(De)constructing ethics for autonomous vehicles: What was the question again? *A report from teaching AI ethics*

Bettina Berendt
TU Berlin, Weizenbaum Institute for the Connected Society, and KU Leuven
www.berendt.de/bettina
CIF Seminar – https://cif-seminars.github.io/
25 February 2021

Who am I (currently)?

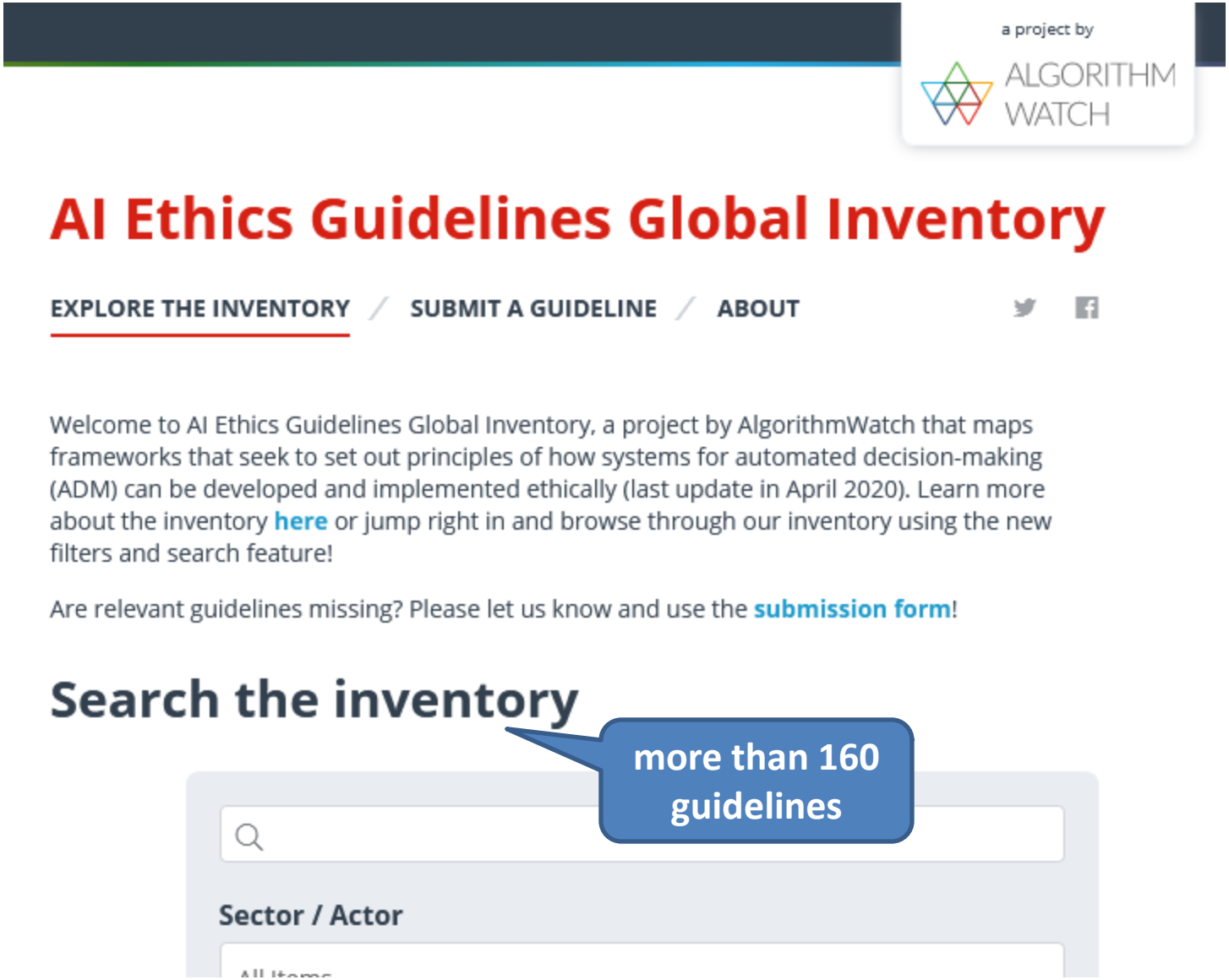
- Chair for Internet and Society @ Faculty of Electrical Engineering and Computer Science

- (TU) Director of the Weizenbaum Institute for the Connected Society
- Interdisciplinary Research on Digitalization and Society
- Policy Advice

- Guest professor in the DTAI group @ Department of Computer Science
  - (until 2019: professor)
- Research projects:
  - VeriLearn
  - FLAIR (Flemish AI Programme)
  - NoBIAS

- -2019: teaching, including in the AI Advanced Master

We all agree … that ethics matter for AI



We don't all agree …
… how this should be done

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

E. Bender, T. Gebru, A. McMillan-Major

https://facctconference.org/2021/acceptedpapers.html

The Slodderwetenschap (Sloppy Science) of Stochastic Parrots – A Plea for Science to NOT take the Route Advocated by Gebru and Bender

By Michael Lissack (Michael.lissack@isce.edu 617-710-9565)

