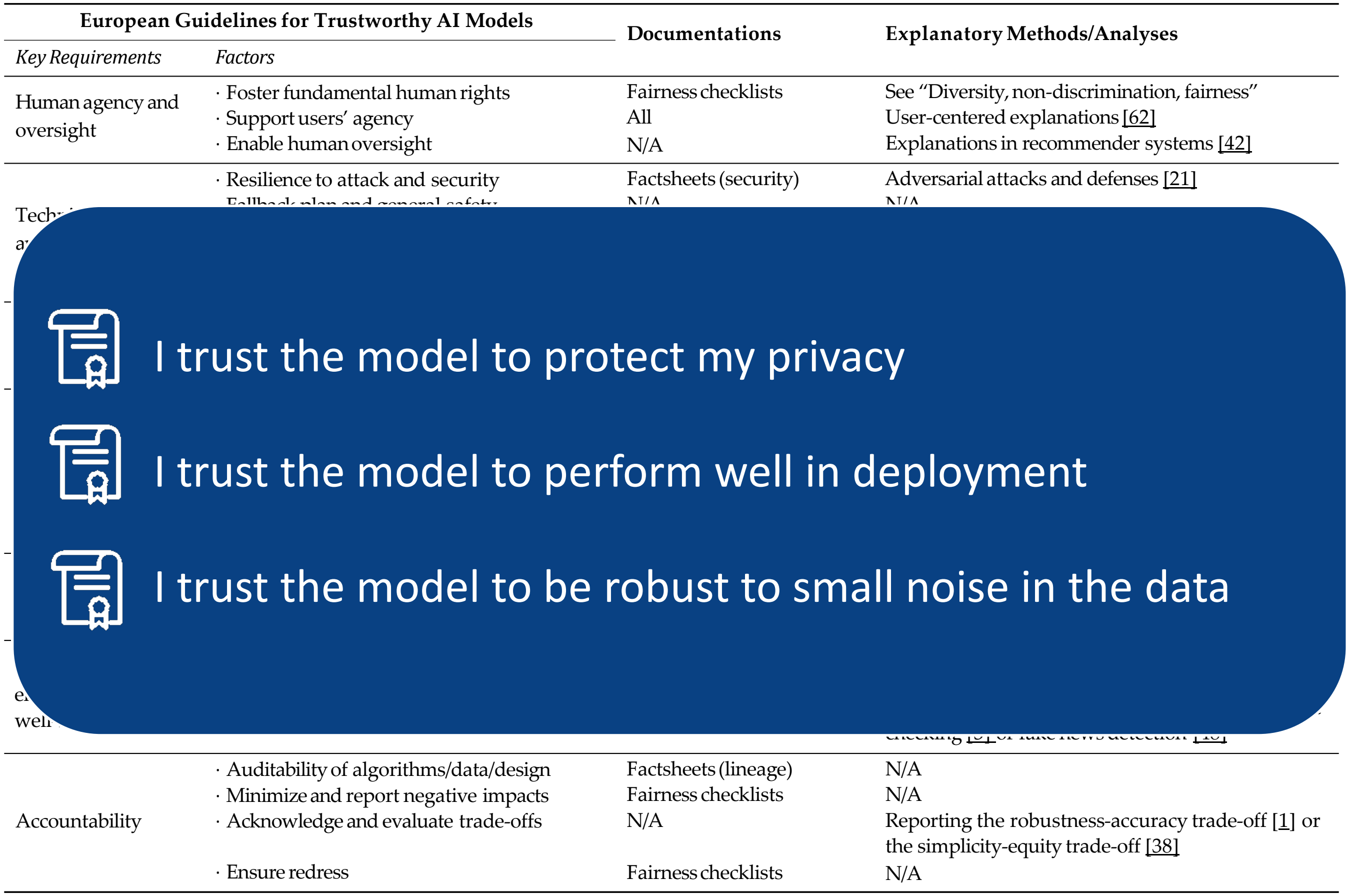
Select procedures:

