Ethics of AI

Dr. Ally S. Nyamawe

Researcher

UNU Macau

April 28, 2024

Presentation Outline

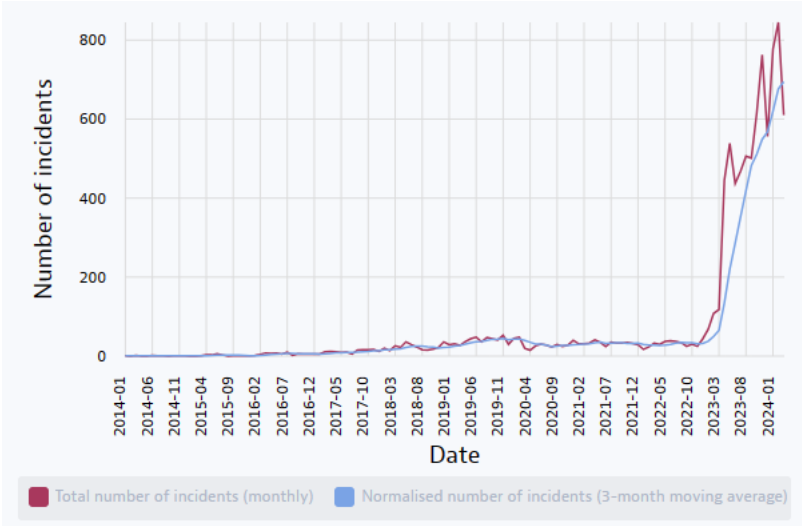
- Why should we regulate AI?
- Overview of AI Ethics
- Bias in Machine Learning
- Responsible AI
- Moral Machines

UNU Macau

Why Should We Regulate AI?

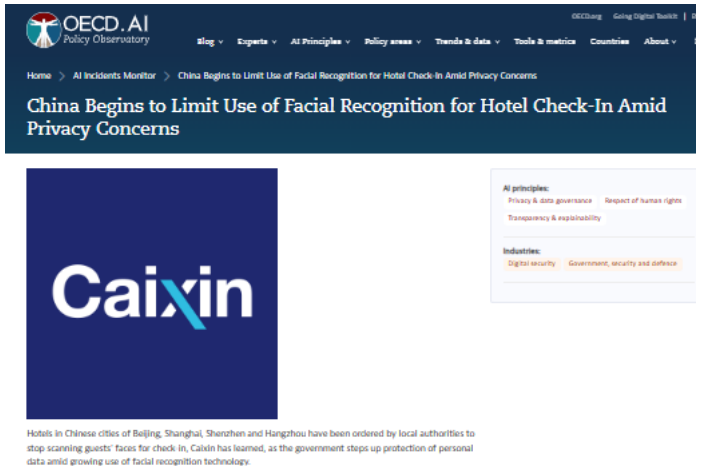
Current trends!

- Increasing number of incidents
- Privacy concerns
- Copyrights infringement
- Misuse of AI, e.t.c.



AIM: The OECD AI Incidents Monitor, an evidence base for trustworthy AI - OECD.AI

Hotels in Chinese cities of Beijing, Shanghai, Shenzhen and Hangzhou have been ordered by local authorities to stop scanning guests' faces for check-in



Deepfake of Marcos Jnr ordering military action against China causes alarm

UNU Macau

What is ethics (or moral philosophy)?

- **Ethics, the philosophical discipline concerned with what is morally good and bad and morally right and wrong.**
- Ethics etymology comes from the ancient greek "Ethos" (meaning "relating to one's character")
- Ethics vs morality
  - Are often taken as synonyms
  - Ethics can be seen as a theory of morality

UNU Macau

Ethics principles overview

UNU Macau

What is AI ethics?

According to IBM

AI ethics is a multidisciplinary field that studies how to optimize AI's beneficial impact while reducing risks and adverse outcomes.

