# CSCM23 DESIGNING-IN TRUST, UNDERSTANDING, AND NEGOTIATION

1920~2020
Swansea University
Prifysgol Abertawe

# MODULE LOGISTICS

The module will be taught in several blocks by the following lecturers.


Bertie Müller


Markus Roggenbach


Adam Wyner


Xiuyi Fan


Siyuan Liu

**Schedule**:

Tuesday 9–11
8-9 & 10-11
8-10

Friday 11–12

**Assessment:**

Coursework (30%)
Handed out: 26 February
Deadline: 28 April

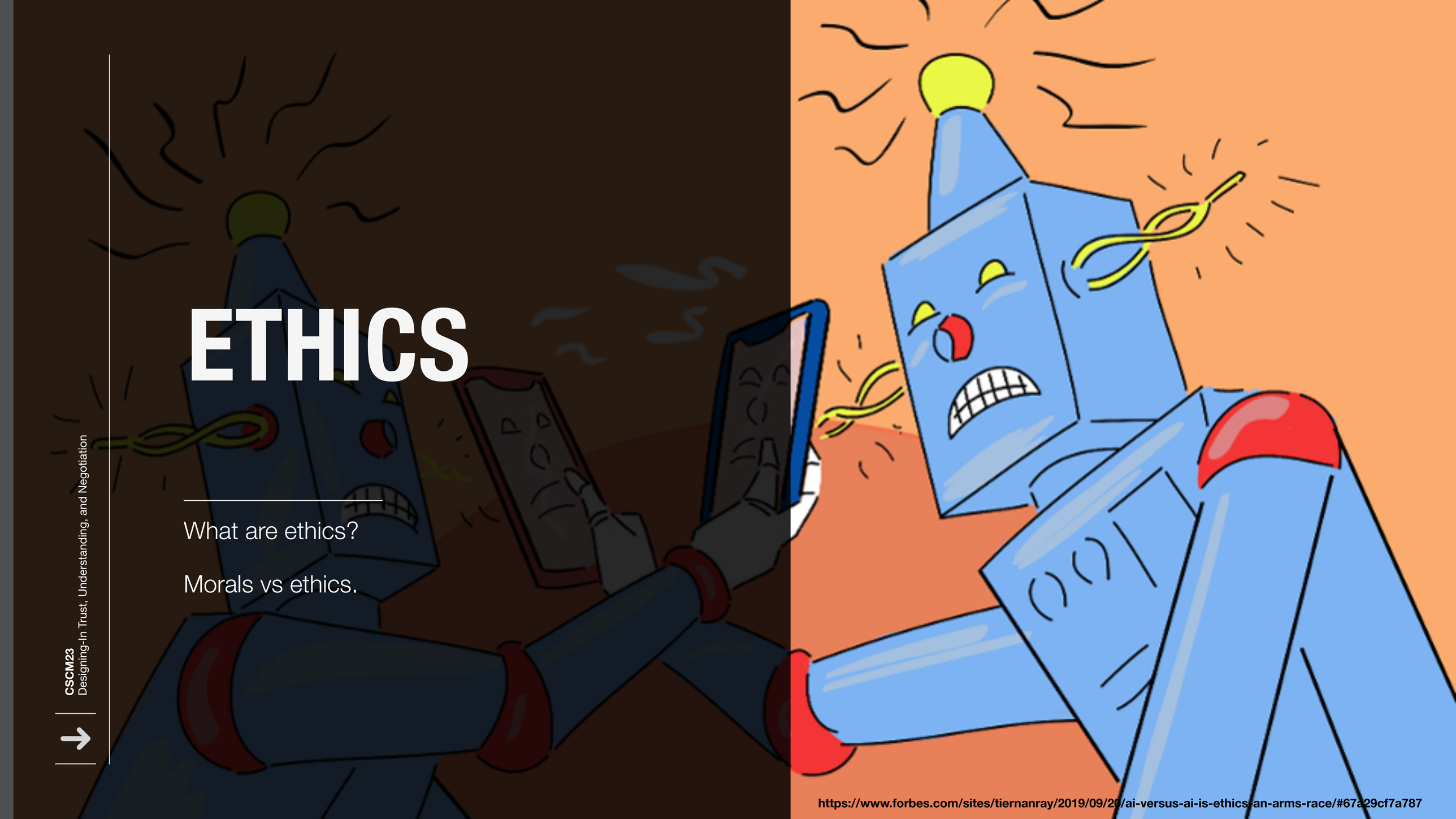
Exam (70%)

- Week 19 25/01 – 29/01 **(BBM) Intro, ethically-aligned design, privacy**
- Week 20 01/02 – 05/02 **(MR) Formal Methods**
- Week 21 08/02 – 12/02 **(MR) Formal Methods**
- Week 22 15/02 – 19/02 **(AW) AI & Law**
- Week 23 22/02 – 26/02 **(AW) AI Principles**
- Week 24 01/03 – 05/03 **(XF) X-AI**
- Week 25 08/03 – 12/03 **(XF) X-AI**
- Week 26 15/03 – 19/03 **(SL) Trust**
- Week 27 22/03 – 26/03 **(SL) Soft Security**
- *Week 29 Easter Recess 29/03 – 02/04*
- *Week 30 Easter Recess 05/04 – 09/04*
- *Week 31 Easter Recess 12/04 – 16/04*
- Week 28 19/04 – 23/04 **(BBM) Transparency, Counterfactuals**
- Week 32 26/04 – 30/04 **(ALL) Revision**

# ETHICS

What are ethics?

Morals vs ethics.



https://www.forbes.com/sites/tiernanray/2019/09/20/ai-versus-ai-is-ethics-an-arms-race/#67a29cf7a787

Cambridge Dictionary

# ethic

*noun* · **UK** /ˈeθ.ɪk/ **US** /ˈeθ.ɪk/

**C2** [ C usually plural ] **a system of accepted beliefs that control behaviour, especially such a system based on morals:**

*the (Protestant) **work** ethic*

*The ethics of journalism are much debated.*

*He said he was bound by a scientist's **code of** ethics.*

*Publication of the article was a **breach of** ethics.*

# moral

*adjective* · **UK** /ˈmɒr.əl/ **US** /ˈmɔːr.əl/

**B2** **relating to the standards of good or bad behaviour, fairness, honesty, etc. that each person believes in, rather than to laws:**

*It's her moral **obligation** to tell the police what she knows.*

*It is not part of a novelist's job to make a moral **judgment**.*

*She was the only politician to condemn the proposed law on moral grounds (= for moral reasons).*

*The Democrats are attempting to **capture the** moral **high ground** (= are trying to appear more honest and good than the other political parties).*

# WHY DO ETHICS MATTER?


## INDUSTRIAL ROBOTS

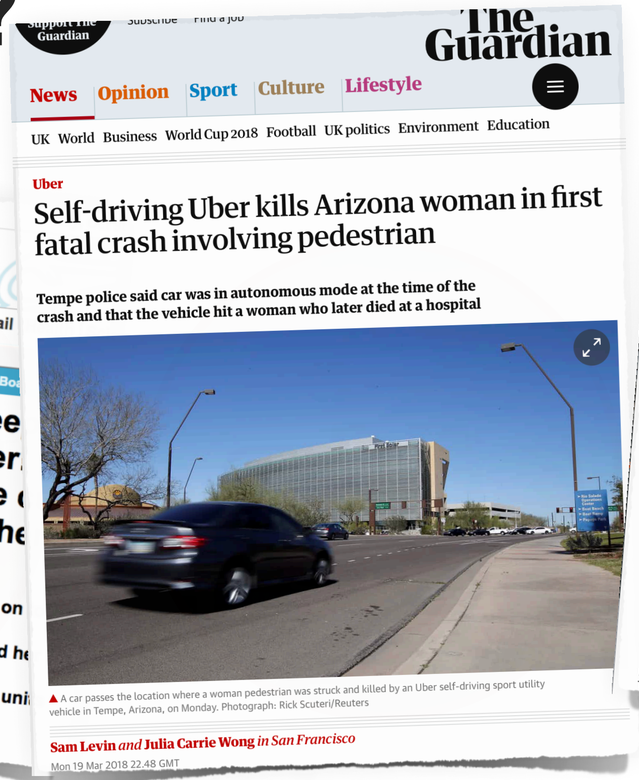
**LONG ESTABLISHED**

MINIMISE RISK OF INJURY & ENHANCE PRODUCTIVITY


## DOMESTIC ROBOTS

**AVAILABLE**

ALLEVIATE YOU FROM DAILY CHORES, JUST LIKE A DISHWASHER?




