SWINBURNE UNIVERSITY OF TECHNOLOGY

COS30019 – Introduction to AI: Week 2

AI Ethics & Responsible AI

SWINBURNE UNIVERSITY OF TECHNOLOGY

AI Ethics & Responsible AI

- **AI Expert: We Urgently Need Ethical Guidelines & Safeguards to Limit Risk of Artificial Intelligence**
  - *https://www.youtube.com/watch?v=YHwP0yYciF8*

Overview

- Ethics
- Ethics of AI
- Emerging AI Ethical Issues
- Responsible AI
  - *Ethics by Design*
- Ethical guidelines and Safeguards

Philosophy – The root of them all

- **Metaphysics**: the study of being/existence (aka. ontology) and the reality of the universe (aka. cosmology)
- **Epistemology**: the study of knowledge and how we know things about the universe (e.g., perception, beliefs and justifications)
- **Axiology (Ethics & Aesthetics)**: the study of moral principles and what constitutes right/wrong conduct
- **Others**:
  - Logic, Political philosophy, etc.

Ethics

- “the study of values - good and bad, right and wrong” & “quality of life impact”
- Meta-Ethics:
  - Studying where our ethics come from
- Normative Ethics:
  - Generating moral standards for right vs. wrong
  - The consequences of our behaviors on others
- Applied Ethics:
  - Examining specific controversial issues (nuclear war, animal rights)

Ethics in Scientific Research/Innovation

- What are some examples of scientific research in which **ethics** play a large role?
  - Stem cell research
  - Cloning/genetically modified food
  - Nuclear technology
  - Animal rights
  - Medical trials
  - Disease research (e.g. biowarfare)
  - …

.

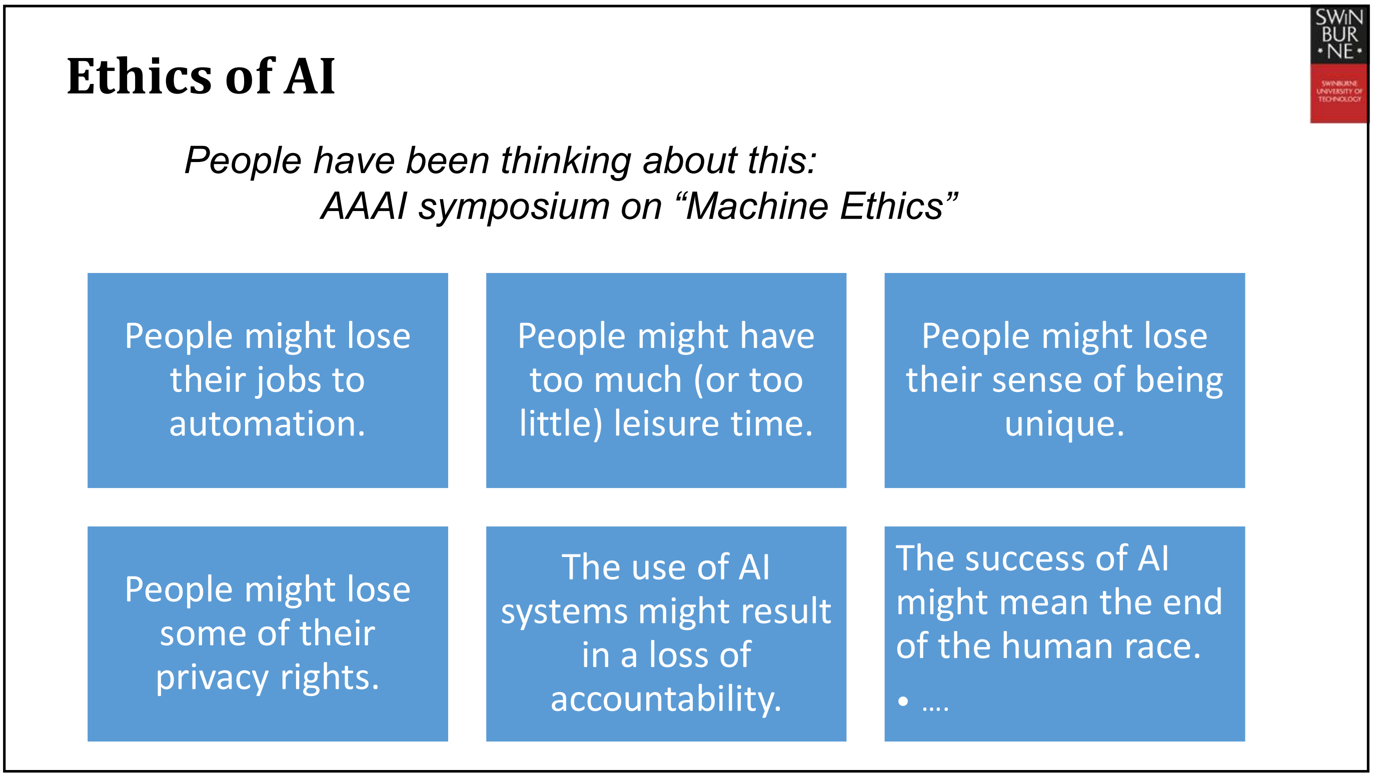
Potentially severe consequences

- New technologies have unintended negative side effects
  - → Scientists and engineers must think about:
    - how they should act on the job
    - what projects should or should not be done
    - and how they should be handled

Ethics in Computer Science/Technology?

- Intellectual property
    - E.g., DRM (digital rights management)
    - Patents/copyright
- Privacy
- Accuracy of Information
- Access and Equality
- …

