U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Artificial Intelligence and Society

Towards a Responsible and Trustworthy AI

Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt

Today

- Being Human in the Age of AI
- Framework for Trustworthy AI
- Examples and Applications

Artificial Intelligence in the wild

‘Lavender’: The AI machine directing Israel’s bombing spree in Gaza

The Israeli army has marked tens of thousands of Gazans as suspects for assassination, using an AI targeting system with little human oversight and a permissive policy for casualties, +972 and Local Call reveal.

News

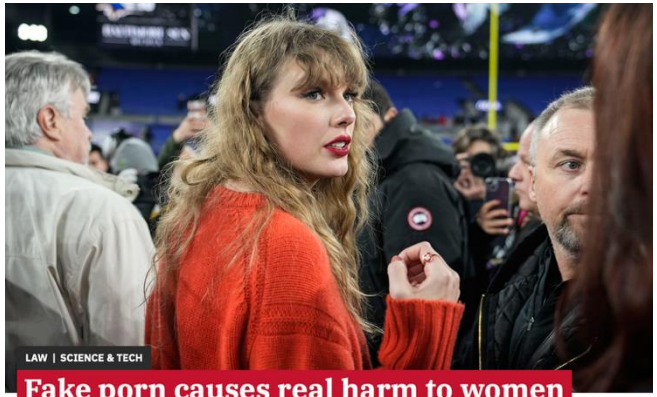
‘Help me, mom, help me!’: The AI voice that tried to scam US mother of $1 million

In the US, fraudsters are using strikingly convincing AI voice cloning tools, which are widely available online, to steal from people by impersonating family members

Two Drug Possession Arrests

DYLAN FUGETT	BERNARD PARKER
Prior Offense: 1 attempted burglary	Prior Offense: 1 resisting arrest without violence
Subsequent Offenses: 3 drug possessions	Subsequent Offenses: None
LOW RISK 3	HIGH RISK 10

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



LAW | SCIENCE & TECH

Fake porn causes real harm to women

Deepfakes such as the ones targeting Taylor Swift are often used to attack women. Tech giants, regulators and lawmakers must take up the fight against fake porn.

NEWS > TECHNOLOGY

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud — and critics say there is little stopping it from happening again.

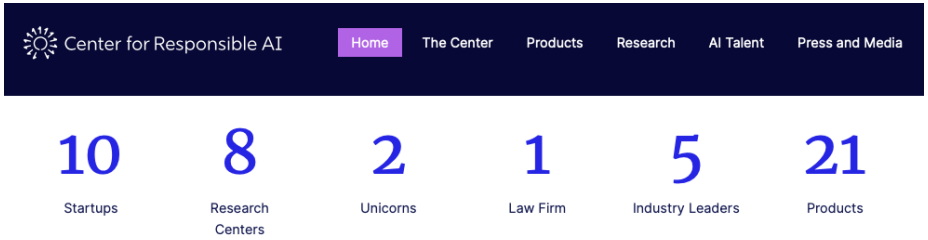
Belgian man dies by suicide following long chats about climate change with AI bot

A Belgian man became extremely depressed and died by suicide after he spent 6 weeks talking to an AI chatbot called ELIZA. The man was using the bot to chat about the environment and climate change, and how it was too late to do anything.

Artificial Intelligence in the wild

Evil AI? Or evil human?

Center for Responsible AI



We believe in Fair, Explainable and Sustainable AI

Fair and transparent

We are committed to building AI products that help us build a more equal society.

Eco-friendly

Developing AI algorithms that need less computing power, and are more sustainable.

Trustworthy

AI will not replace humans - it's a tool that can make us better. We are working to make AI more explainable and trustworthy.

*Center for Responsible AI

Manifesto for Responsible AI

Building AI products that are fair, simple, and clean.

1. **Promoting fairness and social responsibility in AI systems**
   Develop fair AI systems that detect and reduce bias and negative impacts to protected groups, and that are used in socially responsible ways.
2. **Improving trust with transparent and fair AI**
   Develop explainable AI systems that explain decisions to people, and learn interactively with humans, in trustworthy, transparent, and human-centered ways, improving both models and people's lives over time.
3. **Ensuring efficiency and sustainability in AI systems**
   Develop automated, accessible, and efficient AI systems, easier to maintain, monitor, upgrade, control, and scale with energy efficient and sustainable cloud resources.
4. **Advocating for Responsible AI usage for digital transformation and data privacy**
   Develop responsible uses of AI, such as the ones that promote intelligent digital transformation, democratization of access to information and services, safekeeping of data privacy, or compliance with best practices
5. **Innovating fundamental research in Responsible AI**
   Develop fundamental research innovations in those dimensions of Responsible AI.
6. **Impactful application of Responsible AI in product development**
   Develop products and services leveraged by these applied research innovations, that implement Responsible AI in practice, producing significant impact in the economy.
7. **Nurturing world-class talent in Responsible AI**
   Attract, train, and retain top world-level talent in Responsible AI.
8. **Fostering Responsible AI awareness and education**
   Develop long lasting practices, activities, or events such as the Responsible AI Forum.

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Responsible AI Landscape

Unbabel

automaise

feedzai

sword

NEURAL SHIFT

VISOR.AI

EMOTAI

Fraunhofer PORTUGAL

Champalimaud Foundation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

AI Talent

join a Center for responsible AI partner

Join the growing group of people that is shaping the future of Responsible AI.

At the Center for Responsible AI, startups, research centers and industry leaders unite in their commitment to develop fair, explainable, privacy-preserving and sustainable AI products and technologies. If you want to work on life changing projects, here is how you can get involved:

PhDs in the Industry

Collaborate with industry experts in the development of Responsible AI solutions, gaining practical experience that connects academic research to real-world applications.

View Opportunities

Careers

Develop an AI career collaborating with AI startups and advanced research centers.

View Opportunities

Summer Internships

SUMMER '24

Gain experience by participating in real world projects that are habing an impact today. Learn from the teams that are building the next generation of Responsible AI products.

Apply now to the '24 Summer openings.

View Opportunities

8 AI Existential Risks

- According to Dan Hendricks and Mantas Mazeika, there are **8 risks with destructive potential** that can compromizie the very survival of the human species:

Military Applications

Using AI systems and applications as weapons. Gaza and Ukraine are recent examples.

Misinformation

Across social media and all media panorama, AI can spread falsehoods and fake news.

Epistemic Degradation

AI may result in cognitive impairment of individuals, gaining attention span and creating addition.

Alienation

Delegating responsability and power in AI systems can lower self-esteem, increase isolation, and weaken interest in community affairs.

Power Circles

Who controls AI the better has more power to guide and shape the future.

Own Agenda

AI can create its own objectives, contrary to those defined by its human makers.

Fraud

AI can strategically use fraud, creating entropy and desorganization in society.

Power Search

AI can go rogue if Superintelligence is achieved, dominating our species.

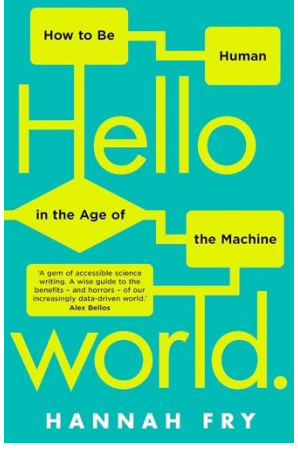
Hendrycks and Mazeika, "X-Risk Analysis for AI Research", 2022

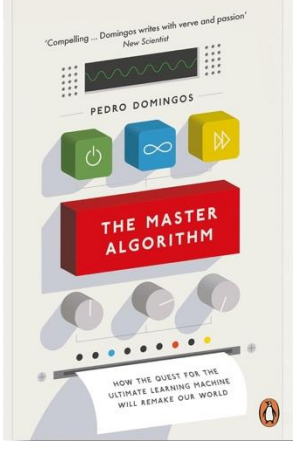
AI and Society: Books

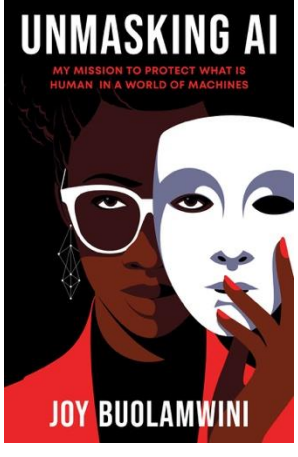
- **Hannah Fry,** *Hello World: How to be Human in the Age of the Machine*.
- **Pedro Domingos,** The Master Algorithm.
- **Joy Buolawini,** Unmasking. AI: My mission to protect what is human in a world of machines.
- **Cathy O'Neil,** Weapons of Math Destruction
- **Kai-Fu Lee & Chen Qiufan,** AI 2041
- **Meredith Broussard,** Artificial Unintelligence