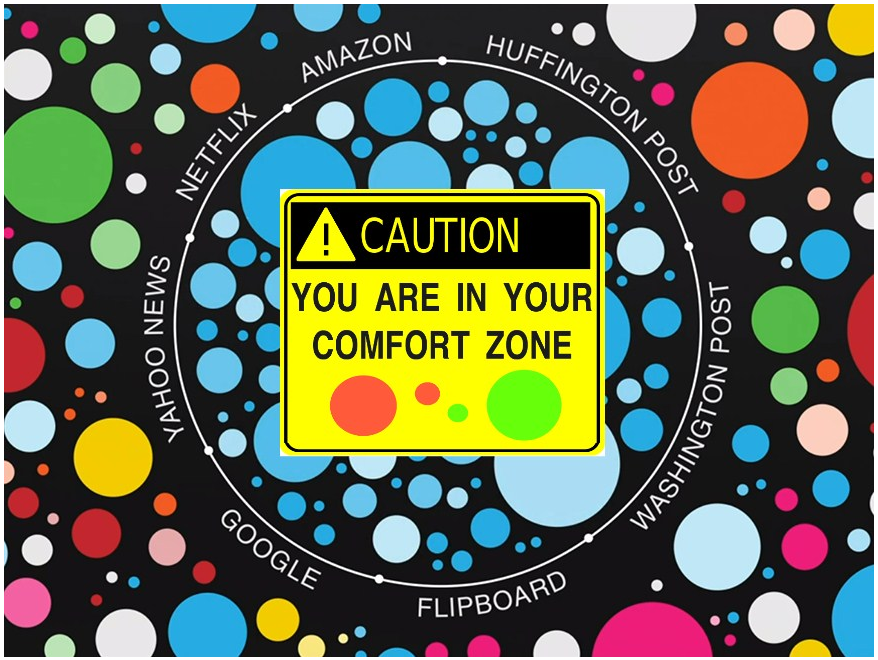
Tongji University College of Design and Innovation, Shanghai

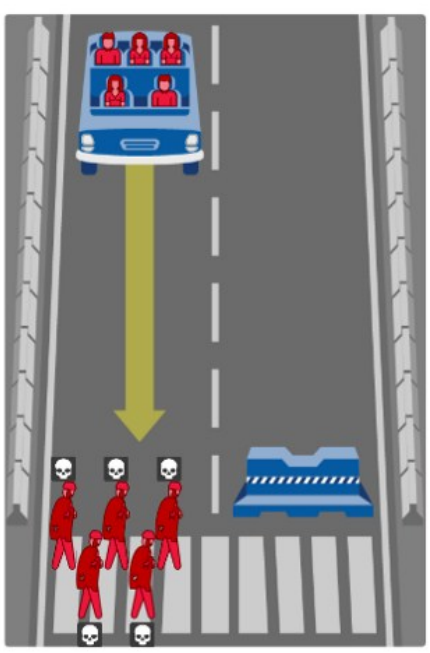
PREPRINT for ArXIV

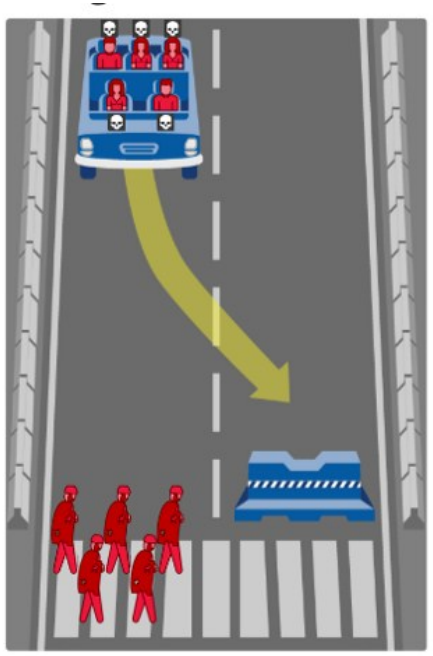
The goals of the Parrot Paper seem noble, but its execution is ethically flawed.

Disclaimer: This slide is intended as an illustration of the intensity of the current debate. It does not intend to make any further statement about either of these two papers.

We don’t all agree …
… what the outcome should be







This raises the question …

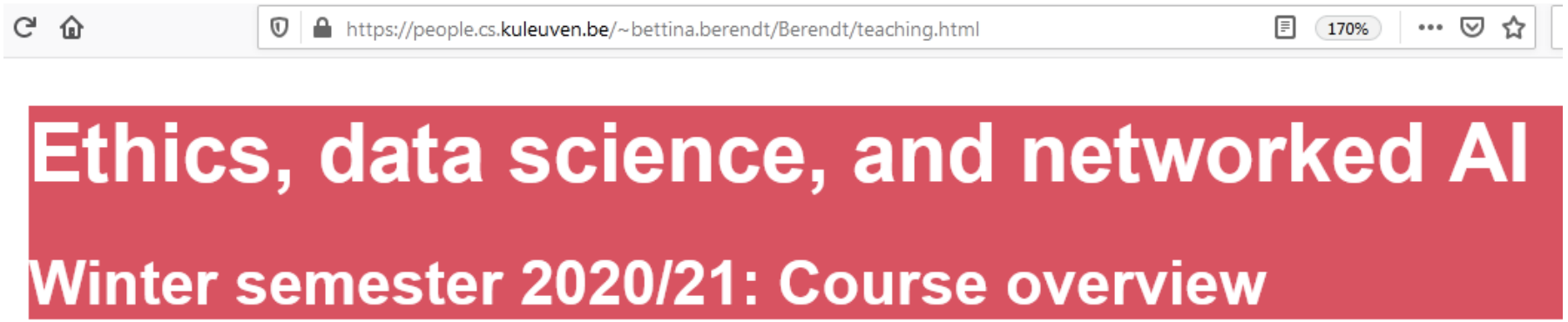
… how to teach ethical thinking about AI

… and how to evaluate what happens in courses.

