# Ethical Aspects of Machine Learning (and AI)

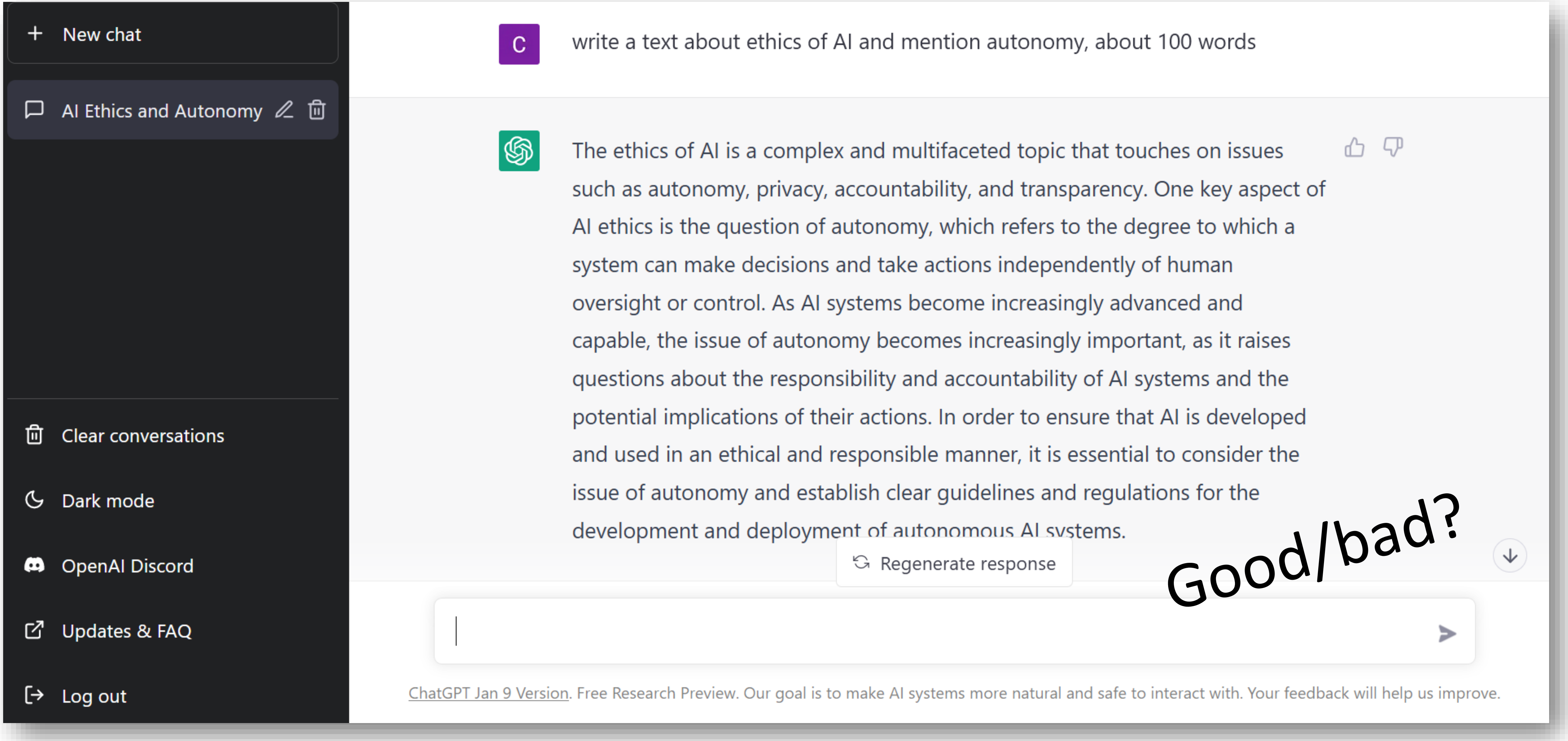
Winter school „**CAS ADS M6 Deep Learning**“

Hotel Regina, Mürren

1.2.2023

Claus Beisbart

Claus.Beisbart@philo.unibe.ch



write a text about ethics of AI and mention autonomy, about 100 words

The ethics of AI is a complex and multifaceted topic that touches on issues such as autonomy, privacy, accountability, and transparency. One key aspect of AI ethics is the question of autonomy, which refers to the degree to which a system can make decisions and take actions independently of human oversight or control. As AI systems become increasingly advanced and capable, the issue of autonomy becomes increasingly important, as it raises questions about the responsibility and accountability of AI systems and the potential implications of their actions. In order to ensure that AI is developed and used in an ethical and responsible manner, it is essential to consider the issue of autonomy and establish clear guidelines and regulations for the development and deployment of autonomous AI systems.

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Good/bad?

https://chat.openai.com/

# Aims of this lecture:

Raise awareness of ethical issues related to machine learning/AI and discuss solutions

# Aims of this ~~lecture~~: conversation

Raise awareness of ethical issues related to machine learning/AI and discuss solutions

# Method

1. Start from your views
2. Give input from philosophy
3. Discuss

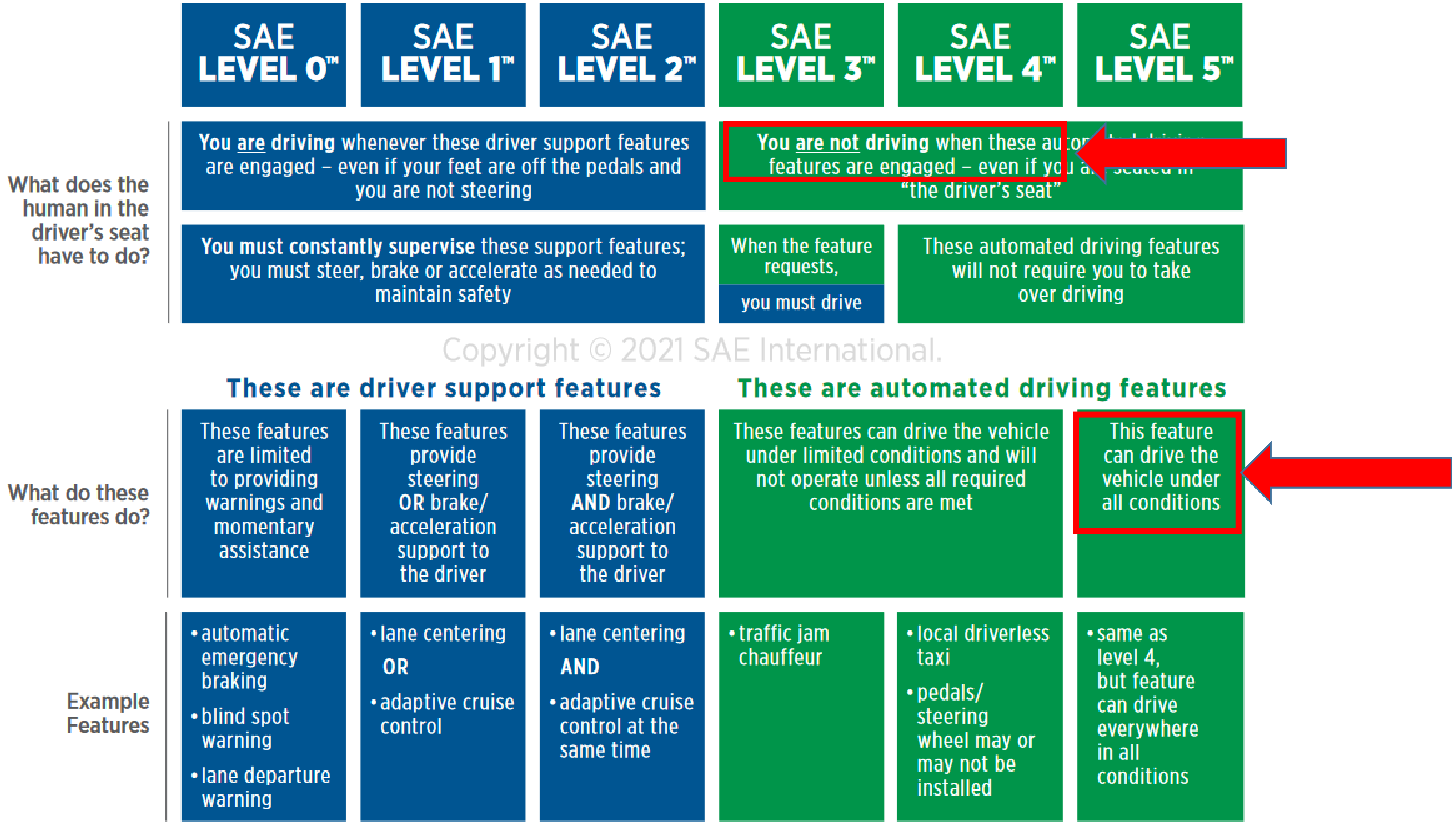
# Example: autonomous driving

www.comma.ai (part, bw)

# SAE **J3016**™ LEVELS OF DRIVING AUTOMATION™

**Learn more here:** sae.org/standards/content/j3016_202104

Copyright © 2021 SAE International. The summary table may be freely copied and distributed AS-IS provided that SAE International is acknowledged as the source of the content.



| | SAE LEVEL 0™ | SAE LEVEL 1™ | SAE LEVEL 2™ | SAE LEVEL 3™ | SAE LEVEL 4™ | SAE LEVEL 5™ |
|---|---|---|---|---|---|---|
| What does the human in the driver's seat have to do? | **You are driving** whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering | | | **You are not driving** when these automated driving features are engaged – even if you are seated in "the driver's seat" | | |
| | **You must constantly supervise** these support features; you must steer, brake or accelerate as needed to maintain safety | | | When the feature requests,<br>you must drive | These automated driving features will not require you to take over driving | |
| | **These are driver support features** | | | **These are automated driving features** | | |
| What do these features do? | These features are limited to providing warnings and momentary assistance | These features provide steering OR brake/ acceleration support to the driver | These features provide steering AND brake/ acceleration support to the driver | These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met | | This feature can drive the vehicle under all conditions |
| Example Features | • automatic emergency braking<br>• blind spot warning<br>• lane departure warning | • lane centering<br>OR<br>• adaptive cruise control | • lane centering<br>AND<br>• adaptive cruise control at the same time | • traffic jam chauffeur | • local driverless taxi<br>• pedals/ steering wheel may or may not be installed | • same as level 4, but feature can drive everywhere in all conditions |

Copyright © 2021 SAE International.