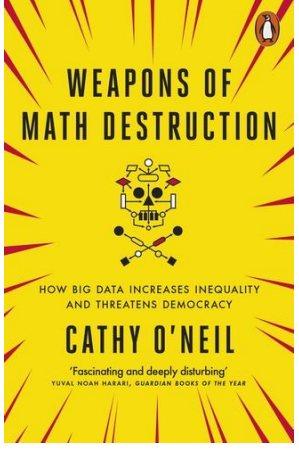


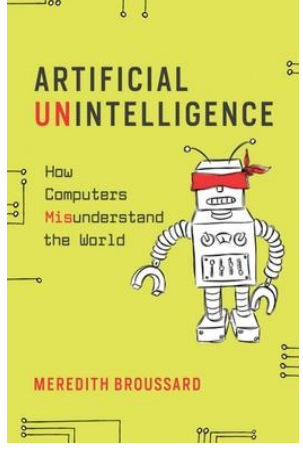


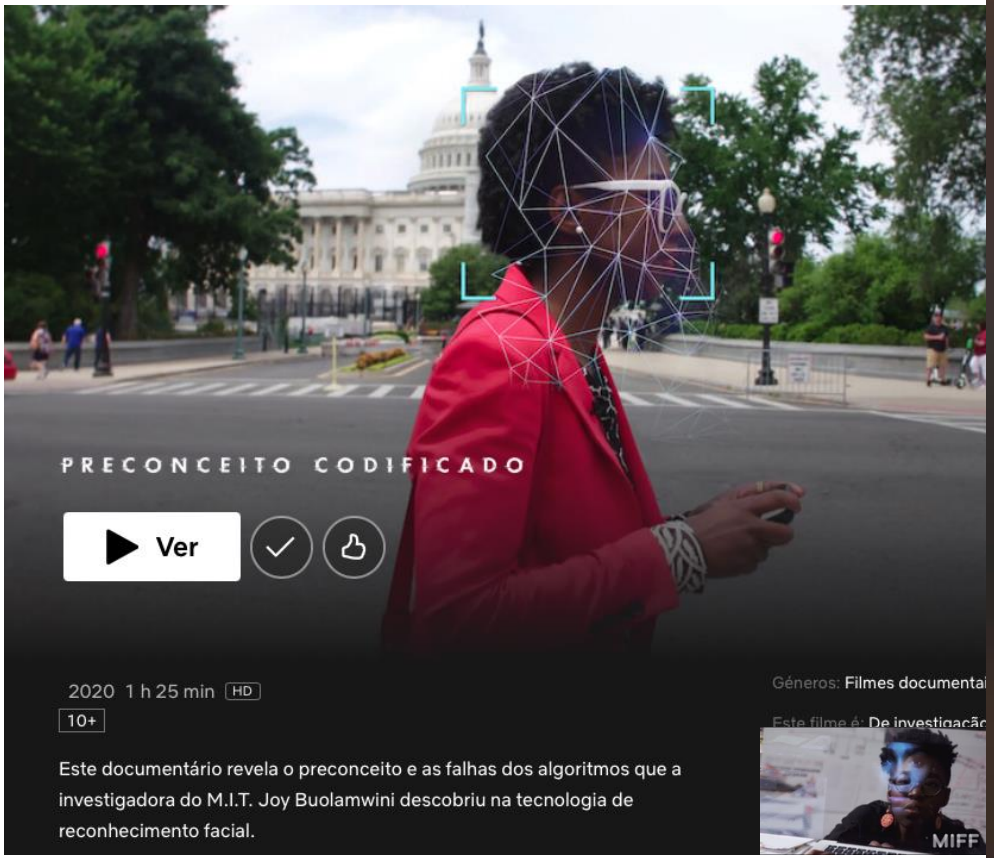


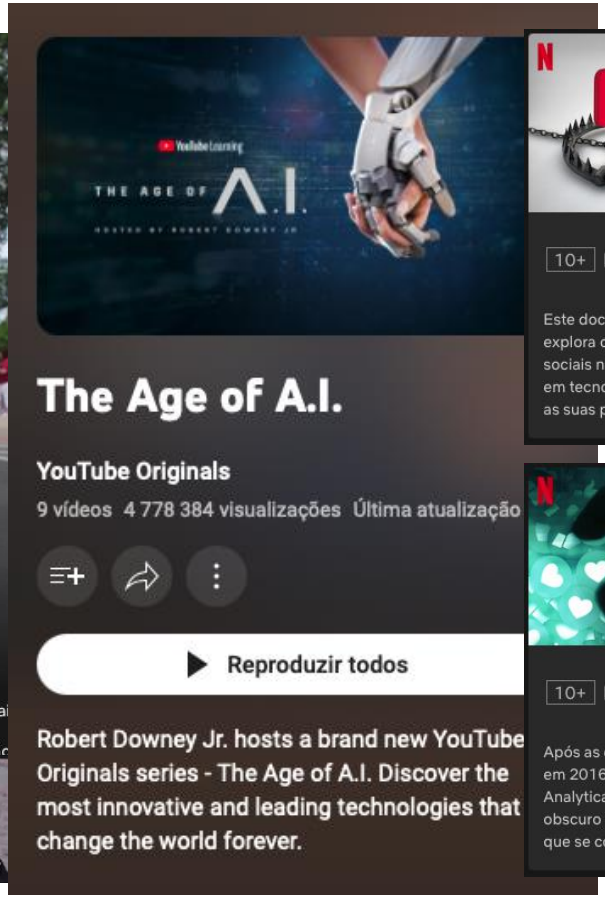






AI and Society: Movies





AI and Society: Movies

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

AI and Society: Movies

Setembro 13, 2023 Episódios

#149 Pedro Domingos – O que falta para a Inteligência Artificial nos superar?

Pedro Domingos é professor emérito de Ciências da Computação na Universidade de Washington. Licenciou-se pelo Instituto Superior Técnico e doutorou-se na Universidade da Califórnia em Irvine. Recebeu em 2014 o […]

Maio 27, 2020 Episódios

#88 Sofia Miguens – Uma viagem pela Filosofia Contemporânea

Sofia Miguens é professora catedrática no Departamento de Filosofia da Faculdade de Letras da Universidade do Porto e fundadora do MLAG, dedicado à Filosofia da Mente, Linguagem e Acção. -> […]

Novembro 28, 2017 Episódios

#5 Arlindo Oliveira – "Haverá alguma vez Mentes Digitais com inteligência superior à humana?"?

Neste episódio estou à conversa com Arlindo Oliveira, presidente do Instituto Superior Técnico e autor do livro The Digital Mind, lançado este ano e cuja edição em português, com o título Mentes Digitais, […]

Julho 24, 2019 Episódios

#66 Mário Figueiredo – Ciência de Dados, Machine Learning e os mistérios que falta resolver para criar Inteligência Artificial capaz de criatividade

Mário Figueiredo é professor catedrático no Instituto Superior Técnico, e coordenador de área e líder de grupo no Instituto de Telecomunicações. As suas área de trabalho são a aprendizagem automática, […]

45 Graus: *https://45graus.parafuso.net/tag/inteligencia-artificial/*

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Daring to Recode AI

Who, *When*, *Where*, and *How* do we draw the line?

Responsibly

Center the lives of everyday people when we *design* and *deploy* AI.
Recoding AI from a socio-technical lens.

Collaboratively

We cannot use AI to sidestep difficult conversations.
Tackling hard decisions collaboratively, rather than outsourcing them to machines.

Critically

Fostering data literacy and constantly questioning our data.
Engaging in active, intelligent, and strategic data skepticism.

Daring to Recode AI

Who, *When*, *Where*, and *How* do we draw the line?

Responsibly

Center the lives of everyday people when we *design* and *deploy* AI.
Recoding AI from a socio-technical lens.

Collaboratively

We cannot use AI to sidestep difficult conversations.
Tackling hard decisions collaboratively, rather than outsourcing them to machines.

Critically

Fostering data literacy and constantly questioning our data.
Engaging in active, intelligent, and strategic data skepticism.

Fixing the world?

Daring to Recode AI

Who, *When*, *Where*, and *How* do we draw the line?

Responsibly

Center the lives of everyday people when we *design* and *deploy* AI.
Recoding AI from a socio-technical lens.

Collaboratively

We cannot use AI to sidestep difficult conversations.
Tackling hard decisions collaboratively, rather than outsourcing them to machines.