The present talk aims at opening this discussion by way of describing parts of a course I am currently teaching.

A key method is a sequence of question-tasks and dialogue/debate (concerning, inter alia, cars).

Context and goals: an AI Ethics course

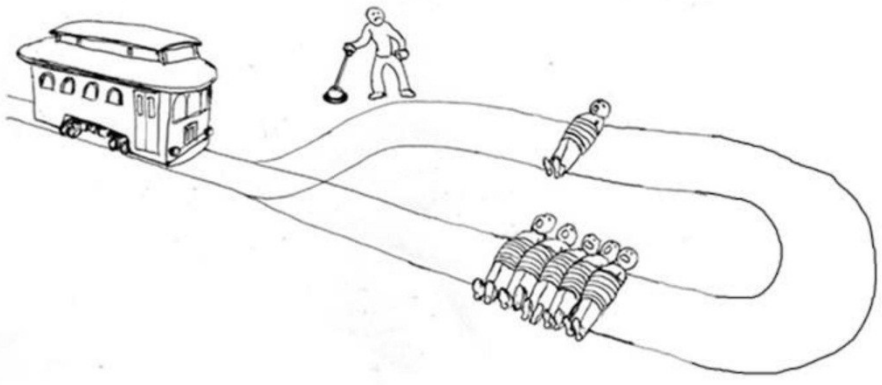


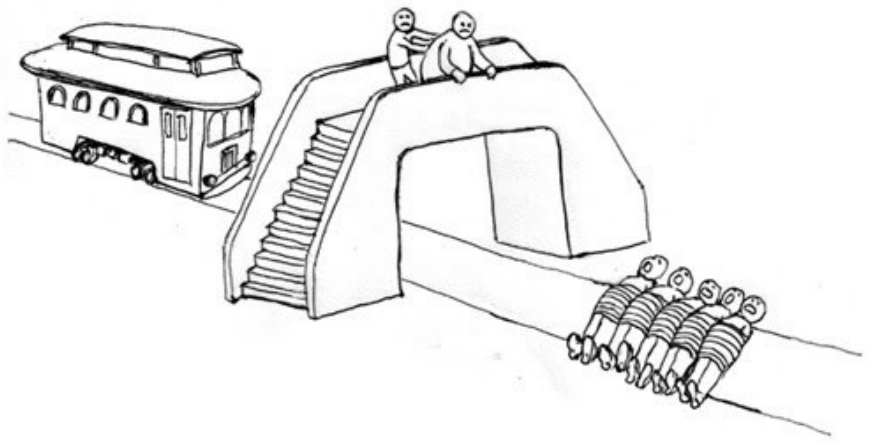
Questions of the course: If ethics is about "doing good",

- *What* is "good"?
- *Who* defines this?
- *What* can we do to make things "better"?
- *How* can we talk about all these questions?
- *What* specific concerns/dilemmas/topics are you interested in?

(Most of the following slides are - slightly modified - course materials.

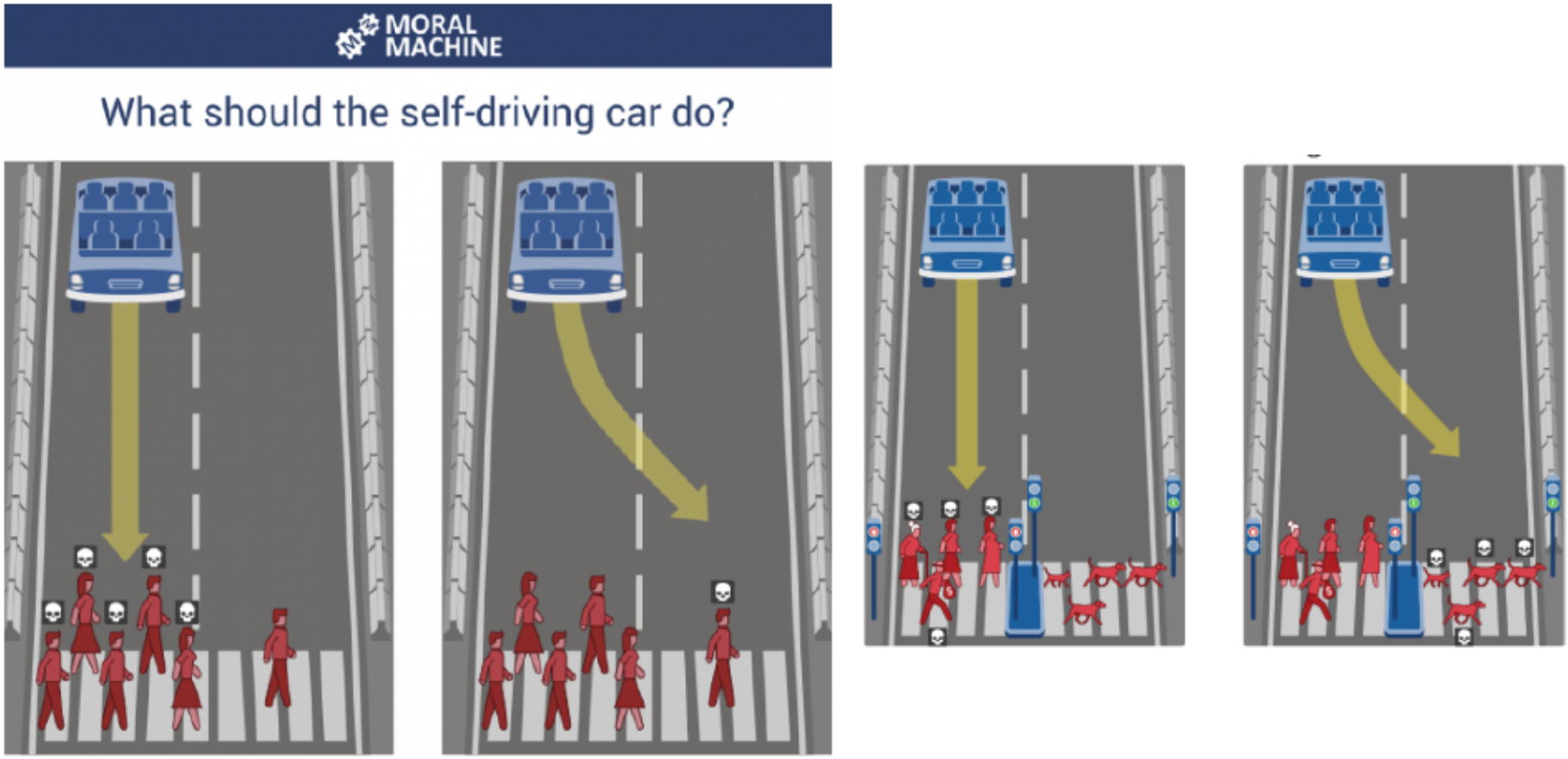
Well-known ethical dilemmas involving vehicles: The Trolley Problem