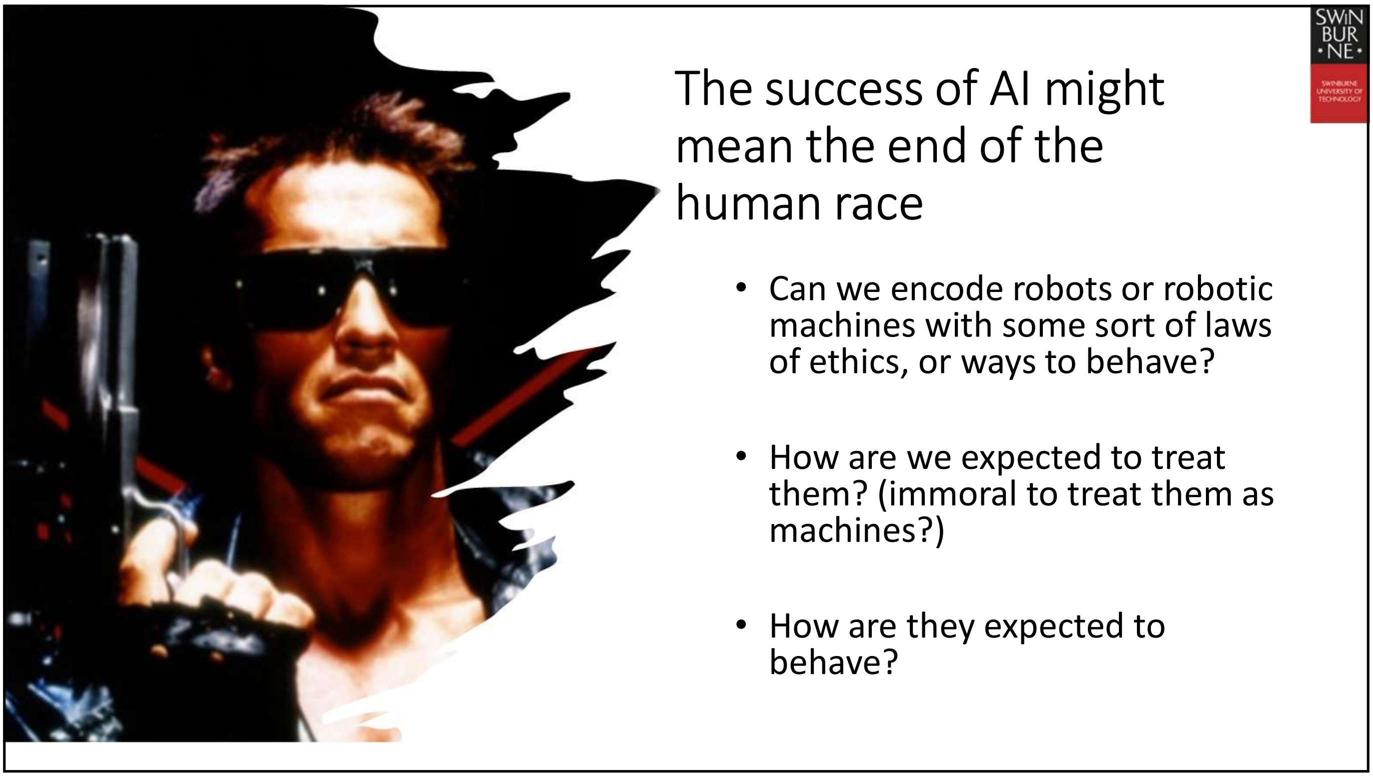
Ethics of AI



Ethics of AI

People have been thinking about this:
AAAI symposium on “Machine Ethics”

- People might lose their jobs to automation.
- People might have too much (or too little) leisure time.
- People might lose their sense of being unique.
- People might lose some of their privacy rights.
- The use of AI systems might result in a loss of accountability.
- The success of AI might mean the end of the human race.
  - ….

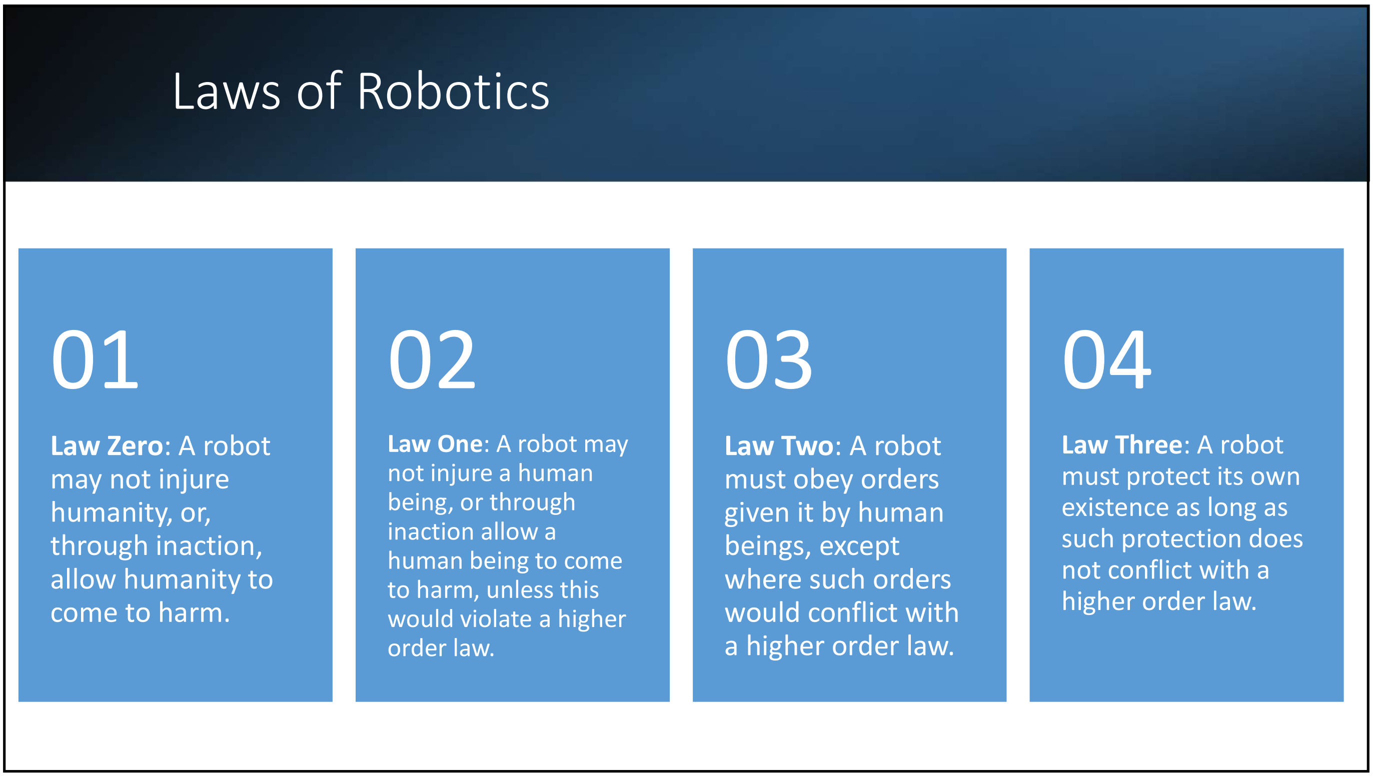


The success of AI might mean the end of the human race

- Can we encode robots or robotic machines with some sort of laws of ethics, or ways to behave?
- How are we expected to treat them? (immoral to treat them as machines?)
- How are they expected to behave?

SWIN BUR NE
SWINBURNE UNIVERSITY OF TECHNOLOGY

11



Laws of Robotics

01

Law Zero: A robot may not injure humanity, or, through inaction, allow humanity to come to harm.

02

Law One: A robot may not injure a human being, or through inaction allow a human being to come to harm, unless this would violate a higher order law.

03

Law Two: A robot must obey orders given it by human beings, except where such orders would conflict with a higher order law.

04

Law Three: A robot must protect its own existence as long as such protection does not conflict with a higher order law.