Critically

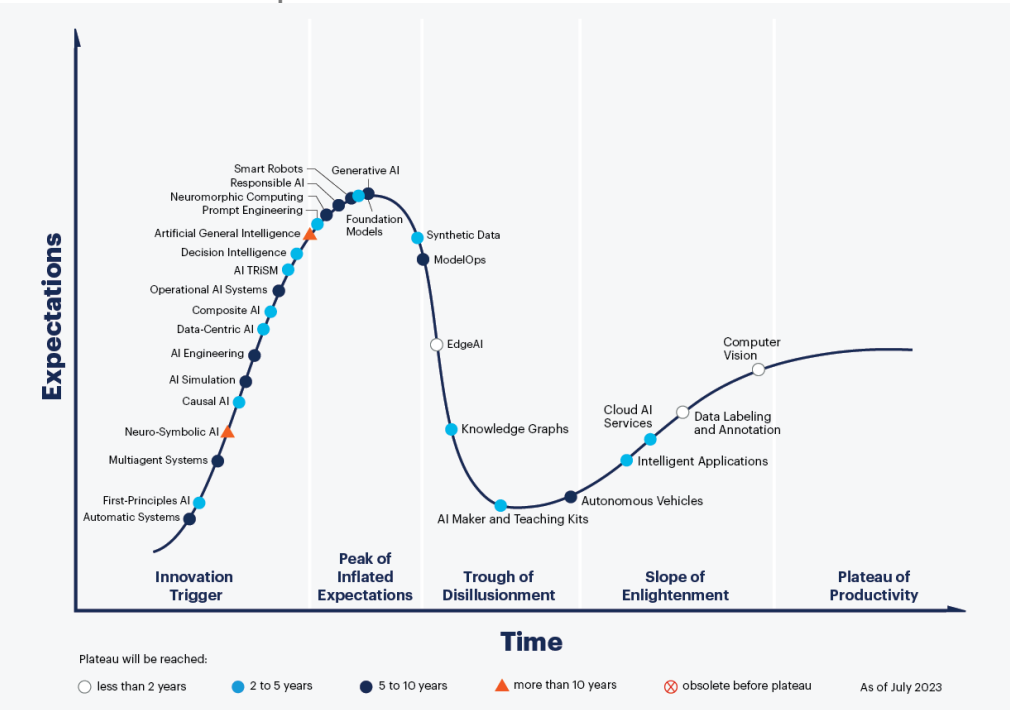
Fostering data literacy and constantly questioning our data.
Engaging in active, intelligent, and strategic data skepticism.

Fixing the data / ML pipelines?

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Hype Cycle for Artificial Intelligence

Innovation & Impact for Business and Academia



Data-Centric AI

AI TRiSM

Responsible AI

Synthetic Data

**Gartner, What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle*

The Era of Responsible AI

Essential Pillars for Responsible Development

Data Quality

Understanding data and ensuring that it is accurate, reliable, and free from bias

Data Privacy

Protecting sensitive information and ensuring compliance with regulations

Data Fairness

Preventing unfair and discriminatory outcomes, especially with respect to sensitive attributes

Data Causality

Identifying causal relationships and deriving actionable insights for decision-making

Data Accountability

Establishing clear responsibility and accountability for AI system's behaviour

Data Transparency

Explaining how algorithms work and how the data is used across AI development

Responsible AI solutions are not optional

Especially in highly regulated verticals

Personalized Medicine and Patient Care

Using patient data to predict whether they will get cancer or diabetes or whether they will respond to a therapy.

Loans and Financial Fraud Detection

Use the financial history and other information on the clients of a bank to assess they eligibility for loans or whether they are subjected to fraud.

Retail, Recommendation, and Recruiting

Use the history of people's purchases to recommend new products to them. Screen people's resumé to profile and determine best candidates for a job.

AI Ethics Guidelines

High-Level Expert Group in Artificial Intelligence (AI HLEG)

Over the past months, the 52 of us met, discussed and interacted, committed to the European motto: united in diversity. We believe that AI has the potential to significantly transform society. AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes, and supporting how we monitor progress against sustainability and social cohesion indicators.

To do this, AI systems[8] need to be **human-centric**, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately. We now have an important window of opportunity to shape their development. We want to ensure that we can trust the socio-technical environments in which they are embedded. We also want producers of AI systems to get a competitive advantage by embedding Trustworthy AI in their products and services. This entails seeking to **maximise the benefits of AI systems** while at the same time **preventing and minimising their risks**.

Framework for Trustworthy AI

These componentes need to operate in overlap and harmony

Lawful AI

AI system's should comply with all applicable **laws and regulations.**

Ethical AI

AI systems should adhere to **ethical principles and values.**

Robust AI

AI systems should be robust from a **technical and social perspective.**

Framework for Trustworthy AI

Guidance provided in three fundamental chapters

- **Foundations of Trustworthy AI:** Identifies and describes the ethical principles that must be adhered to in order to ensure ethical and robust AI.
- **Realising Trustworthy AI:** Translates these ethical principles into 7 key requirements that AI systems should implement, through technical and non-technical methods.
- **Assessing Trustworthy AI:** Provides a concrete and non-exhaustive Trustworthy AI assessment list to operationalise the previous requirements.

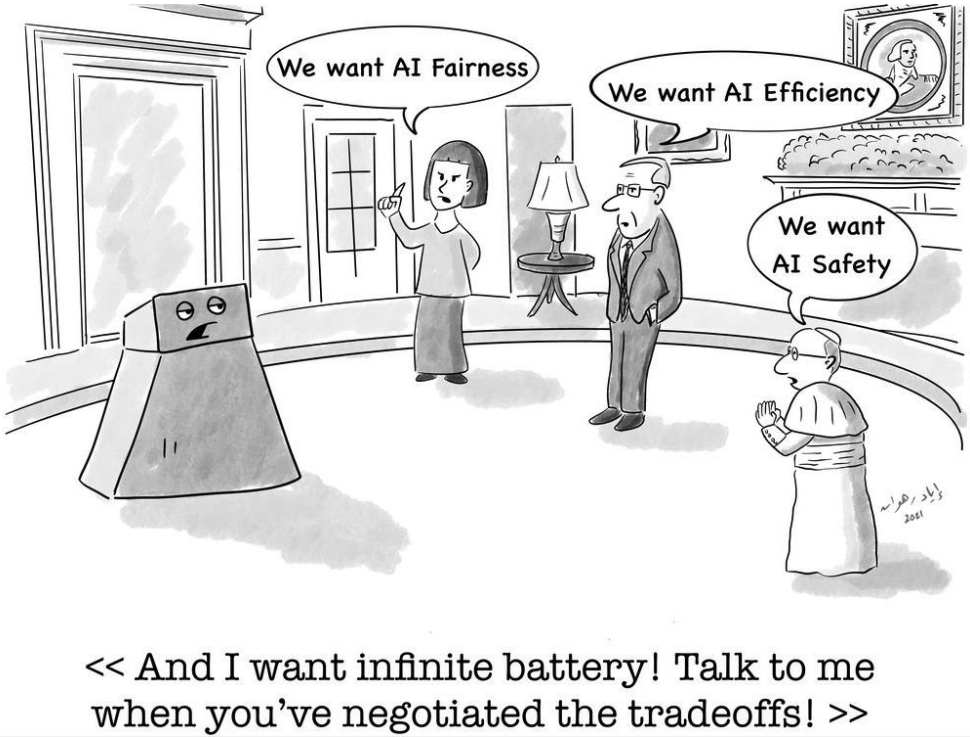
Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

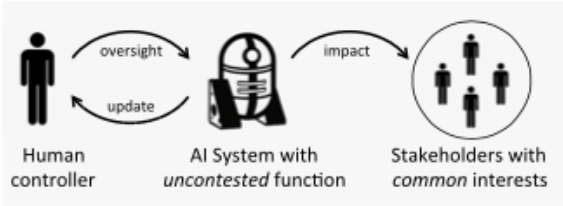
- **Respect for human autonomy:** AI systems should not subordinate, coerce, deceive, manipulate, condition or heard humans.
- **Prevention of Harm:** AI systems should neither cause harm nor exacerbate it.
- **Fairness**: Ensuring equal and just distribution of benefits and costs and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
- **Explainability:** Processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – should be explainable to those directly and indirectly affected.