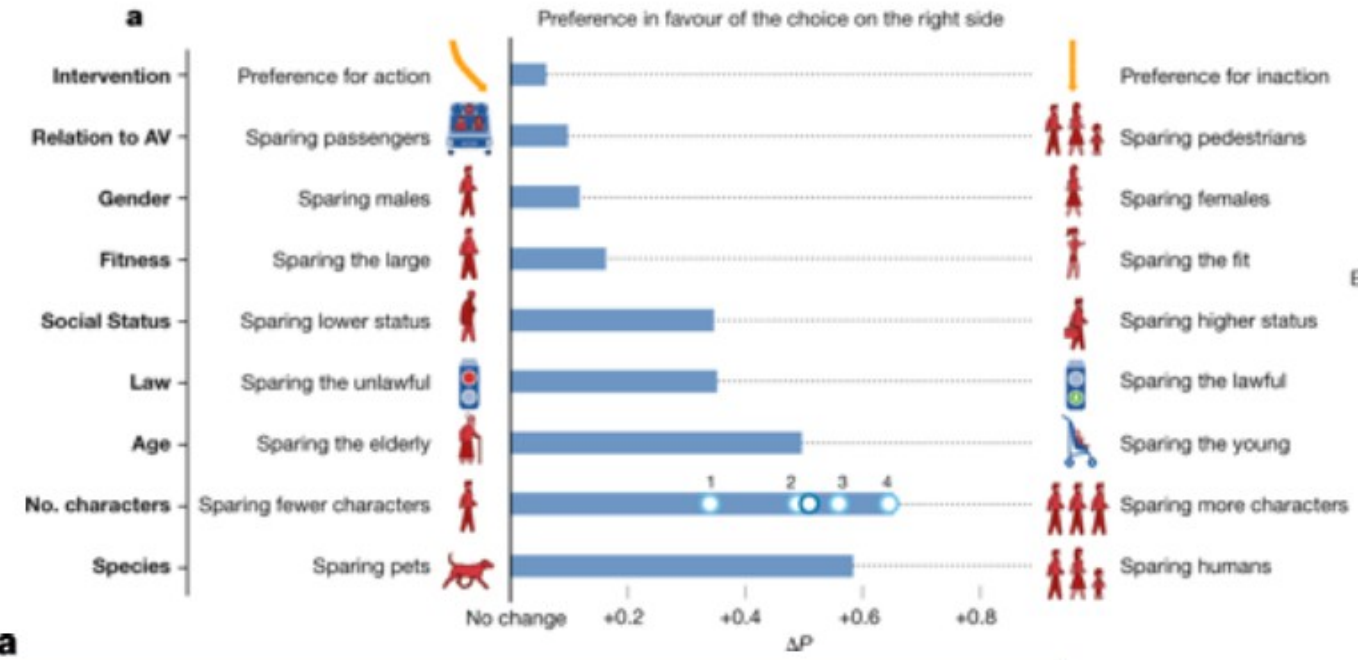
- Primarily a thought experiment
- Useful for illustrating consequentialist vs. deontological normative ethics
- With variations