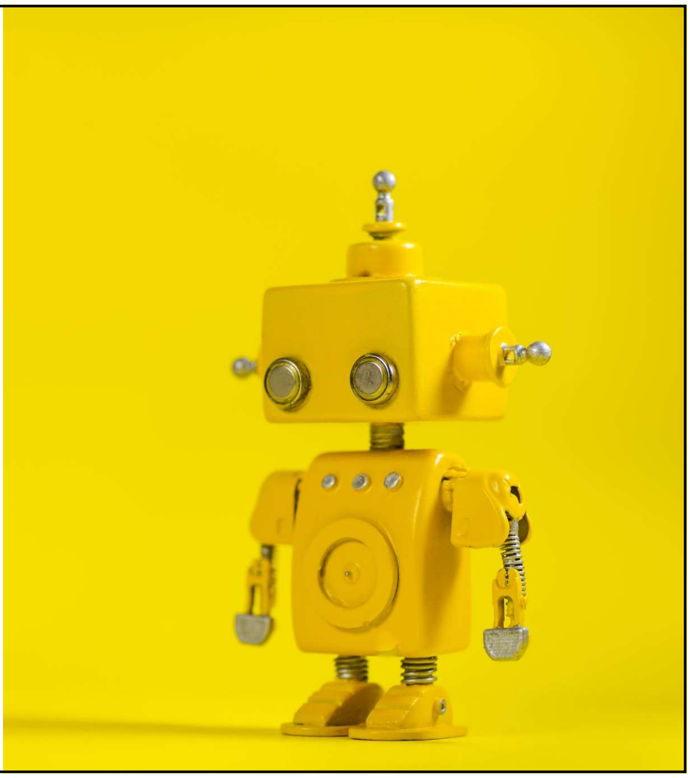
12

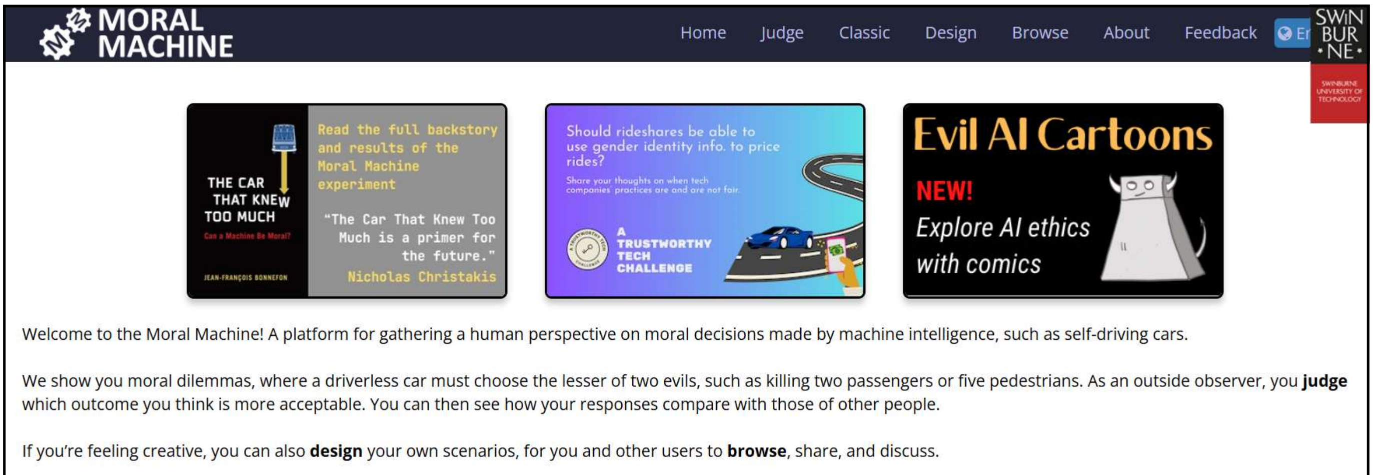
Robot Rights

- *Robot rights are like animal rights* –

  David J. Calverly

- Examples
  - Can you abuse a robot? E.g., to release your own stress?





Read the full backstory and results of the Moral Machine experiment

"The Car That Knew Too Much is a primer for the future." Nicholas Christakis

Should rideshares be able to use gender identity info. to price rides?

Share your thoughts on when tech companies' practices are and are not fair.

Explore AI ethics with comics

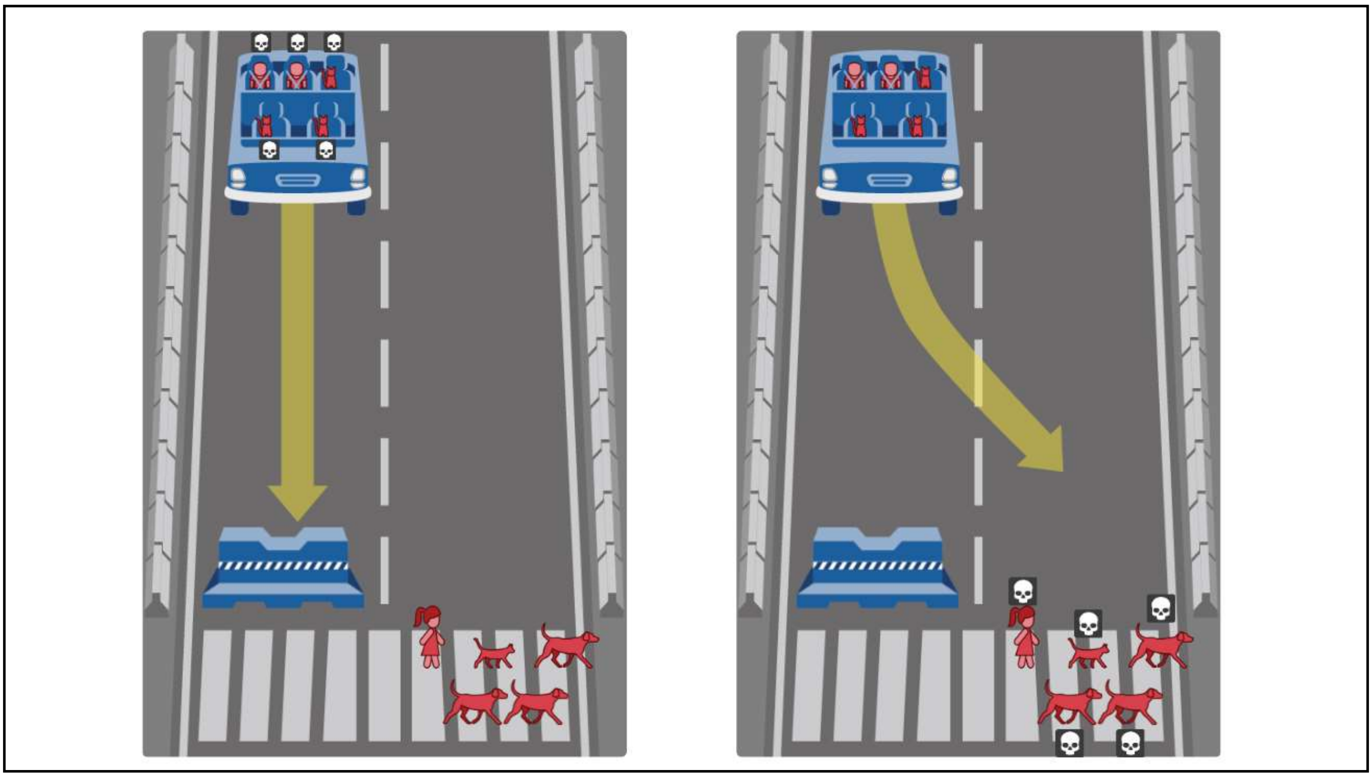
Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