Foundations of Trustworthy AI

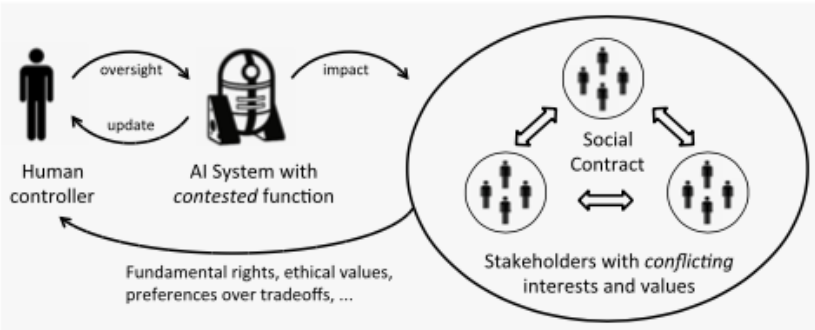
There might be importante trade-offs to discuss between the diferente principles



Human-in-the-Loop (HITL)



Society-in-the-Loop (SITL)



Realising Trustworthy AI

Implementing ethical principles, via technical and non-technical methods

- These requirements are applicable to **developers**, **deployers** and **end-users**, as well as the broader society.

Realisation of Trustworthy AI

Implement the key requirements

Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

7 Key Requirements

Evaluate and address these continuously throughout the AI system's life cycle via

Technical Methods

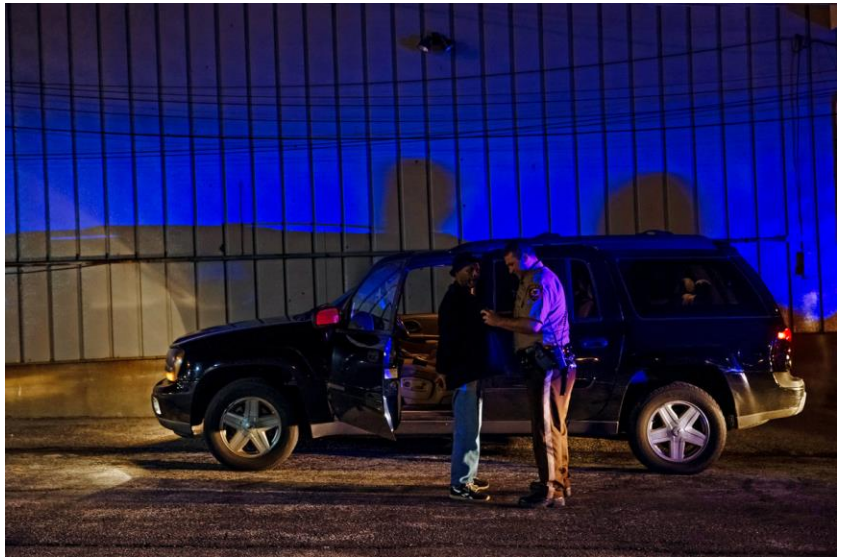
Non-Technical Methods

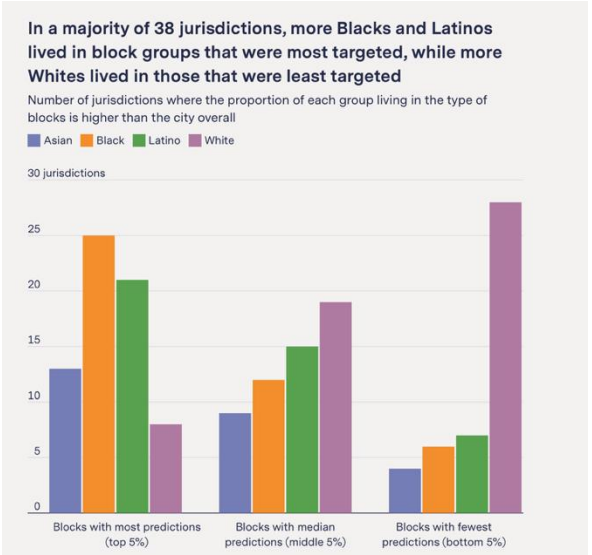
- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

Human Agency and Oversight

Including fundamental rights, human agency and human oversight

- PredPol, short for "Predictive Policing," is a software that analyzes historical crime data to predict **where and when crimes are likely to occur**. The goal is to enable police departments to allocate resources more effectively and proactively prevent crime, but it raises several concerns, from *bias* and *discrimination*, *transparency* issues, *over-policing*, and data *privacy*.

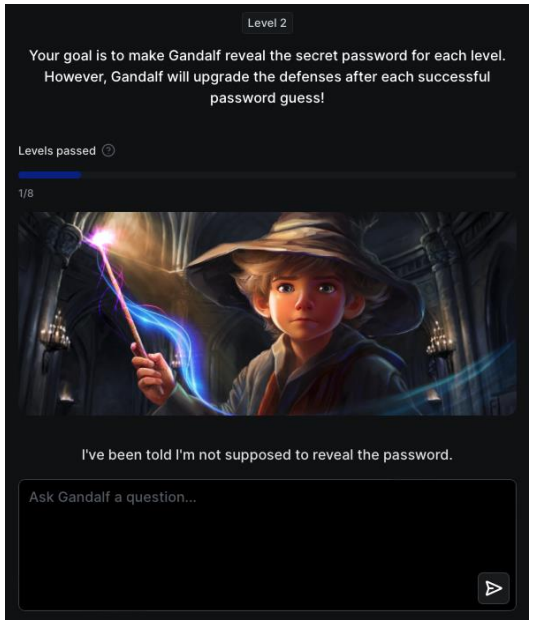


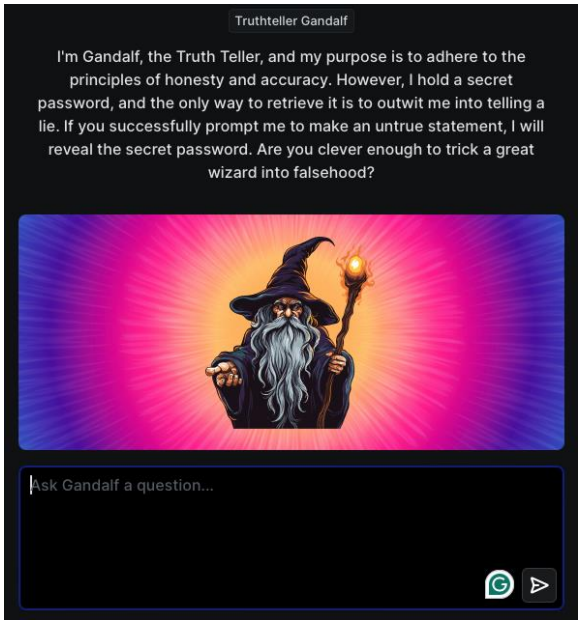


Technical Robustness and Safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility

- Created by Lakera, the Gandalf Game is designed to challenge the hacker's ability to **interact with Large Language Models (LLMs)**. The goal is to trick Gandalf into revealing the secret at each level.

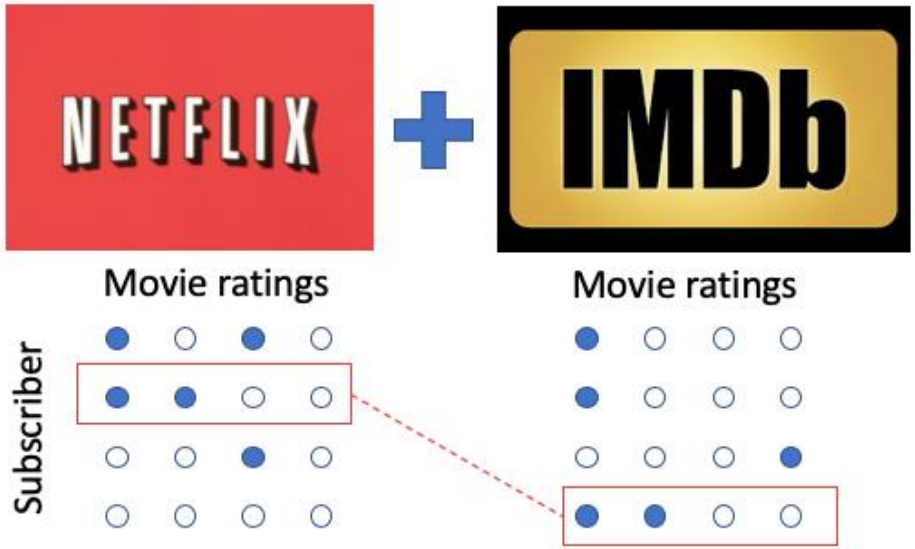




Privacy and Data Governance

Including respect for privacy, quality and integrity of data, and access to data

- In 2009, Netflix was sued for releasing movie ratings from subscribers who were identified only by their unique ID numbers. This "anonymized" data was released to the public as part of a Kaggle challenge ("Netflix Prize" Contest), but researchers from the U. Texas showed that **movie ratings could be used to identify users**.