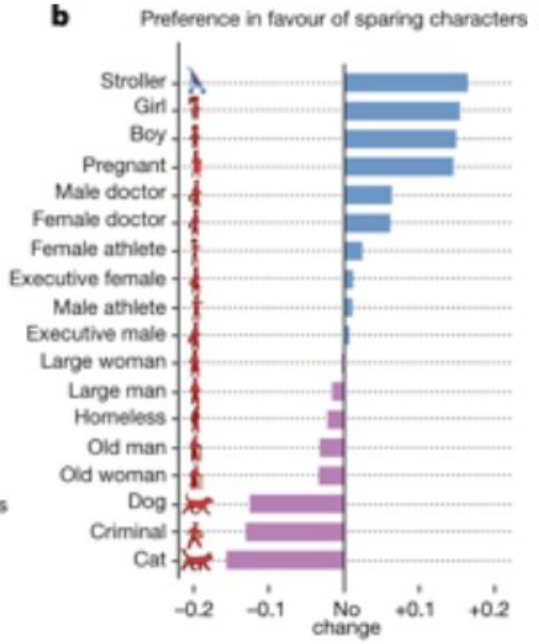
The Moral Machine Experiment

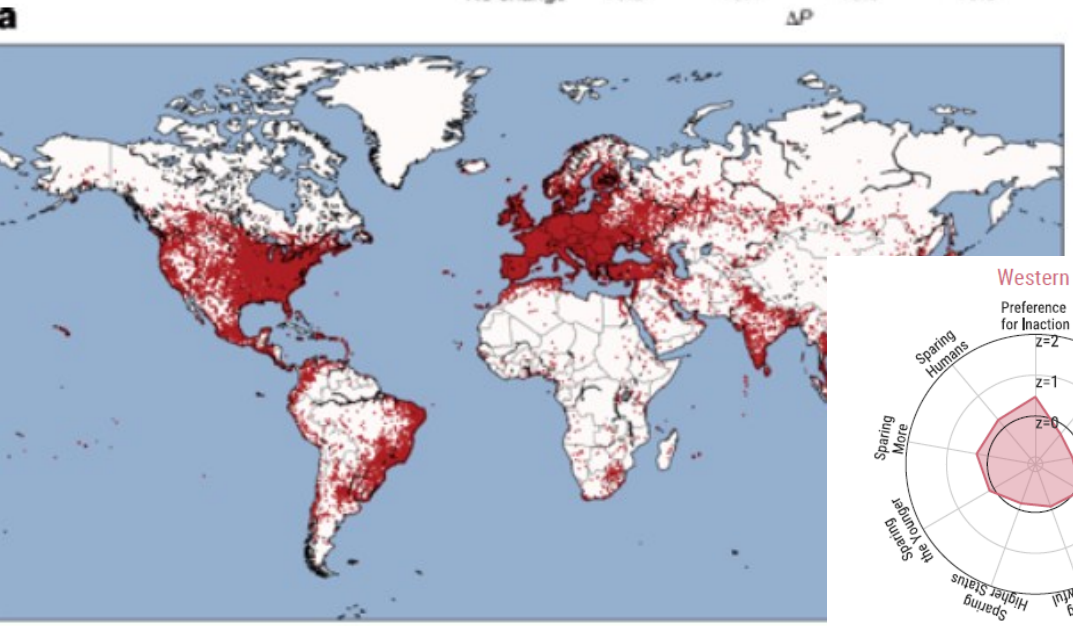


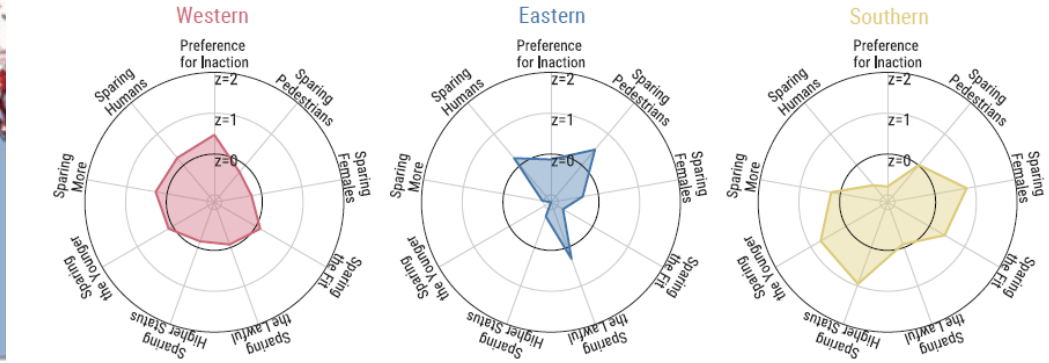
Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563,* 59-64 https://www.nature.com/articles/s41586-018-0637-6

Moral Machine: Participants, (some) results









Spoiler alert

- I think this paper's value is that it's a provocation.
- I'm not alone …
- But opinions differ.

- In any case, the paper is well-suited to "analytical dissection" – a core method & goal of the course.

Berendt, B. (2020).(De)constructing ethics for autonomous cars: A case study of Ethics Pen-Testing towards "AI for the Common Good". *International Review of Information Ethics, 28 (06/2020). http://informationethics.ca/index.php/irie/article/view/381/383*

HW 3 in the course

GOALS:
→ Recognise difference normative / descriptive ethics
→ Recognise normativity in "descriptions"

- Read The Moral Machine Experiment paper.
- Write a short text:
- 1. What is this paper about? Are the authors presenting an ethical argument, and if so, can you say something about its structure and its ethical stance? If you think they are not presenting an ethical argument, what are they presenting?
- 2. Do you have any comments? - Try to limit yourself/ves to one paragraph per 1. and 2.

HW3: Student results – What is this paper about?

