Responsible Data Science

Session 2: 26.04.2023, 13.30 – 16.45 h
MA Seminar, SoSe 2024, Hasso-Plattner Institute

Dr. Simon David Hirsbrunner

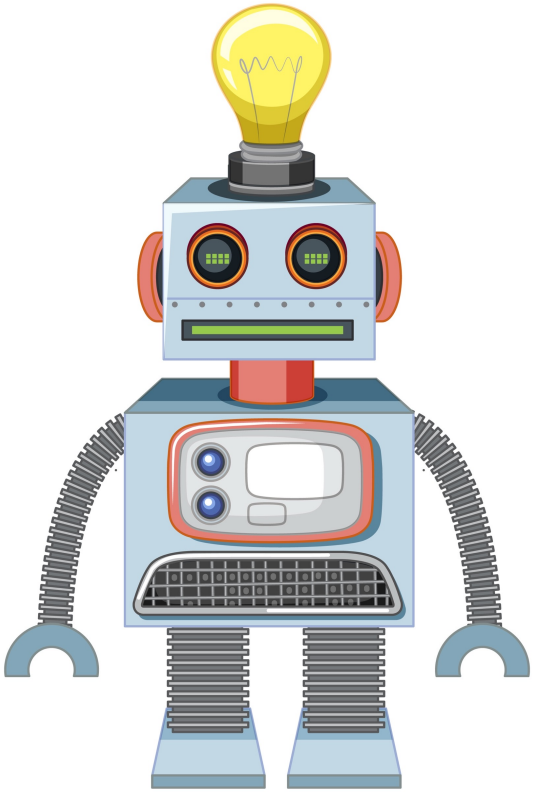
Today

13h30 Introduction and getting to know each other
13h45 Key concepts in ethics
14h30 Break
14h40 The universe of AI Ethics
15h30 Break
15h45 Introduction to Value-Sensitive Design
16h00 Exercise
16h30 Distribution of presentation topics
16h45 End

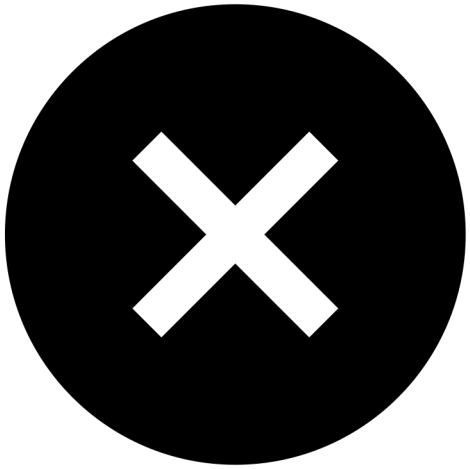
Getting to know each other

Tell us a few words about you.

If you could teach a robot one ethical value, which one would it be?



-- Illustrating pictures removed by lecturer for copyright reasons ---



AI Ethics?

No speculative doomsday summoning science

Key concepts in ethics research

Ethics (from ethos gr. for „way of living“)

1. **Descriptive ethics**
   Description of behaviors, norms of behavior, and behavioral attitudes and value judgments.

2. **Normative ethics**
   Justification, criticism, or justification of behaviors, behavioral norms, and behavioral attitudes and value judgments.

3. **Applied ethics**
   Domain-specific ethical questions (medicine, technology, sustainability, etc.)

4. **Meta-ethics**
   Clarification of the basics of communication and understanding of behavioral norms and value judgments.

(Werner 2021: 6ff)