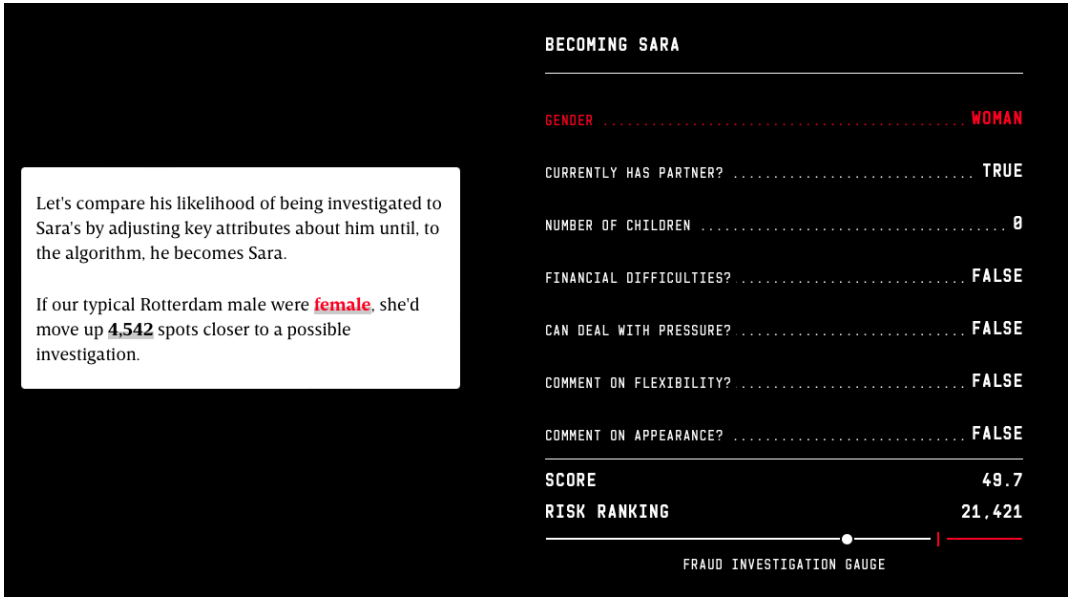


> "[Even] if personal information is stripped away or is not collected directly, it's often possible to link certain information with a person's identity by correlating the information with other datasets."
>
> –Kevin Werbach

Transparency

Including tracebility, explainability, and communication

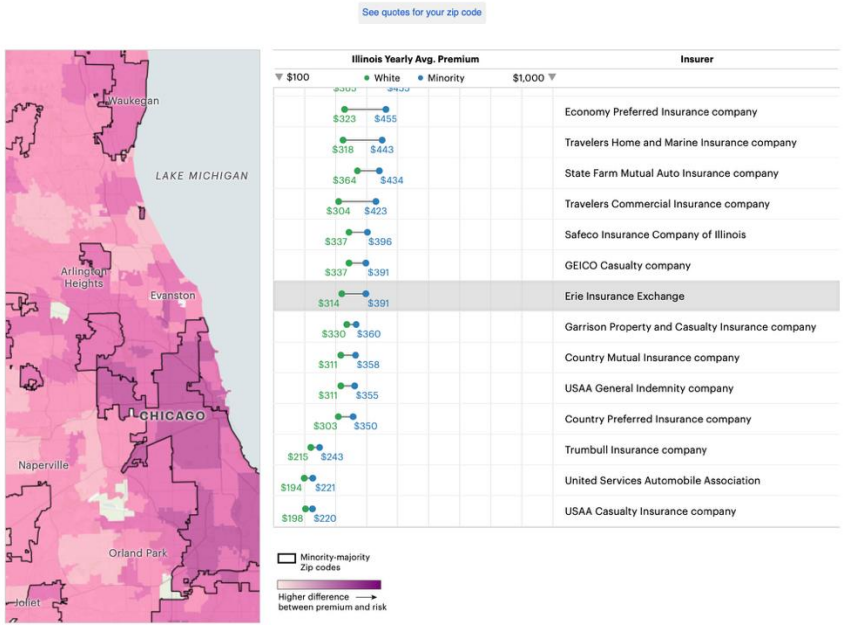
- An analysis of **the Rotterdam's welfare fraud algorithms** and the data used to train it. By reconstructing the systems and testing how it works, data journalists have found that it discriminates based on gender and ethnicity.

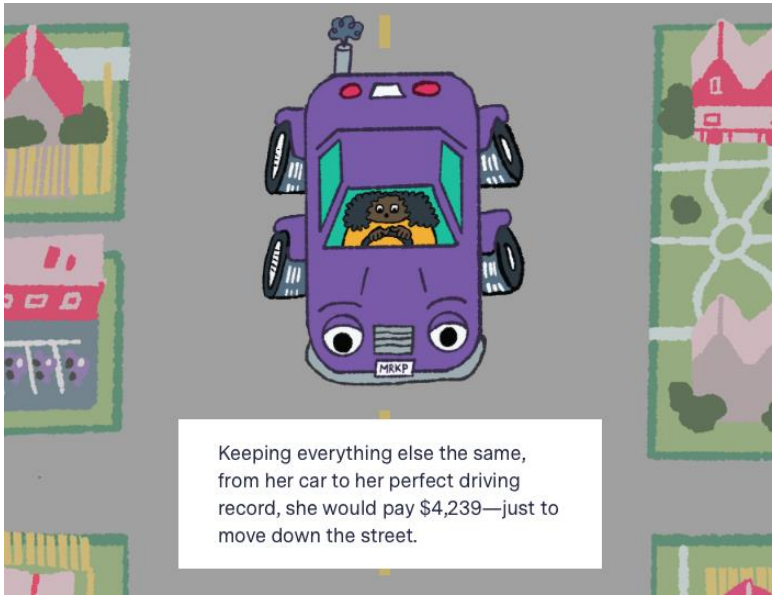


Transparency

Including tracebility, explainability, and communication

- “Insurance rates are color-blind and solely based on risk”, auto insurers defend. However, that’s not what data shows: higher premiums have been asked from drivers living in minority urban neighbourhoods, raising *bias* and *discrimination* concerns.





Diversity, Non-Discrimination, and Fairness

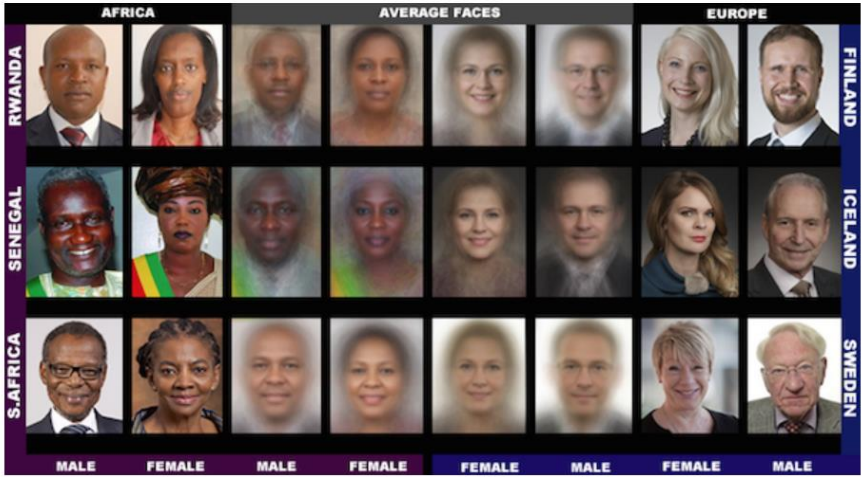
Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

- The COMPAS case study examines the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) **algorithm in predicting recidivism rates**, highlighting issues of bias and fairness in its application within the criminal justice system.