- ☐ ALL
- ☐ ALL STARS FOR RWY 04
- ☐ ALL SIDS FOR RWY 04
- ☑ ALL STARS FOR RWY 22
- ☐ ALL SIDS FOR RWY 22
- ☐ BAVO1F (SID 22)
- ☐ BAVO1H (SID 04)
- ☐ BAXE1F (SID 22)
- ☐ BAXE1H (SID 04)
- ☐ BILN2F (SID 22)
- ☐ BILN3H (SID 04)
- ☐ HLV5F (SID 22)
- ☐ HLV5H (SID 04)
- ☐ NETI1F (SID 22)
- ☐ NETI1H (SID 04)
- ☐ REGL1F (SID 22)
- ☐ REGL1H (SID 04)
- ☐ BAVO1T (STAR 22)
- ☐ BAVO1W (STAR 04)
- ☐ BAXE1T (STAR 22)
- ☐ BAXE1W (STAR 04)
- ☐ BILN2T (STAR 22)
- ☐ BILN2W (STAR 04)
- ☐ HLV2T (STAR 22)
- ☐ HLV5W (STAR 04)
- ☐ REGL1T (STAR 22)
- ☐ REGL1W (STAR 04)
- ☐ TUSI1T (STAR 22)
- ☐ TUSI1W (STAR 04)
- ☐ D04_MORUV (APPR 04)
- ☐ D04_ODRAN (APPR 04)
- ☐ D04_OTA (APPR 04)
- ☐ D04_POLOM (APPR 04)
- ☐ D22_BOGTU (APPR 22)
- ☐ D22_EKMIT (APPR 22)
- ☐ D22_OTA (APPR 22)



Example: Aircraft collision detection

Contracts in AI



European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	· Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	· Support users' agency	All	User-centered explanations [62]
	· Enable human oversight	N/A	Explanations in recommender systems [42]
Techn… a…	· Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [21]
	· Fallback plan and general safety	N/A	N/A
e… well…			checking [5] or fake news detection [10]
Accountability	· Auditability of algorithms/data/design	Factsheets (lineage)	N/A
	· Minimize and report negative impacts	Fairness checklists	N/A
	· Acknowledge and evaluate trade-offs	N/A	Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38]
	· Ensure redress	Fairness checklists	N/A

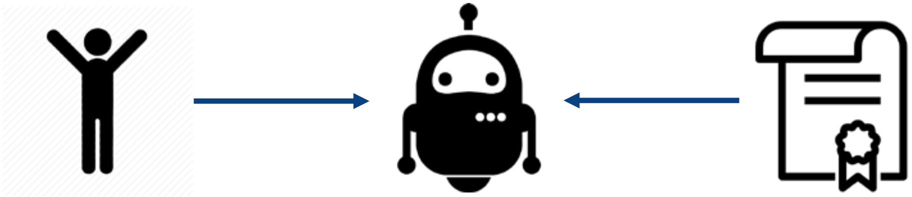
I trust the model to protect my privacy

I trust the model to perform well in deployment

I trust the model to be robust to small noise in the data

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 Yoav Goldberg. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2021)*, 2021.

Human-AI trust reframed



H (*human*) trusts M (*machine*) if…

- H believes that M will fulfill a particular set of contracts that are in H’s best interests;
- H accepts vulnerability to M's actions;

So that H can...

- anticipate the impact of M’s decisions on H

therefore making the interaction more predictable, enabling collaboration.

THE UNIVERSITY OF MELBOURNE

Trustworthiness and trust

Trustworthy AI

An AI model/agent is **trustworthy** if:

- It can fulfill its set of contracts

This is independent of trust:

- *Trust* does not imply *trustworthiness*.
- *Trustworthiness* does not imply *trust*