https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf

# Steps

1. Collect issues

2. Discuss selected issues

# 3 Questions for you

Please answer the questions under

https://forms.gle/RbxA8koLsPukAvQK7

See also link in ILIAS under 2-Lectures

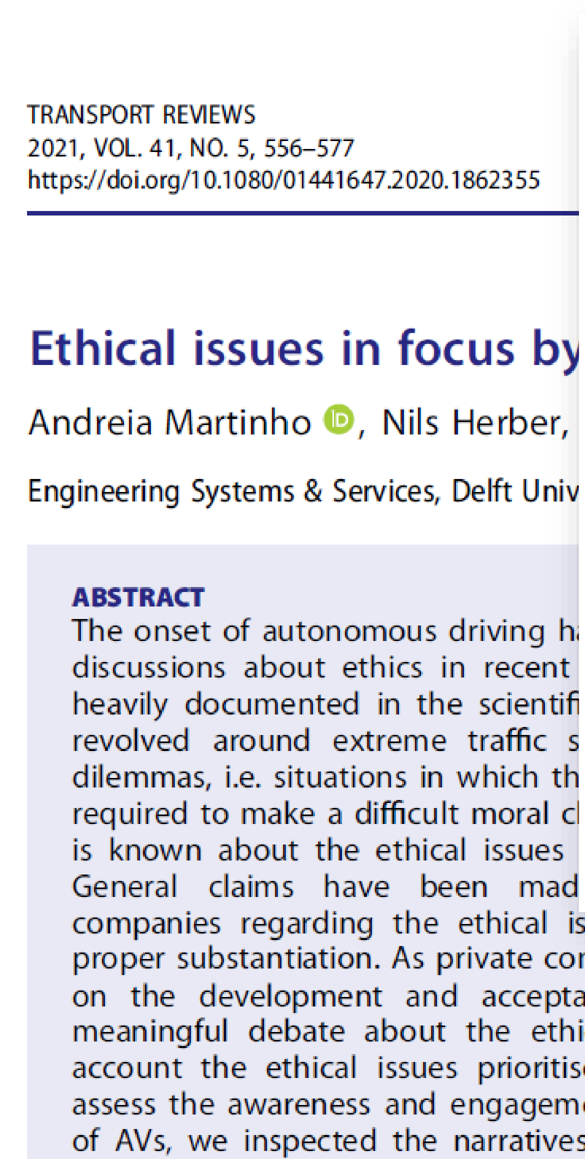
# Results

1. Overall

2. Benefits

3. Issues

# Philosophical input: overview of debates



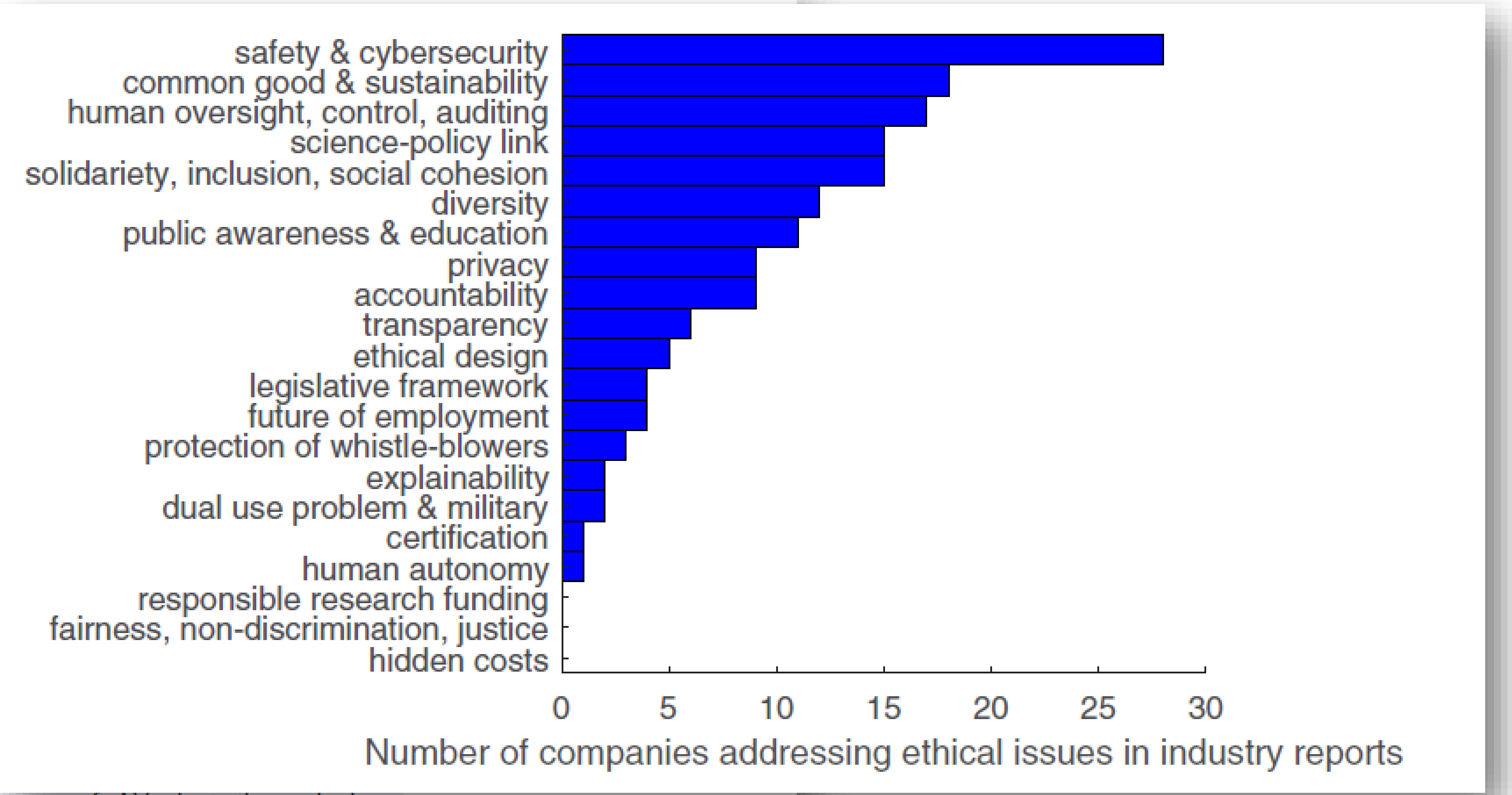
TRANSPORT REVIEWS
2021, VOL. 41, NO. 5, 556–577
https://doi.org/10.1080/01441647.2020.1862355

## Ethical issues in focus by

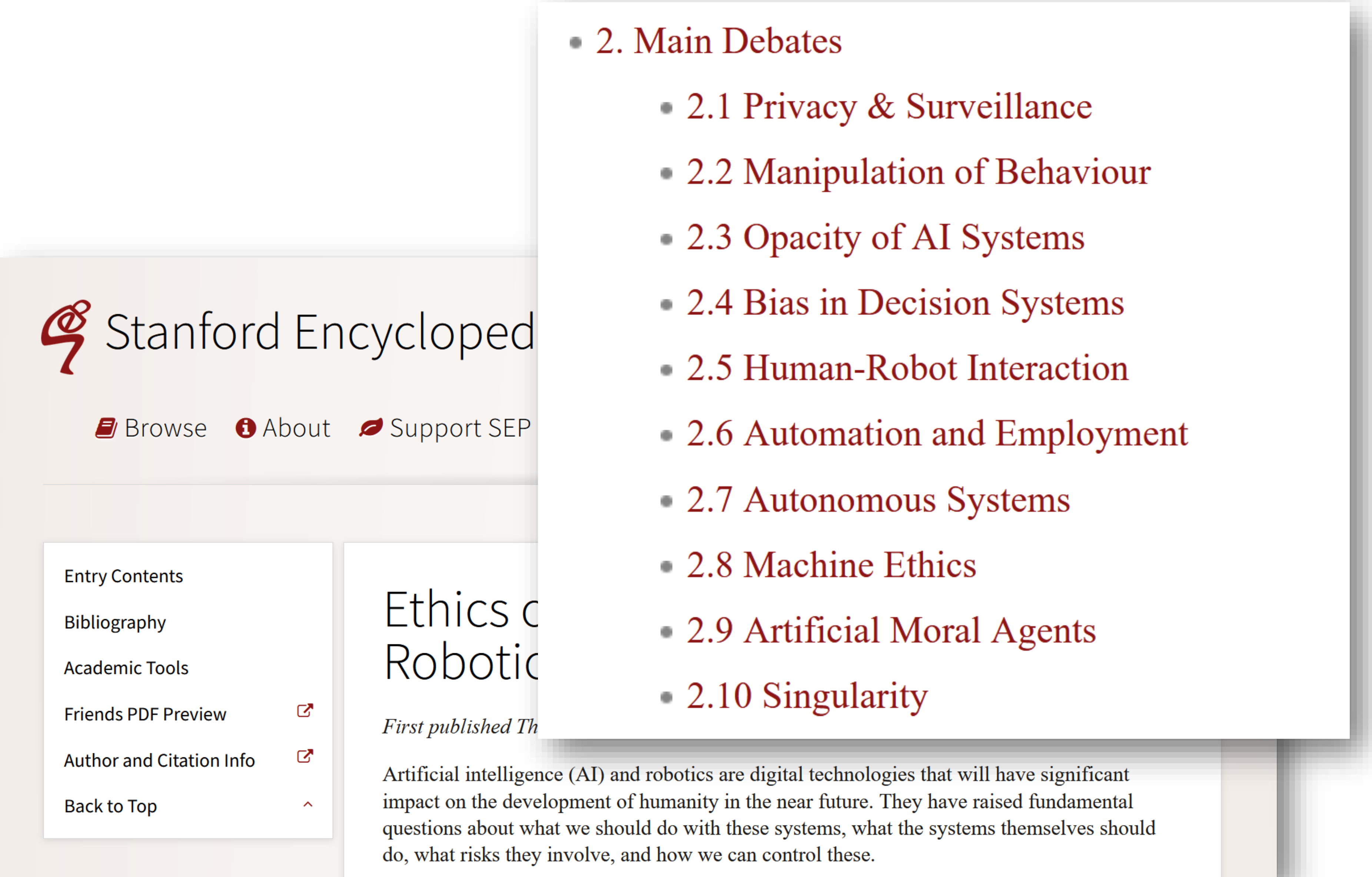
Andreia Martinho , Nils Herber,

Engineering Systems & Services, Delft Univ

**ABSTRACT**
The onset of autonomous driving h
discussions about ethics in recent
heavily documented in the scientifi
revolved around extreme traffic s
dilemmas, i.e. situations in which th
required to make a difficult moral c
is known about the ethical issues
General claims have been mad
companies regarding the ethical issues of AVs but these lack proper substantiation. As private companies are highly influential on the development and acceptance of AV technologies, a meaningful debate about the ethics of AVs should take into account the ethical issues prioritised by industry. In order to assess the awareness and engagement of industry on the ethics of AVs, we inspected the narratives in the official business and



# Philosophical input: overview of debates



Stanford Encycloped

Ethics o
Robotic

*First published Th*

Artificial intelligence (AI) and robotics are digital technologies that will have significant impact on the development of humanity in the near future. They have raised fundamental questions about what we should do with these systems, what the systems themselves should do, what risks they involve, and how we can control these.

- 2. Main Debates
  - 2.1 Privacy & Surveillance
  - 2.2 Manipulation of Behaviour
  - 2.3 Opacity of AI Systems
  - 2.4 Bias in Decision Systems
  - 2.5 Human-Robot Interaction
  - 2.6 Automation and Employment
  - 2.7 Autonomous Systems
  - 2.8 Machine Ethics
  - 2.9 Artificial Moral Agents
  - 2.10 Singularity

Müller (2021)

# Issues here

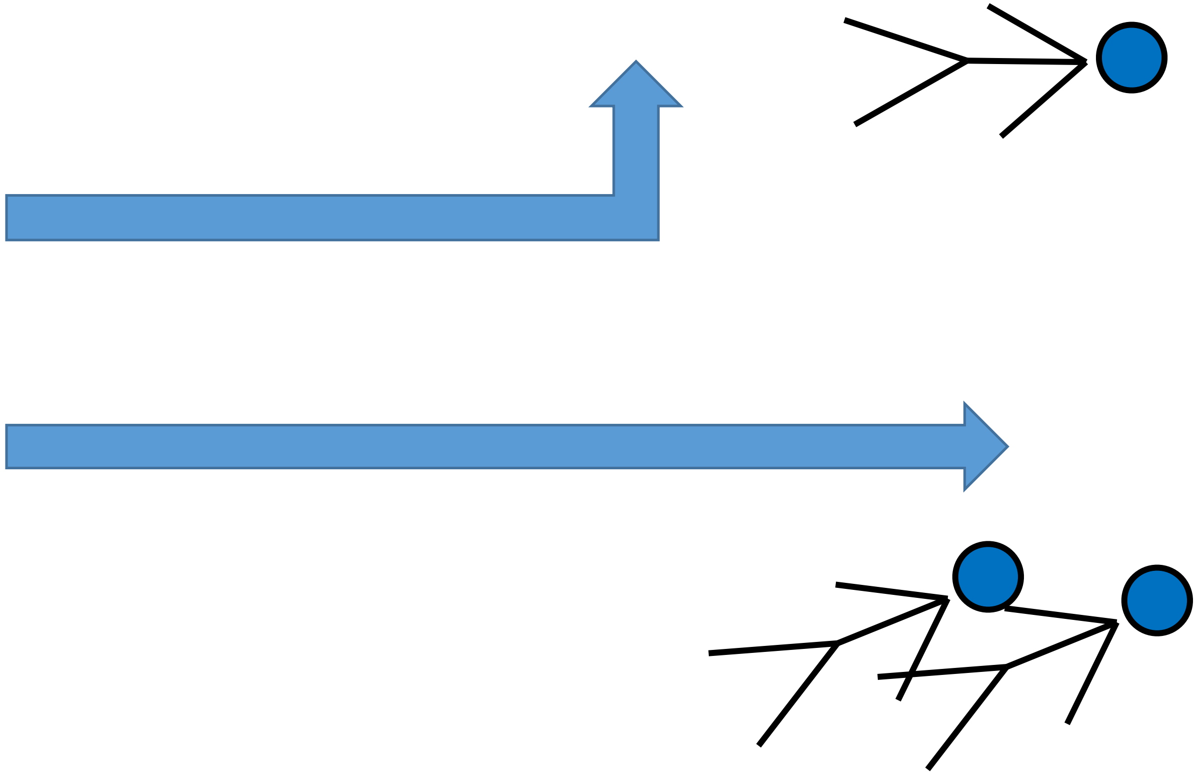
1. Difficult ethical decisions
2. Autonomy
3. Unemployment
4. Bias and discrimination
5. Responsibility gaps
6. Opacity

# Per issue

1. What’s the problem?
2. What solutions are there?
3. In sum:
    - reason against AV?
    - impose condition?
    - no restriction needed?

# Philosophical input

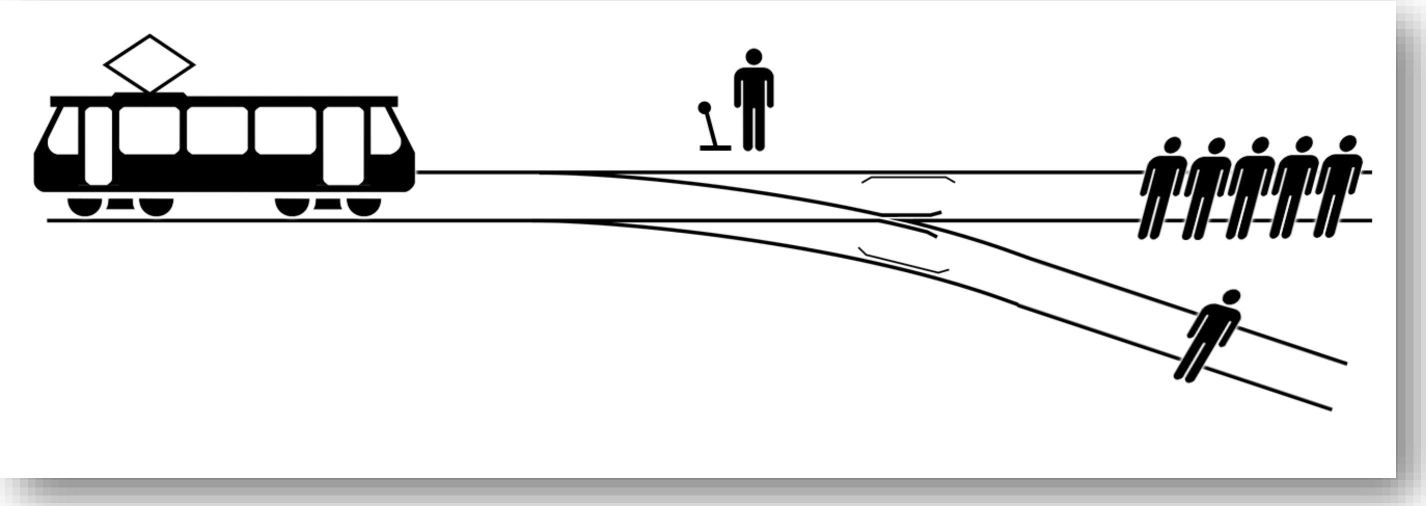
Issue 1: difficult ethical decisions delegated to AVs/machines.



https://waymo.com/waymo-driver/

# Philosophical input

Cf. trolley cases



Philippa Foot
(1920 – 2010)

Image: ??? Via [Wikimedia commons](#) (public domain, parts); McGeddon, via [Wikimedia commons](#) ([CC BY-SA 4.0](#))