## AUTONOMOUS VEHICLES

**THE FUTURE**

WILL THEY MAKE TRAVELLING MORE ENJOYABLE AND SAFER?


## SMART DEVICES

**UBIQUITOUS**

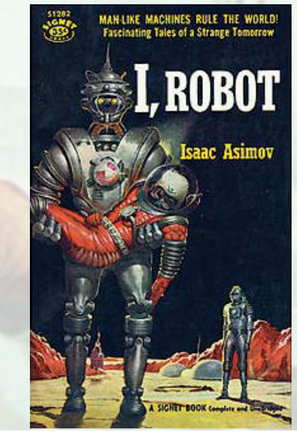
CAN WE IMAGINE A WORLD WITHOUT THEM?

ETHICS

# ROBOT ETHICS

Laws that govern acceptable use and behaviour of robots.

Assumption: Robots can act **autonomously**.




# LAWS OF ROBOTICS

1. **A robot may not injure a human being or, through inaction, allow a human being to come to harm.**
2. **A robot must obey orders given it by human beings except where such orders would conflict with the First Law.**
3. **A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.**

*The rules were introduced in Isaac Asimov's 1942 short story "Runaround" (included in the 1950 collection I, Robot), although they had been foreshadowed in a few earlier stories. The Three Laws, quoted as being from the "Handbook of Robotics, 56th Edition, 2058 A.D."*

ETHICS

# EPSRC

Regulating robots in the real world

In September 2010, experts drawn from the worlds of technology, industry, the arts, law and social sciences met at the joint EPSRC and AHRC Robotics Retreat to discuss robotics, its applications in the real world and the huge amount of promise it offers to benefit society.

**Five rules:**
**Principles for designers, builders and users of robots**

EPSRC
Engineering and Physical Sciences Research Council

1. Robots should not be designed as weapons, except for national security reasons.
2. Robots should be designed and operated to comply with existing law, including privacy.
3. Robots are products: as with other products, they should be designed to be safe and secure.
4. Robots are manufactured artefacts: the illusion of emotions and intent should not be used to exploit vulnerable users.
5. It should be possible to find out who is responsible for any robot.

# ❶ ROBOTS SHOULD NOT BE DESIGNED AS WEAPONS, EXCEPT FOR NATIONAL SECURITY REASONS.

- Tools have more than one use.
- We allow guns to be designed which farmers use to kill pests and vermin but killing human beings with them (outside warfare) is clearly wrong.
- Knives can be used to spread butter or to stab people. In most societies, neither guns nor knives are banned but controls may be imposed if necessary (e.g. gun laws) to secure public safety.
- Robots also have multiple uses.
- Although a creative end-user could probably use any robot for violent ends, just as with a blunt instrument, we are saying that robots should never be designed solely or even principally, to be used as weapons with deadly or other offensive capability.
- This law, if adopted, limits the commercial capacities of robots, but we view it as an essential principle for their acceptance as safe in civil society.

**Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.**

# ❷ ROBOTS SHOULD BE DESIGNED AND OPERATED TO COMPLY WITH EXISTING LAW, INCLUDING PRIVACY.

- We can make sure that robot actions are designed to obey the laws humans have made.

- **1:… Of course no one is likely deliberately set out to build a robot which breaks the law.**

  - **But designers are not lawyers** … building robots which do their tasks as well as possible will sometimes need to be balanced against protective laws and accepted human rights standards.

  - **Privacy** is a particularly difficult issue, …, a robot used in the care of a vulnerable individual may well be usefully designed to collect information about that person 24/7 and transmit it to hospitals for medical purposes. But the benefit of this must be balanced against that person's right to privacy and to control their own life e.g. refusing treatment. Data collected should only be kept for a limited time; …. Robot designers have to think about how laws like these can be respected during the design process (e.g. by providing off-switches).

# ❷ ROBOTS SHOULD BE DESIGNED AND OPERATED TO COMPLY WITH EXISTING LAW, INCLUDING PRIVACY.

- **2:… designed to make it clear that robots are just tools, designed to achieve goals and desires that humans specify.**
  - Users and owners have responsibilities as well as designers and manufacturers.
  - Sometimes it is up to designers to think ahead because robots may have the ability to learn and adapt their behaviour.
- **But users may also make robots do things their designers did not foresee.**
  - **Sometimes it is the owner's job to supervise the user** (e.g. if a parent bought a robot to play with a child).
  - But if a robot's actions do turn out to break the law, **it will always be the responsibility, legal and moral, of one or more human beings**, not of the robot (We consider how to find out who is responsible in law 5).

**Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.**

# ❸ ROBOTS ARE PRODUCTS: AS WITH OTHER PRODUCTS, THEY SHOULD BE DESIGNED TO BE SAFE AND SECURE.

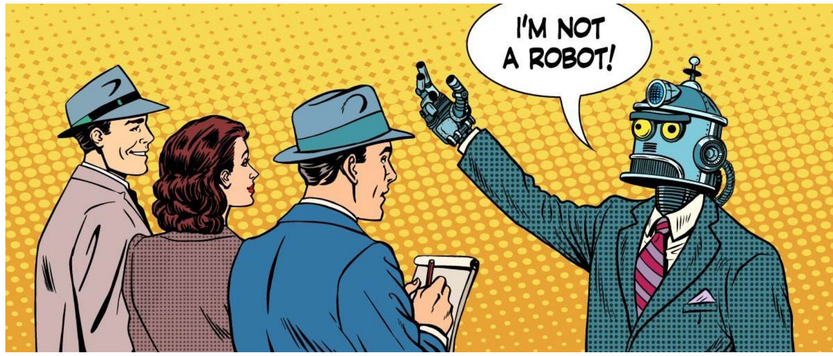
- **Robots are simply not people**. They are **pieces of technology** their owners may certainly want to protect … but we will always value **human safety** over that of machines … so that **people** can **trust** and **have confidence** in them.

- This is not a new problem in technology. We already have rules and processes that guarantee that, e.g. household appliances and children's toys are safe to buy and use. …

- This still leaves a debate open about how far those who own or operate robots should be allowed to protect them from e.g. theft or vandalism, say by built-in taser shocks. **The group chose to delete a phrase that had ensured the right of manufacturers or owners to include "self defence" capability into a robot.** In other words we do not think a robot should ever be "armed" to protect itself. This actually goes further than existing law, where the general question would be whether the owner of the appliance had committed a criminal act like assault without reasonable excuse.