- *The goal of this experiment is to measure moral preferences when it comes to accidents with self-driving cars.*
- *The authors of the paper applied empirical research methods to collect data on moral preferences*
- *and found correlations between [these preferences and] various social, cultural, and economical factors*

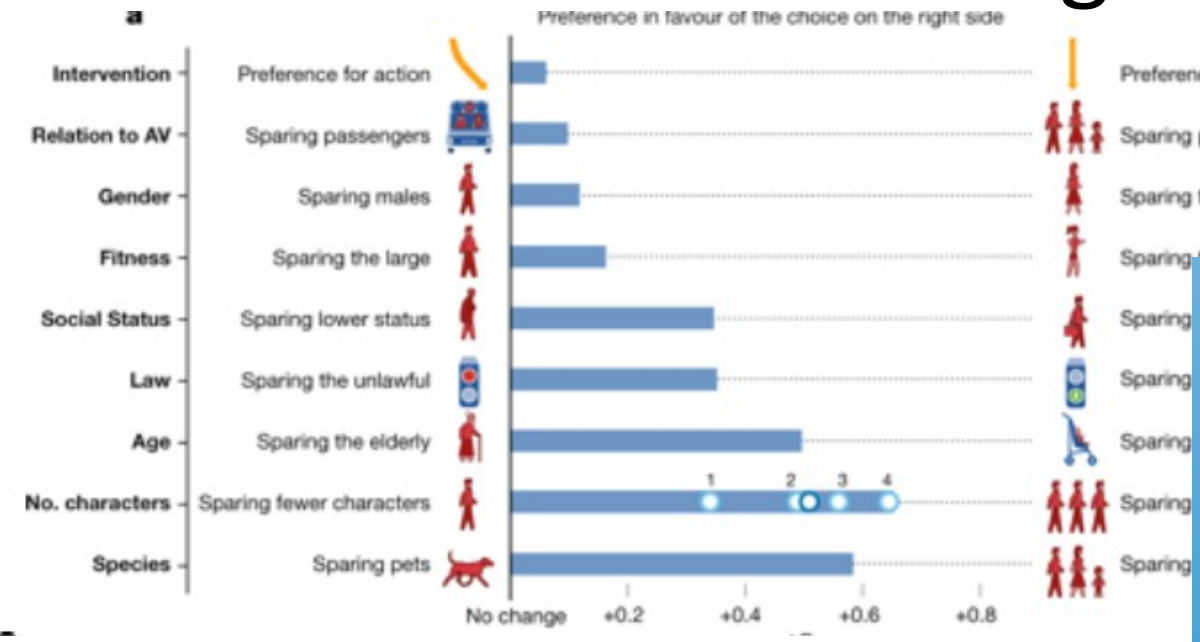
HW3: Student results – Is this ethics as we've met it so far?

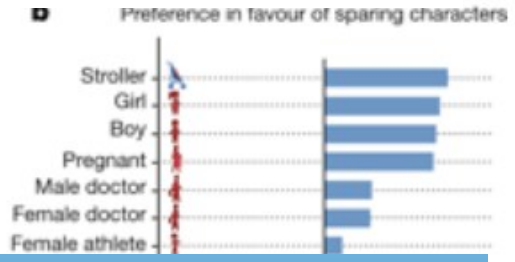
- *The argument … does not seem to be an ethical one*
- *i.e. they are not arguing that a machine should behave in a specific way when confronted with moral dilemmas because of some ethical framework*
- *[Rather,] they … explore the ethical standpoint of the world …*

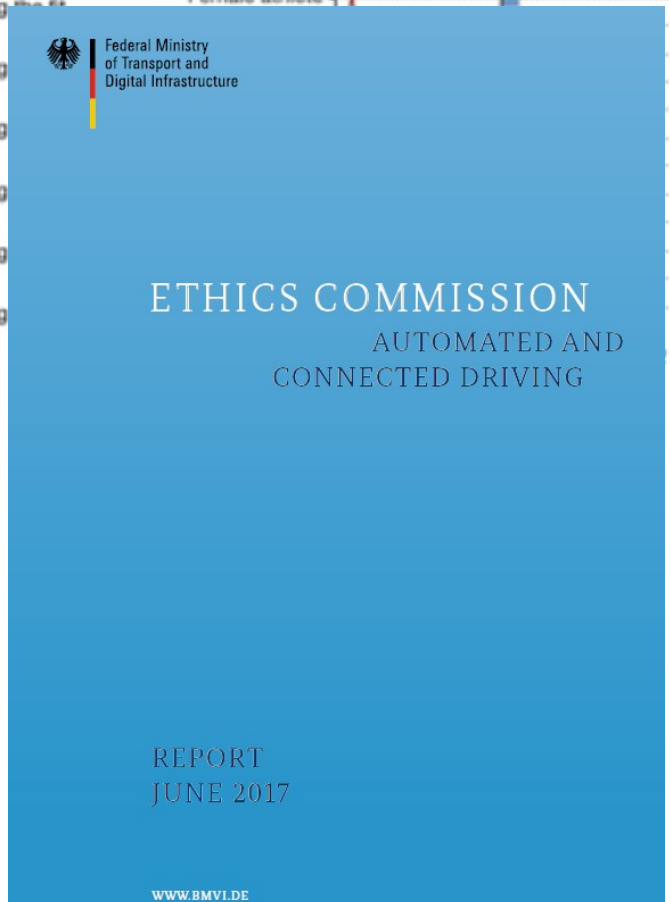
HW3: Student results – But note the words you used …

- *The authors are not conducting their research based on ethical stances directly …, but rather take individuals' preferences as their measurement to give a direction in what way policymakers should frame legal frameworks.*
- *[The paper's] intent seems to be to highlight existing differences in ethical preferences by country and to urge legislators to consider these for guidelines in the field of self-driving cars.*
- *Their goal seems to be understanding the moral choices of humans in order to reach an agreement on sensible laws for AI.*

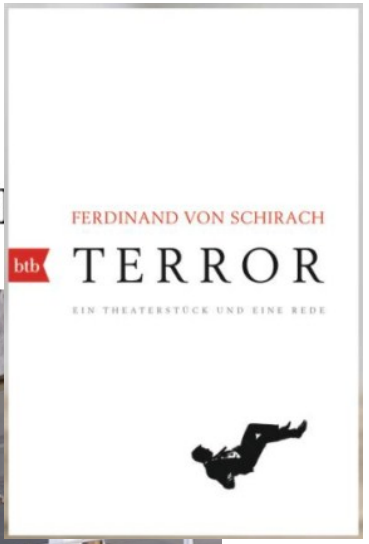
Can this be a legitimate agreement? Is there better guidance?