Warranted and unwarranted trust

Warranted trust = trust is *caused by* trustworthiness

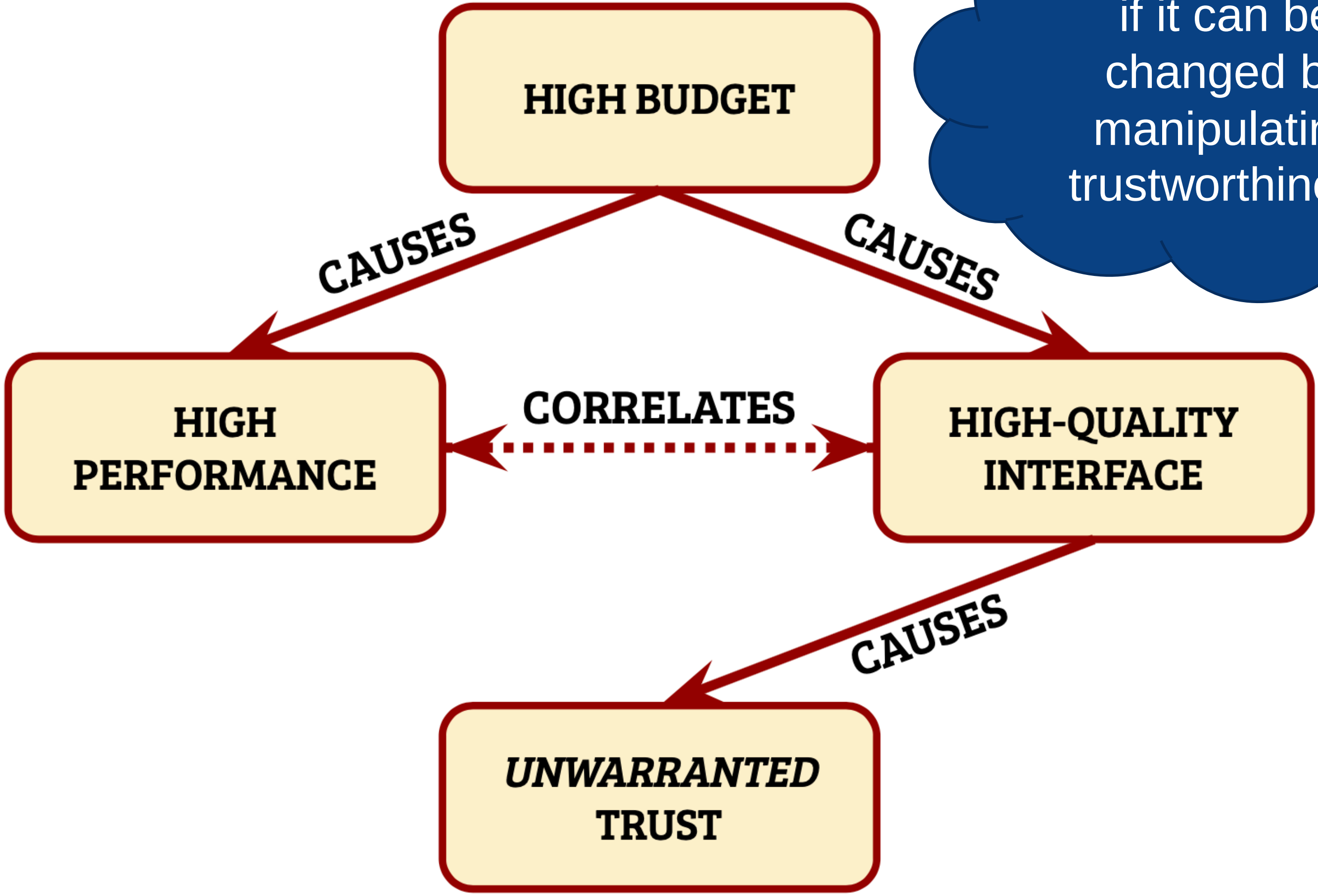
Unwarranted trust = trust is caused by something else

	Trusted	Distrusted
Trustworthy	Warranted Trust*	Unwarranted Distrust
Not trustworthy	Unwarranted Trust	Warranted Distrust**

* If caused by trustworthiness

** If caused by lack of trustworthiness

Warranted trust example



Trust is warranted if it can be changed by manipulating trustworthiness

Desirable outcomes of trust

We should pursue:

- Warranted trust
- Warranted distrust

We should try to avoid:

- Unwarranted trust
- Unwarranted distrust

Unwarranted trust is not caused by trustworthiness, therefore:

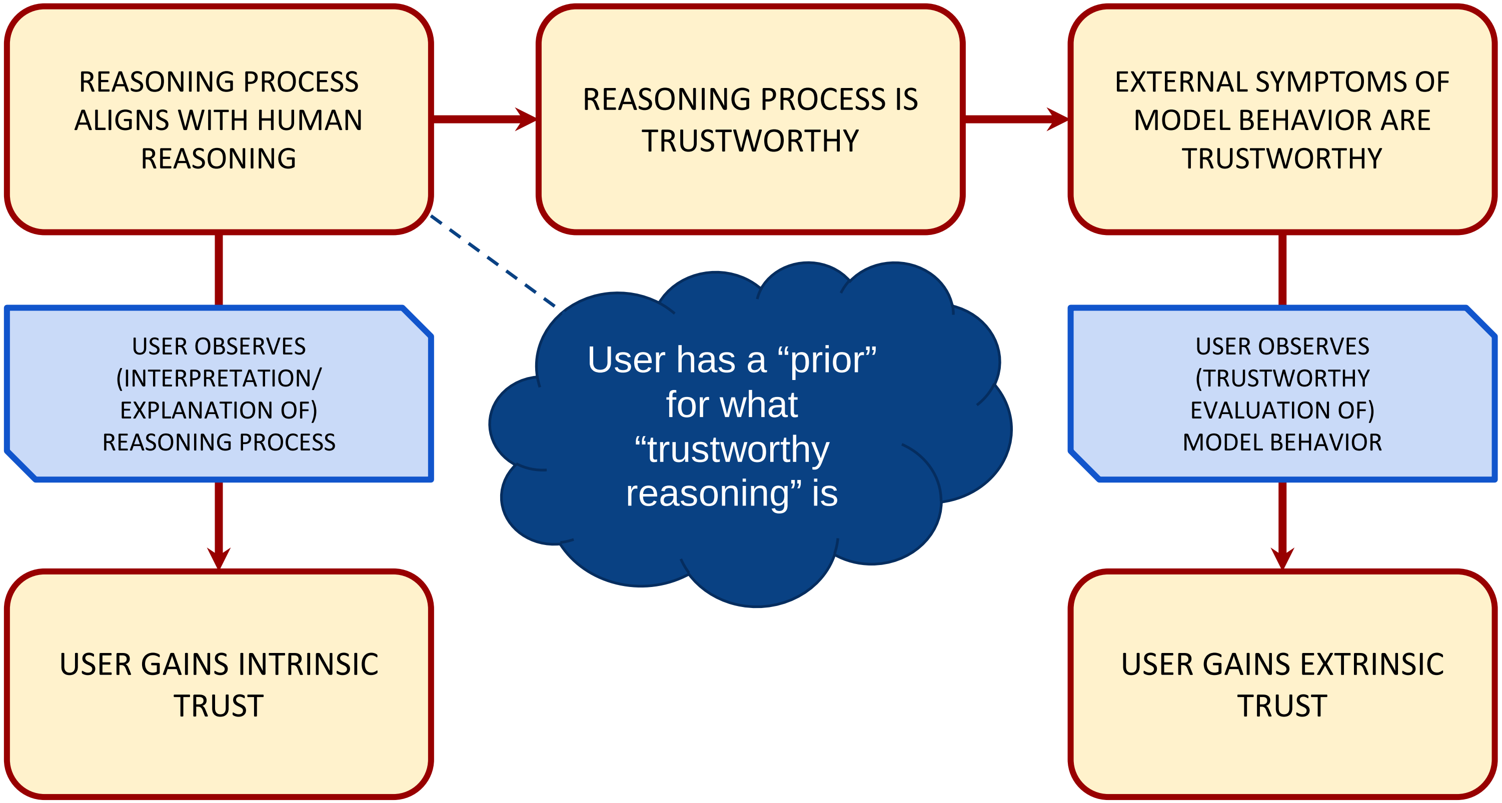
we cannot rely on it to result in proper anticipation