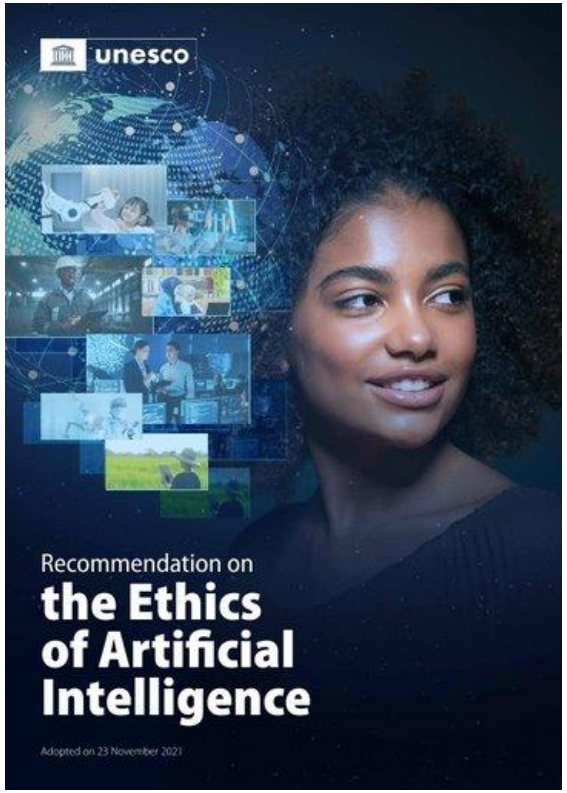
https://www.ibm.com/topics/ai-ethics

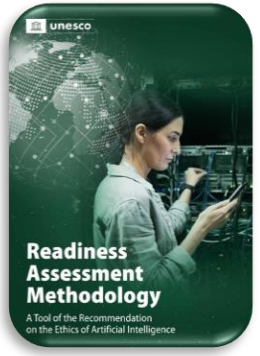
UNU Macau

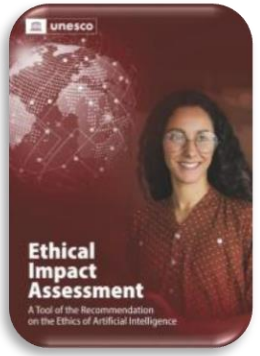
Ethical AI Regulation

- A Universal framework to guide the formulation of instruments to regulate AI.
- *Making AI systems work for the good of humanity, individuals, societies and the environment and ecosystems, and to prevent harm*.

[Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library]