**Robots are products.
They should be designed using processes which assure their safety and security.**

# ❹ ROBOTS ARE MANUFACTURED ARTEFACTS: THE ILLUSION OF EMOTIONS AND INTENT SHOULD NOT BE USED TO EXPLOIT VULNERABLE USERS.

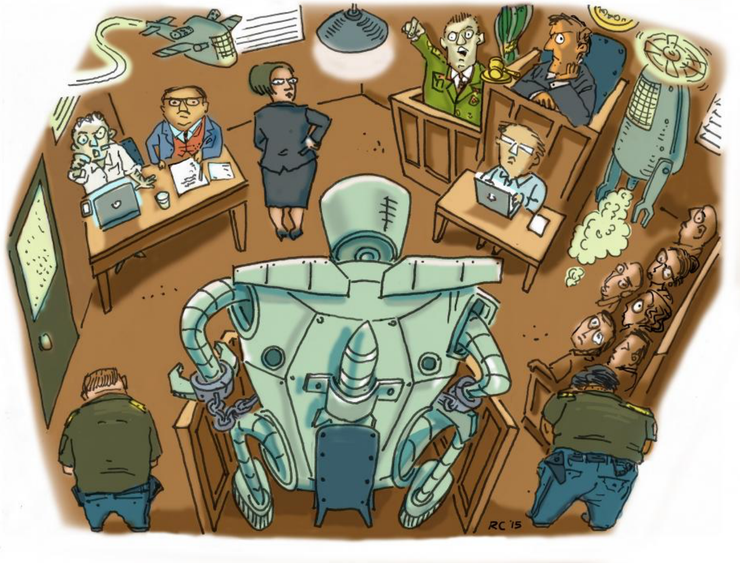



- One of the great promises of robotics is that robot toys **may give pleasure, comfort and even a form of companionship to people** who are not able to care for pets, whether due to rules of their homes, physical capacity, time or money.
  - [If] a user becomes attached to such a toy, manufacturers [could] claim the robot has needs or desires that could unfairly cost the owners or their families more money.
  - The legal version of this rule was designed to say that although it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence, **anyone who owns or interacts with a robot should be able to find out what it really is and perhaps what it was really manufactured to do**.
- **Robot intelligence is artificial**, and … the best way to protect consumers was to remind them of that by guaranteeing a way for them to "lift the curtain" (to use the metaphor from The Wizard of Oz).

**Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.**

# ❺ IT SHOULD BE POSSIBLE TO FIND OUT WHO IS RESPONSIBLE FOR ANY ROBOT.

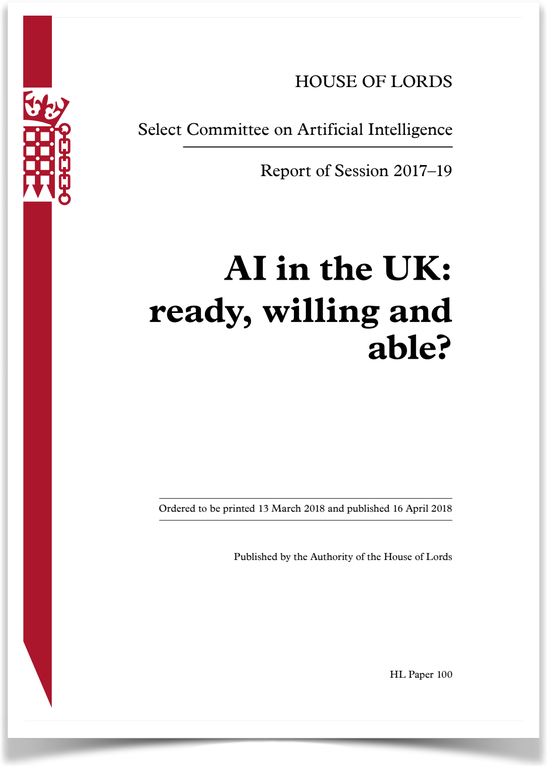
- … a practical framework for what all the rules above already implicitly depend on:
  **a robot is never legally responsible for anything**.
  - It is a tool. If it malfunctions and causes damage, a human will be to blame. Finding out who the responsible person is may not however be easy. …
- Responsibility might be practically addressed in a number of ways.
  - E.g., one way forward would be a licence and register (just as there is for cars) that records who is responsible for any robot. …
- Importantly, it should still **remain possible for legal liability to be shared or transferred** e.g. both designer and user might share fault where a robot malfunctions during use due to a mixture of design problems and user modifications. …



**The person with legal responsibility for a robot should be attributed.**

# SELECT COMMITTEE

Principles for an AI code




1. Artificial intelligence should be developed for the common good and benefit of humanity.

2. Artificial intelligence should operate on principles of intelligibility and fairness.

3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities.

4. All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence.

5. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence.

# EUROPEAN COMMISSION

European Commission High-Level Expert group on Artificial Intelligence (2019)


- **Principles:**

(i) Respect for human autonomy

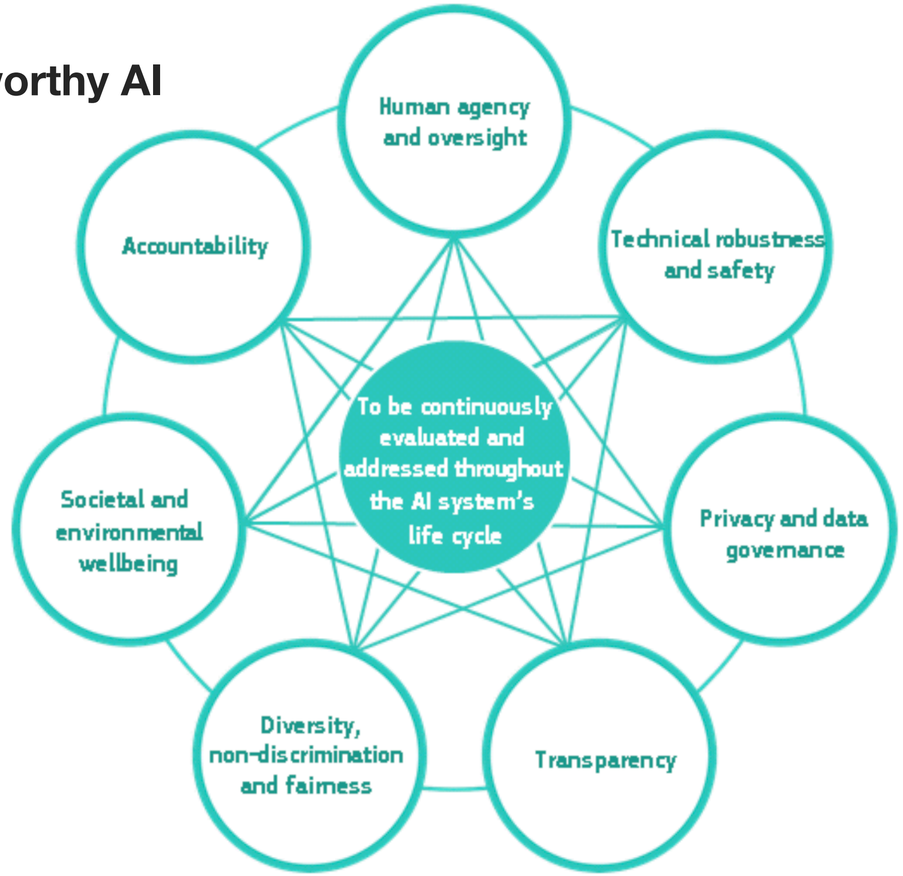
(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- **Realising Trustworthy AI**

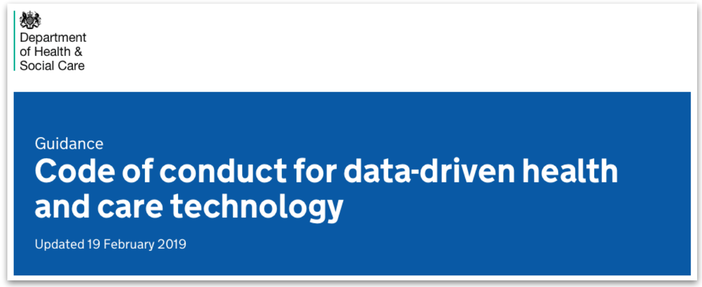
7 key requirements (non-exhaustive) include **systemic**, **individual** and **societal** aspects



ETHICS

# GOOGLE & OTHERS

Sidebar Text Lorem Ipsum, Dolor Sit Amet: Duis mollis, est non commodo luctus, nisi erat porttitor ligula, eget lacinia odio sem nec elit.