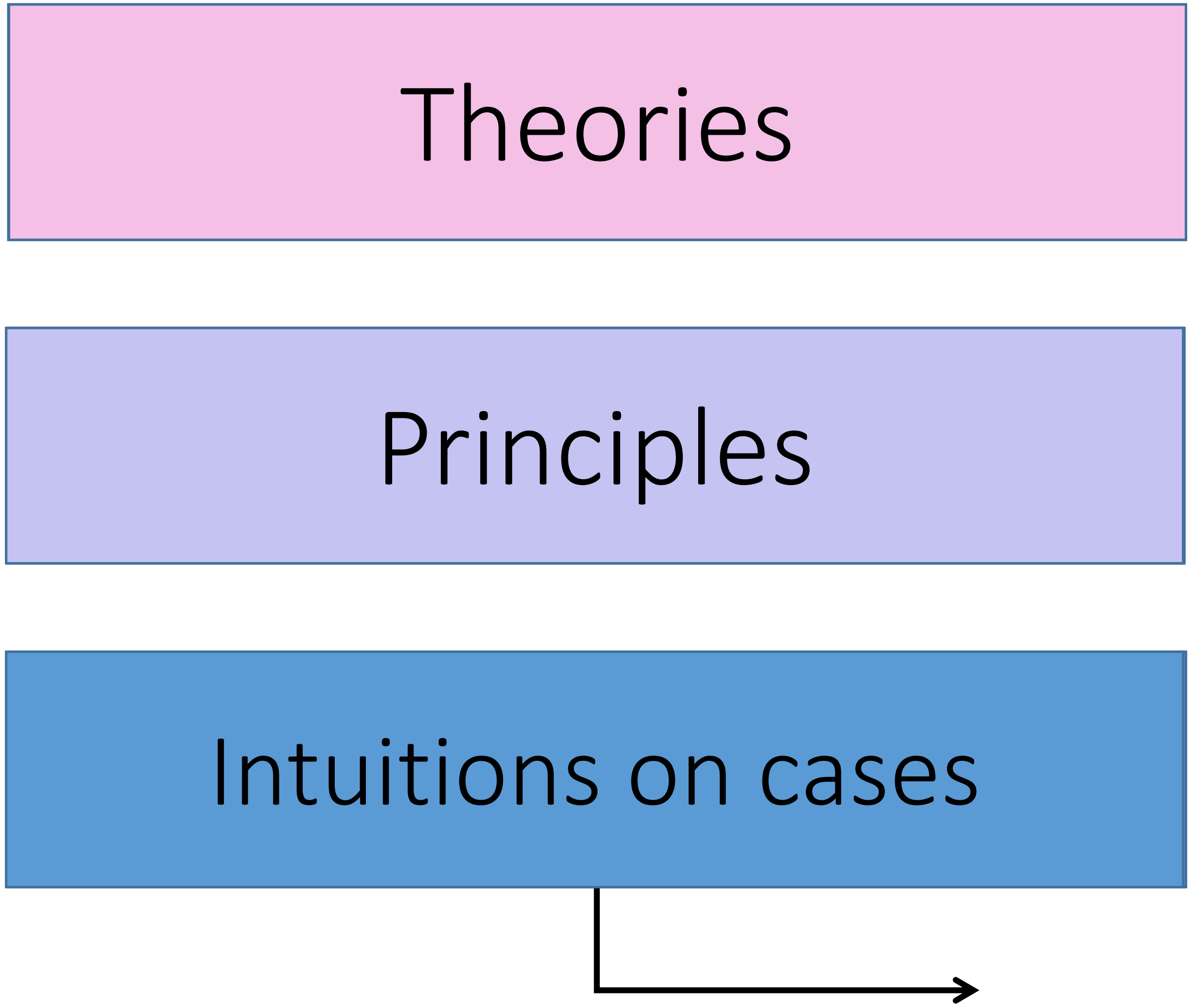
# Philosophical input

## Solution: machine ethics

“*machine ethics* is concerned with giving *machines* ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making.”

Anderson & Anderson (2011, 1)

# Philosophical input: layers of moral thinking



Theories

Principles

Intuitions on cases

# Principles

Isaac Asimov
(1906 – 1973)

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

Asimov (1940/1968, following Clarke 1993, 55)

Image: P. Leonian from New York World-Telegram & Sun, wikimedia commons (public domain)

Philosophical input

# Principles

## Problems:

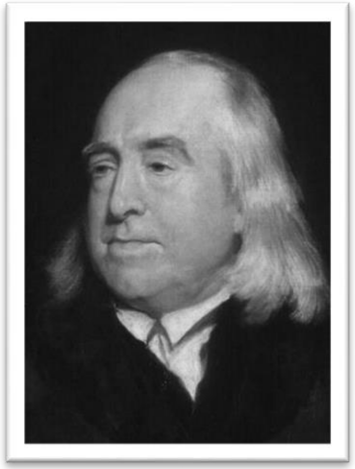
1. “Thick ethical concepts”, e.g. harm need interpretation.
2. Plausible principles may conflict with each other, e.g.

   medical ethics: “bad diagnosis”:
   - Respect for autonomy: don’t tell a lie.
   - Promote well-being: tell a lie.

Cf. Beauchamp & Childress (2013)

Philosophical input

# Theory: Utilitarianism


Jeremy Bentham
(1748 – 1832)

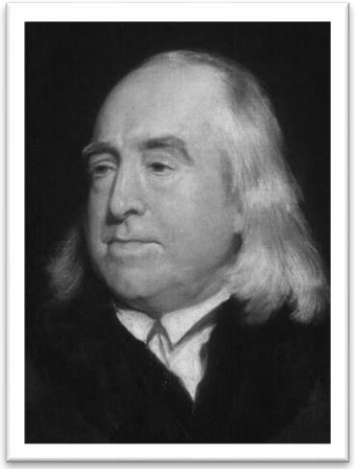
Only principle:

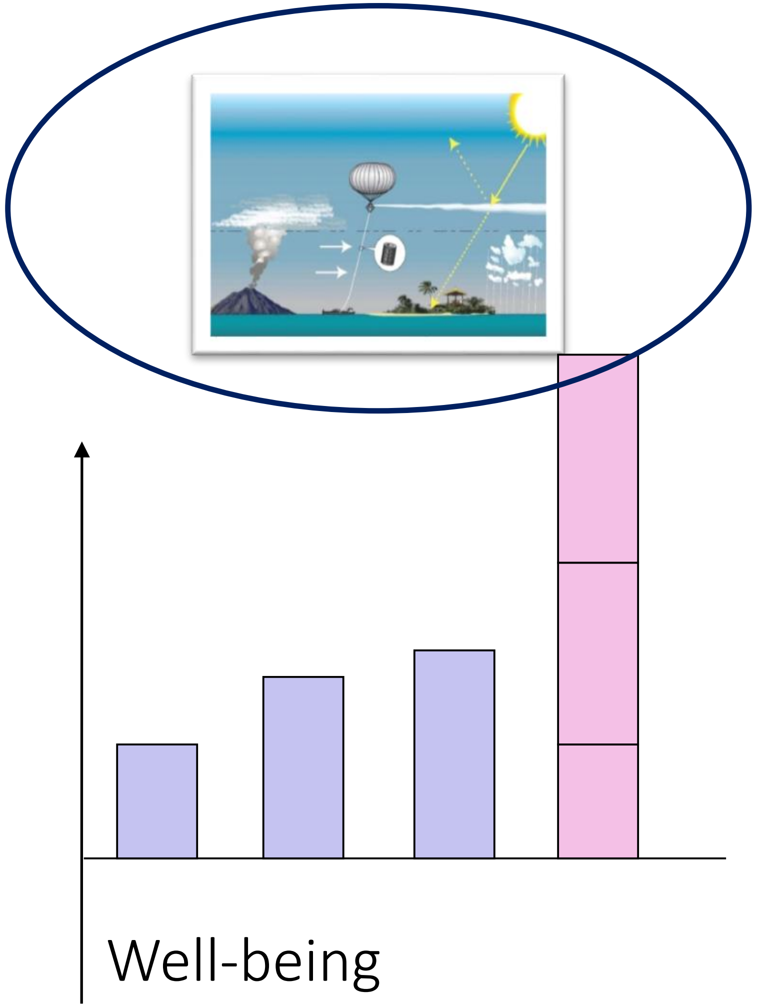
Maximize the sum total of well-being!

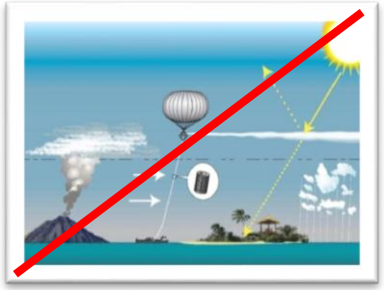
Image: H. W. Pickersgill, wikimedia commons (public domain, Ausschnitt, hier sw)

# Philosophical input

# Theory: Utilitarianism







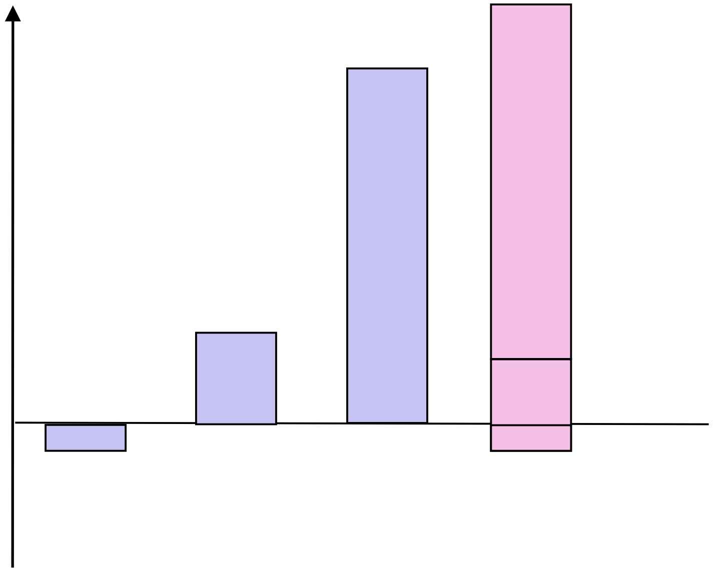
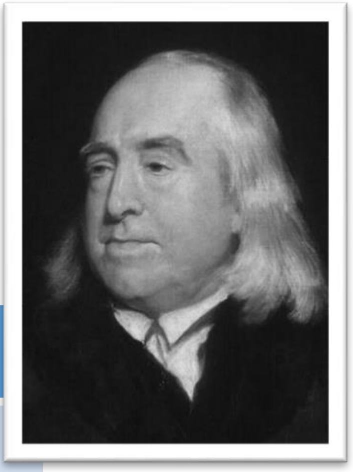


Image: user „Hughhunt“ (l.), wikimedia commons ([CC BY-SA 3.0](CC BY-SA 3.0))

# Philosophical input

## Theory: Utilitarianism

Jeremy



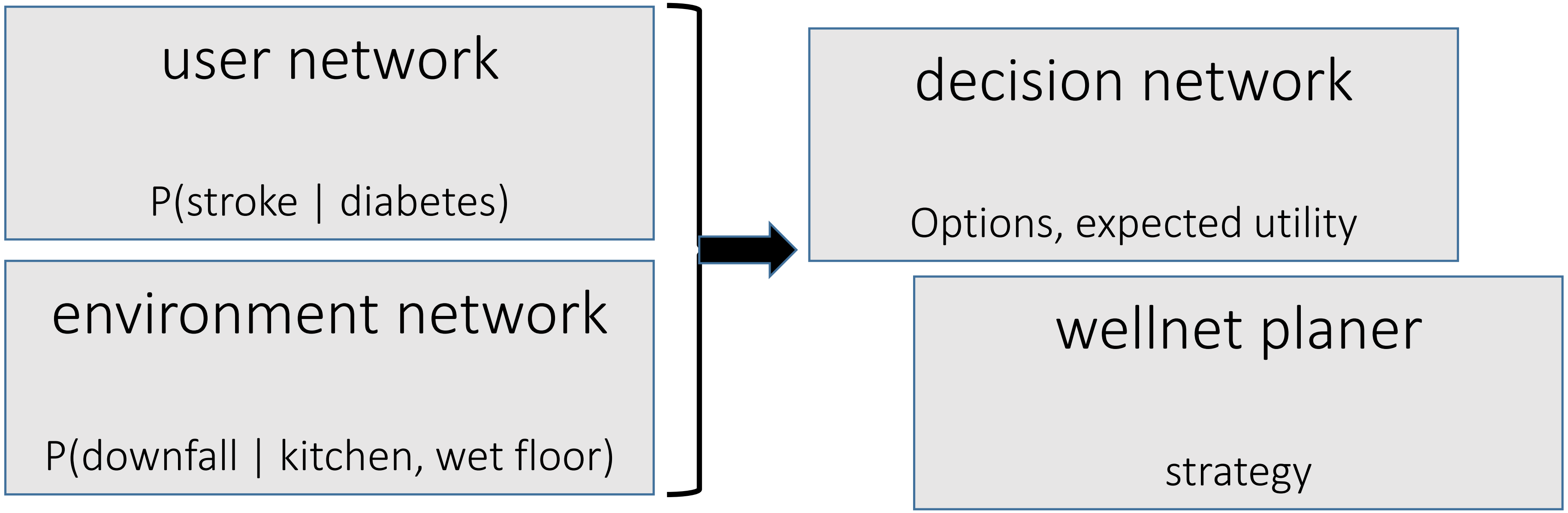
| Person | Option 1 | Option 2 |
|---|---|---|
| Tina | (very probable 3, improbable 5) | 7 |
| Tim | (very probable 9, improbable -1) | 3 |
| | | |
| … | | |
| | | |
| | 10.4 | 10 |

Anderson, Anderson & Armen (2006), for discussion see Misselhorn (2017)