Back to – Laws of Robotics

- How should they be implemented??
  - https://www.moralmachine.net/



15

16

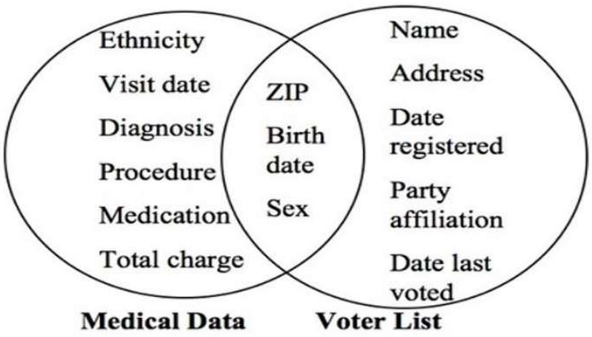
Emerging AI Ethical Issues

- **Lack of Transparency**
- **Bias and Discrimination**
- **Privacy Concerns**
- **Ethical Dilemmas**
- **Security Risks**
- **…**

William Weld vs Latanya Sweeney

Massachusetts Group Insurance Commission (1997): Anonymized medical history of state employees

Latanya Sweeney (MIT grad student): $20 – Cambridge voter roll



Ethnicity
Visit date
Diagnosis
Procedure
Medication
Total charge

ZIP
Birth date
Sex

Name
Address
Date registered
Party affiliation
Date last voted

Medical Data
Voter List

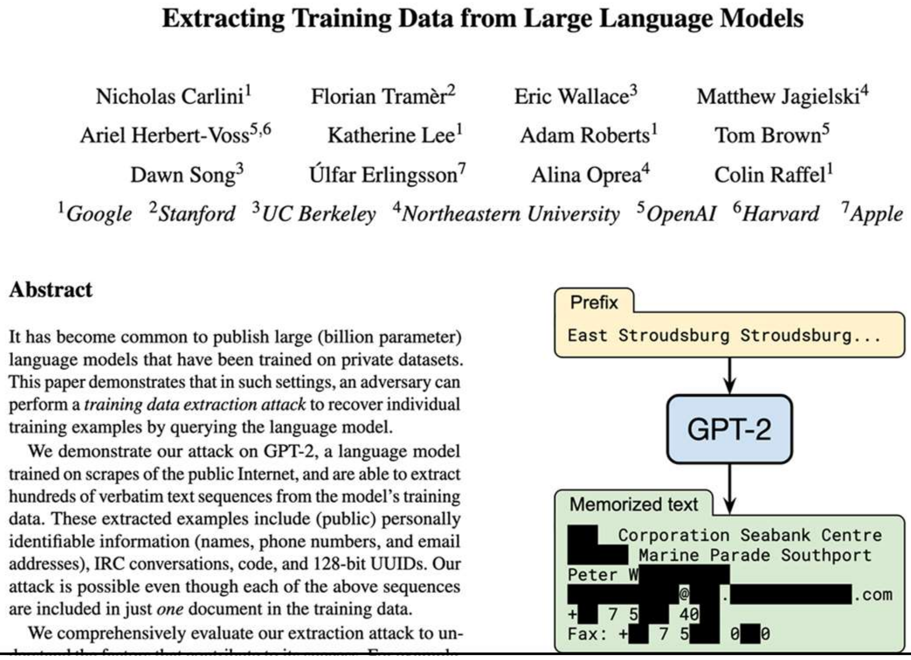
born July 31, 1945
resident of 02138

64%

Uniquely identifiable with ZIP + birth date + gender (in the US population)

Golle, “Revisiting the Uniqueness of Simple Demographics in the US Population”, WPES 2006

Recent Privacy Attack on Large Language Models



Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Abstract

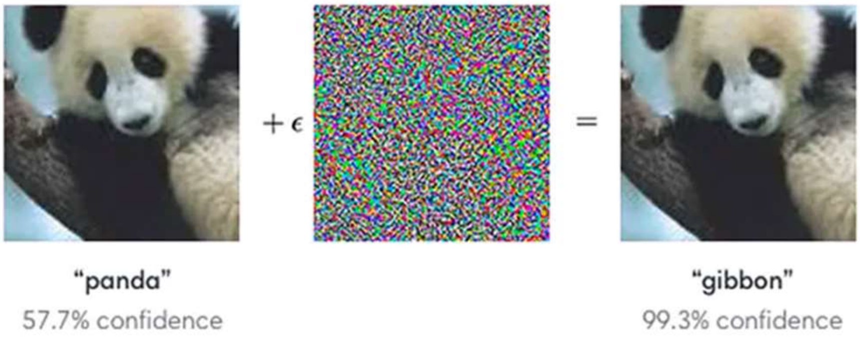
It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model.

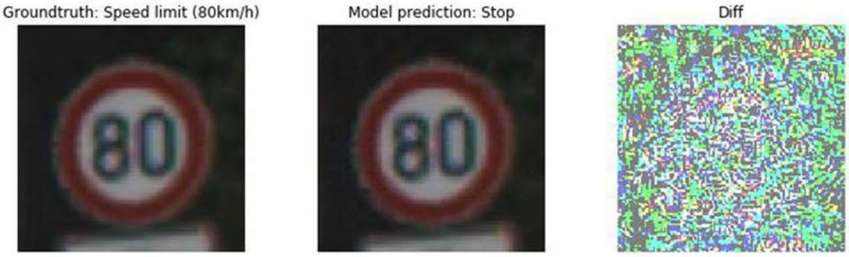
We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just *one* document in the training data.

We comprehensively evaluate our extraction attack to un-

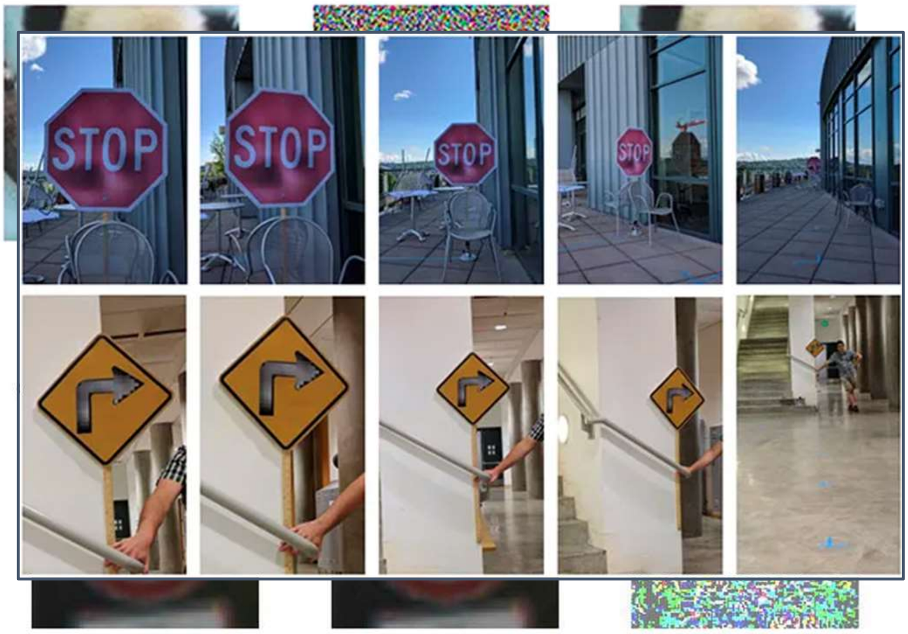
```
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40
Fax: +[redacted] 7 5[redacted] 0[redacted]0
```