THE UNIVERSITY OF
MELBOURNE

Intrinsic and extrinsic trust

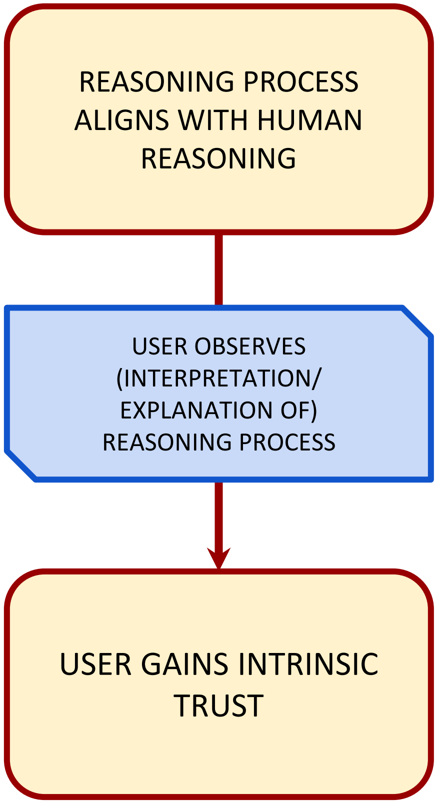
Intrinsic and extrinsic trust

What **causes** warranted trust?


REASONING PROCESS ALIGNS WITH HUMAN REASONING
REASONING PROCESS IS TRUSTWORTHY
EXTERNAL SYMPTOMS OF MODEL BEHAVIOR ARE TRUSTWORTHY
USER OBSERVES (INTERPRETATION/ EXPLANATION OF) REASONING PROCESS
User has a "prior" for what "trustworthy reasoning" is
USER OBSERVES (TRUSTWORTHY EVALUATION OF) MODEL BEHAVIOR
USER GAINS INTRINSIC TRUST
USER GAINS EXTRINSIC TRUST

Warranted intrinsic trust

What **causes** warranted intrinsic trust?


REASONING PROCESS ALIGNS WITH HUMAN REASONING
USER OBSERVES (INTERPRETATION/ EXPLANATION OF) REASONING PROCESS
USER GAINS INTRINSIC TRUST

Examples

We trust a medical specialist when they explain the various factors that led to their diagnosis, citing respectable studies to justify their claims.

We trust an AI-based credit-scoring model because we have an explanation of the important features for each decision and advice how to change the decision.

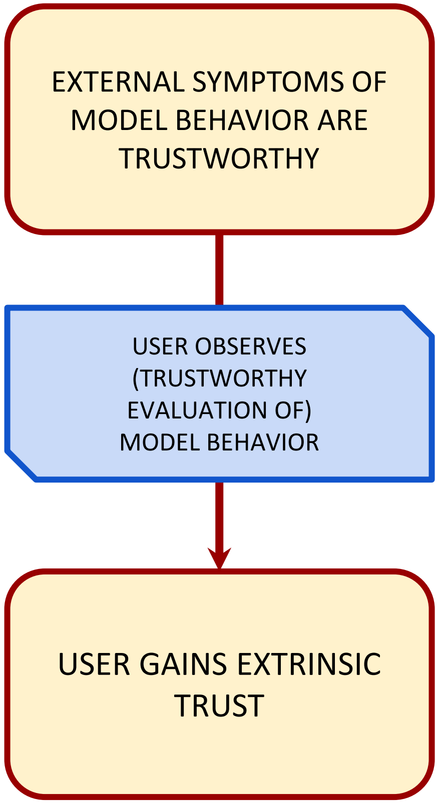
Warranted extrinsic trust

Examples

We trust a medical specialist because they have passed several examinations of their competence and have a long history of making correct diagnosis for us.

We trust an AI-based credit-scoring model because we have seen the results on test data and have seen it work well in deployment

What **causes** warranted extrinsic trust?