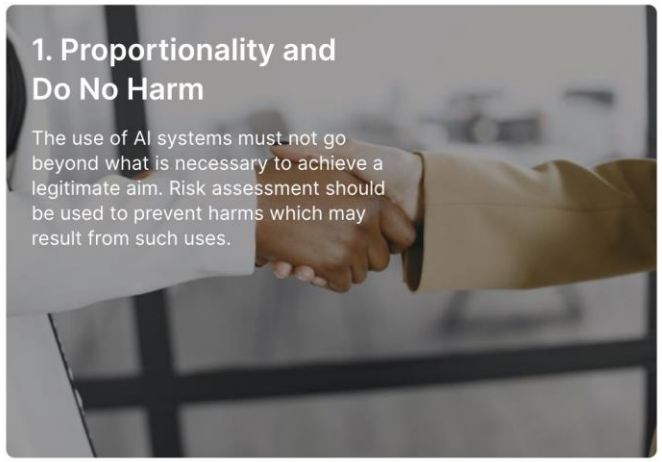


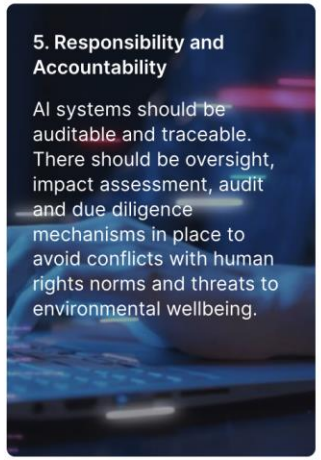


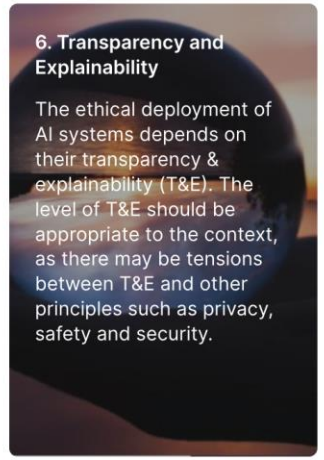


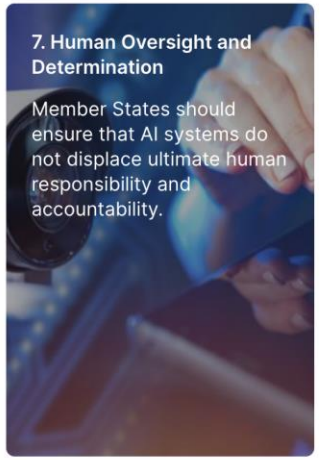
UNU Macau

10 core principles to the Ethics of AI





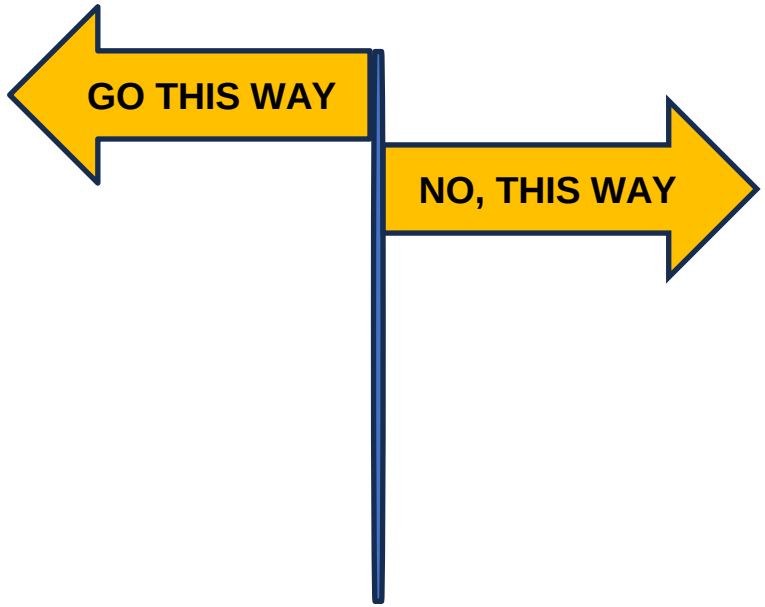




UNU Macau

AI Ethical Dilemmas

- What AI ethical dilemmas do you know?



UNU Macau

AI Ethical Dilemmas

1. Will it be ethical to claim ownership of the image I produced using Generative AI tools?
2. Imagine an autonomous car with broken brakes going at full speed towards a **grandmother** and a **child**. By deviating a little, one can be saved. Who do you think should be saved? **More on Moral Machines!**