Adversarial Attacks on Machine Learning Models



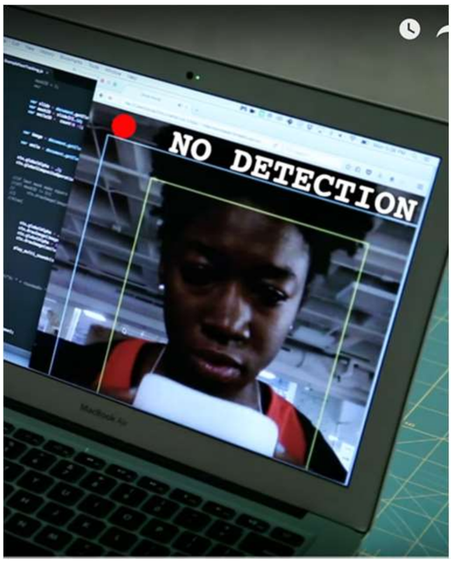


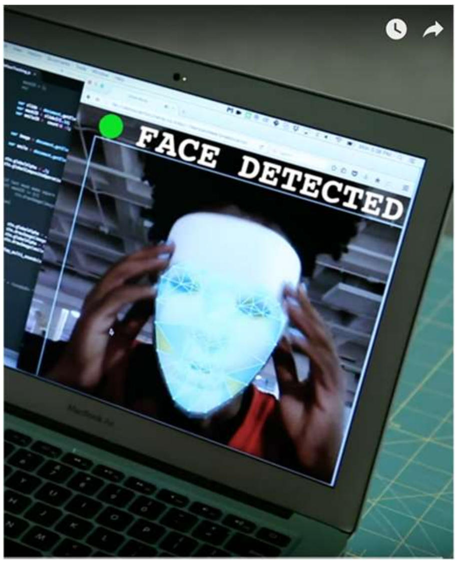
Adversarial Attacks on Machine Learning Models



- Goodfellow et al. Explaining and harnessing adversarial examples, ICLR'15
- Eykholt et al. Robust Physical-World Attacks on Deep Learning Models, CVPR'18
- Rauschmayr et al. Amazon SageMaker Debugger: A system for real-time insights into machine learning model training, MLSys'21

The Coded Gaze [Joy Buolamwini 2016]

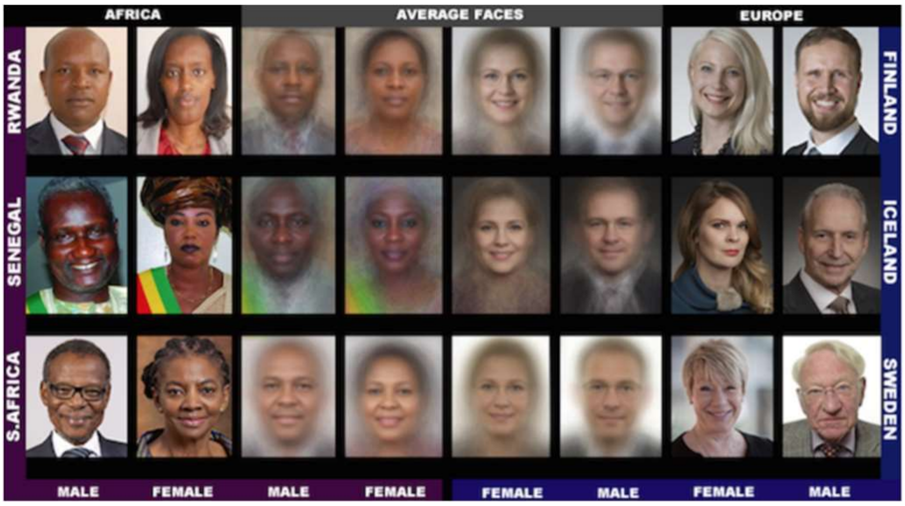




Face detection software:
Fails for some darker faces

https://youtu.be/162VzSzzoPs

Gender Shades [Joy Buolamwini & Timnit Gebru, 2018]

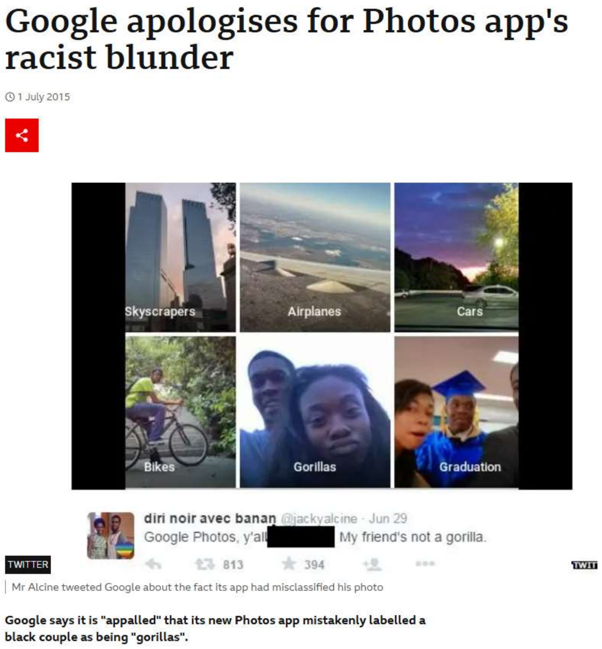


- Facial analysis software:

Higher accuracy for light skinned men

- Error rates for dark skinned women: 20% - 34%

Algorithmic Bias



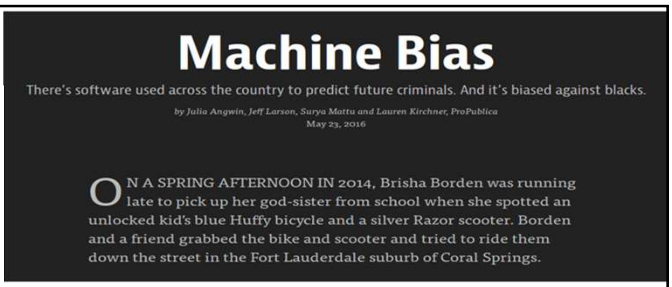
Google apologises for Photos app's racist blunder

1 July 2015

diri noir avec banan @jackyalcine · Jun 29
Google Photos, y'all [redacted] My friend's not a gorilla.