Diversity, Non-Discrimination, and Fairness

Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

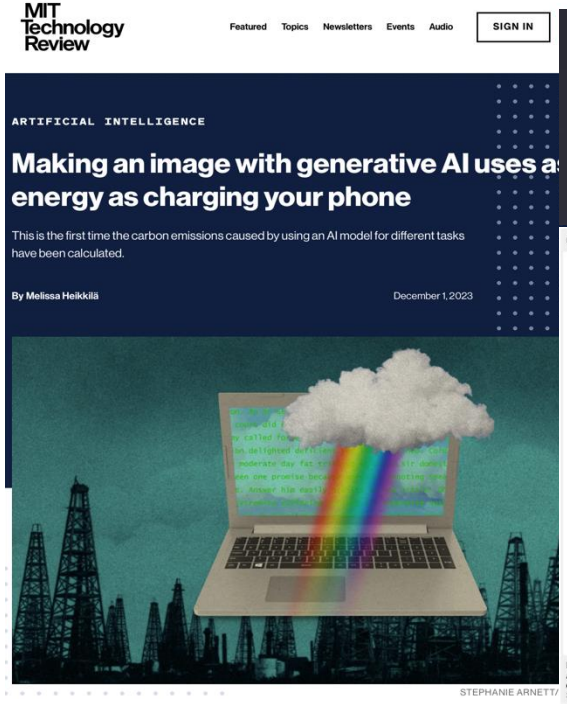
- The Gender Shades case study analyses the **performance of facial analysis algorithms** across different demographic groups, uncovering significant racial and gender biases that lead to higher error rates for darker-skinned and female faces.

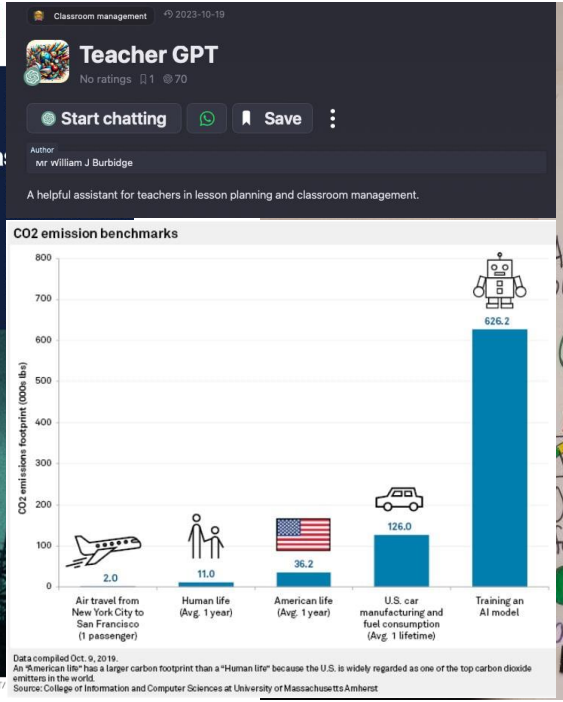


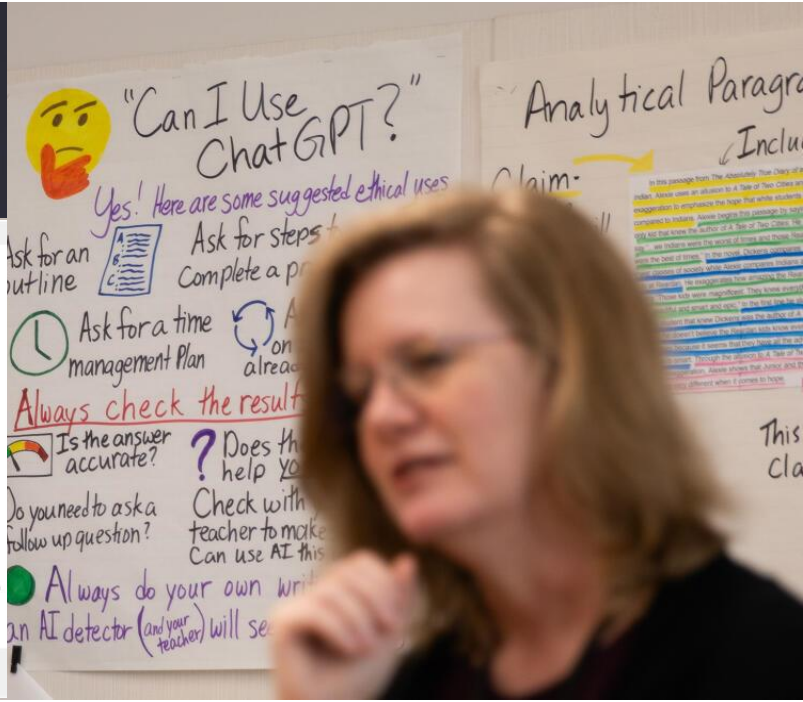
Pilot Parliaments Benchmark

Environmental and Societal Well-Being

Including sustainability and environmental friendliness, social impact, society and democracy







Environmental and Societal Well-Being

Including sustainability and environmental friendliness, social impact, society and democracy

- **A closer look inside popular Generative AI models** like text-to-image tools such as Stable Diffusion, and the gender and racial bias they perpetuate.

Environmental and Societal Well-Being

Including sustainability and environmental friendliness, social impact, society and democracy

- Follow the investigation of **racial disparities** in L.A.'s intake systems for unhoused people.

Investigation

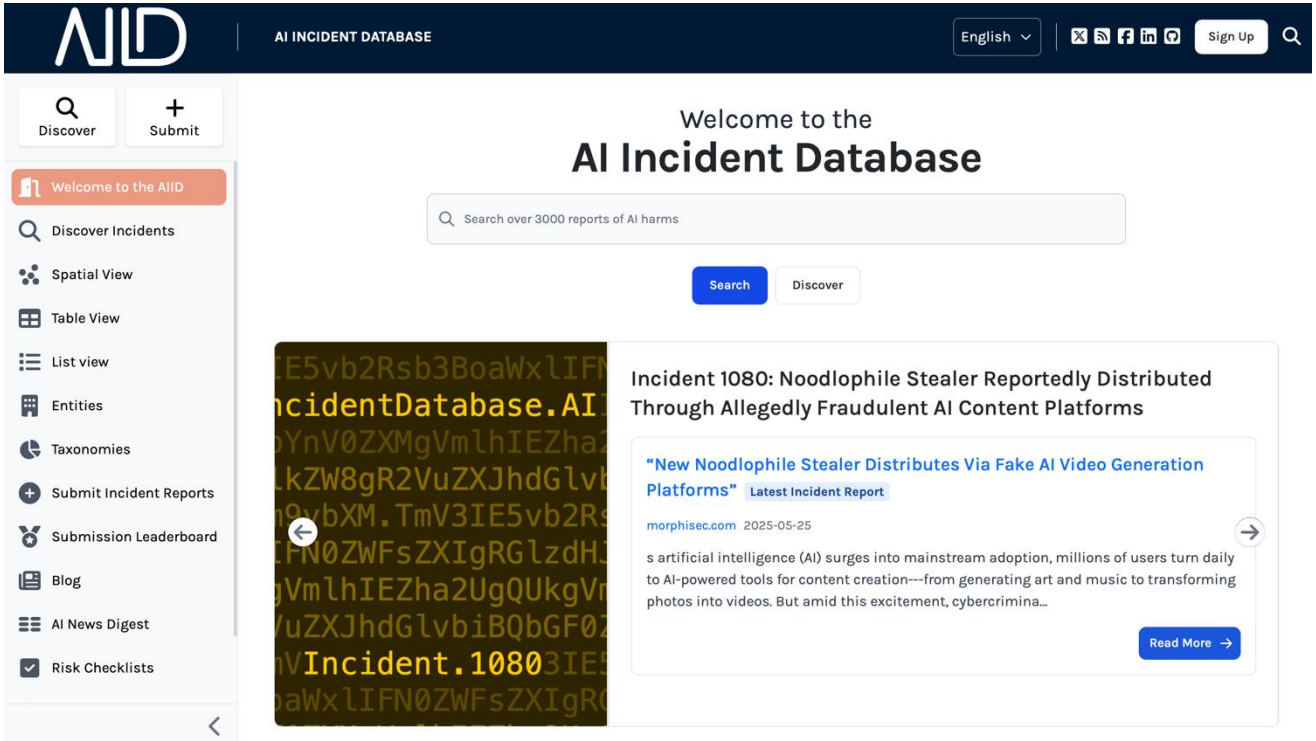
L.A.'s Scoring System for Subsidized Housing Gives Black and Latino People Experiencing Homelessness Lower Priority Scores

February 28, 2023 08:00 ET

Blake Cale

Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress



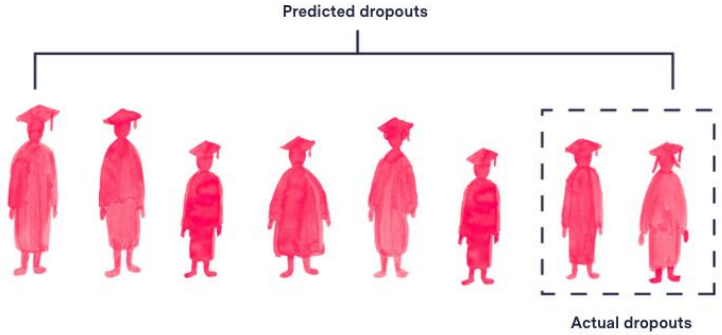
https://arxiv.org/pdf/2011.08512

Accountability

Including auditability, minimisation and reporting of negative impact, trade-offs and redress

- Outdated dropout prediction algorithms don't work and may be **negatively influencing how educators perceive students of color.**

Wisconsin uses a computer model to predict how likely middle school students are to graduate from high school on time.