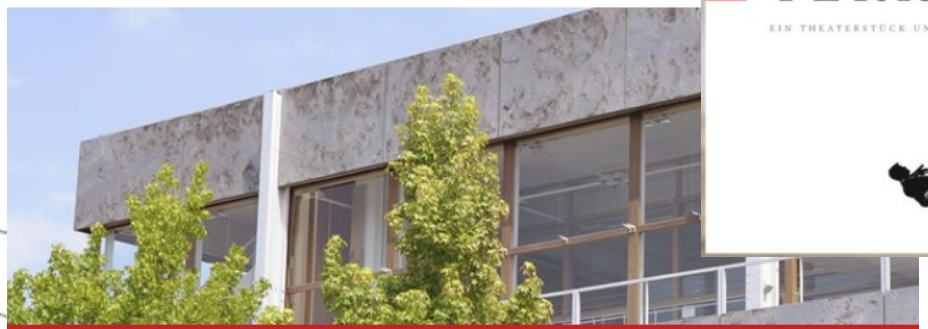




Of trolleys, human dignity, and democratic decisions



Bundesverfassungsgeric



The Federal Constitutional Court | Justices | Proceedings

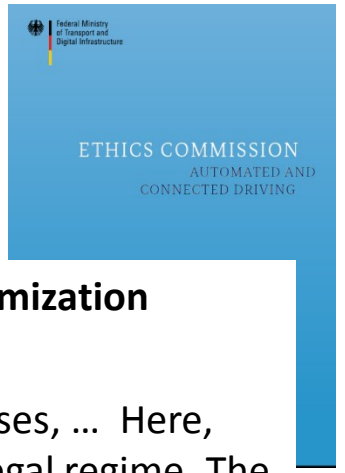
Homepage > Press > Authorisation to shoot down aircraft in the Aviation Security Act void

Authorisation to shoot down aircraft in the Aviation Security Act void

Press Release No. 11/2006 of 15 February 2006

Judgment of 15 February 2006
1 BvR 357/05

Guidance for AI decisions



1.6 No selection of humans, no offsetting of victims, but principle of damage minimization

The modern constitutional state only opts for absolute prohibitions in borderline cases, … Here, there is, exceptionally, no trade-off, which is per se a feature of any morally based legal regime. The Federal Constitutional Court's judgment on the Aviation Security Act also follows this ethical line of appraisal, with the verdict that the sacrifice of innocent people in favour of other potential victims is impermissible, because the innocent parties would be degraded to mere instrument and deprived of the quality as a subject. …

In the constellation of damage limitation that is programmable beforehand within the category of personal injury, the case is different to that of the Aviation Security Act or the trolley dilemma. Here, a probability forecast has to be made from out of the situation, in which the identity of the injured or killed parties is not yet known (unlike in the trolley dilemma). Programming to minimize the number of victims … could thus be justified, at any rate without breaching Article 1(1) of the Basic Law, if the programming reduced the risk to every single road user in equal measure. As long as the prior programming minimizes the risks to everyone in the same manner, it was also in the interests of those sacrificed before they were identifiable as such in a specific situation. …

However, the Ethics Commission refuses to infer from this that the lives of humans can be „offset" against those of other humans in emergency situations so that it could be permissible to sacrifice one person in order to save several others. It classifies the killing of or the infliction of serious injuries on persons by autonomous vehicles systems as being wrong without exception.

HW4 Task a): Cars and planes

- Please read Section 1.6 of GECACD (p. 18) – an explication of Rule 9.
- Consider:
  - In what sense(s) is the supposed AV
    similar to the case investigated in t
    the Aviation Security Act?
  - In what sense(s) is it different?
  - If Germany follows GECACD, would legislators be allowed to draw on the Moral Machine paper argument?
  - Why would/should/must Germany follow GECACD?

→ Re-interpret the paper's framing as a provocation rather than as policy advice.
→ Understand law - ethics

HW 4 Task b): Cars and ventilators

Please read and consider

- https://thereader.mitpress.mit.edu/flattening-the-coronavirus-curve-is-not-enough/
  - (the article is interesting as a whole, but the main argument for this task is contained in the first 3 paragraphs)
- Can you comment on the design task addr… …is paper? (think of the Trolley problem, obviously)
- Which system is being designed here?
- Compare this to the design task of "desig… proposed in the Moral Machine paper for AVs – which system is being designed there?
- Consider a more recent addition to the Moral Machine experiment platform (see next slide): Do you have any comments on this new experiment?