Mr Alcine tweeted Google about the fact its app had misclassified his photo

Google says it is "appalled" that its new Photos app mistakenly labelled a black couple as being "gorillas".



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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

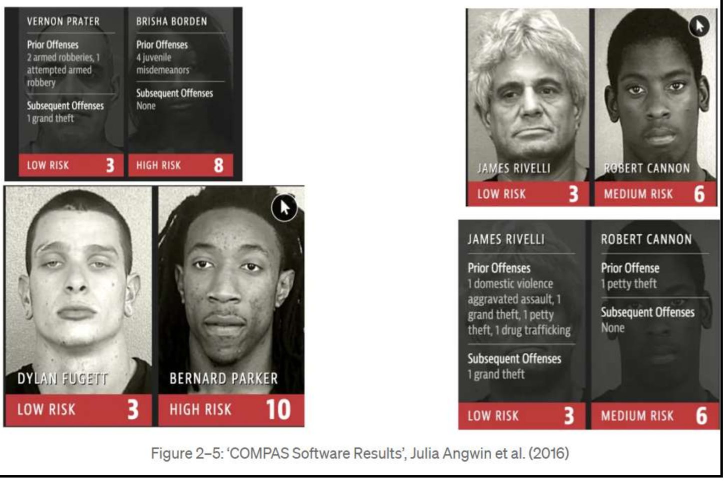
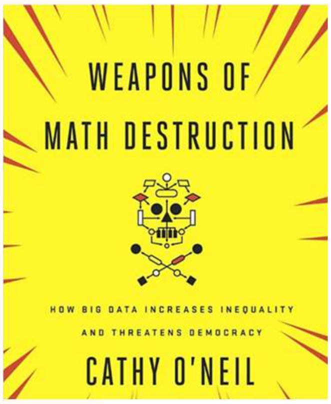
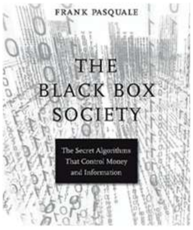


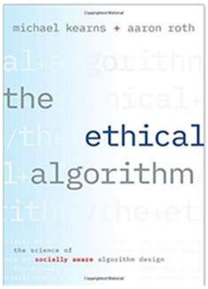
Figure 2–5: 'COMPAS Software Results', Julia Angwin et al. (2016)

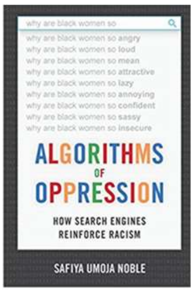
Algorithmic Bias

- Ethical challenges posed by AI systems
- Inherent biases present in society
- Reflected in training data
- AI/ML models prone to amplifying such biases


WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL




michael kearns + aaron roth
the ethical algorithm
the science of socially aware algorithm design


why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE

Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

27

Bias, Discrimination & Machine Learning

Isn’t bias a technical concept?

Selection, sampling, reporting bias, Bias of an estimator, Inductive bias

Isn’t discrimination the very point of machine learning?

Unjustified basis for differentiation

[Barocas & Hardt 2017]

28

Types of Harm

Harms of allocation

withhold opportunity or resources

Harms of representation

reinforce subordination along the lines of identity, stereotypes

[Cramer et al 2019, Shapiro et al., 2017, Kate Crawford, "The Trouble With Bias" keynote NeurIPS'17]

Regulated Domains

Credit (Equal Credit Opportunity Act)

Education (Civil Rights Act of 1964; Education Amendments of 1972)

Employment (Civil Rights Act of 1964)

Housing (Fair Housing Act)

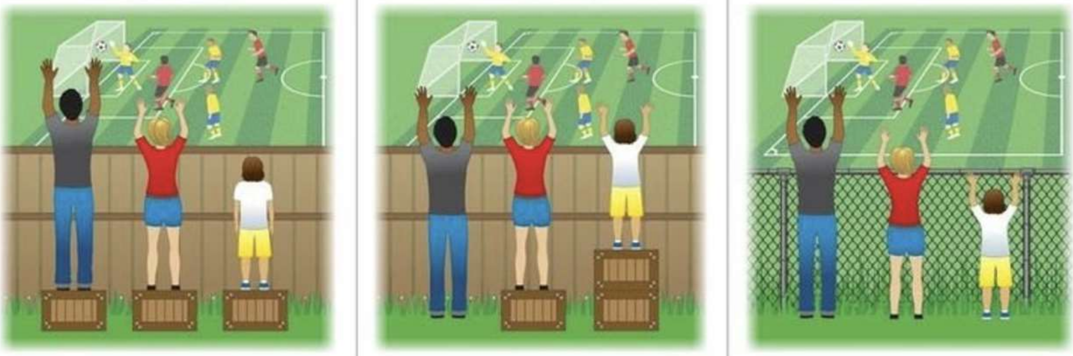
'Public Accommodation' (Civil Rights Act of 1964)

What about **Healthcare**? **Justice Systems**?

…

Extends to marketing and advertising; not limited to final decision

[Barocas & Hardt 2017]



Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

[https://www.reddit.com/r/GCdebatesQT/comments/7qpbpp/food_for_thought_equality_vs_equity_vs_justice/]

Fairness

GDPR

Privacy

Transparency

CALIFORNIA CONSUMER PRIVACY ACT OF 2018

Explainability

AI Teams Lack Visibility into Their Models

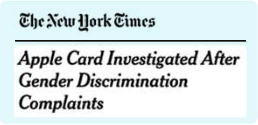
- **Model Transparency**

MIT Technology Review

Facebook whistleblower Frances Haugen's testimony at the Senate today ***raised serious questions about how Facebook's algorithms work…***

- **Model Decay**

- **Model Bias**



The New York Times

Apple Card Investigated After Gender Discrimination Complaints

- **Model Compliance**

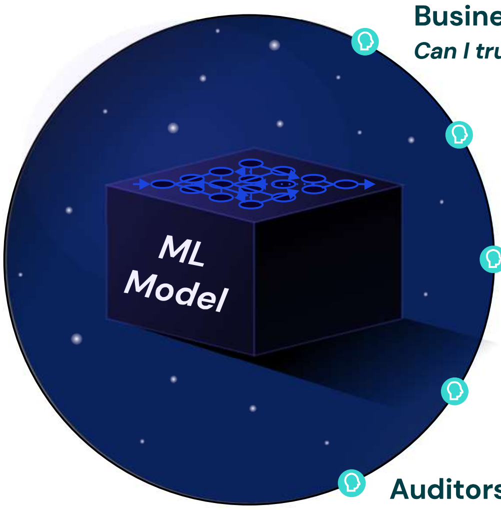
> *"On Artificial Intelligence,* ***trust is a must****,* ***not a nice to have****.*
> **- EU Commission**

Most ML Models are Opaque

No Explanations
of model behavior

No Understanding
of feature impact and fairness

No Monitoring
to catch potential bias or drift



Business User
Can I trust our AI?

Customer Support
How do I answer this complaint?