1 Exercise 1: trolley problem

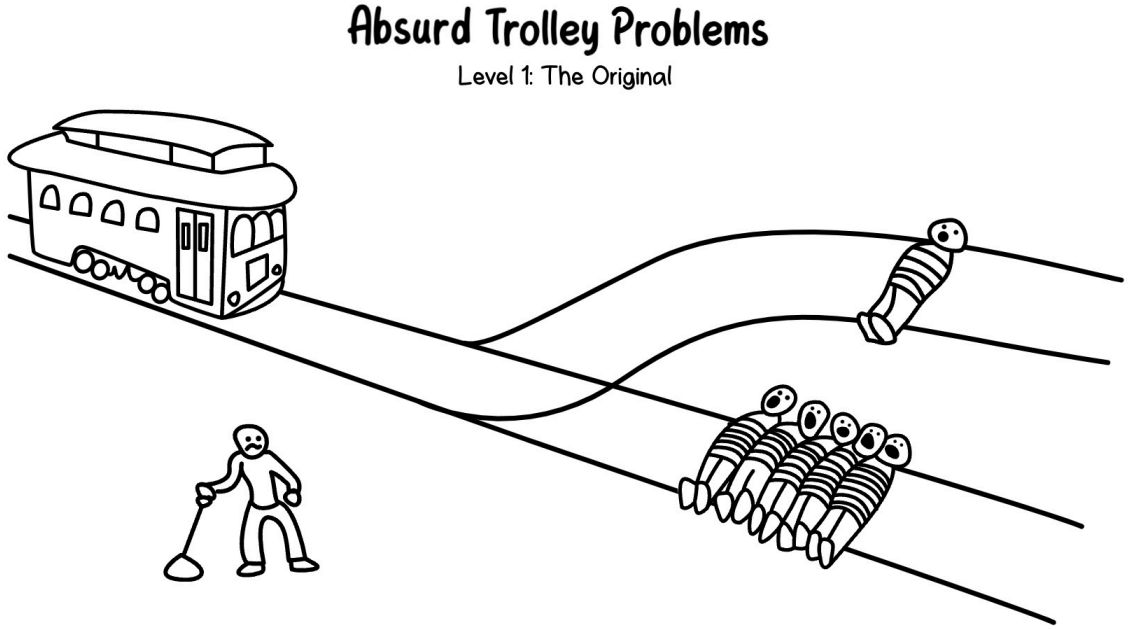
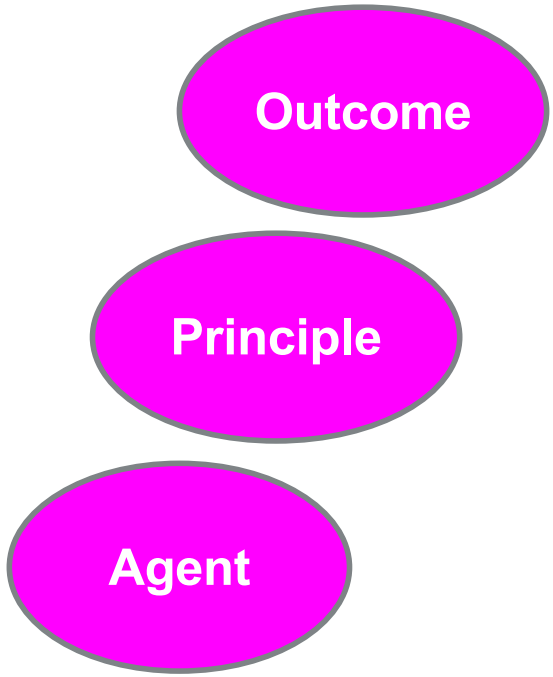


Image source: https://neal.fun/absurd-trolley-problems/

Ethics schools

1. **Consequentialist ethics**
   (consequences matter)
2. **Deontological ethics**
   (duty matters)
3. **Virtue ethics**
   (character matters)



(Henning 2019: 45ff; Stanford Encyclopedia of Philosophy)

Consequentialist ethics

The moral status of an action **depends exclusively on its consequences.** An action is **morally permissible** precisely **if its consequences have a certain value** compared to the available alternatives.
(Henning 2019, 45 ff)

e.g. Classical Utilitarianism

An action is morally permissible exactly *if the sum of the resulting pleasure of the affected parties minus the sum of the resulting suffering of the affected parties* is at least as large as the corresponding sum of any other available action.
(Bentham in Henning 2019, 45 ff)

Advantages of utilitarist ethics

- Impartiality
- Clarity
- Efficiency

(Werner 2021, 119)

Challenges of utilitarism in practice e.g. assigning values in AI and robotics

Example

Imagine a robot is walking to the post office to post a letter. It walks along a path by a stream. Suddenly a toddler chases a duck which hops into the stream. The toddler slips and falls into the water which is one meter deep. The toddler is in imminent danger of drowning. The robot is waterproof. Should it enter the water and rescue the toddler or should it post the letter?