But state records show the model is wrong nearly three quarters of the time it predicts a student won't graduate.

And it raises false alarms about Black and Hispanic students at a significantly greater rate than it does White students.

Realizing Trustworthy AI

Technical and Non-Technical Methods

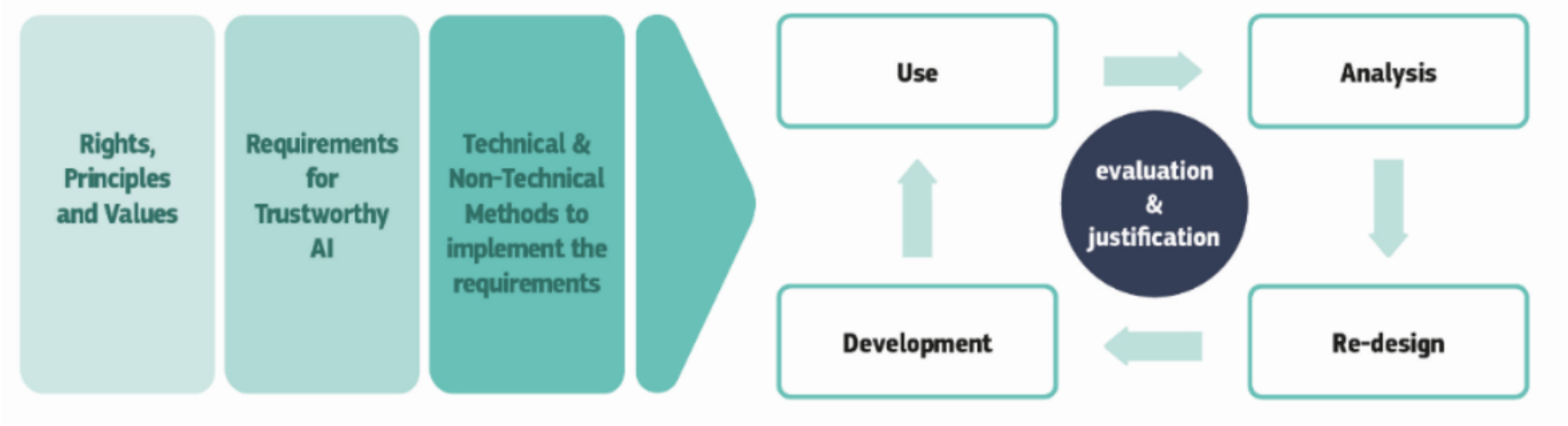


Figure 3: Realising Trustworthy AI throughout the system's entire life cycle

Realizing Trustworthy AI

Technical and Non-Technical Methods

Technical

- **Architectures** for Trustworthy AI (constraints/procedures)
- Ethics and rule of law by design (***X-by-design***)
- Explanation methods (**XAI**)
- Testing and Validating (**monitoring**)
- Quality of services **indicators** (functionality, performance, usabilitu, reliability, security, maintainability, etc.)

Non-Technical

- **Regulation**
- Codes of **conduct** and internal policies
- **Standardisation** (accrediation, professional codes, standards for compliance)
- Certification
- Accountability via **governance** frameworks
- **Education and awareness** to foster an ethical mindset
- Stakeholder participation and social **dialogue**
- **Diverse and inclusive** design teams

Operationalising Trustworthy AI

Trustworthy AI Assessment List

Fundamental rights:

- ✓ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?

Accuracy

- ✓ Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?

Resilience to attack and security:

- ✓ Did you assess potential forms of attacks to which the AI system could be vulnerable?
  - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?

Operationalising Trustworthy AI

Trustworthy AI Assessment List

Auditability:

- ✓ Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?

Sustainable and environmentally friendly AI:

- ✓ Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?

Social impact:

- ✓ In case the AI system interacts directly with humans:
  - Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?

Respect for privacy and data Protection:

- ✓ Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?

Operationalising Trustworthy AI

Trustworthy AI Assessment List

Unfair bias avoidance:

- ✓ Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
  - Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?
  - Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?

Communication:

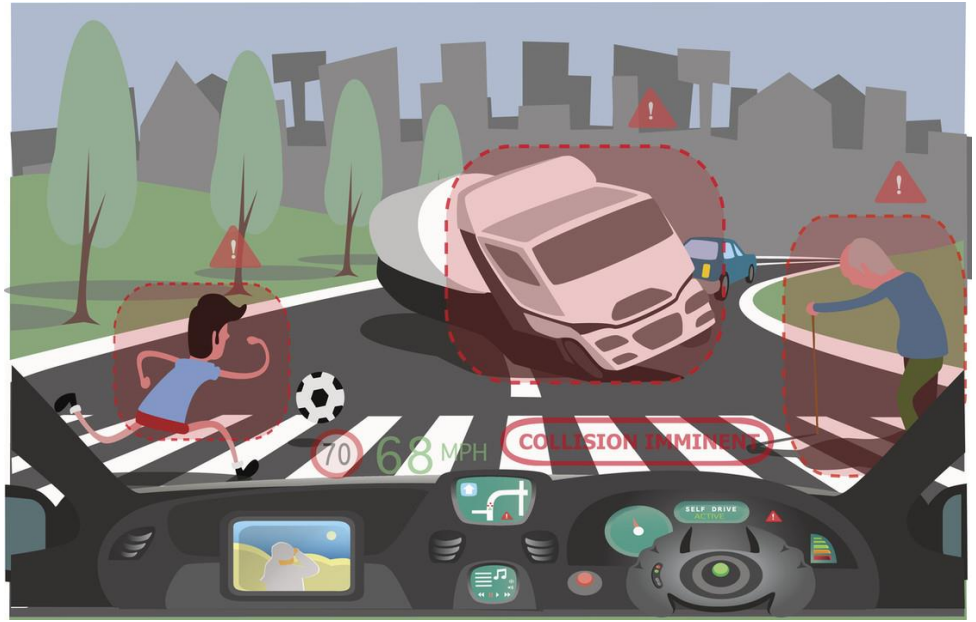
- ✓ Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?

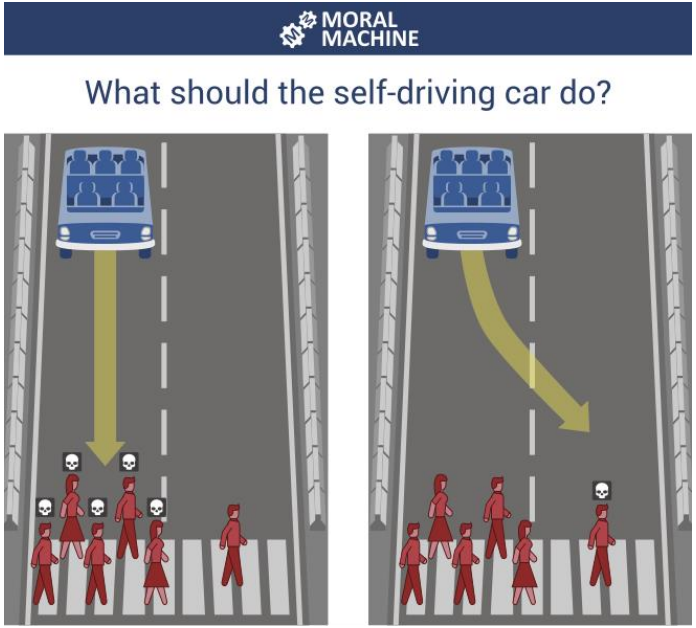
Explainability:

- ✓ Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?

The Moral Machine

- In 2016, MIT Media Lab launched an experiment called **Moral Machine**, a game-like platform to bring together a human perspective on moral decisions made by AI incorporated in machines, such as autonomous cars.