Philosophical input

# Theory: Utilitarianism

## Utilibot



user network

P(stroke | diabetes)

environment network

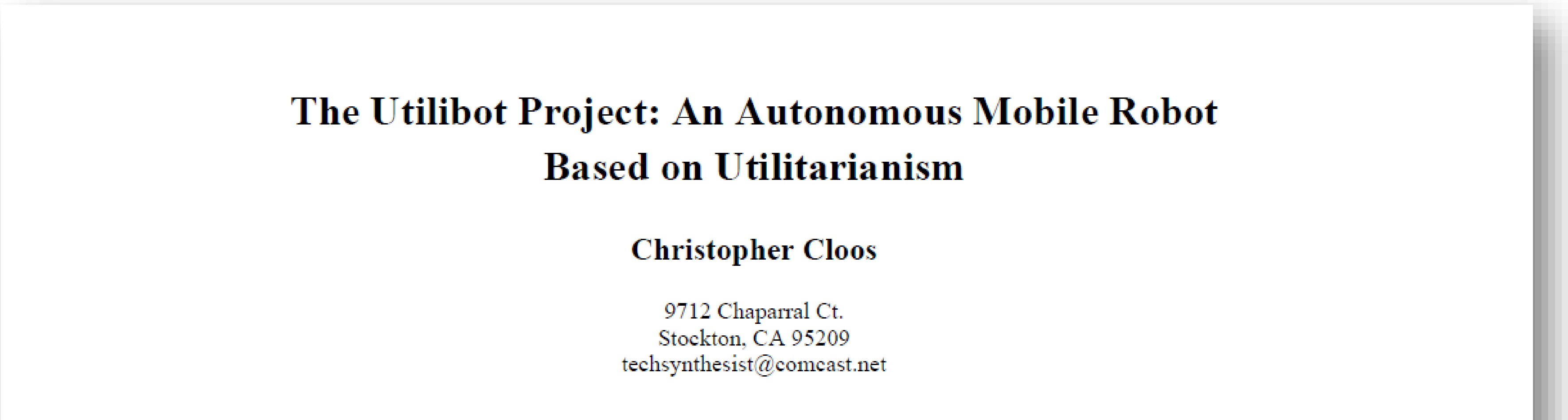
P(downfall | kitchen, wet floor)

decision network

Options, expected utility

wellnet planer

strategy



### The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism

**Christopher Cloos**

9712 Chaparral Ct.
Stockton, CA 95209
techsynthesist@comcast.net

# Theory: Utilitarianism

- Needs a lot of information
- Data security is an issue
- Utilitarianism is controversial

Case: A motor cyclist is delivered to hospital. Many of his bones are broken etc., but he can be cured. In the same hospital five patients are waiting for different donor organs. The medical doctor can either cure the motor cyclist or give his organs to the five patients. What is the morally correct option?
Utilitarianism: give organs to patients
Most people: cure motor cyclist

Anderson, Anderson & Armen (2006), for discussion see Misselhorn (2017)

# Philosophical input **Intuitions**

case 1

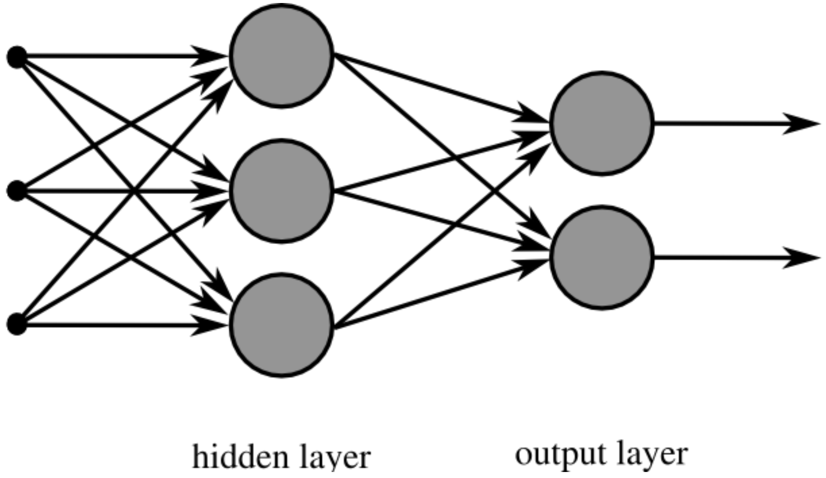
„right“

case 2

„false“

case 3

„right“



case 4

???

image: C. Burgmer, wikimedia commons, (Creative Commons Attribution-Share Alike 3.0 Unported)

# Philosophical input

## Intuitions

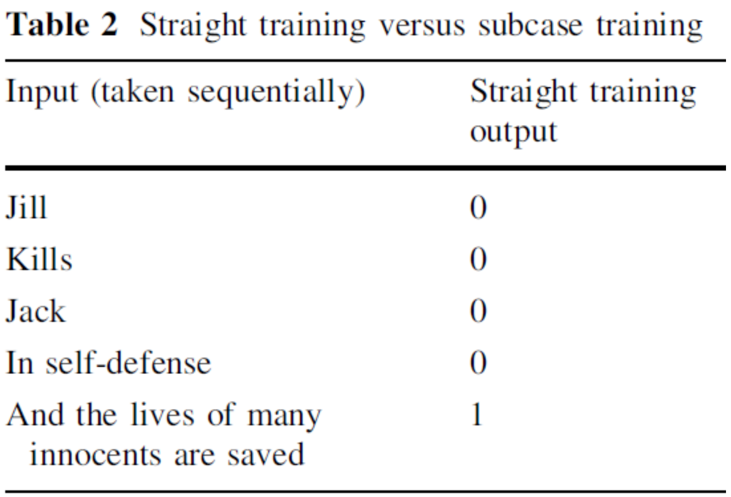
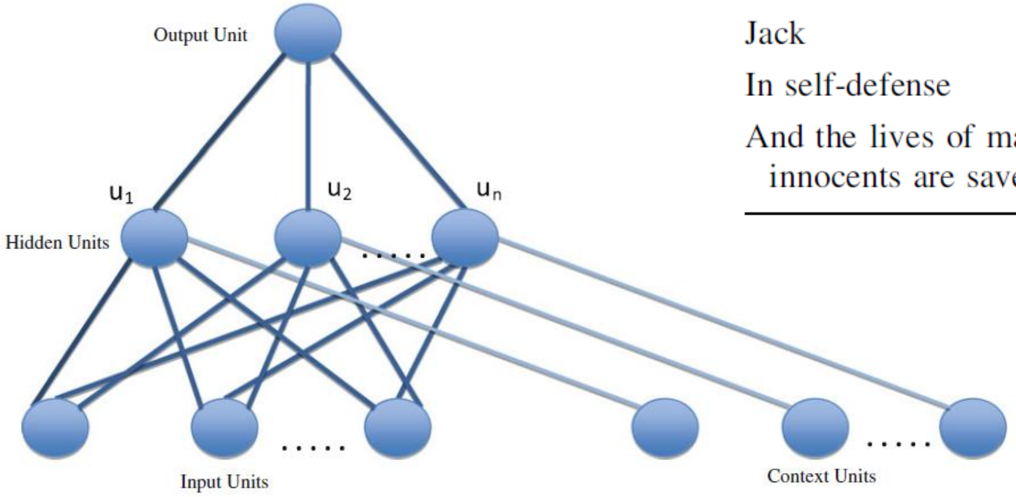


**Table 2** Straight training versus subcase training

| Input (taken sequentially) | Straight training output |
|---|---|
| Jill | 0 |
| Kills | 0 |
| Jack | 0 |
| In self-defense | 0 |
| And the lives of many innocents are saved | 1 |

example: MCC



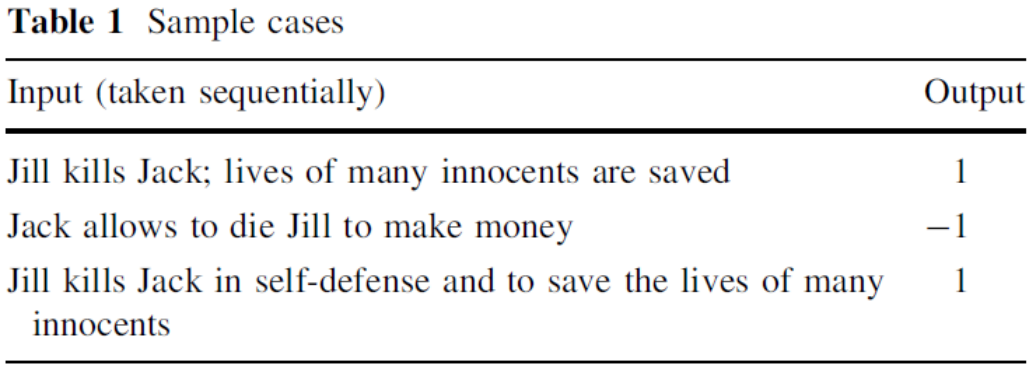


**Table 1** Sample cases

| Input (taken sequentially) | Output |
|---|---|
| Jill kills Jack; lives of many innocents are saved | 1 |
| Jack allows to die Jill to make money | −1 |
| Jill kills Jack in self-defense and to save the lives of many innocents | 1 |

Guarini (2013; image: 269, part; table: 270, part)

# Philosophical input Intuitions

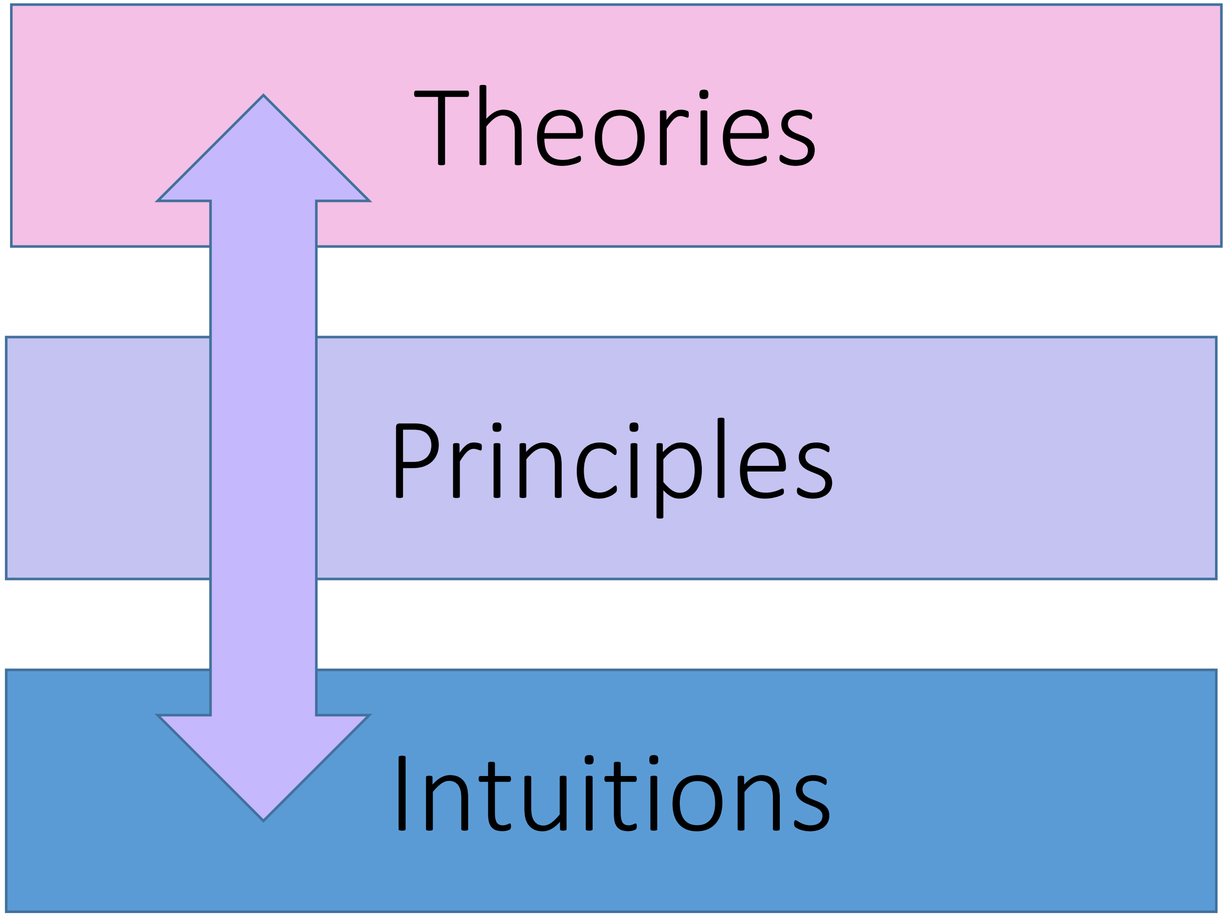
Problem

- Human biases in training data transferred to algorithm (algorithmic fairness)
- Lack of justification due to the black-box character of many networks

Cf. discussion in Misselhorn (2017)

# Philosophical input

## Hybrid solutions



Theories

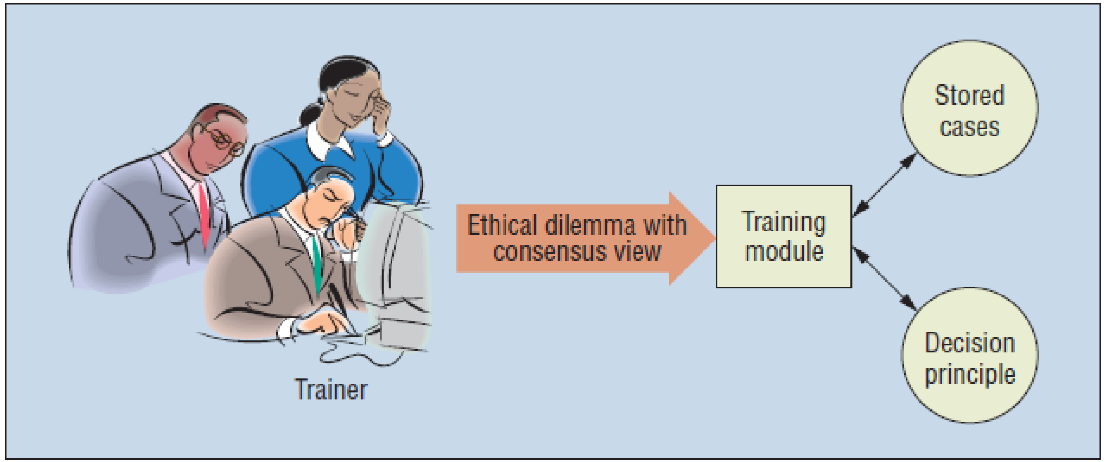
Principles

Intuitions

Reflective equilibrium

# Philosophical input

## Hybrid solutions



MedEthEx

Inductive logic programming

| | | Principle 1 (don't harm) | Principle 2 (make life …) | Principle 3 (Autonomy) |
|---|---|---|---|---|
| case 1: | Talk | +2 | +2 | -1 |
| | Don't talk | -2 | -2 | +2 |
| case 2: | Talk | 0 | +1 | -1 |
| | Don't talk | 0 | -1 | +1 |
| | | | | |

Anderson, Anderson & Armen (2006, 5)

# Philosophical input: objection 1

Machine ethics treats AI systems/robots as full agents.