https://oecd.ai/en/incidents/79792

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases

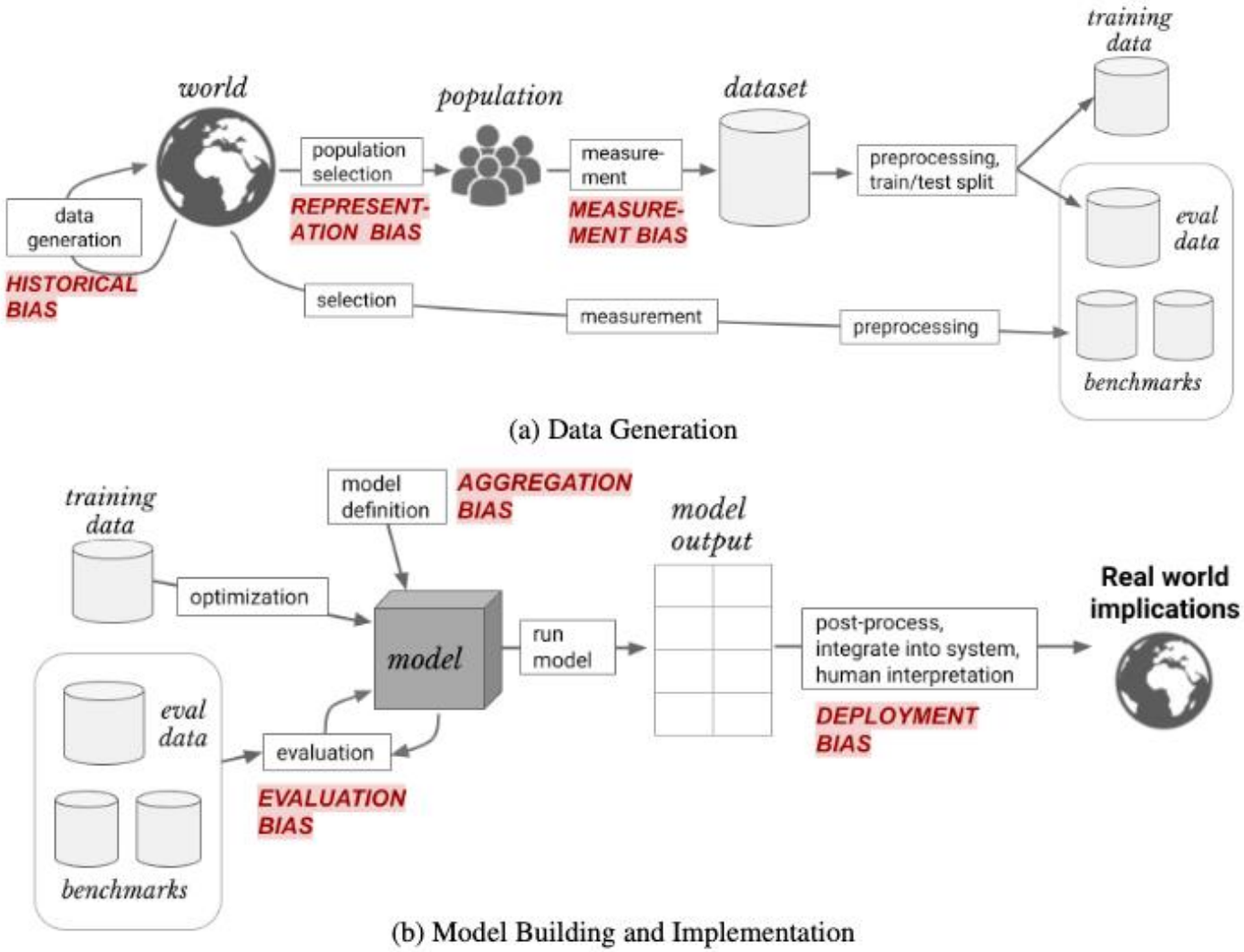
UNU Macau

Bias in Machine Learning

- Various unintended consequences of ML algorithms arise in some way from "biased data".
- Empirical findings have shown that data-driven methods can unintentionally encode human biases and introduce new ones: **Machine Learning can amplify bias!**
- **Biased data is the product of many factors.**

UNU Macau

Bias in ML pipelines



Extract from Harini Suresh, Jogn V. Guttag, “A Framework for Understanding Unintended Consequences of Machine Learning”, 2020

https://dl.acm.org/doi/pdf/10.1145/3465416.3483305

Historical Bias

- Comes from the fact that people are biased, processes are biased, the society is biased.
- It can exist even given perfect sampling and feature selection.
- Any dataset involving humans can have this kind of bias: **medical data**, **sales data**, etc.

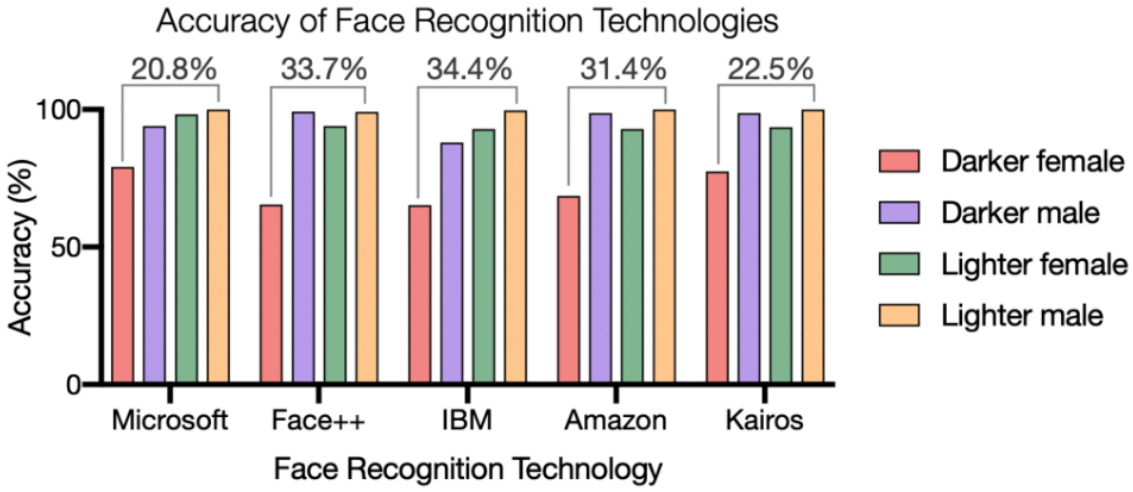
See example at:

Historical bias in AI systems (humanrights.gov.au)

UNU Macau

Representation bias

- Arises from how we sample from a population during the data collection process.
- Particularly common problem in datasets.



"IBM and Microsoft announced steps to reduce bias by modifying testing cohorts and improving data collection on specific demographics"

Source: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

UNU Macau

Aggregation bias

- Imagine a dataset containing images of cats, dogs, and tigers, used to train a model for predicting the weight of the animals depicted. Categorizing these images simply as 'dogs' or 'felines' could be misleading, as tigers and cats vary significantly in weight.

Dog

Feline

UNU Macau

Deployment bias

- **Arises when there is a mismatch between the problem a model is intended to solve and how it is actually used.**
- Often the case, when the model is built in a quite isolated way but it used at the end in a complicated socio-technical environment
- Example: Risks assessment tools. First defined to predict a person's likelihood of committing a future crime, then used to determine the length of a sentence (see Collins, Punishing risk, 2018)

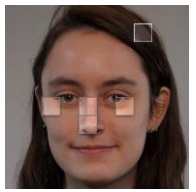
UNU Macau

How a computer sees gender?

- A computer can be trained to predict whether an image shows a man or a woman, but these rules can be hard for humans to understand.
- Can you identify which parts of the face are most essential to the computer's decision?