(Bartneck et al. 2021: 24)

Challenges of utilitarism in practice e.g. assigning values in AI and robotics

Value of toddler = +1,000,000

Value of letter = +1

(Bartneck et al. 2021: 24)

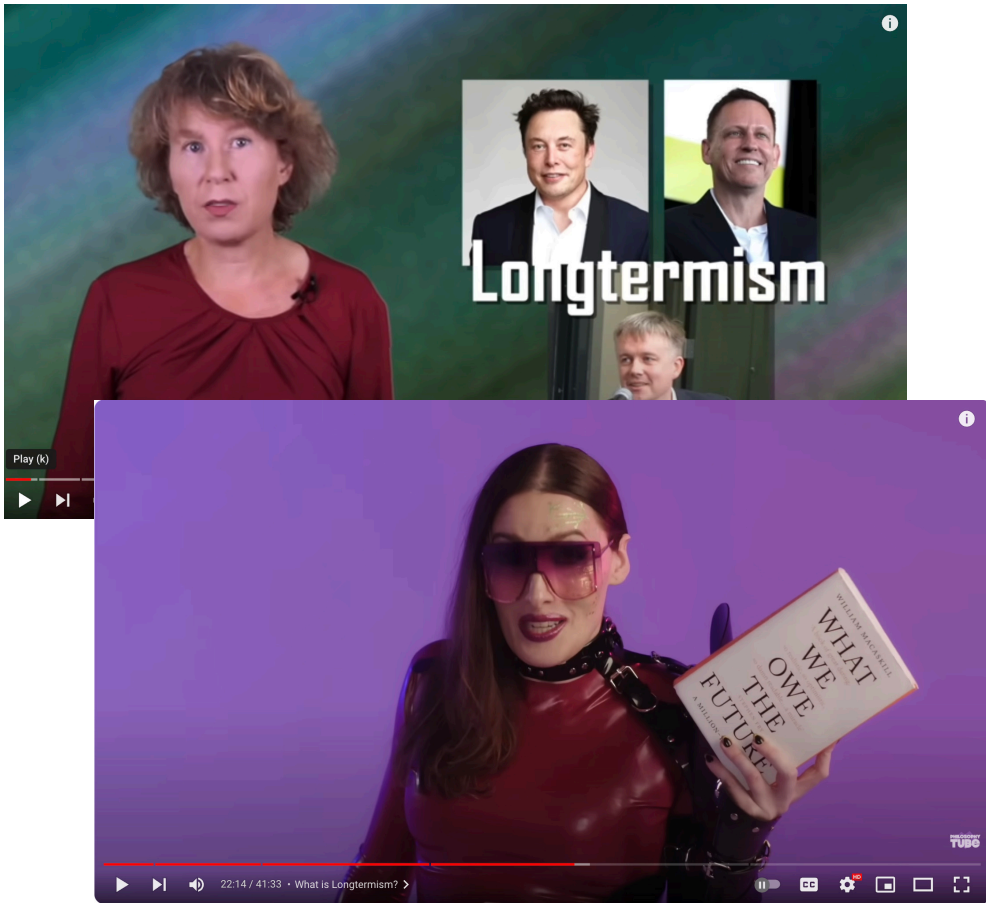
Challenges of utilitarism in practice e.g. assigning values in AI and robotics