EXTERNAL SYMPTOMS OF MODEL BEHAVIOR ARE TRUSTWORTHY
USER OBSERVES (TRUSTWORTHY EVALUATION OF) MODEL BEHAVIOR
USER GAINS EXTRINSIC TRUST

Increasing trust in AI

Increasing intrinsic trust

- Explainability
  - ➢ Simplicity
  - ➢ Transparency
  - ➢ Explanation

Intrinsic = understanding the reasoning
Extrinsic = understanding the behaviour

Increasing extrinsic trust

- *By proxy*: a trusted expert judges the AI model
- *Post-deployment data*: examples where contracts are upheld after deployment in the real environment
- *Test sets*: examples distributed in a particular way

THE UNIVERSITY OF MELBOURNE

Use, misuse, disuse, abuse: unwarranted trust and distrust

Factors that determine use of automation

According to Parasurman and Riley (1997), there are three main factors that determine whether someone will use AI/automation:

Mental workload	Cognitive overhead	Trust (!)

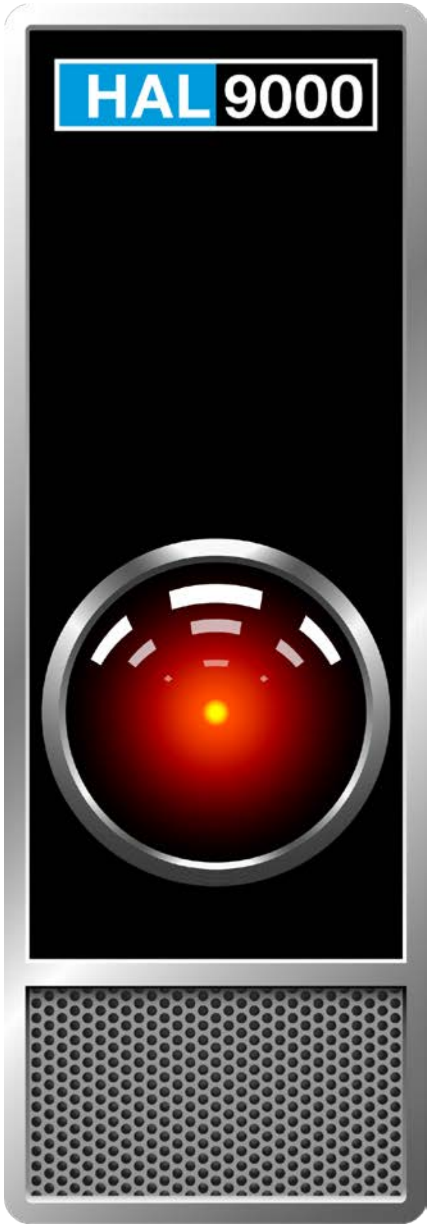
Misuse of automation

Definition: Using automation when it should not be used.

Cause: Unwarranted trust, due to:

- Overreliance on automation (e.g. high mental workload)
- Decision biases from heuristic decision making
- Human monitoring errors (e.g. unclear error messages, high false alarm)
- Machine monitoring errors
- Automation bias

Impact: Issues caused by automation and not detected by human (e.g. complacency)



Disuse of automation

Definition: Not using automation when it should be used.

Cause: Unwarranted distrust, due to:

- Human monitoring errors (low false alarm rate*)*
- Machine monitoring errors
- Human bias

Impacts: Disabling/ignoring alarms, leading to issues not detected by human

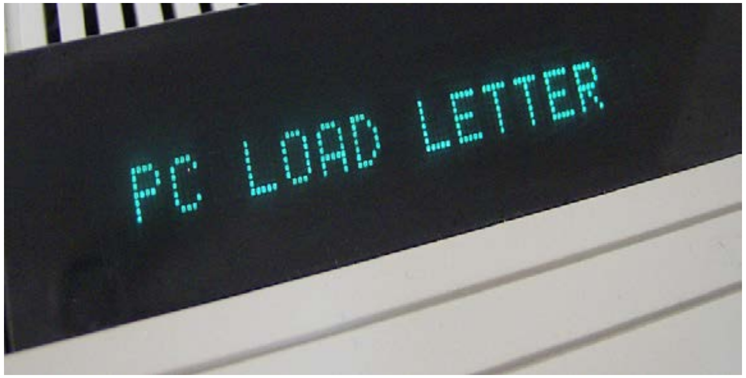
Abuse of automation

Definition: Deploying automation when it should not be used (e.g. designing without considering the operator).