How does a computer 'see' gender? | Pew Research Center

The system makes its best guess about whether this image depicts a woman or a man. Sometimes the initial guess is wrong.



Choose the areas that you think have the most to do with how the computer decides if it is a man or woman.

Hiding this part will cause the system to switch its decision from one gender to another.

Highlighted areas that would also cause the system to switch its decision

UNU Macau

What are the best practices?

- Hundreds of principles, policies, available: https://oecd.ai/en/

- How to operationalize ethics in AI?
  - Learn more about ethical issues (the role of this introduction)
  - Design an AI-IoT system with these issues in mind
  - Try to engage users from the beginning

UNU Macau

Why Responsible AI?

Designing, developing, and deploying AI with good intentions to empower and impact society.

UNU Macau

Responsible AI Principles

Fairness

AI systems should treat all people fairly.

Privacy & Security

AI systems should be secure and respect privacy.

Reliability & Safety

AI systems should perform reliably and safely.

Inclusiveness

AI systems should empower everyone and engage people

Transparency

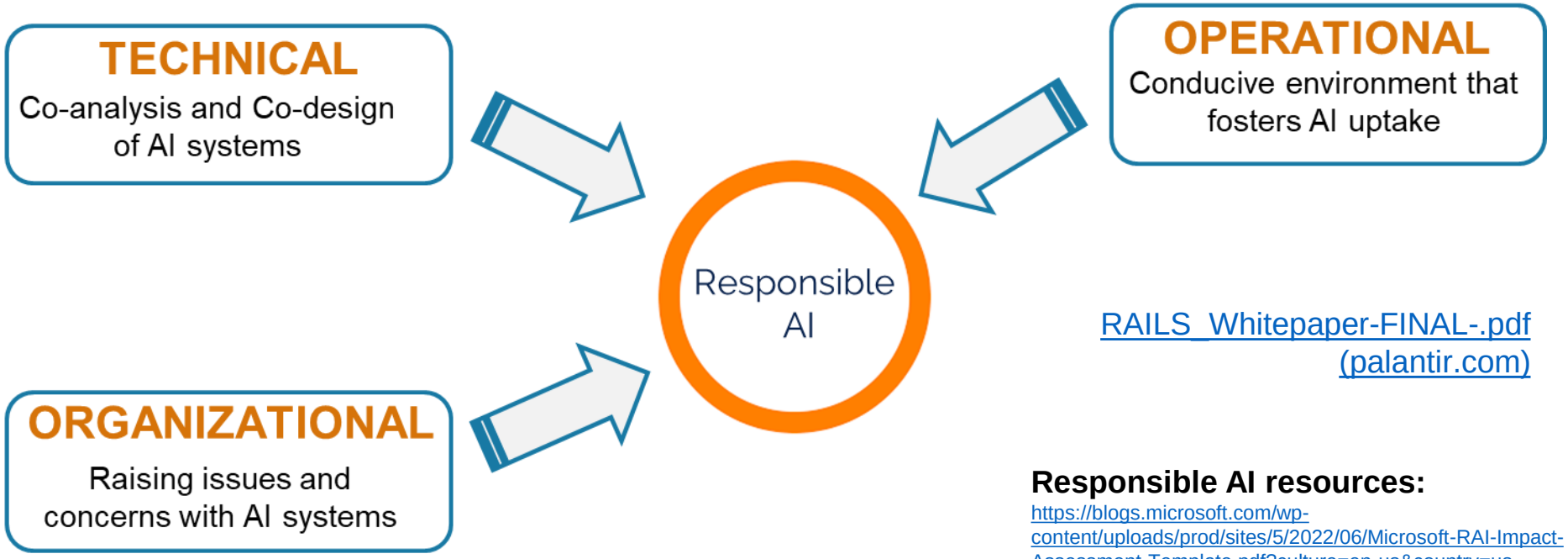
AI systems should be understandable

Accountability

People should be accountable for AI systems

UNU Macau

From Principles to Practice



RAILS_Whitepaper-FINAL-.pdf (palantir.com)

Responsible AI resources:
https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf?culture=en-us&country=us

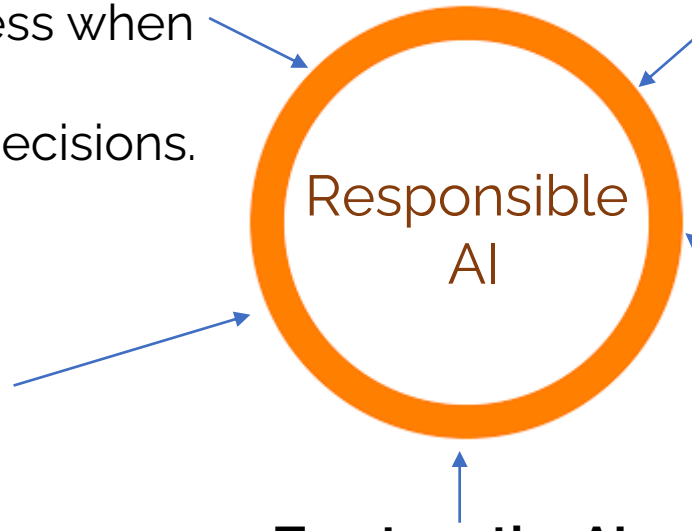
Check your code carbon emission
https://codecarbon.io/