Google

1. **Be socially beneficial.**
2. **Avoid creating or reinforcing unfair bias.**
3. **Be built and tested for safety.**
4. **Be accountable to people.**
5. **Incorporate privacy design principles.**
6. **Uphold high standards of scientific excellence.**
7. **Be made available for uses that accord with these principles.**

*"Many technologies have multiple uses. We will work to limit potentially harmful or abusive applications. As we develop and deploy AI technologies, we will evaluate likely uses in light of the following factors:*

- *Primary purpose and use: the primary purpose and likely use of a technology and application, including how closely the solution is related to or adaptable to a harmful use*
- *Nature and uniqueness: whether we are making available technology that is unique or more generally available*
- *Scale: whether the use of this technology will have significant impact*
- *Nature of Google's involvement: whether we are providing general-purpose tools, integrating tools for customers, or developing custom solutions"*

**AI applications Google will not pursue:**

- In addition to the above objectives, we will not design or deploy AI in the following application areas:

- **Technologies that cause or are likely to cause overall harm.** Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.

- **Weapons** or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.

- **Technologies that gather or use information for surveillance** violating internationally accepted norms.

- **Technologies whose purpose contravenes widely accepted principles of international law and human rights.**

- We want to be clear that while we are **not developing AI for use in weapons**, we will **continue our work with governments and the military in many other areas**. These include cybersecurity, training, military recruitment, veterans' healthcare, and search and rescue. These collaborations are important and we'll actively look for more ways to augment the critical work of these organizations and keep service members and civilians safe.

ETHICS

# IMAGINE …

# ETHICAL BUMPS

There are pitfalls along the way. Boston's *Street Bump* smartphone app was hailed as a big data triumph that allowed the city to quickly identify potholes and then fix them without having workers patrol the streets. Anyone who downloaded the app onto their smartphone would automatically notify the City of potholes as they drove around. Boston proudly claimed that the data gave them real-time information to fix problems and to plan long-term investments. On the other hand, on its own, *Street Bump* actually produced a map of potholes that systematically favors young, affluent areas where more people own smartphones. This is a core problem of "found data"—as opposed to data gathered from a fair sample—it can contain systematic biases and it takes careful thought to spot and correct them.

ETHICS

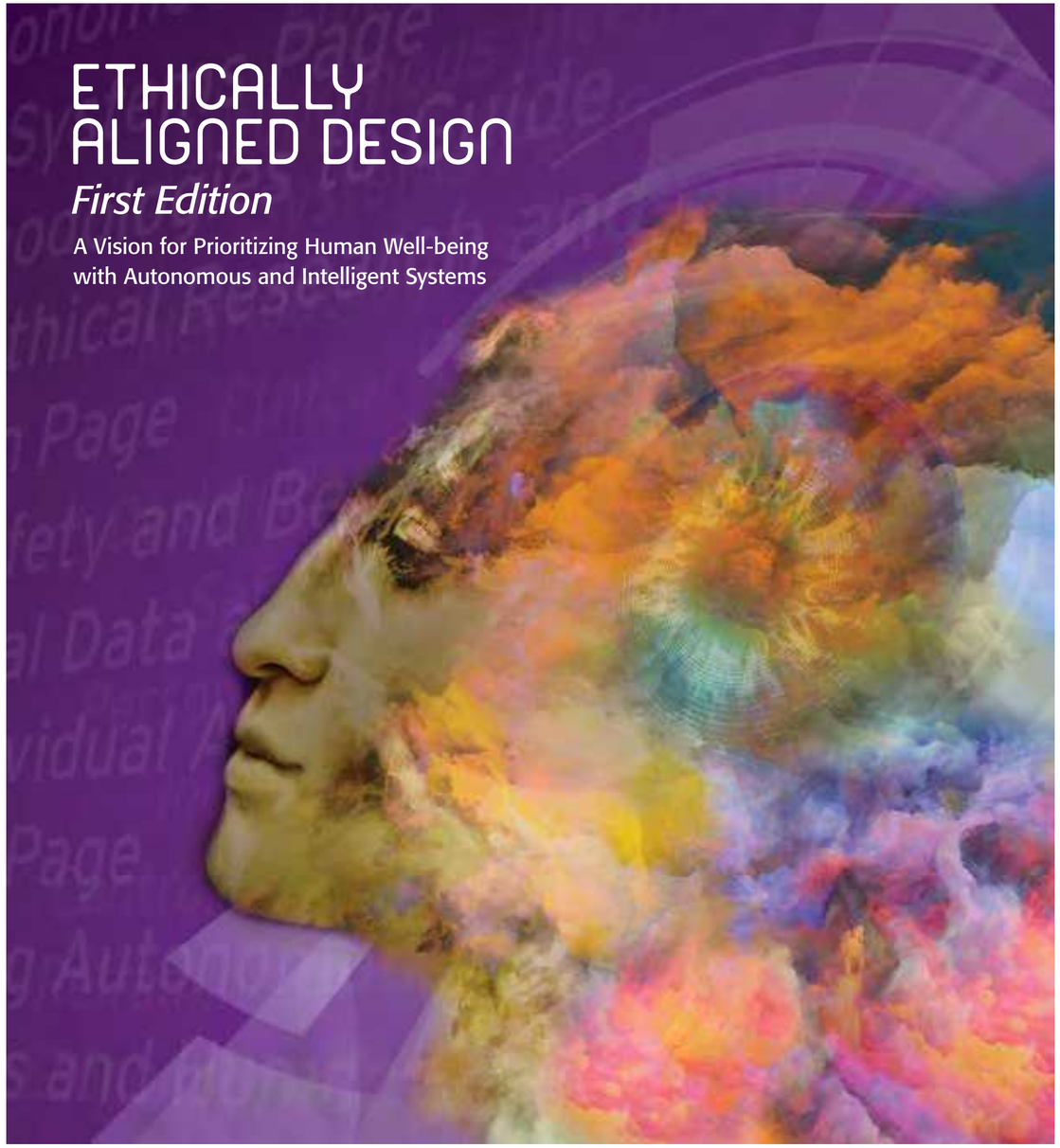
# ETHICAL DESIGN

IEEE initiative on Ethically Aligned Design


PwC: Responsible AI


- Diversity and interdisciplinarity in teams and methods.
- Explainability of decision processes
  - E.g., timeline documents certifying quality of data and algorithms.
- Respectful of human rights.
- Beneficial for humanity.
- Purposeful.