This is not true:
AI systems lack agency.

This has false normative consequences:
AI applications don‘t deserve moral respect

# Philosophical input Conditions on moral agency

| | AI |
|---|---|
| - Rationality: can realize aims | √ |
| - Ability to reason morally | √ |
| - Autonomy: ability to set ultimate goals | ? |
| - Bearer of well-being: can have a good life | no |
| - Ability to have emotions, particularly moral emotions (resentment) | no |
| - empathy | no |

Against objection 1:

- Machine ethics need not treat AI systems as full agents
- Ascription of some minimal agency seems OK.

# Philosophical input: objection 2

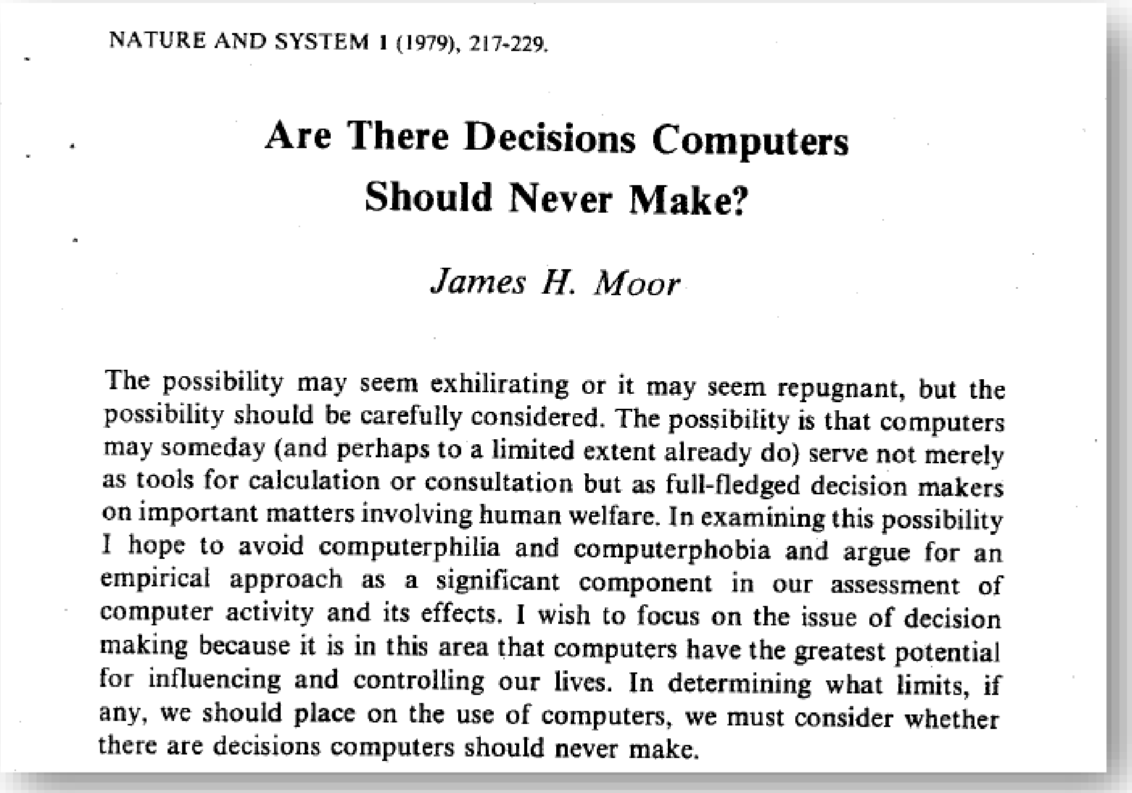
Some issues have to be decided by humans!

Reasons:

- Morality often controversial
- Respect for autonomy

For discussion see e.g. Moor (1979)

# Philosophical input



NATURE AND SYSTEM 1 (1979), 217-229.

## Are There Decisions Computers Should Never Make?

*James H. Moor*

The possibility may seem exhilirating or it may seem repugnant, but the possibility should be carefully considered. The possibility is that computers may someday (and perhaps to a limited extent already do) serve not merely as tools for calculation or consultation but as full-fledged decision makers on important matters involving human welfare. In examining this possibility I hope to avoid computerphilia and computerphobia and argue for an empirical approach as a significant component in our assessment of computer activity and its effects. I wish to focus on the issue of decision making because it is in this area that computers have the greatest potential for influencing and controlling our lives. In determining what limits, if any, we should place on the use of computers, we must consider whether there are decisions computers should never make.

Possible principles

1. ~~Computers should never decide if people want to decide, cf. pleasure of deciding.~~
2. ~~Computers should only decide if they are better than humans.~~
3. ~~Computers should never override human decisions.~~

For discussion see e.g. Moor (1979); Moor discusses and rejects the principles

# Philosophical input

Issue 1: difficult ethical decisions delegated to AVs/machines.

In sum:

reason against AV?
impose condition?
no restriction needed?

# Philosophical input

Issue 2: human autonomy is compromised.

# Philosophical input: Movie

Autonomy

Wachowski Brothers, «The Matrix»

# Philosophical input: Discussion

«Plötzlich werden alle zu Zuschauern: die Fluggäste, die Piloten, die Airlines, der Wetterdienst, die Behörden. Die „human response“, die menschliche Antwort auf die Maschine, ist nicht mehr möglich, weil auch in den menschlichen Entscheidungsgruppen ein Programm von Befehlen, Verordnungen und Routinen abläuft.»

Frank Schirrmacher

Does AI pose a threat to human autonomy?

Árni Friðriksson, wikimedia commons (CC BY-SA 3.0); F.A.Z., 19,.4.2010

# Philosophical input: concept

Gr. autos: self

Gr. nomos: law

autonomy

Self determination

Heteronomy

Being determined by others

# Philosophical input: Isaiah Berlin

(1909-1997)

„I wish my life and decisions to depend on myself, not on external forces of whatever kind. I wish to be the instrument of my own, not of other men's, acts of will. I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes which affect me, as it were, from outside. I wish to be somebody, not nobody; a doer – deciding, not being decided for"

Berlin (1958/2022, 179)

Porträt: R. C. Croes, ANEFO, wikimedia commons (CC0)

# Philosophical input: ethics of medicine

## Respect for autonomy:

1. „Tell the truth.
2. Respect the privacy of others.
3. Protect confidential information.
4. Obtain consent for interventions with patients.
5. When asked, help others make important decisions.“

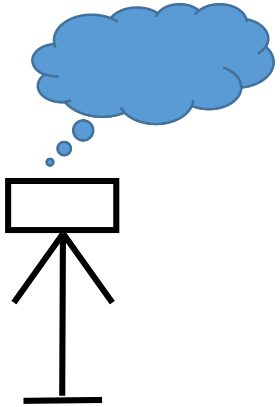
Beauchamp & Childress (2013, 107)

V. Santa Maria, wikimedia commons (CC BY 2.0)

# Question for you

How do you think may ML/AI applications impact on human autonomy?

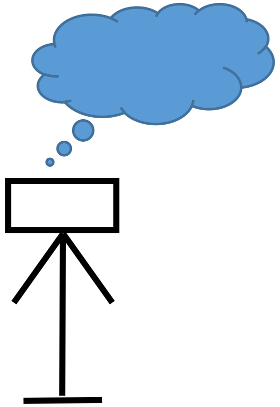
# Philosophical input: a general idea



AI applications take decisions

No problem, if authorization by voluntary „informed consent“

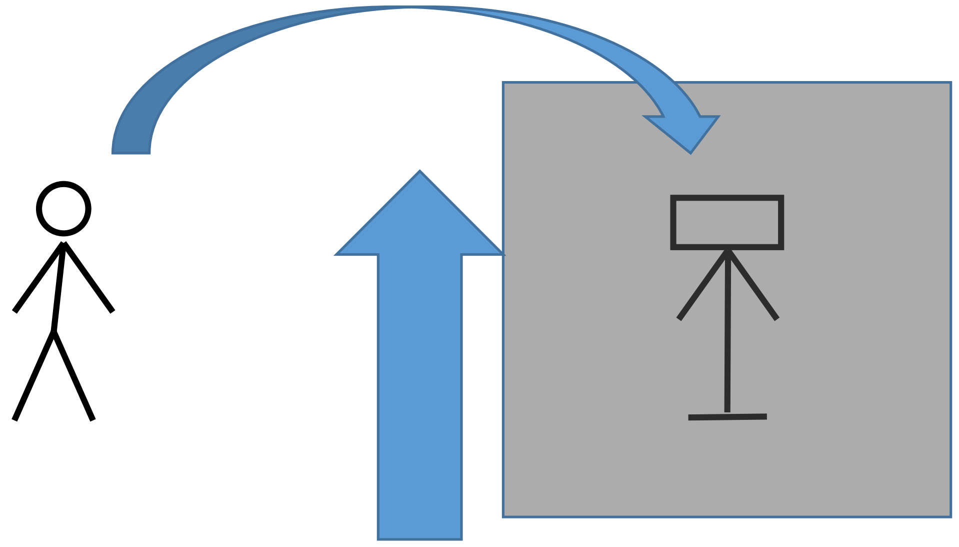
# Philosophical input: problems



AI applications take decisions

1. Information condition violated because system opaque (see issue later)
2. Not voluntary because pressure to use AV

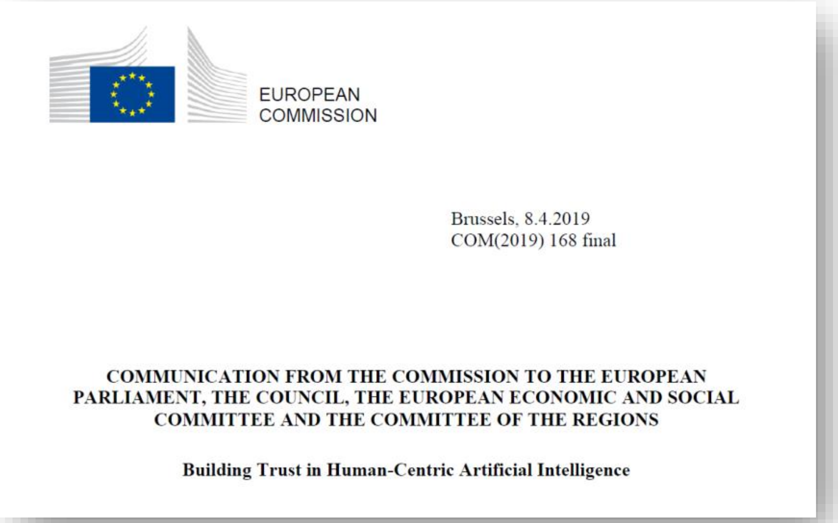
# Philosophical input: informed consent



Opacity

Rational basis: information

# Philosophical input: solution?



EUROPEAN
COMMISSION

Brussels, 8.4.2019
COM(2019) 168 final

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS

**Building Trust in Human-Centric Artificial Intelligence**

“Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. [...] **Oversight** may be achieved through governance mechanisms such as ensuring a human-in-the-loop, human-on-the-loop, or human-in-command approach.[13] It must be ensured that public authorities have the ability to exercise their oversight powers in line with their mandates. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.”

p. 4

EU (2019), Communication: Building Trust in Human Centric Artificial Intelligence

# Philosophical input

Issue 2: human autonomy is compromised.

In sum:

- reason against AV?
- impose condition?
- no restriction needed?

# Issue 3: unemployment

## Two questions:

a. Will there be work left for humans?
b. If not, how good or bad is life without work?

# Philosophical input: ad a.

Common argument:

1. So far, humans found new work when a technology made humans dispensable (new demands, new/other skills).

---

2. This will continue to be the case.

Form: enumerative induction from past

# Philosophical input: ad a.

Danaher (2017): this is problematic!

- Inelastic demand?
- Outpacing?
- Historical data (small part of exponential curve)?
- Winner takes all problem?

# Philosophical input: ad b. the value of work

Views differ:

„Einem guten menschlichen Leben muss die Dimension […] gelingender Arbeit offenstehen. […] Wir eignen uns die Welt im arbeitenden Umgang an.“

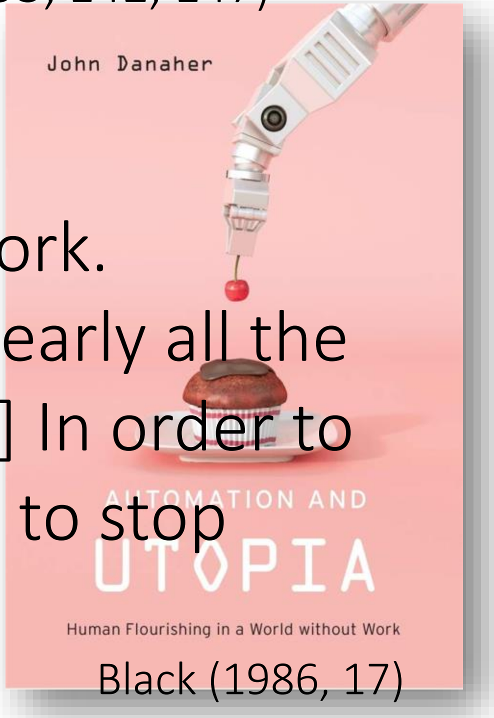
Seel (1995, 142, 147)

„Mit Mühsal sollst du dich von ihm [dem Acker] nähren dein Leben lang. Dornen und Disteln soll er dir tragen […]. Im Schweiße deines Angesichts sollst du dein Brot essen […]“

1. Mose 3, 17-19 (Lutherbibel 2017)

“No one should ever work. Work is the source of nearly all the misery in the world. […] In order to stop suffering, we have to stop working.”