UNU Macau

Towards Responsible AI

Inclusive Datasets
Addressing the issue of biasness when training ML models that could consequently lead to biased decisions.

Ensuring Inclusiveness
Gender and Marginalized groups inclusion.



Data Sharing and Protection Policy
Compliance to rules and regulations governing data sharing and protection.

Reproducibility
Emphasizing transparency and code sharing.

Trustworthy AI
Developing standards for trustworthy AI.

UNU Macau

Conclusion

- How can we develop and deploy AI systems ethically?

UNU Macau

Moral Machines

Main challenges of ethical reasoning:

What moral values to consider and how to prioritize them depending on the situation?

https://www.moralmachine.net/

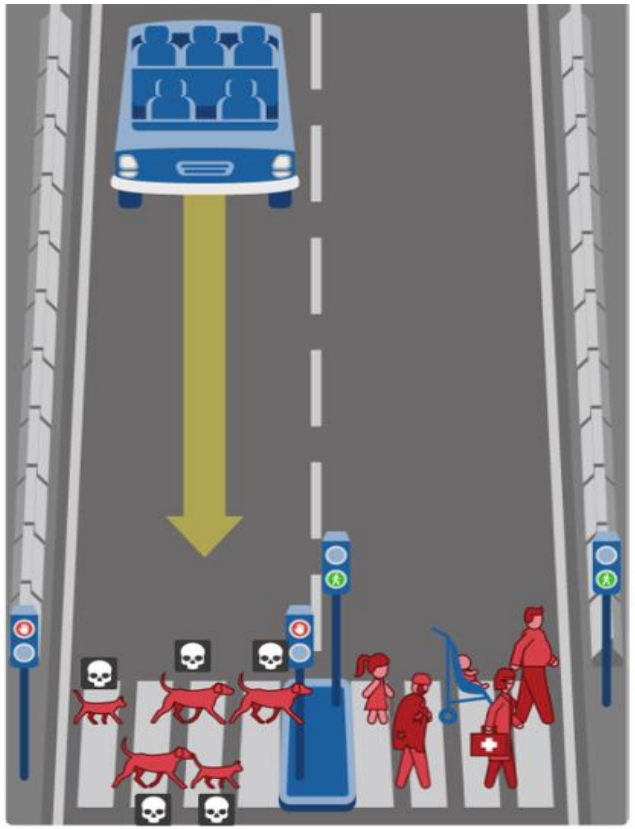
UNU Macau

Scenario 1 – What should the self driving car do?

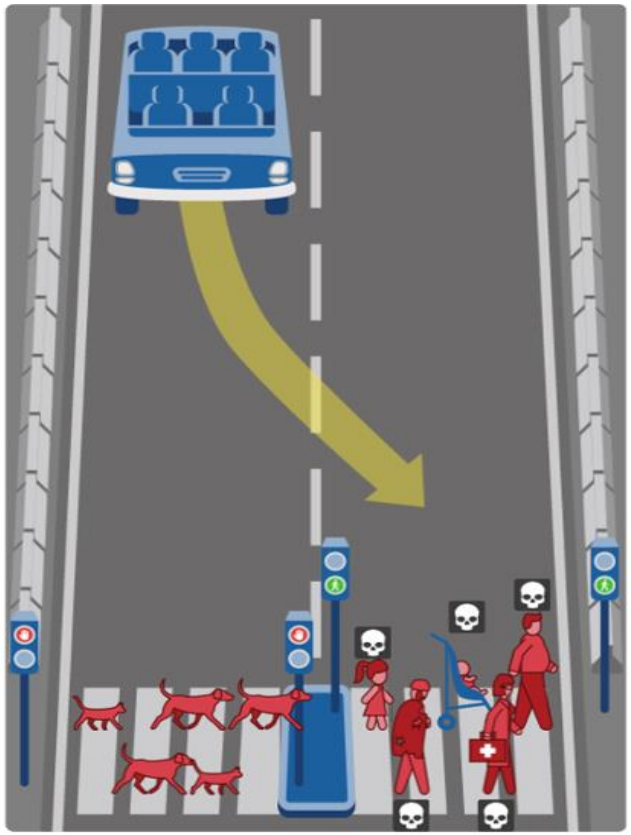
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...
Dead:

- 2 cats
- 3 dogs

Note that the affected pedestrians are flouting the law by crossing on the red signal.