Cause: Unwarranted trust from the *designer*, due to:

- Distrust in human operators
- Automation bias
- Arrogance

Impact: Mismatch in human-automation interface, lack of *situation awareness* from the operator



Example: Therac-25

Therac-25 was a radiation therapy machine, controlled by software

Outcome Therac-25 gave six patients huge overdoses of radiation, leading to their deaths.

Causes Software errors from

- Misuse: Unwarranted trust from radiographers? Error codes were meaningless to operators: e.g. “Malfunction 16”
- Disuse(?): Hardware interlocks removed from earlier Therac versions but not replaced by software.
- Abuse: Designing Therac-25 with little input from radiographers; arrogance from designers when burns and early deaths were reported.

A radiation therapy machine (not Therac-25!).

THE UNIVERSITY OF MELBOURNE

Trust and ethics in AI

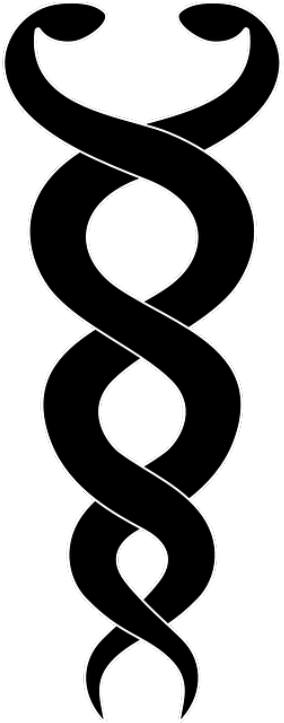
Trust and ethics

Trust

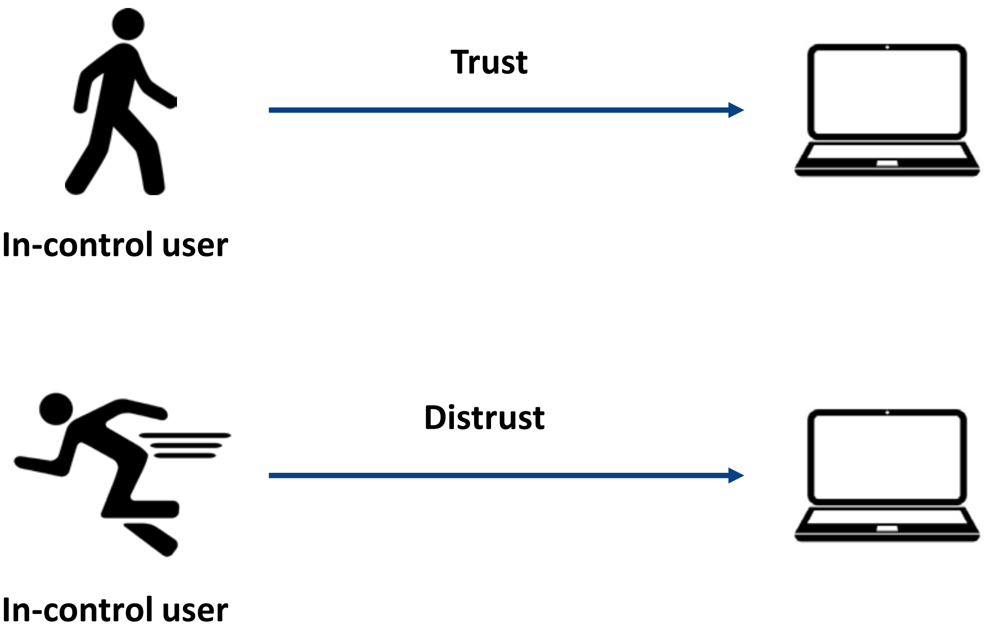
≠

Ethics

But! They are closely related and cannot be separated.