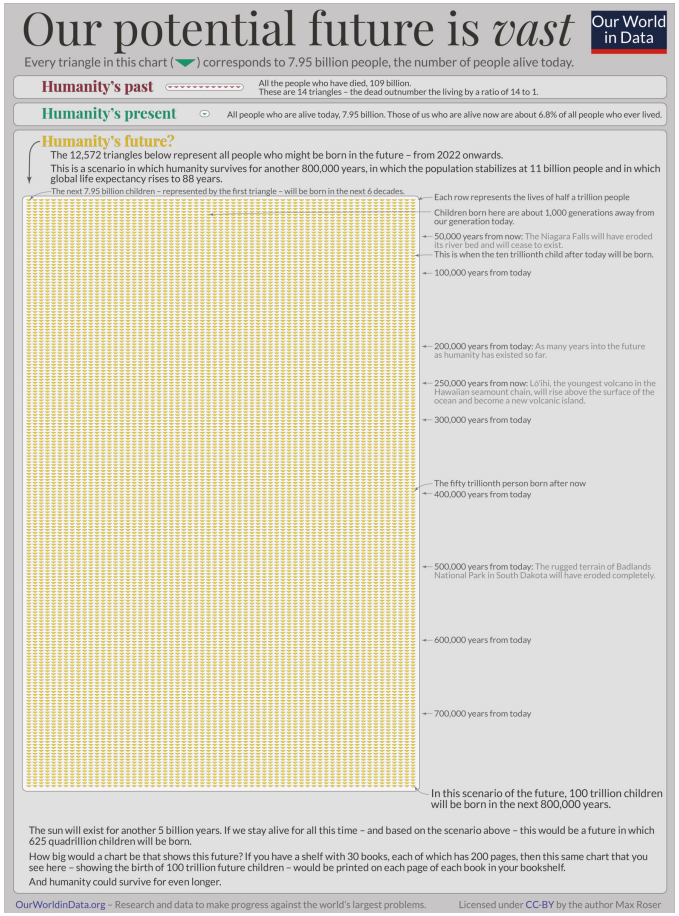
Value of toddler = +1,000,000

Value of letter**s** = **+1,000,001**

Robot drives a post truck and is supposed to deliver 1,000,001 letters.
--> In this case, a recourse on simple rules ("children lives go first" or "never depart from your primary duties" might be clearer and more reliable)

What about future life? Longtermism





YouTube channels: https://youtu.be/Lm0vHQYKI-Y?si=uW9-ed3iBI-PScwr (PhilosophyTube),
https://youtu.be/B_M64BSzcRY?si=uD3INuEg-n0v3YUi (Sabine Hossenfelder)
Image source: Max Rosler 2022 / https://ourworldindata.org/the-future-is-vast

The sun will exist for another 5 billion years. If we stay alive for all this time – and based on the scenario above – this would be a future in which 625 quadrillion children will be born.

How big would a chart be that shows this future? If you have a shelf with 30 books, each of which has 200 pages, then this same chart that you see here – showing the birth of 100 trillion future children – would be printed on each page of each book in your bookshelf.

And humanity could survive for even longer.

Deontological ethics

In respect to the moral status of an action, what counts are **motivations and moral principles, not consequences.**

Deon = **duty** (greek)

Categorical Imperative
Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.

(Kant in Henning 2019, 80)

Example for deontological ethics in AI: Asimov's robot laws

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 it receives from human beings** except when such orders conflict with Law 1.
3. A robot **must protect its own existence** as long as such protection does not conflict with Laws 1 and 2.

0. **No robot may harm humanity** or through inaction allow humanity to come to harm.

(Asimov 2004)

Contractualism

A framework for determining **just principles of social organization** that are fair, impartial, and chosen by individuals in a hypothetical situation of equality. It emphasizes the importance of ensuring that the principles chosen are acceptable to all members of society, regardless of their particular circumstances or interests (i.e. moral judgements taken under a **„veil of ignorance"** in which the personal position in a situation is unknown).

(see Rawls 1971)

Virtue ethics

Being virtuous is a necessary condition for living a happy life. Ethics of the Greek antiquity, building on ideas by Plato and Aristotle. Virtues include bravery, moderation, justice, prudence, generosity.

(Henning 2019, 128)