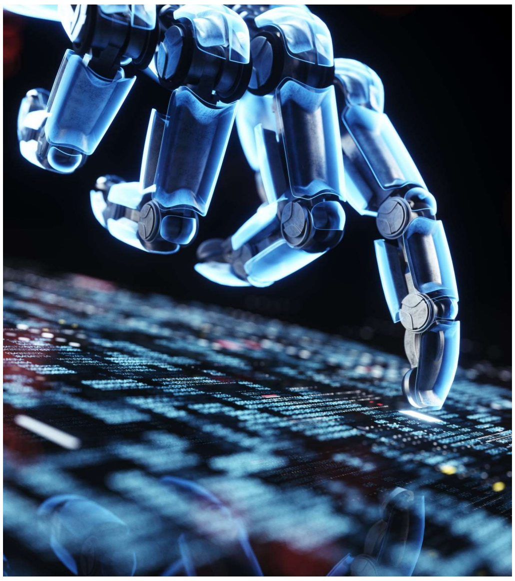
IT & Operations
How do I monitor & debug?

Data Scientists
How does this model work?

Auditors & Regulators
Are these decisions fair?

Responsible AI

RESPONSIBLE AI: WHY CARE?



- AI systems act autonomously in our world
- Eventually, AI systems will make *better* decisions than humans
- **AI is designed, is an artefact**
- We need to make sure that the purpose put into the machine is the purpose which we really want

TAKING RESPONSIBILITY

- Responsibility / Ethics **in** Design
  - Ensuring that development processes take into account ethical and societal implications of AI as it integrates and replaces traditional systems and social structures
- Responsibility /Ethics **by** Design
  - Integration of ethical abilities as part of the behaviour of artificial autonomous systems
- Responsibility /Ethics **for** Design(ers)
  - Research integrity of researchers and manufacturers, and certification mechanisms

TAKING RESPONSIBILITY

- Responsibility / Ethics **in** Design
  - Ensuring that development processes take into account ethical as it nal

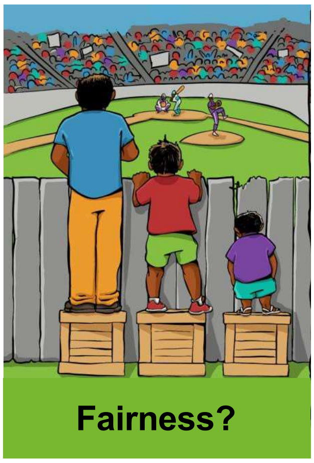
Can we guarantee that behaviour is ethical?

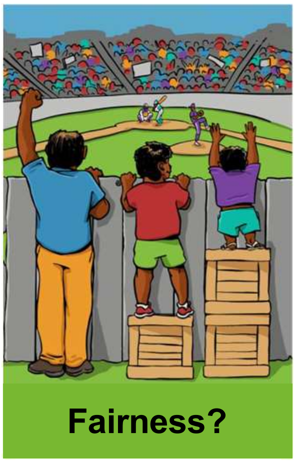
- Responsibility /Ethics **by** Design
  - Integration of ethical abilities as part of the behaviour of artificial autonomous systems
- Responsibility /Ethics **for** Design(ers)
  - Research integrity of researchers and manufacturers, and certification mechanisms

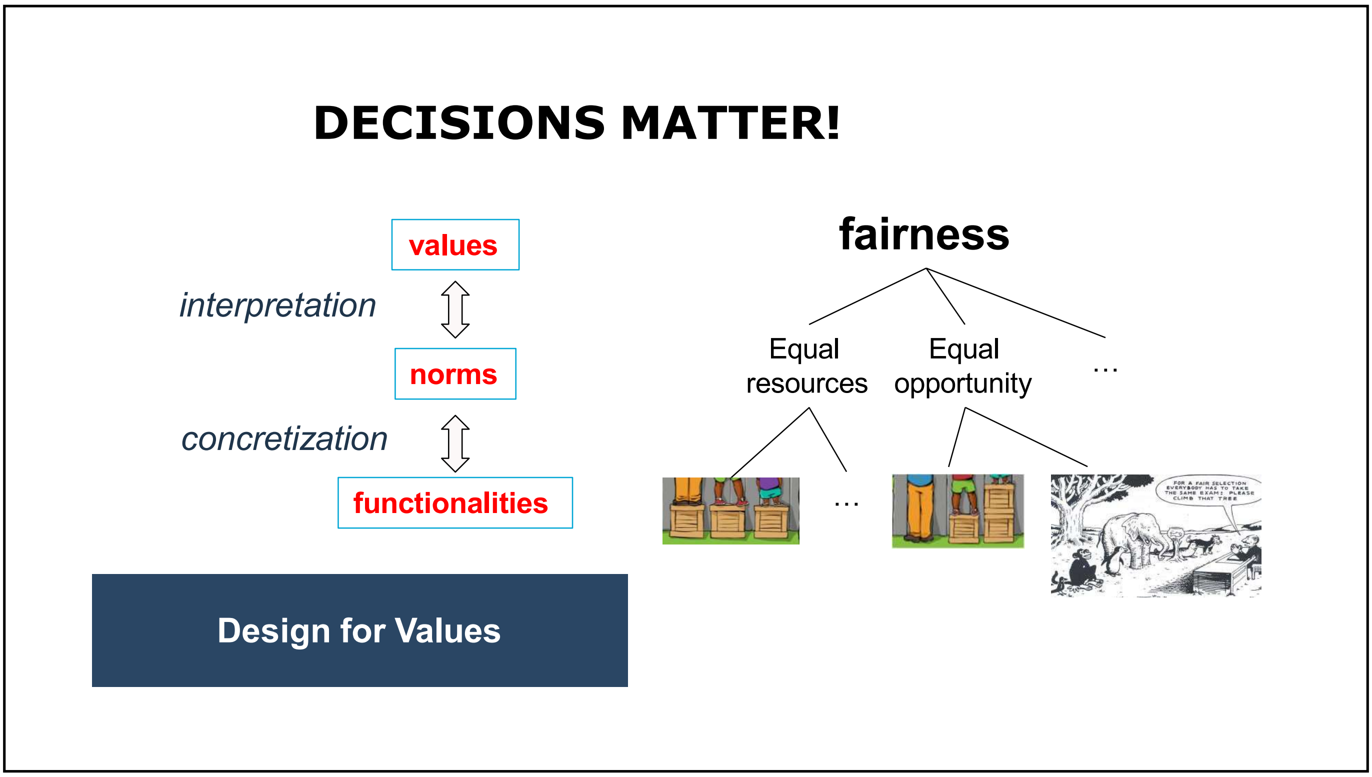
ETHICS BY DESIGN

- Can AI artefacts be built to be verifiably ethical?
  - What does that mean?
  - What is needed?
- Which values?
- Whose values?
- Which ethical rules?
- **Which interpretation?**

VALUES IN CONTEXT







DECISIONS MATTER!