IEEE
Advancing Technology for Humanity






# Table of Contents

## II. General Principles

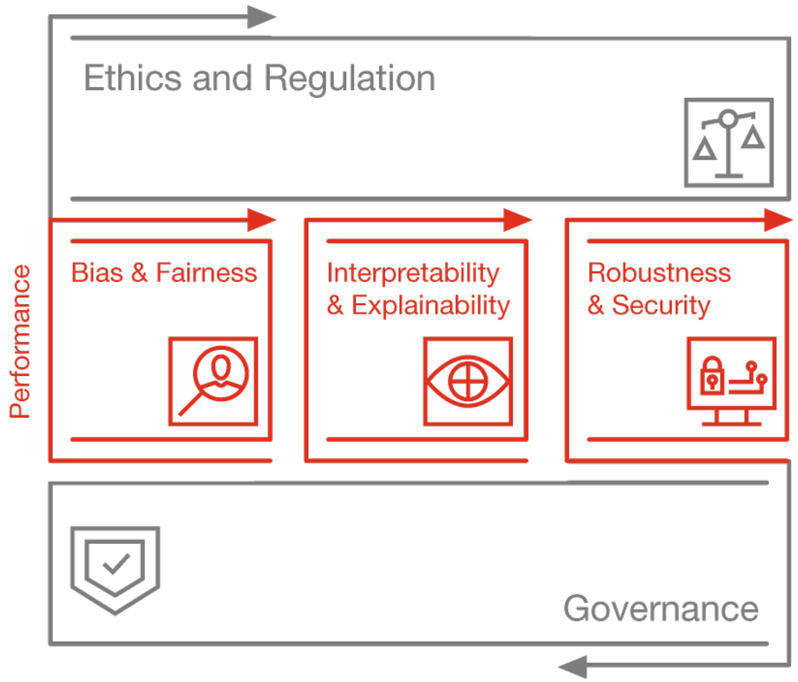
The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

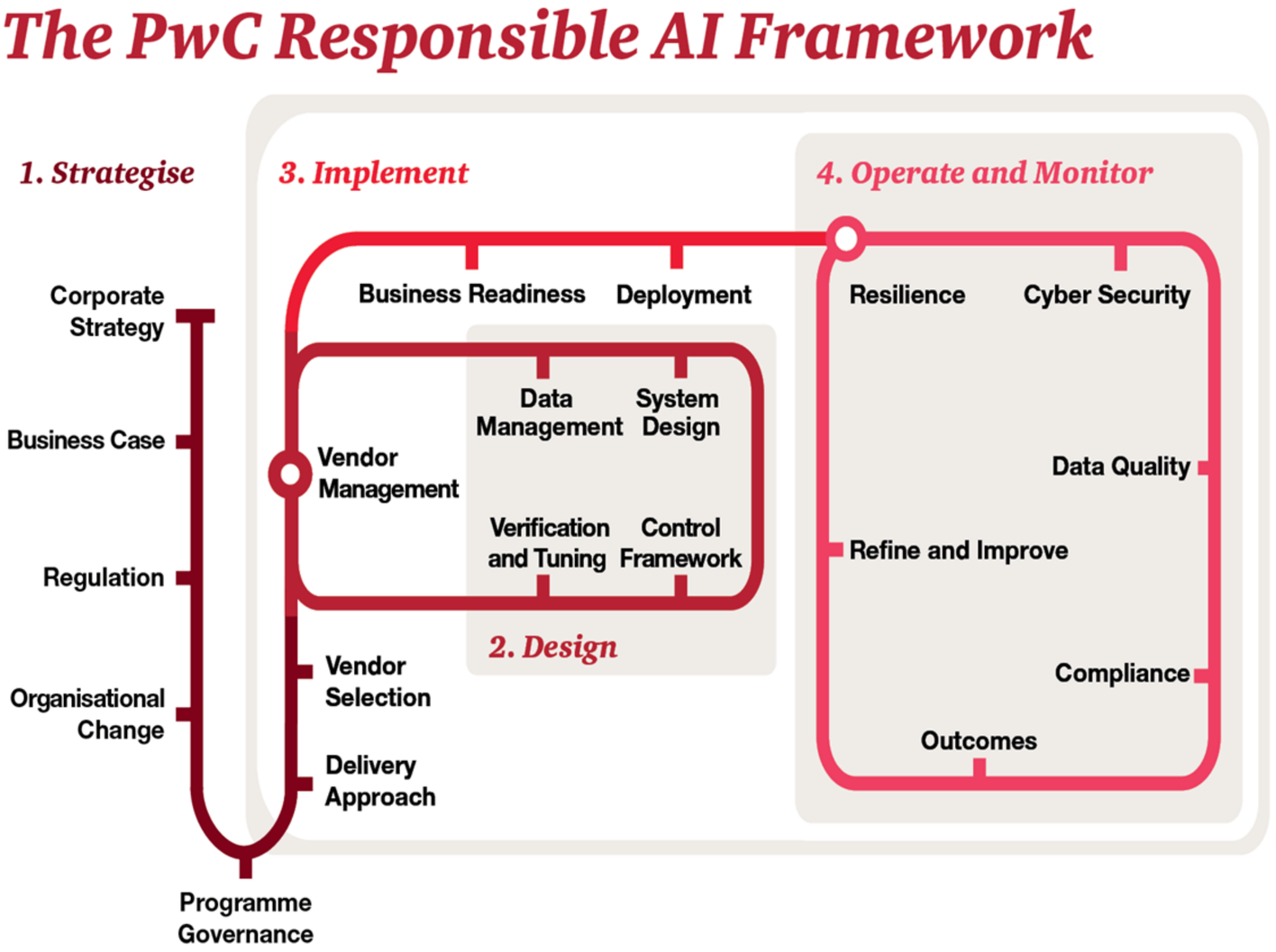
1. **Human Rights**
   A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being**
   A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency**
   A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness**
   A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency**
   The basis of a particular A/IS decision should always be discoverable.
6. **Accountability**
   A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse**
   A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence**
   A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

# PWC

Responsible AI

https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html



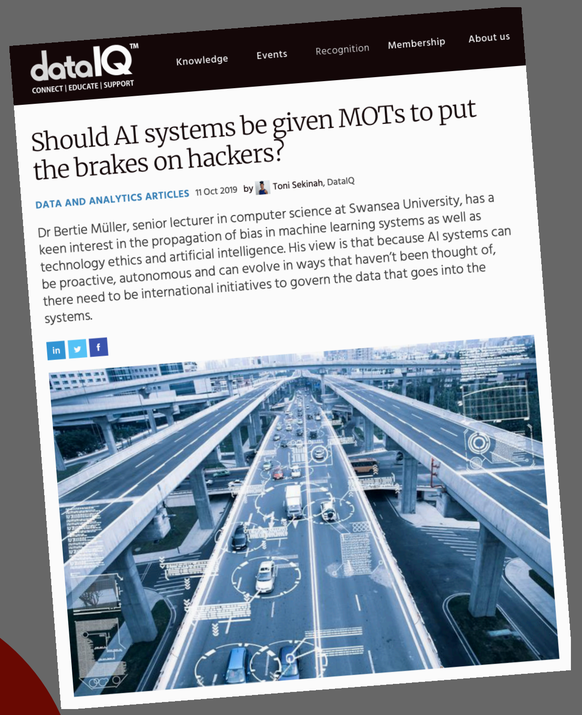



ETHICS

# ETHICAL APPROVAL

Approval as a tick box exercise carried out once in the systems development lifecycle is not enough for autonomous systems …