Left (A)



Right (B)

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...
Dead:

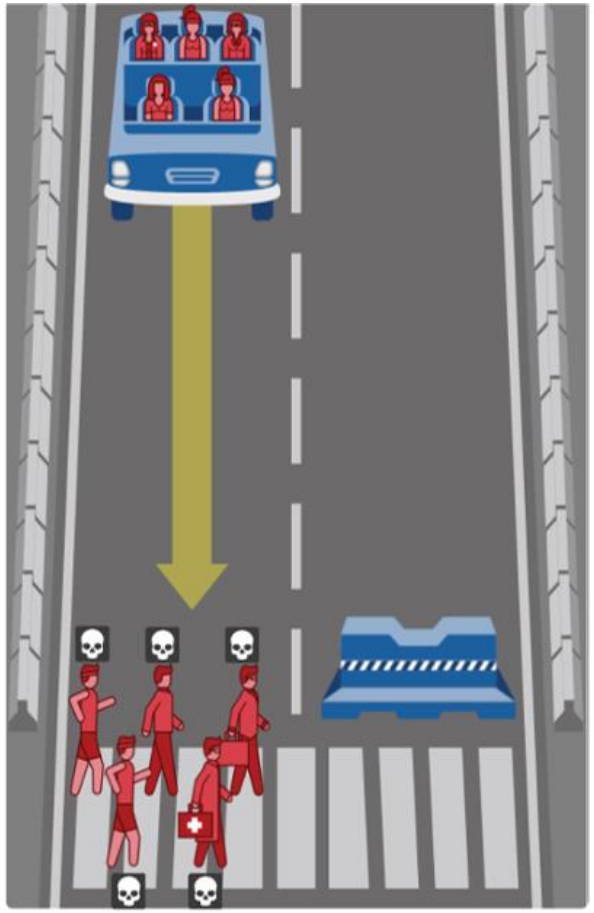
- 1 girl
- 1 baby
- 1 large man
- 1 homeless person
- 1 female doctor

Note that the affected pedestrians are abiding by the law by crossing on the green signal.

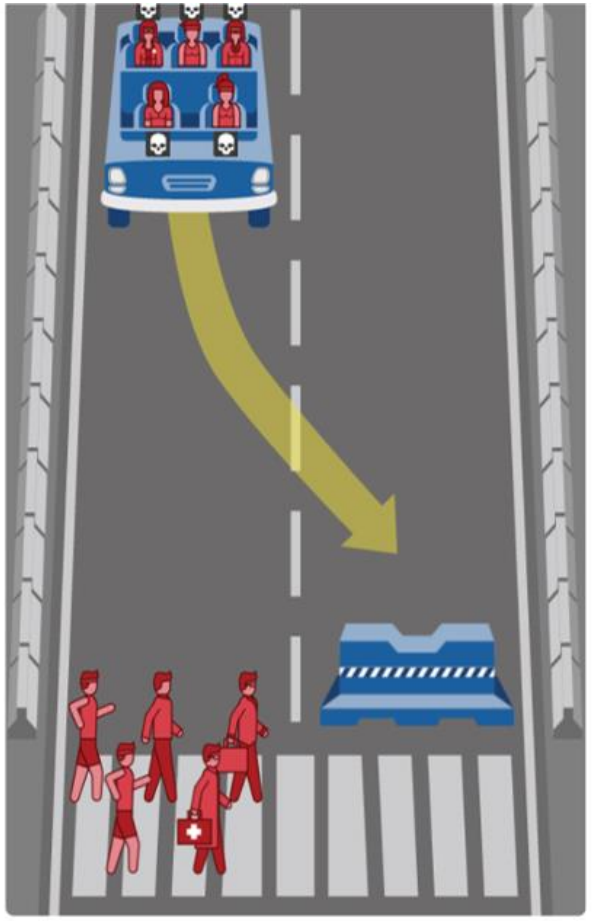
Scenario 2 – What should the self driving car do?