User trust



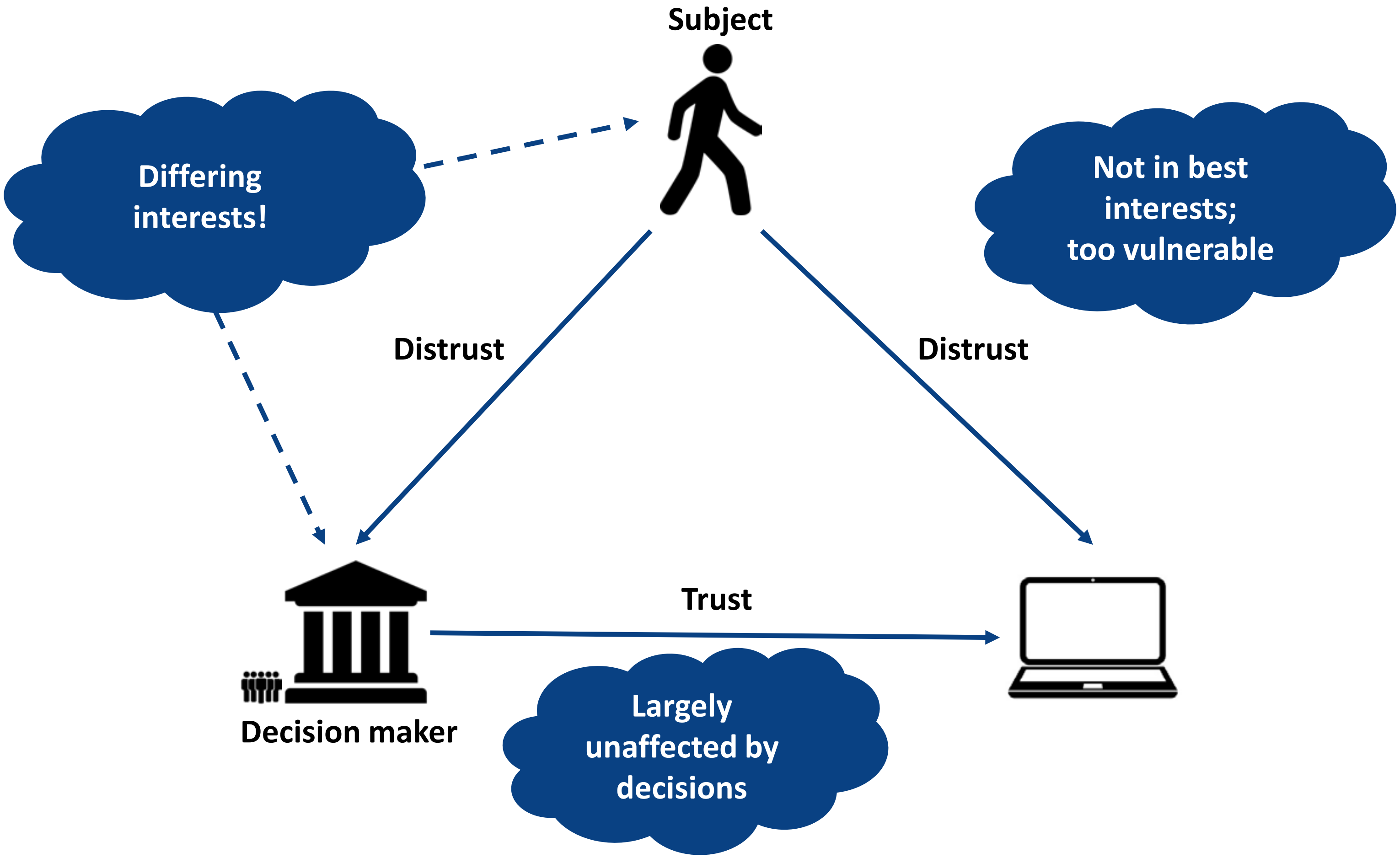
Trust

In-control user

Distrust

In-control user

Ethical issues in AI



Subject

Differing interests!

Not in best interests; too vulnerable

Distrust

Distrust

Trust

Decision maker

Largely unaffected by decisions

Trust, machines, and ethics: summary

Trust

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

Contractual trust

Warranted and unwarranted trust and distrust

Causes of trust

- Intrinsic (reasoning)
- Extrinsic (behavior)

Key takeaways

Be explicit about which contracts hold for your models/systems

Trust is only (ethically) desirable if it is warranted

Distrust is desirable if it is warranted

Incorrectly calibrated trust leads to real problems

Ethical issues in AI emerge from different interests between people, and therefore different levels of trust

Thank you

—