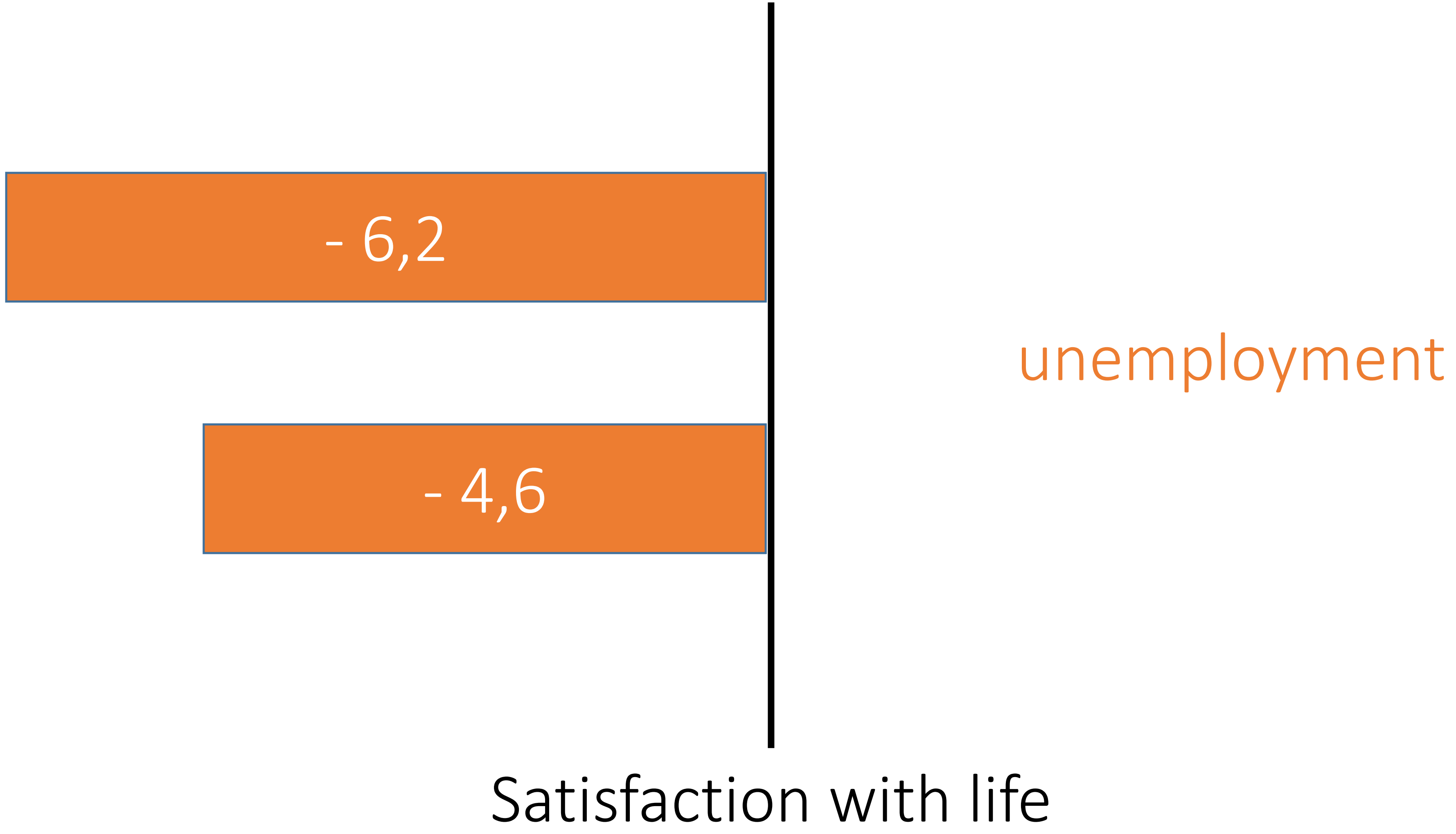
Black (1986, 17)


John Danaher

AUTOMATION AND UTOPIA

Human Flourishing in a World without Work

# Data: SOEP (Germany)



Data: Schröder (2020, 104)

# Philosophical input: ad b. the value of work

Sense of accomplishment part of meaningful live?

Proposals Danaher & Nyholm (2021):

- Give products a human touch
- Humans consider themselves to be parts of hybrid systems (humans and machines)

Philosophical input

# Issue 3: unemployment

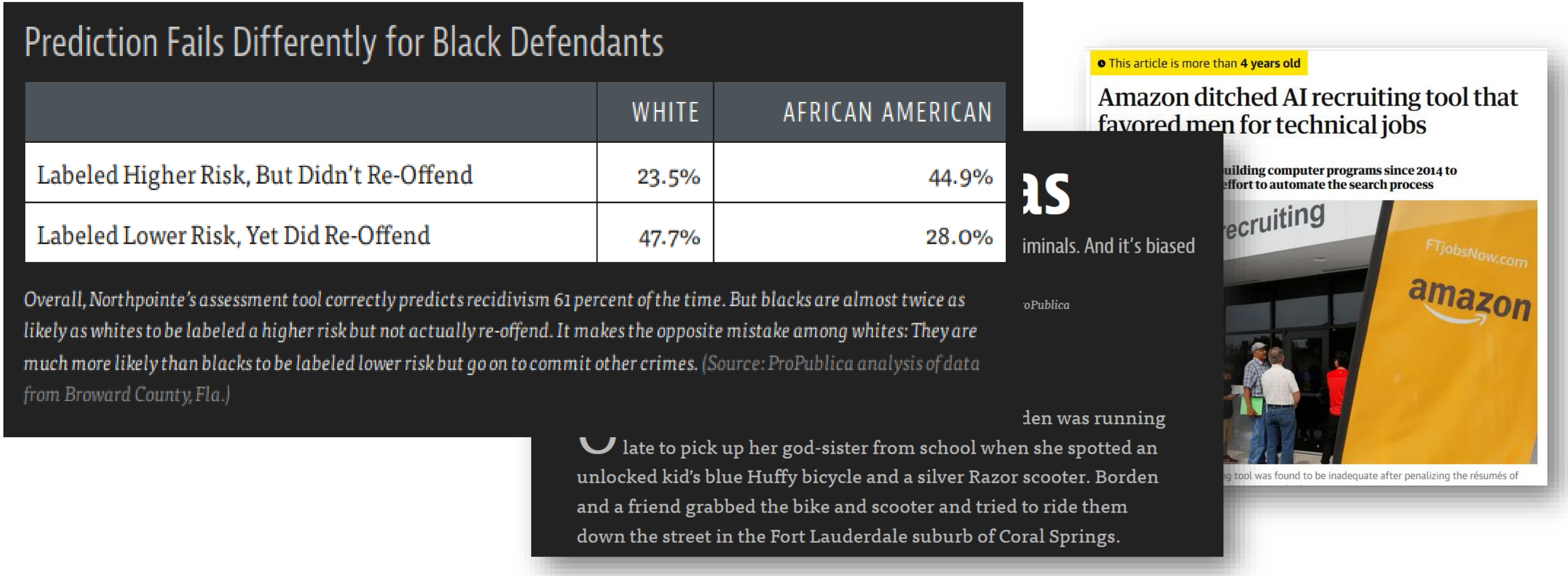
In sum:

reason against AV?
impose condition?
no restriction needed?

Philosophical input

Really bias? Discrimination?

# Issue 4: bias and discrimination



## Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

as

iminals. And it's biased

oPublica

den was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

This article is more than **4 years old**

## Amazon ditched AI recruiting tool that favored men for technical jobs

**uilding computer programs since 2014 to**
**effort to automate the search process**

g tool was found to be inadequate after penalizing the résumés of

The Guardian, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Philosophical input: some ideas on justices

- “fairness through blinding”: Don’t use variables such as gender, race, etc.

  problem:

  other variables may be correlated with race

- “group fairness” by “statistical parity”: the fraction of people who obtain a benefit should be the same for every group

  problems:

  accuracy and lack of individual fairness
- “individual fairness”: people with similar characteristics should obtain same chance of a good

Lepri et al. (2018, pp. 615-618)

# Philosophical input: general point

Impossibility theorem:

Several different conditions that sound somehow reasonable cannot be satisfied for all distributions

Question: which condition is the relevant one?
Aristotle: different kinds of justice:
e.g. distributive, retributive justice

# Issue 4: bias and discrimination

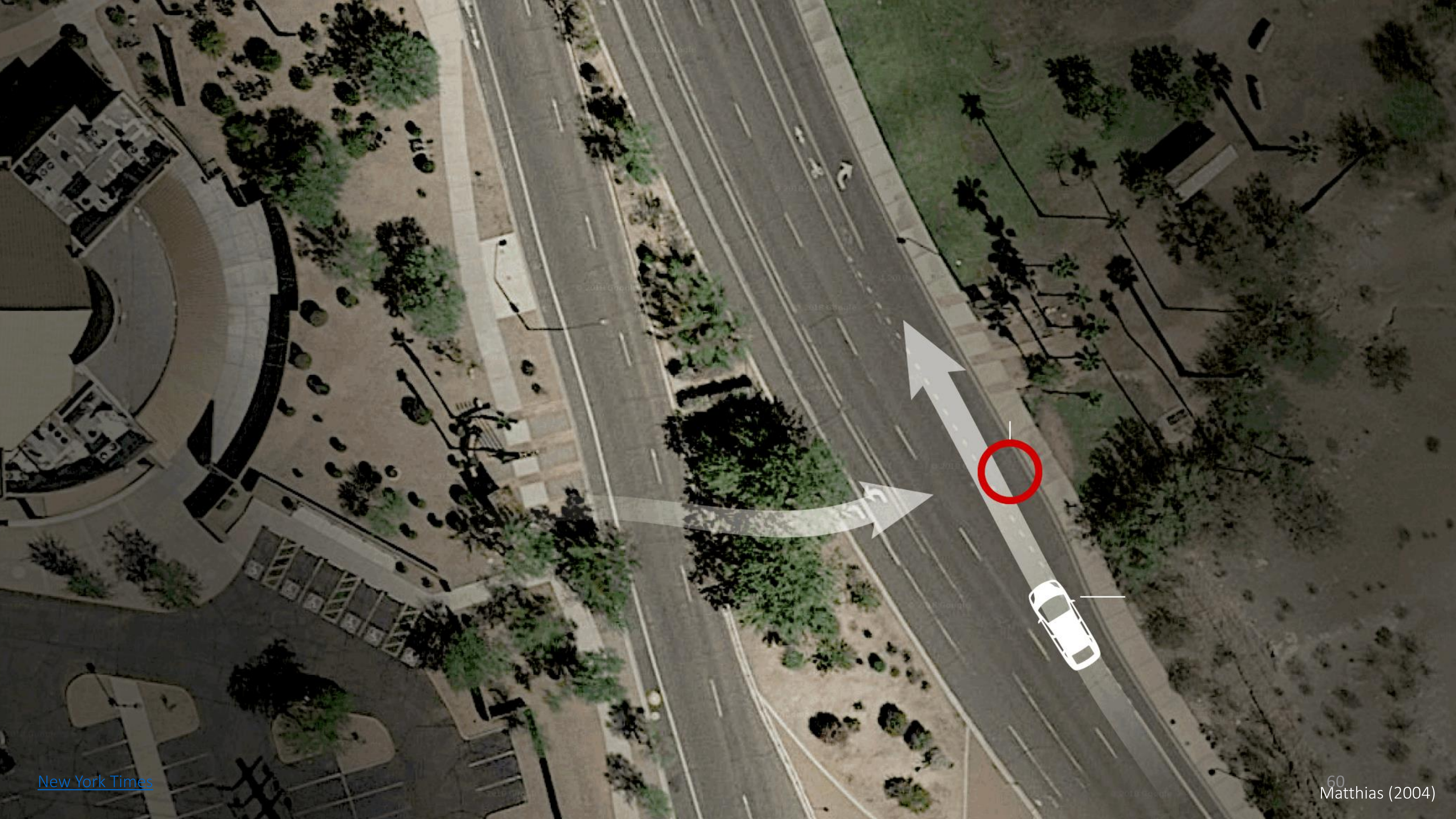
In sum:

reason against AV?

impose condition?

no restriction needed?