Ethics of care

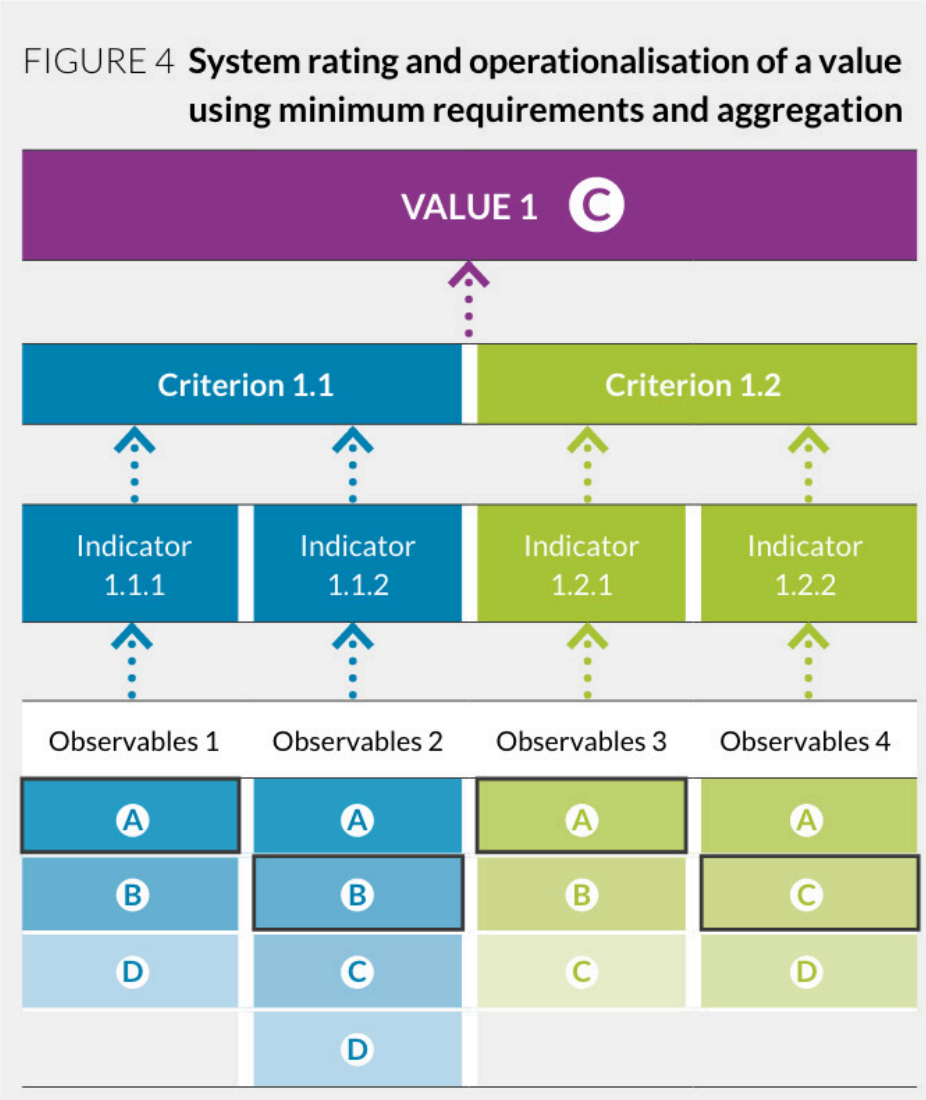
An ethical tradition implying that there is moral significance in the fundamental elements of relationships and dependencies in human life.

Creating relationships and taking care of each other is not just an abstract moral duty, but it is an essential quality of human and other beings in the world.

(Internet Encyclopedia of Philosophy, https://iep.utm.edu/care-ethics/, Gillian 1982; Noddings 1984)

From principles to practice

(AIEIG 2020; VDE Spec)



Consider context and design options

How does socio-technical context matter?

How to design around ethical *hard cases?*

(Crawford and Calo 2016)

The wide world of AI ethics

Discussion and exercise on:
Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9): 389–399. https://www.nature.com/articles/s42256-019-0088-2

Image source: https://annajobin.com/about

Sample of considered documents

- private companies (n = 19; 22.6%)
- governmental agencies (n = 18; 21.4%)
- academic and research institutions (n = 9; 10.7%)
- intergovernmental or supranational organizations (n = 8; 9.5%)
- non-profit organizations and professional associations / scientific societies (n = 7 each; 8.3% each)
- private sector alliances (n = 4; 4.8%)
- research alliances (n = 1; 1.2%)
- science foundations (n = 1; 1.2%)
- federations of worker unions (n = 1; 1.2%)
- political parties (n = 1; 1.2%)

Jobin et al. 2019

Identified principles / values

- Transparency
- Justice, fairness and equity
- Non-maleficence
- Responsibility and accountability
- Privacy
- Beneficience
- Freedom and autonomy
- Trust
- Sustainability
- Dignity
- Solidarity

Jobin et al. 2019

Exercise 2: group A

1. Map the universe of values in AI
(according to Jobin et al. 2019)

2. In your opinion, what ethical values / principles are missing in the global landscape of AI ethics?

Discuss in your group.

Exercise 2: group B

1. Map the universe of values in AI (according to Jobin et al. 2019)

2. In your opinion, how might generative AI change the weight given to particular values / principles?

Discuss in your group.

See you on 17.05.!

simon.hirsbrunner@uni-tuebingen.de

-- Illustrating pictures removed by lecturer for copyright reasons ---

Bibliography

The entire bibilography for the course can be found on Github here: https://github.com/simonsimson/responsible-data-science/blob/main/slides/Bibliography-of-the-entire-course.pdf

Image sources

Most sources are cited on the relevant slide. Slide 1: © Adobe Stock / kras99, slide 37: genewolf CC BY-ND 2.0