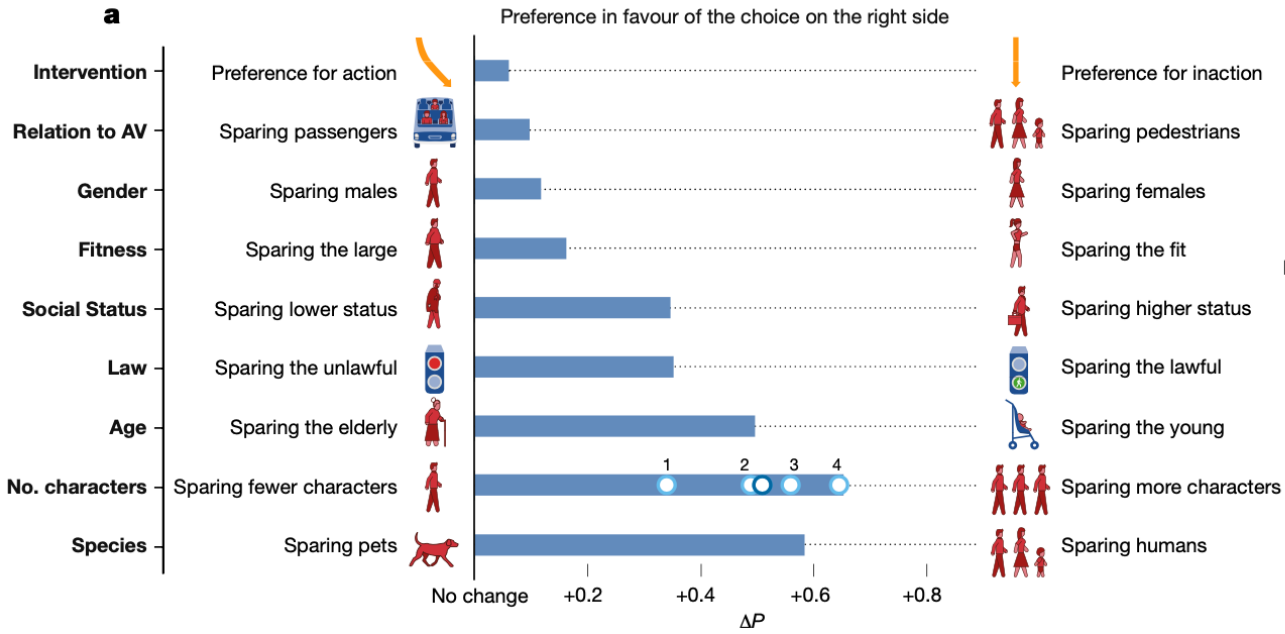


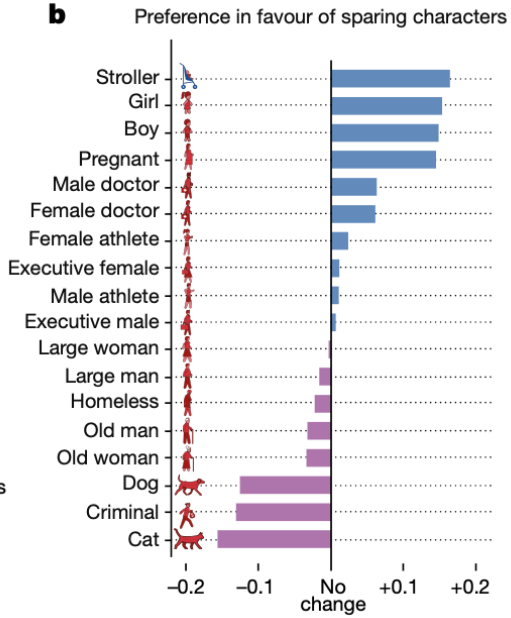


https://www.moralmachine.net

The Moral Machine

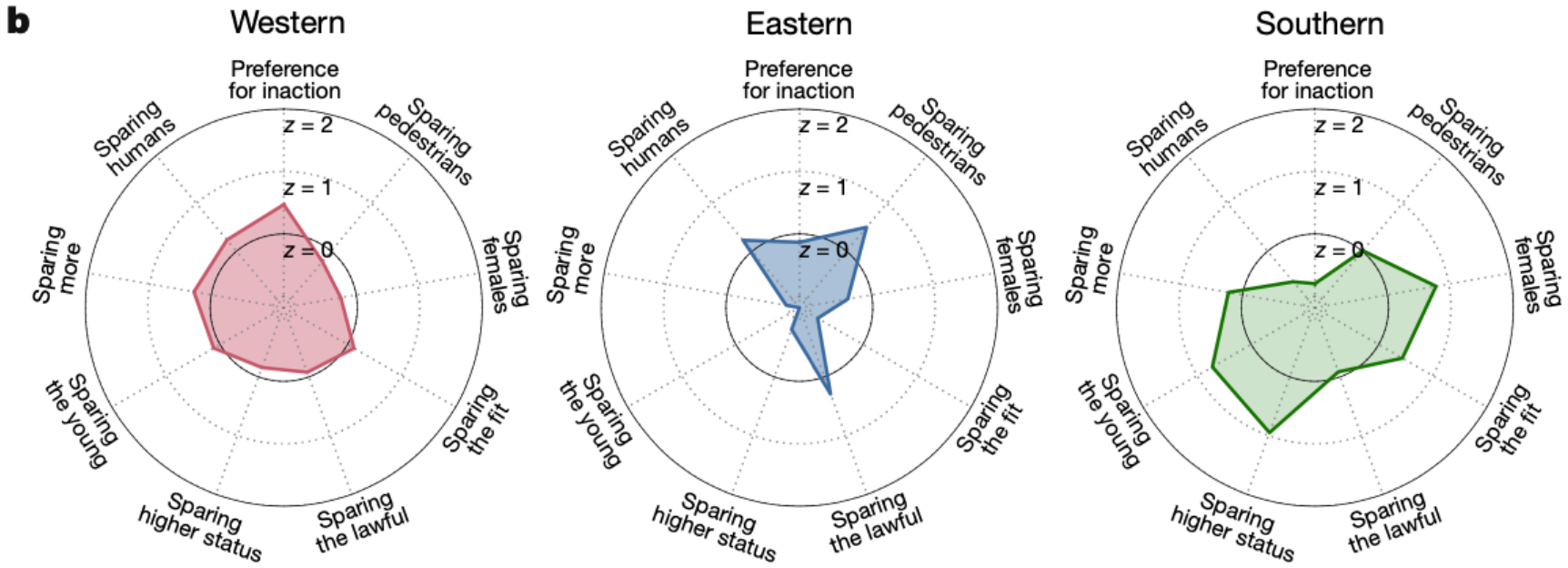
- The Moral Machine collected 40 million decisions from 4 million participants in 233 countries





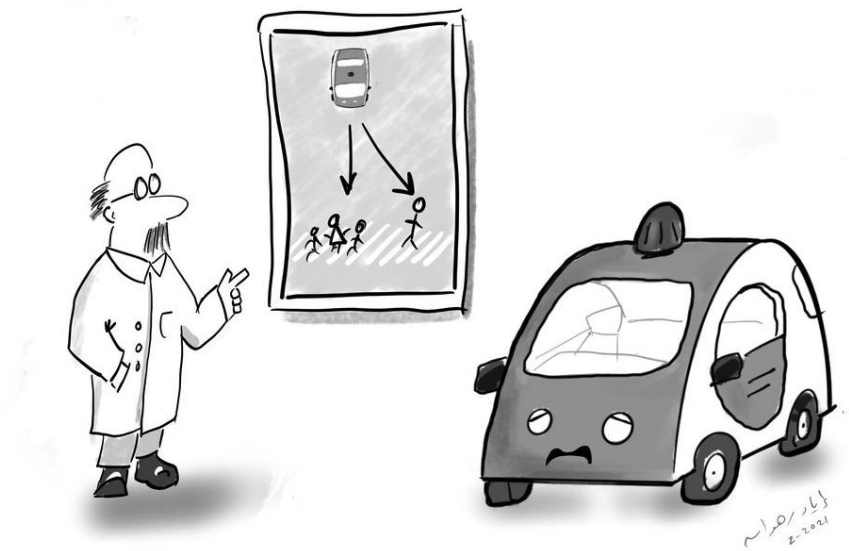
The Moral Machine

- The Moral Machine collected 40 million decisions from 4 million participants in 233 countries



The Moral Machine

- "Only humans can solve ethical dilemmas. AIs will simply maximize the objectives we give them. It is up to us, humans to specify the objectives and constraints that guide machine behaviour."



<< Please don't make me choose!
Just tell me what to do. >>

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Artificial Intelligence and Society

Towards a Responsible and Trustworthy AI

Miriam Seoane Santos
LIAAD, INESC TEC, FCUP, University of Porto
miriam.santos@fc.up.pt