- Dynamic Technology “M.O.T.”
- **“We need to have something like an MOT of governance, of ethical approval that is repeated maybe once a year, or more often depending on how critical a system is, how sensitive the data is, how vulnerable the system is.”**




**Thresholds**
Definition of ethical constraints

**Weightings**
Dynamic weighting of contextual factors

**Law**
Consideration of general legal requirements and restrictions

**Contract**
Agreement/ consent on permissible use of data

**Adequacy**
Decisions checked for accuracy & contractual compliance

ETHICS

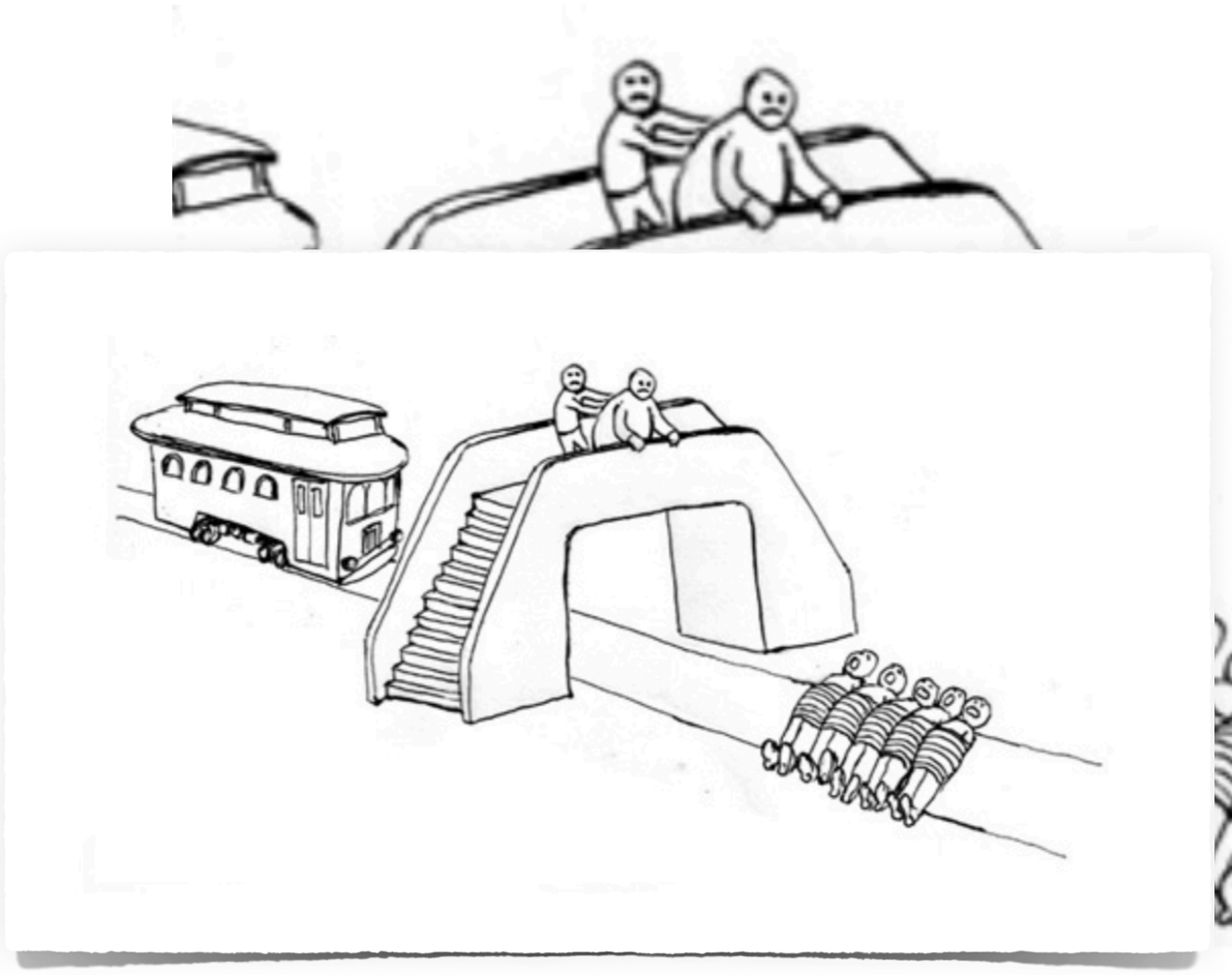
# UTILITARIANISM

The right actions are the one that increase the utility in society.

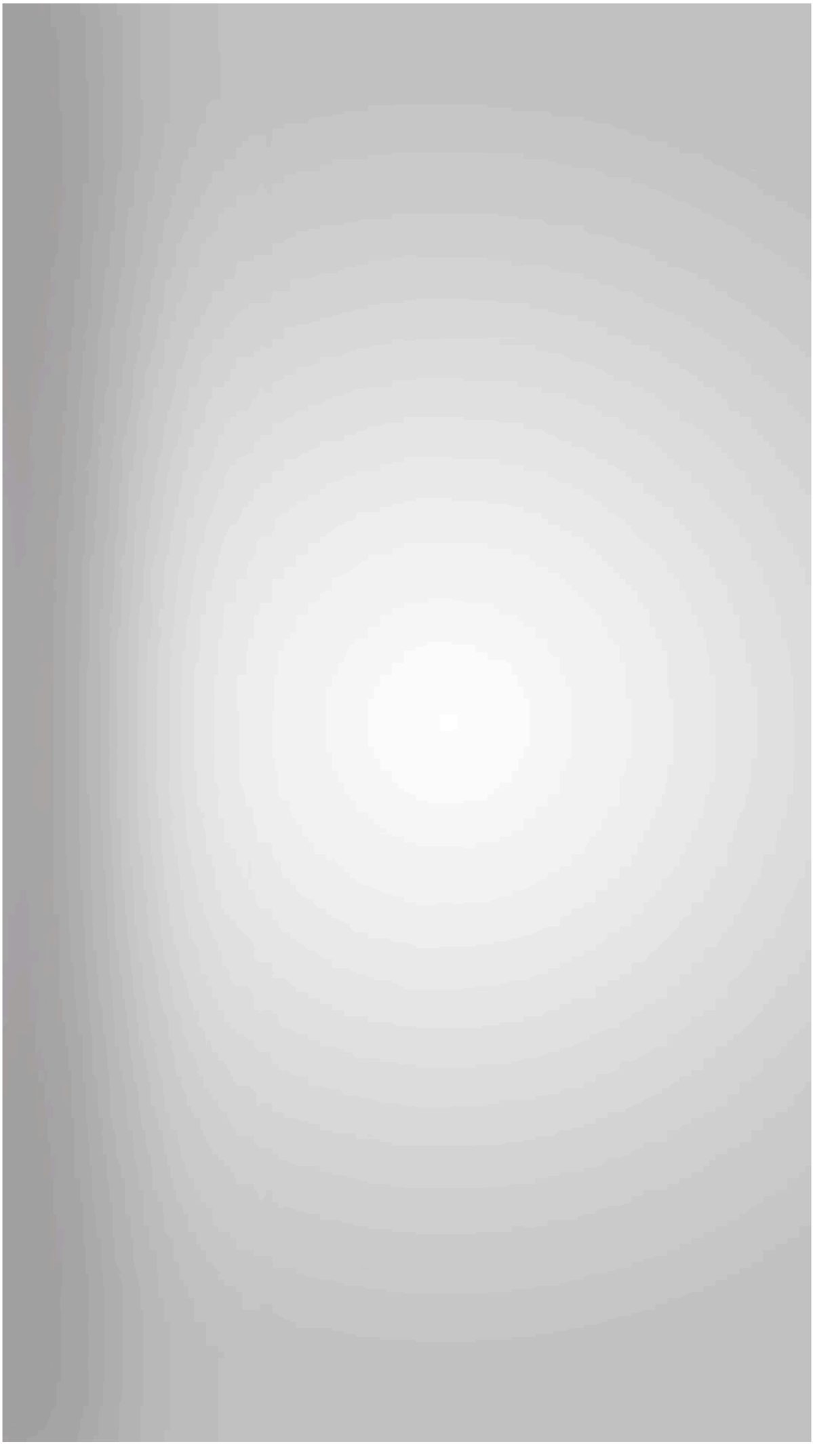
Jeremy Bentham (1748-1832) and John Stuart Mill (1806-1873)