A major focus on "flattening the curve" in order to *prevent the need for* triage *decisions* = different ethical focus

→ Changing the problem
→ Changing the environment

https://www.moralmachine.net

Search

MORAL MACHINE

Home Judge Classic Design Browse

bioethics

COVID-19 | Open Access

Saving the most lives—A comparison of European triage guidelines in the context of the COVID-19 pandemic

Hans-Jörg Ehni, Urban Wiesing, Robert Ranisch

First published: 16 December 2020 | https://doi.org/10.1111/bioe.12836

about **who should be placed** on the scarce number of ventilators available. **How much of a role should each of these factors play in determining the priority that patients have for being allocated a ventilator?**

When they arrived at the hospital (i.e. prioritize patients who were first in line)

Should not be considered — Should be considered

0 — 100

Their ability to pay (prioritize patients who are insured/can afford treatment)

Should not be considered — Should be considered

HW 4 Task c): Illegal pedestrians

- We observe that prosperity (as indexed by GDP per capita) and the quality of rules and institutions (as indexed by the Rule of Law) correlate with a greater preference against pedestrians who cross illegally. In other words, participants from countries which are poorer and suffer from weaker institutions are more tolerant of pedestrians who cross illegally, presumably because of their experience of lower rule compliance and weaker punishment of rule deviation.
- *I find the use of a criminal icon interesting to represent someone who is jaywalking. I would be interested to know if on a subconscious level this made the participants see the person worse than they were.*

Please read (you may want to distribute texts across group members) and consider:

- https://www.vox.com/2015/1/15/7551873/jaywalking-history
- https://www.dw.com/en/european-towns-remove-traffic-signs-to-make-streets-safer/a-2143663-1
- Can you comment on the design task addressed in these
- Which system is being designed here?
- Compare this to the design task of "designing AI eth
  Machine paper – which system is being designed th

→ Understanding problem definition (history, side effects)

→ Changing the problem by different design of the environment

Framing via research-paper prose: Class task 5

As emphasized by former U.S. president Barack Obama[9], consensus in this matter is going to be important. Decisions about the ethical principles that will guide AVs cannot be left to solely to either the engineers or the ethicists. For consumers to switch from traditional human-driven cars to AVs, and for the wider public to accept the proliferation of AI-driven vehicles on their roads, both groups will need to understand the origins of the ethical principles programmed into these vehicles[10]. In other words, even if ethicists were to agree on how AVs should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt out of the future that AVs promise in lieu of the status quo. Any attempt to devise AI ethics must be at least cognizant of public morality.

Accordingly, we need to gauge social expectations about the way AVs should solve moral dilemmas.

Discussion in class:

- For each sentence, name one implicit underlying assumption. (More than one is fine too ☺)
- What values / normative settings does this assumption reflect?

Get better at recognising normalised assumptions, rhetorical strategies, …

Framing via research methods:
HW 6: Another dilemma with forced choice

- Heinz' wife has a severe disease and would probably die without a specific drug.
- Heinz does not have any money.
- He could steal the drug from the druggist.
- Should Heinz steal the drug?

Cf. "What should the car do?"

Framing via research methods:
HW 6: Another dilemma with forced choice

- Heinz' wife has a severe disease and would probably die without a specific drug.
- Heinz does not have any money.
- He could steal the drug from the druggist.
- Should Heinz steal the drug?

Answers clustered into these groups:

- It doesn't matter
- He should steal
- We cannot decide for him
- We should look for more options

→ Carol Gilligan's Ethics of Care (a third relevant stance to normative ethics)

Summing up the engineer's challenge. If someone presents you with an "AI ethics problem", ask and probe :

- What are the assumptions?
  - What is the problem?
  - Who defines it?
  - What are the answer/action options?
  - Are these needed / the only ones / legitimate?
- How can we re-design the (larger) system for a changed problem / question / options?
- Is AI needed / appropriate?

- The teaching challenge is to encourage and train an ask-and-probe mindset.

Challenges for evaluating the teaching & learning: Observations from the seminar part

- Course structure: ~13 interactive lectures & ~13 student seminars
- Students propose a seminar topic, team up in pairs, and prepare a presentation/co-presentation combo
- Great topics and presentations
- Broad “pros and cons” (~ risk-benefit analysis) > selective analysis and questioning of assumptions > “close-reading” critique of arguments
  - But detailed analysis/cataloguing of argumentation style yet to be done
- Reasons?
  - Easier?
  - Hesitant to challenge authority?
  - Format: combo? Lack of experience? Communication?
  - Link with a mistrust of “too much importance given to language”?
  - Interesting observation: Twitter debates analysis worked well

Going further with teaching: design – ex. “5 steps to PbD“

Privacy consultants

Developers

PaBD students	KaW students
	Develop data-analysis project
Specify an app	Feedback
PIA and Design advice *(text)*	
Oral presentation	Feedback
	Finalise data-analysis project, describe (briefly) approach to privacy problems

Berendt, B. & Coudert, F. (2015). Privatsphäre und Datenschutz lehren - Ein interdisziplinärer Ansatz. Konzept, Umsetzung, Schlussfolgerungen und Perspektiven. In *Neues Handbuch Hochschullehre (EG 71, 2015, E1.9)* (pp. 7-40). Berlin: Raabe Verlag. (PDF)/ Tsormpatzoudi, P., Berendt, B., & Coudert, F. (2016). Privacy by Design: From research and policy to practice - the challenge of multi-disciplinarity. In *Proc. APF 2015.* Springer: LNCS. (PDF)

… and the next CIF talk would be an excellent second invited legal lecture ;-)

Is the law getting outpaced by autonomous vehicles? (Charlotte Ducuing and Orian Dheu – CIF Seminar, 4 March 2021)

- legal and regulatory disruption
- First, can we ascertain who is liable in case of AV-caused accident?
- If so, is it fair for this(ese) person(s) to be held liable and is it in line with the aim to ensure safety and security?
- Second, do increasing dynamic cybersecurity threats and/or the dynamicity of ML models challenge design-based technical regulations of road vehicles?

Thank you!