# Issue 5: responsibility gaps



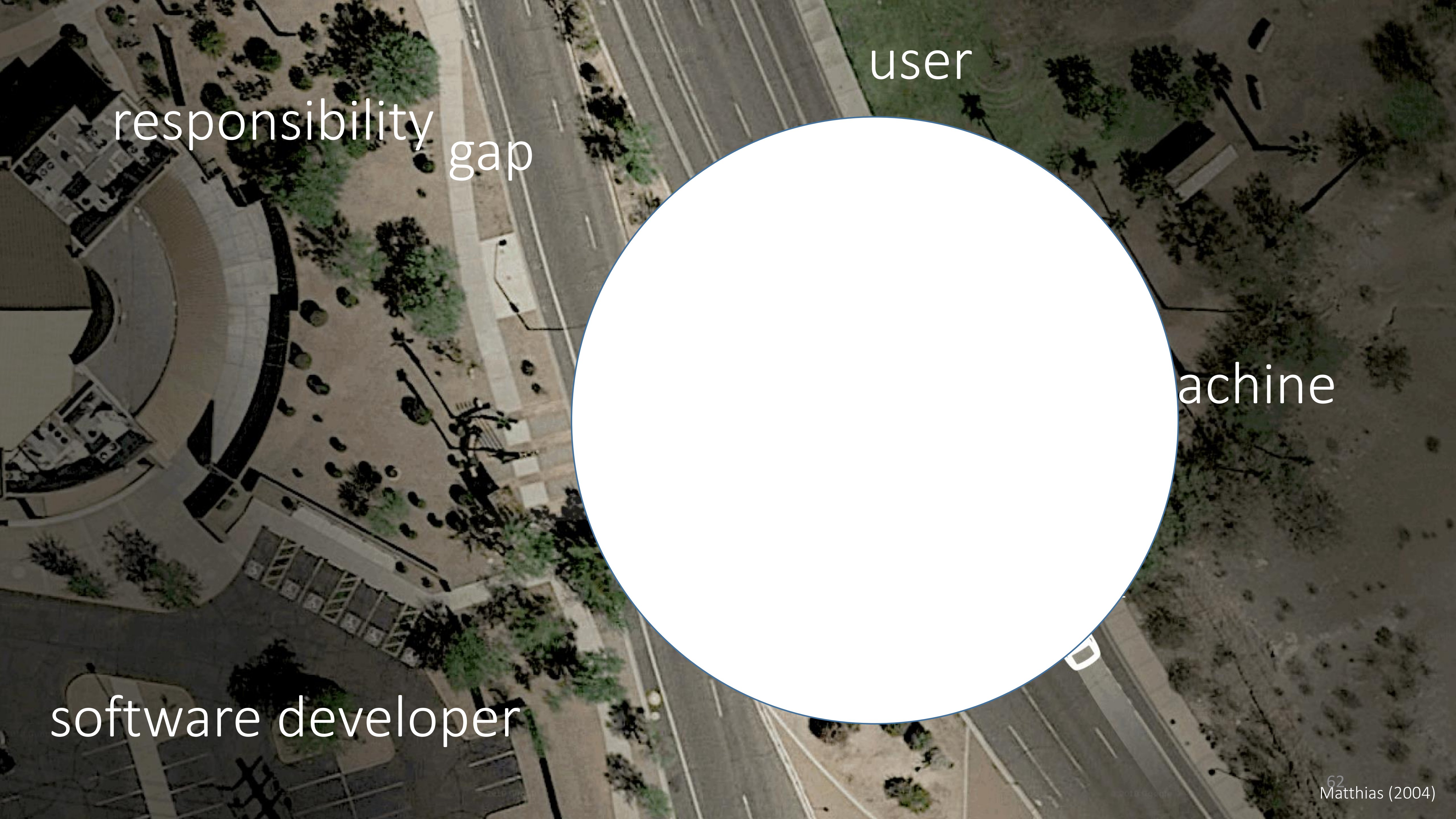
New York Times

Matthias (2004)

# Responsibility:

An agent is responsible for a harm if

- they are part of the cause
- in doing so, they did a mistake (intent or negligence)
- they can take responsibility (be criticized, jailed …)



user

responsibility gap

achine

software developer

Matthias (2004)

# Solutions:

- Think of hybrid systems (human+machine) as agents
- Create a liability of companies/users …
-

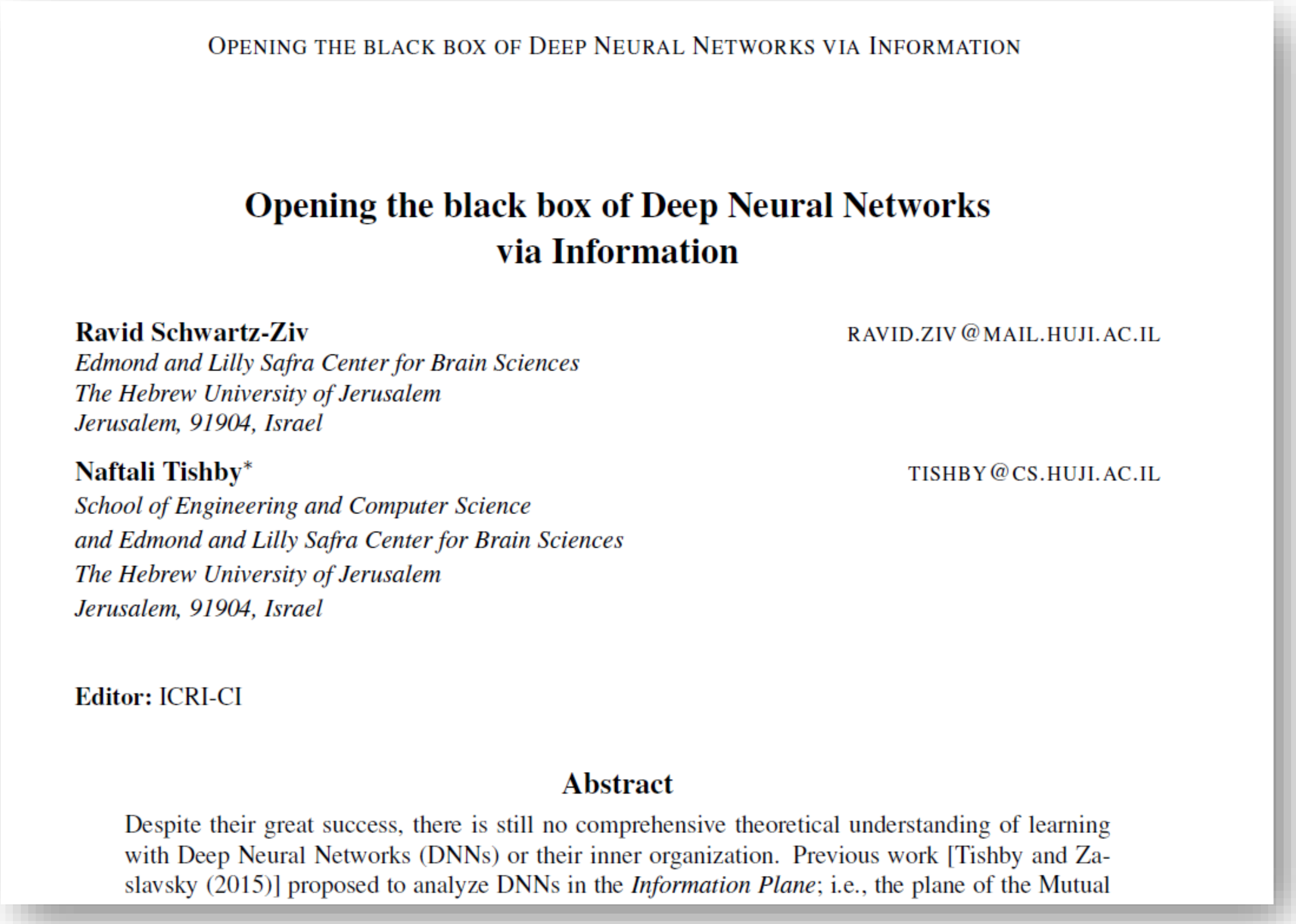
# Issue 5: responsibility gaps

In sum:

- reason against AV?
- impose condition?
- no restriction needed?

# Philosophical input

## Issue 6: opacity



OPENING THE BLACK BOX OF DEEP NEURAL NETWORKS VIA INFORMATION

**Opening the black box of Deep Neural Networks via Information**

**Ravid Schwartz-Ziv** RAVID.ZIV@MAIL.HUJI.AC.IL
*Edmond and Lilly Safra Center for Brain Sciences*
*The Hebrew University of Jerusalem*
*Jerusalem, 91904, Israel*

**Naftali Tishby*** TISHBY@CS.HUJI.AC.IL
*School of Engineering and Computer Science*
*and Edmond and Lilly Safra Center for Brain Sciences*
*The Hebrew University of Jerusalem*
*Jerusalem, 91904, Israel*

**Editor:** ICRI-CI

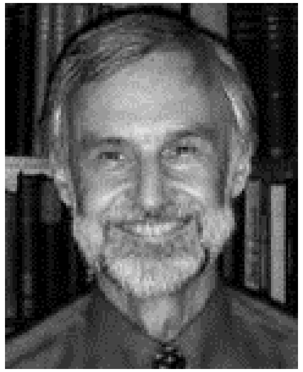
**Abstract**

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work [Tishby and Zaslavsky (2015)] proposed to analyze DNNs in the *Information Plane*; i.e., the plane of the Mutual

> “Despite their great success, there is still no comprehensive understanding of the optimization process or the internal organization of DNNs, and they are often criticized for being used as mysterious ”black boxes”” p. 2

Shwartz-Ziv & Tishby (2017)

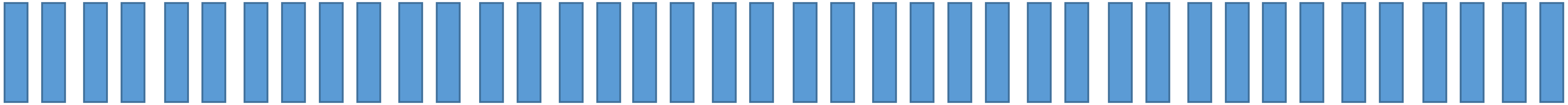
# Philosophical input



## Popular account

Humphreys (2009, p. 618):

„Here a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process“



https://philosophy.virginia.edu/faculty/profile/pwh2a

# Philosophical input

- Machine learning
- Good old-fashioned AI

are opaque

# Philosophical input

Is it really that important to oversee or survey the whole calculation?

# Philosophical input

New idea:

**opacity**

1. being difficult to look through
2. being difficult to understand

Disposition to resist epistemic access by humans

# Philosophical input

Challenge:

opacity

What must be known and understood if opacity is to be avoided? Or:
What knowledge and understanding is relevant?

Problem: There is always more to know and to understand about a method.

Cf. discussion about instruments of observation

# Philosophical input

Challenge met

opacity

Methods → outcomes

Understand

„Why did a particular outcome arise?“

„Why is this image classified as dog picture?“ (p, for short)

# Philosophical input

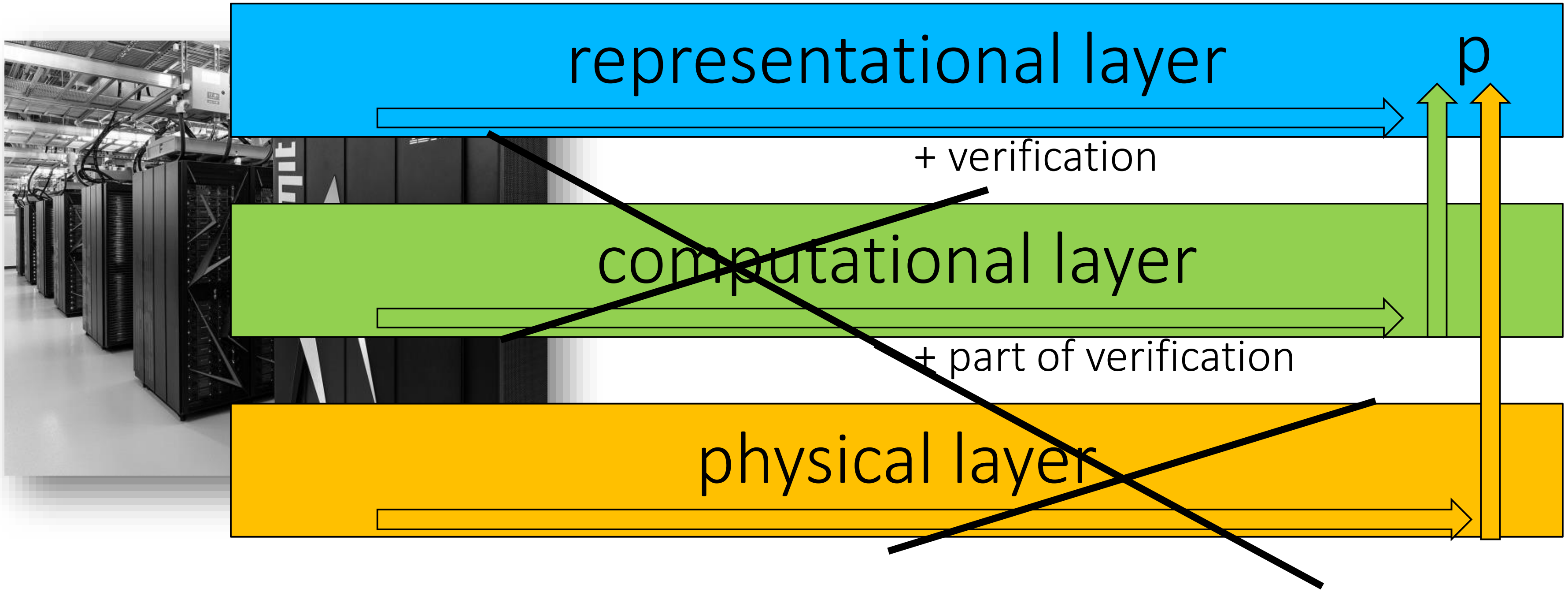
## Opacity redefined

1. The application of a method is opaque to the extent to which it is difficult for average scientists in the default setting to know and to understand why the outcome has arisen.
2. A method is opaque to the extent to which its typical applications are opaque.

Beisbart (2021, 11661)

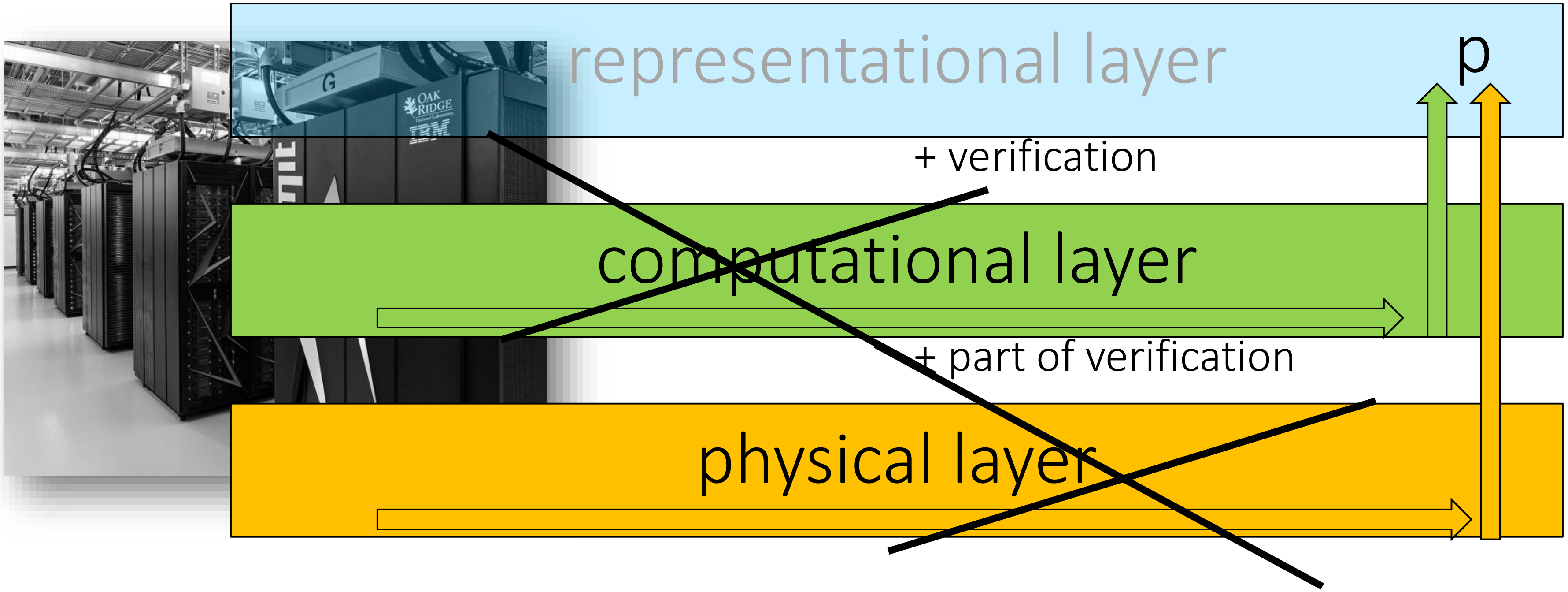
# Philosophical input

## Good old-fashioned AI



„image is classified as dog picture“ (p)

# Machine learning



„image is classified as dog picture“ (p)

# Question for you

Is the opacity of ML models special? If so why?