**Act-Utilitarianism**
Acting driven by utilitarianism:
Morally correct actions are those that directly produce the greatest overall good, everyone considered

**Rule-utilitarianism**
Morally correct action is the one covered by a rule that if generally followed would produce the most favourable balance between good and evil, everyone considered (rules must be followed constantly even if they are locally not the best choice)

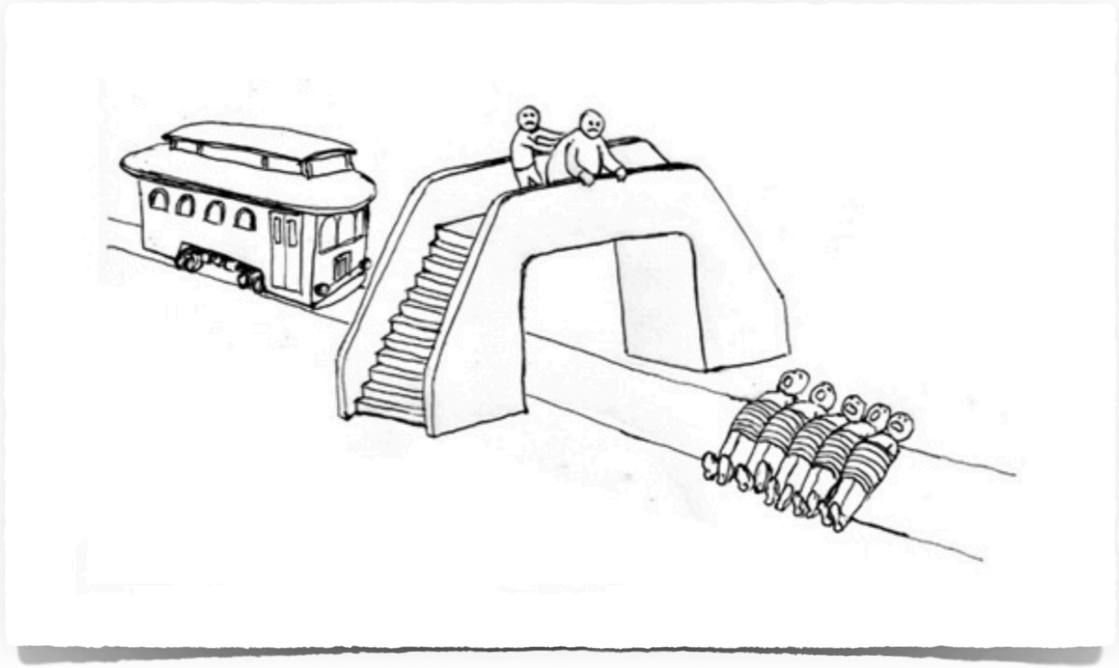


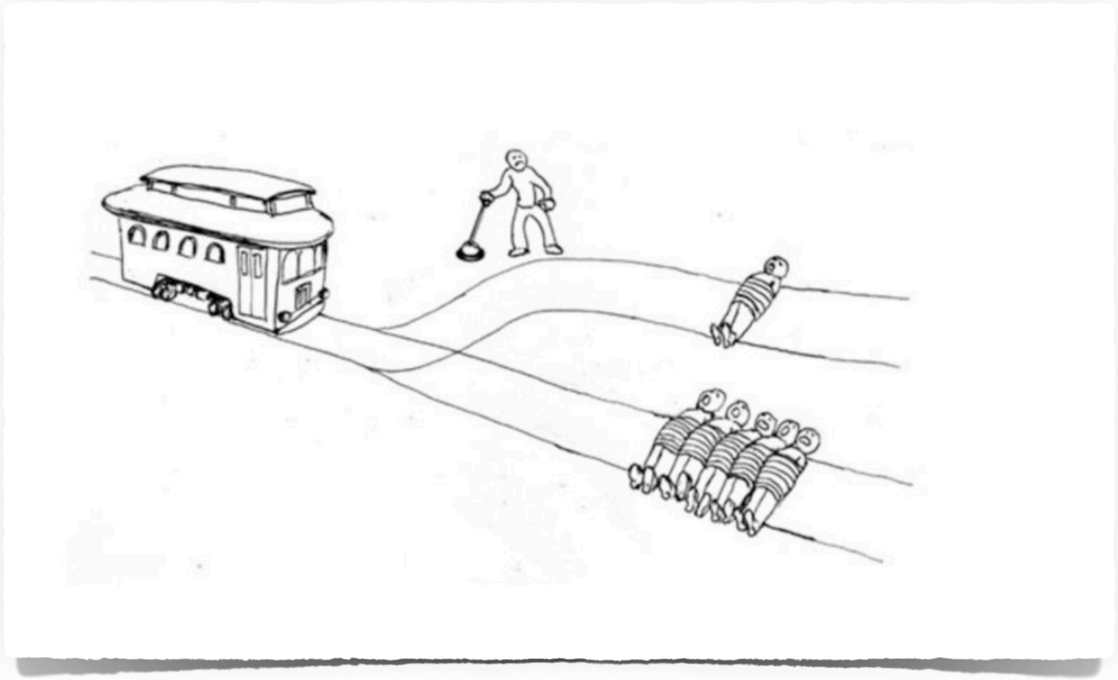
- The Trolley Problem

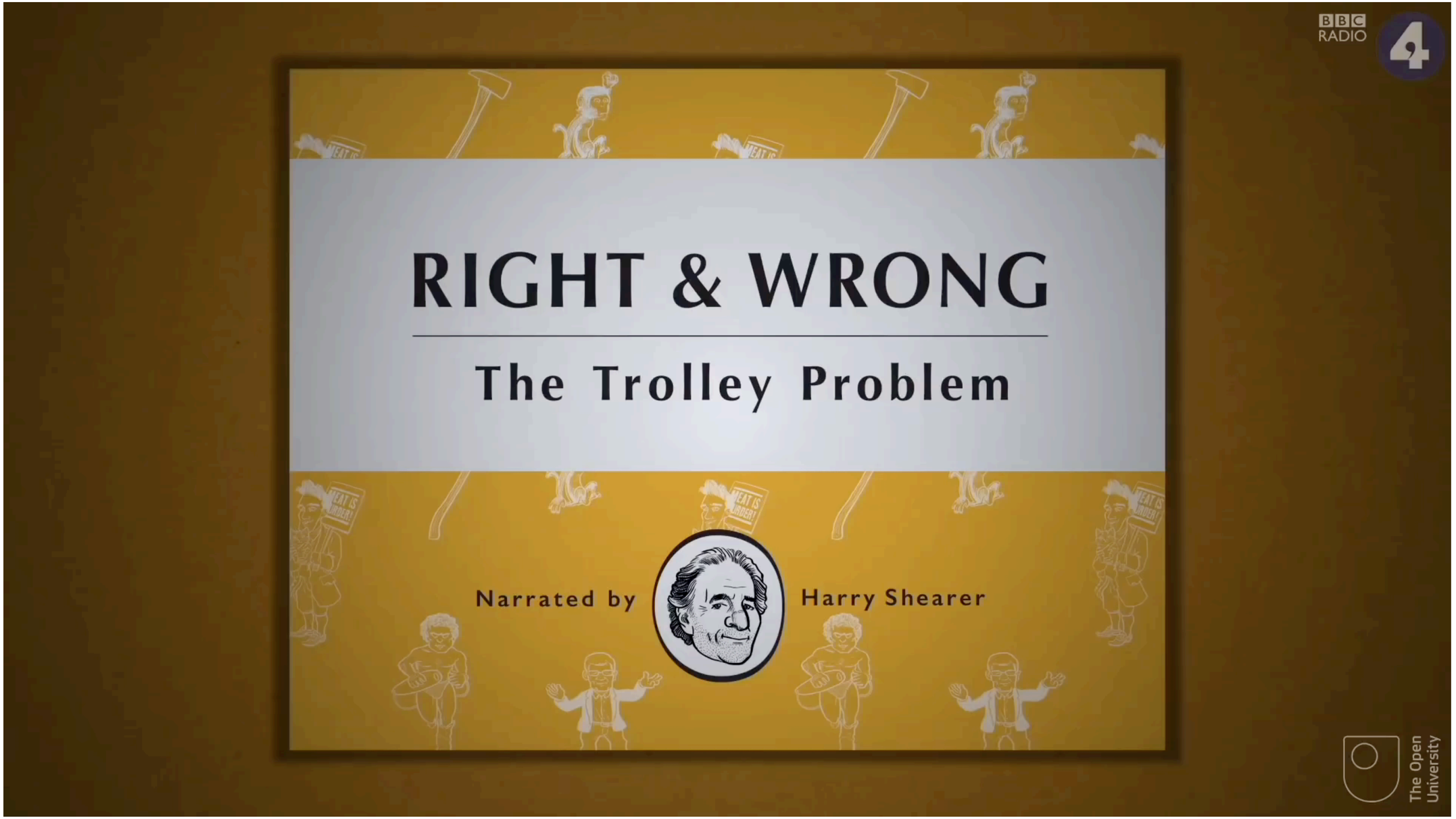


ETHICS

# IS THERE ANY DIFFERENCE BETWEEN THESE SCENARIOS?



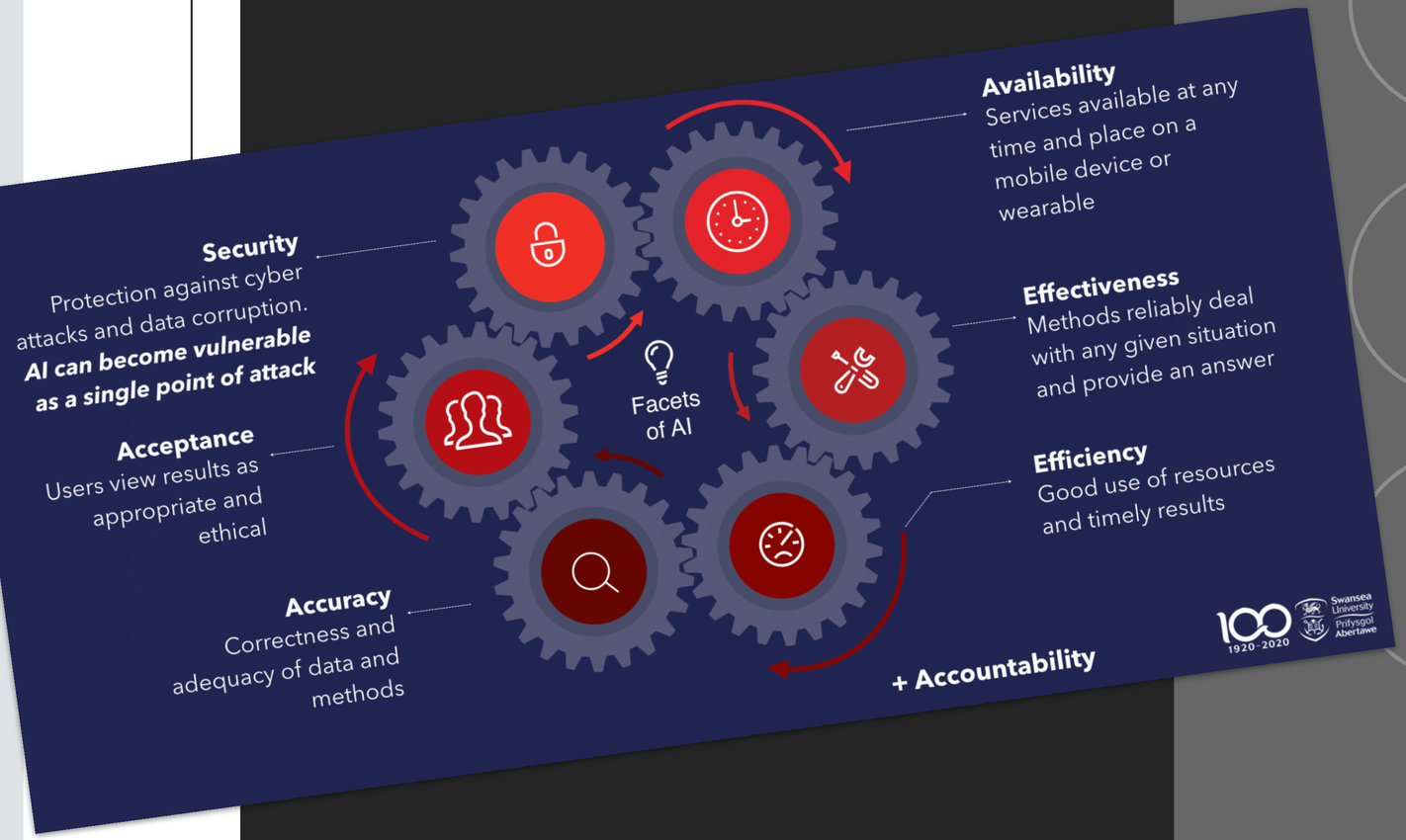



BBC RADIO 4
RIGHT & WRONG
The Trolley Problem
Narrated by
Harry Shearer
The Open University

RE³ APPROACH

# CURRENT RESEARCH AIMS

**RE³ = Reliable, Responsible, & Resilient**




## Learning & Reasoning

Combining machine learning with rule-based systems of reasoning, e.g., multi-agent systems.

Partial explainability

## Counterfactuals

Explainability of decisions by automatic application of counterfactual reasoning.

Experimental explainability

## Transparency

Use of tagged datasets and results to retain information about, e.g., the provenance, bias

Theory of propagation of bias.

## Dynamic Assurances

Autonomy requires new processes for deploying and monitoring systems that rely on automation and learning to ensure accuracy and long-term reliability in changing contexts.

Data governance, standards, regulation

OUTLOOK

# SO WHAT SHOULD YOU EXPECT FROM THIS MODULE?

- **Reliable**
  - **Formal methods** can lead to provably correct systems.
  - **Diversity** in design leads to fewer unexpected behaviours.
- **Responsible**
  - **Legal and ethical compliance** by design
  - **Explainability** and understanding of decision-making processes
- **Resilient**
  - Safe and **secure** systems
  - **Robust** systems design