values

interpretation

norms

concretization

functionalities

Design for Values

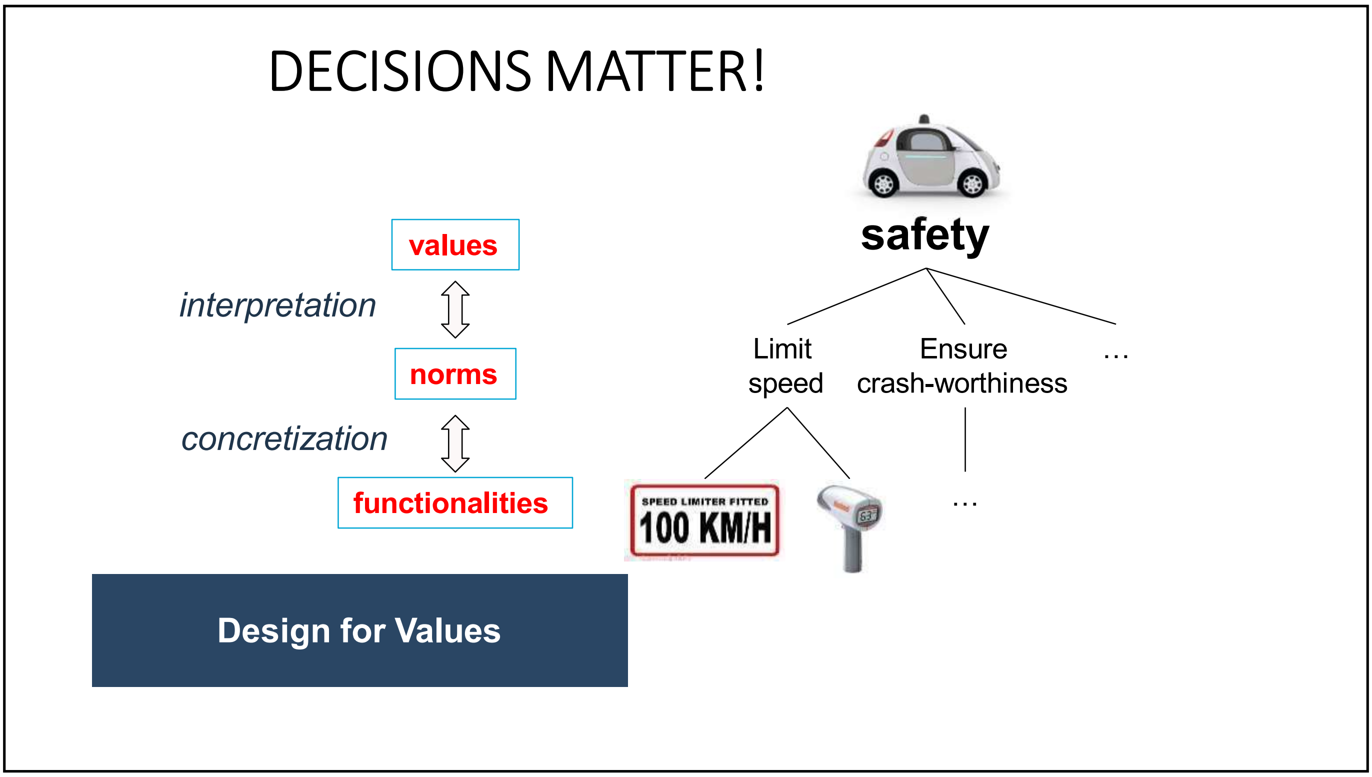
fairness

Equal resources

Equal opportunity

…

…
FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE



DECISIONS MATTER!

values

interpretation

norms

concretization

functionalities

Design for Values

safety

Limit speed

Ensure crash-worthiness

…

GUIDELINES – BE OPEN AND EXPLICIT

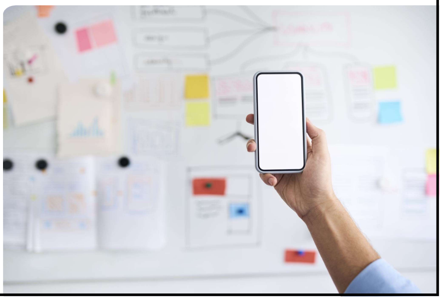
- Question your options and choices
- Motivate your choices
- Document your choices and options
- Compliance
  - External monitoring and control
  - Norms and institutions

- Engineering principles for policy
  - Analyze – synthetize – evaluate - repeat

https://medium.com/@virginiadignum/on-bias-black-boxes-and-the-quest-for-transparency-in-artificial-intelligence-bcde64f59f5b

ASK YOURSELF

- Who will be affected?
- What are the decision criteria we are optimizing for?
- How are these criteria justified?
- Are these justifications acceptable in the context we are designing for?
- How are we training our algorithm?
  - Does training data resemble the context of use?



Ethical Guidelines & Safeguards

- Initiatives by governments and regulators:
  - EU Artificial Intelligence Act
  - Australian Government's *Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings*
    - *https://consult.industry.gov.au/ai-mandatory-guardrails*
  - The US White House's *Blueprint for an AI Bill of Rights*:
    - https://www.whitehouse.gov/ostp/ai-bill-of-rights/
  - EC's High-level expert group on artificial intelligence (HLEG)'s *Ethics Guidelines for Trustworthy AI*:
    - https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
  - The U.S. Department of Commerce's National Institute of Standards and Technology (NIST)'s *Artificial Intelligence Risk Management Framework*:
    - https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
  - …

AI Trustworthiness Assurance

- Certification of AI Trustworthiness?
  - Assure that AI won't cause harm
  - Assure that AI systems follow the ethical guidelines
- Ethics-Based Auditing (EBA):
  - Ethics-Based Auditing to Develop Trustworthy AI:
    - https://doi.org/10.1007/s11023-021-09557-8
- Trustworthy AI Posture (TAIP):
  - Continuous automation approach for autonomous ethics-based audit of AI systems, by Guy Lupo, Bao Quoc Vo and Natania Locke:
    - https://doi.org/10.1016/B978-0-44-315991-6.00015-7

Summary

- Ethics has long been studied in philosophy
  - Values, moral principles, right vs. wrong
  - Plays an important role in Scientific Research, Innovation & Technology
- Ethics of AI has become critical as AI has become ubiquitous in everyday life
  - Critical AI ethical issues are emerging
- Responsible AI is a paradigm in AI design and development
  - *Ethics by design*
- AI Trustworthiness Assurance

References

1. AI Expert: We Urgently Need Ethical Guidelines & Safeguards to Limit Risk of Artificial Intelligence: *https://www.youtube.com/watch?v=YHwP0yYciF8*
2. Krishnaram Kenthapadi. **Responsible AI in Industry (Tutorial)**; https://sites.google.com/view/ResponsibleAITutorial
3. Virginia Dignum. **Responsible AI**. https://www.informatics-europe.org/images/ECSS/ECSS2019/Slides/ECSS2019_Dignum.pdf
4. **The Responsible AI Institute**: https://www.responsible.ai/

References/Readings

1. Sweeney, L., 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, *10*(05), pp.557-570.
2. Ohm, P., 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, *57*, p.1701.
3. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U. and Oprea, A., 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)* (pp. 2633-2650).
4. https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1