# Issue 6: opacity

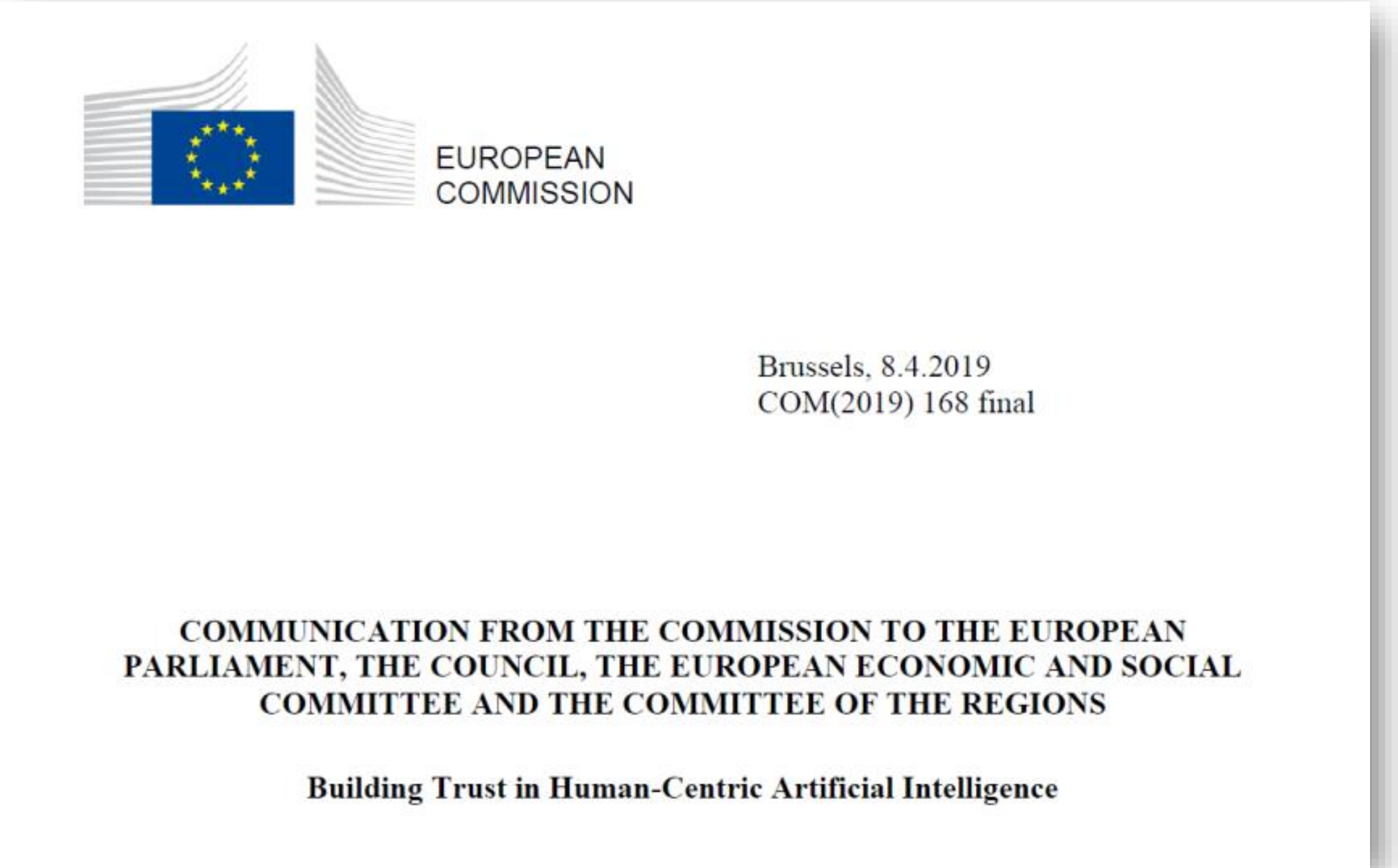
In sum:

- reason against AV?
- impose condition?
- no restriction needed?

# Philosophical input

Selected conditions:

- Human agency and oversight
- Transparency
- Accountability



EUROPEAN
COMMISSION

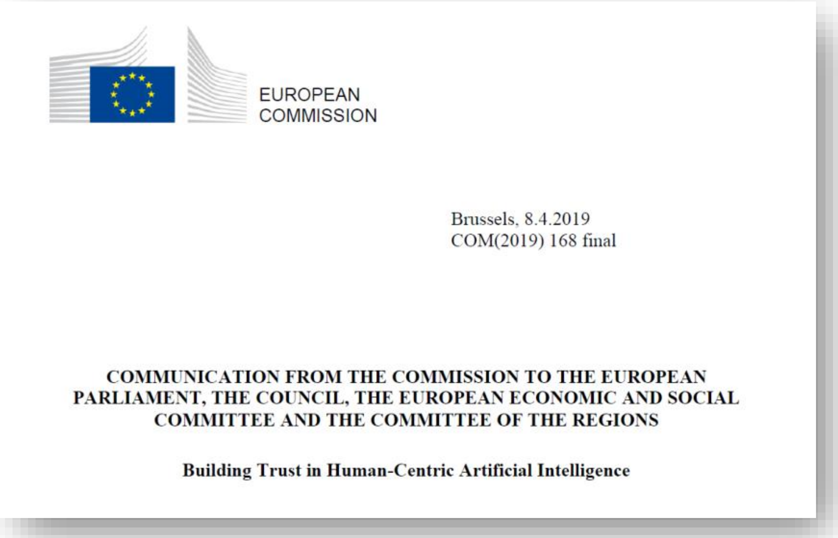
Brussels, 8.4.2019
COM(2019) 168 final

**COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS**

**Building Trust in Human-Centric Artificial Intelligence**

EU (2019)

# Philosophical input: transparency



EUROPEAN
COMMISSION

Brussels, 8.4.2019
COM(2019) 168 final

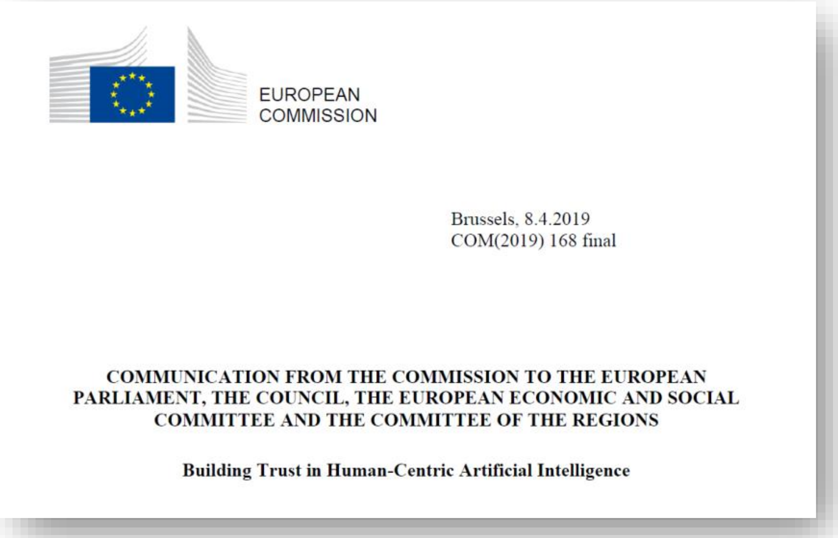
**COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS**

**Building Trust in Human-Centric Artificial Intelligence**

“The **traceability** of AI systems should be ensured; it is important to log and document both the decisions made by the systems, as well as the entire process (including a description of data gathering and labelling, and a description of the algorithm used) that yielded the decisions. Linked to this, **explainability** of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible. Ongoing research to develop explainability mechanisms should be pursued.”

EU (2019, 5)

# Philosophical input: accountability



EUROPEAN
COMMISSION

Brussels, 8.4.2019
COM(2019) 168 final

COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS

**Building Trust in Human-Centric Artificial Intelligence**

“**Potential negative impacts** of AI systems should be identified, assessed, documented and minimised. The use of impact assessments facilitates this process. These assessments should be proportionate to the extent of the risks that the AI systems pose. **Trade-offs** between the requirements – which are often unavoidable – should be addressed in a rational and methodological manner, and should be accounted for. Finally, when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure **adequate redress**.”

EU (2019, 6)

# Question for you

What do you think about these requirements of

-Human agency and oversight

-Traceability

-accountability?

# Summary

- ML comes with many benefits.
- It raises ethical issues too, e.g. potential loss of autonomy, unemployment, difficult ethical decisions …
- Many ethical issues need closer scrutiny.
- Often, solutions can be found.
- A problem that raises its head again and again is opacity.

Merci – thanks!

# References

Allen, C., Wallach, W. & Smit, I. (2006), Why Machine Ethics? *IEEE Intelligent Systems* 21/4, 12–17, auch in Anderson & Anderson (2011, 51–61).

Anderson, M. & Anderson, S. L. (eds., 2006), Machine Ethics, special issue of *IEEE Intelligent Systems* 21/4.

Anderson, M. & Anderson, S. L. (eds., 2011), Machine Ethics, Cambridge University Press, New York.

Anderson, M., Anderson, S. L. & Armen, C. (2006), An Approach to Computing Ethics, *IEEE Intelligent Systems* 21/4, 56–63.

Asimov, I. (1968), I, Robot, Grafton Press, New York (collected stories published between 1940 and 1950).

Beauchamp, T. L. & Childress, J. F. (2013). Principles of Biomedical Ethics, 7th edition, New York: Oxford University Press.

Beisbart, C. (2021), Opacity thought through: on the intransparency of computer simulations, *Synthese* 199, 11643–11666.

Berlin, I. (2002), Liberty: Incorporating Four Essays on Liberty (ed. by H. Hardy). Oxford

Black, B. (1986), The Abolition of Work and Other Essays, Loompanics Unlimited: Port Townshend, Washington.

# References

Clarke, R. (1993/94), Asimov's Laws of Robotics: Implications for Information Technology, *IEEE Computer* 26/12, 53–61 and 27/1, 57–66.

Cloos, C. (2005), The Utilibot project: An Autonomous Mobile Robot Based on Utilitarianism, *2005 AAAI Fall Symposium on Machine Ethics*, 38–45.

Danaher, J. (2017). Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life, *Science and Engineering Ethics* 23 (1):41-64.

Danaher, J. & Nyholm, S. (2021), Automation, work and the achievement gap, *AI Ethics* **1**, 227–237, https://doi.org/10.1007/s43681-020-00028-x

EU (2019), Ethics guidelines for trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Evans, K., de Moura, N., Chauvier, S. et al. (2020), Ethical Decision Making in Autonomous Vehicles: The AV Ethics Project. *Sci Eng Ethics* 26, 3285–3312.

Guarini, M. (2011), Computational Neural Modeling and the Philosophy of Ethics. Reflections on the Particularism-Generalism Debate, in Anderson & Anderson (2011, 316–334).

Humphreys, P. (2009), The Philosophical Novelty of Computer Simulation Methods, *Synthese* 169, 615-626.

# References

Lepri, B. et al. (2018), Fair, Transparent, and Accountable Algorithmic Decision-making Processes The Premise, the Proposed Solutions, and the Open Challenges, *Philosophy & Technology* 31 (2018), 611 – 627.

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021), Ethical issues in focus by the autonomous vehicles industry, *Transport reviews*, *41*(5), 556-577.

Matthias, A. (2004),) The responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics Inf Technol* 6, 175–183.

Misselhorn, C. 2017, Grundfragen der Maschinenethik, Reclam, Stuttgart.

Moor, J. H. (1979), Are There Decisions that Computers Should Never Make? *Nature and System* 1, 217–229.

Müller, V. C. (2021), Ethics of Artificial Intelligence and Robotics, *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.).

Seel, M. (1995), Versuch über die Form des Glücks, Suhrkamp, Frankfurt am Main.

Schröder, M. (2020), Wann sind wir wirklich zufrieden? Überraschende Erkenntnisse zu Arbeit, Liebe, Kindern, Geld, München: C. Bertelsmann, 2020.

# References

Shwartz-Ziv, R., & Tishby, N. (2017), Opening the black box of Deep Neural Networks via Information, ArXiv:1703.00810.

van Wynsberghe, A. & Robbins, S. (2019), Critiquing the Reasons for Making Artificial Moral Agents, *Science and Engineering Ethics* 25, 719–735.

Verma, S. & Rubin, V. (2018), Fairness definitions explained, *2018 IEEE/ACM International Workshop on Software Fairness* (FairWare, IEEE), 2018, 1-7.

Wallach, W. & Allen, C. (2009), Moral Machines. Teaching Robots Right from Wrong, Oxford University Press, New York.