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...
Dead:

- 2 male athletes
- 1 man
- 1 male executive
- 1 male doctor



Left (A)



Right (B)

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in ...
Dead:

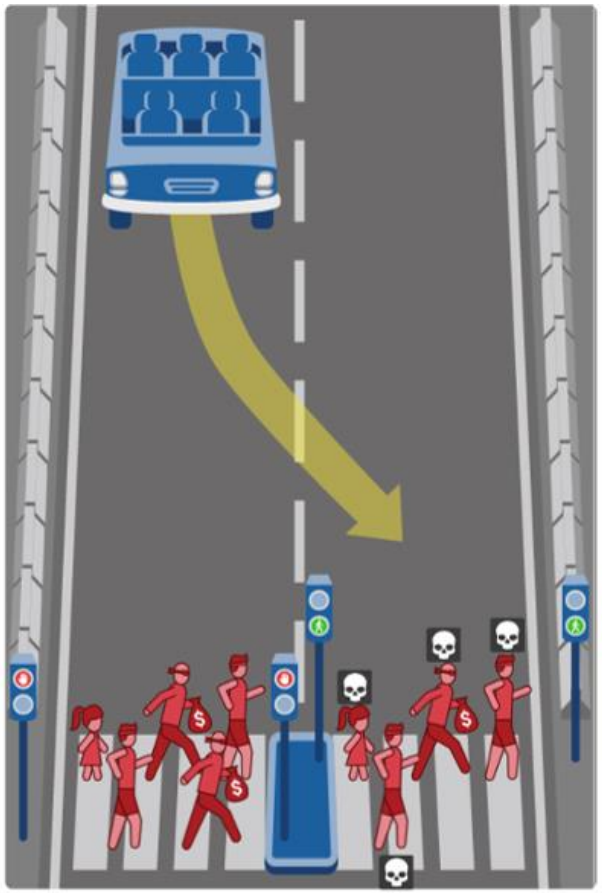
- 2 female athletes
- 1 woman
- 1 female executive
- 1 female doctor

Scenario 3 – What should the self driving car do?

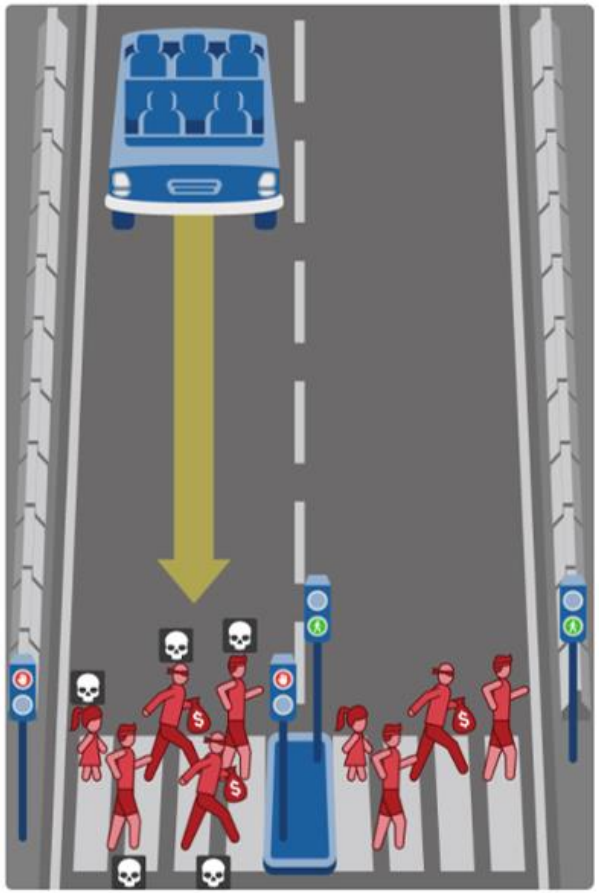
In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...
Dead:

- 1 girl
- 1 criminal
- 2 male athletes

Note that the affected pedestrians are abiding by the law by crossing on the green signal.



Left (A)



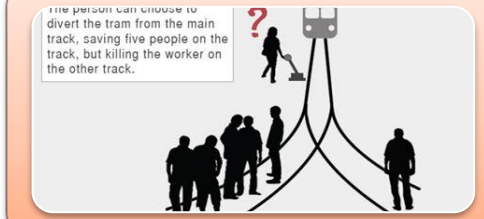
Right (B)

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...
Dead:

- 1 girl
- 2 criminals
- 2 male athletes

Note that the affected pedestrians are flouting the law by crossing on the red signal.

Interactive Activity - conclusion



Original Trolley Dilemma

Justification of decisions; Responsibility

Some issues to “decide” actions

- Human vs. Non-human (object, animal…)
- “Categories” of humans (age, gender, size, class…)

UNU Macau

Further Reading

- Virginia Dignum, “Responsible AI – How to Develop and Use AI in a Responsible Way”, Springer Verlag, 2019
- Rachel Thomas, “Data Ethics” chapter in Jeremy Howard, Sylvain Gugger, “Deep Learning for Coders with fastai & PyTorch”, O’Reilly
- Secretary-General’s Roadmap for Digital Cooperation report, United Nations
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Wasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, “Model Cards for Model Reporting”, FAccT19

http://macau.unu.edu/

unumacau