Seminar Ethics in NLP – Ethics Intro

Master-Seminar – E4N: Ethics in Natural Language Processing (IN2107, IN4428)

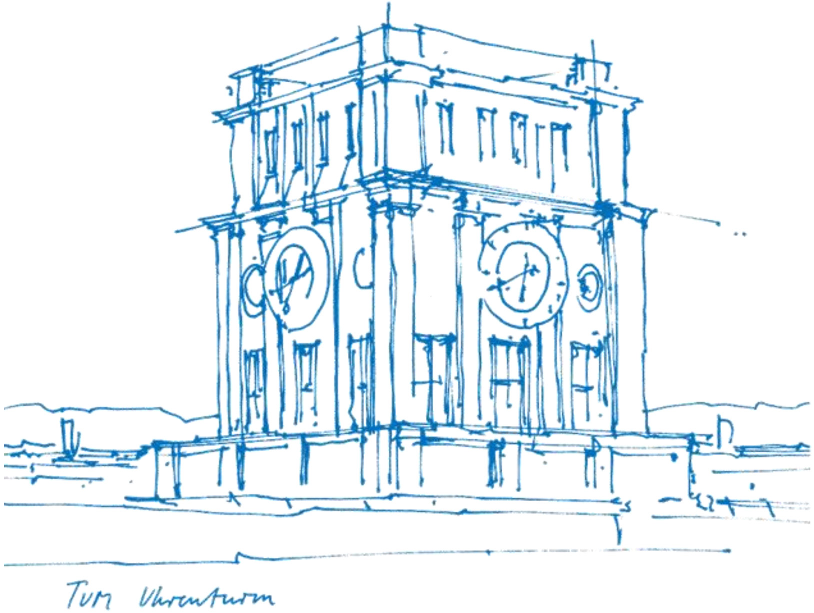
Tobias Eder, M.A. M.Sc.

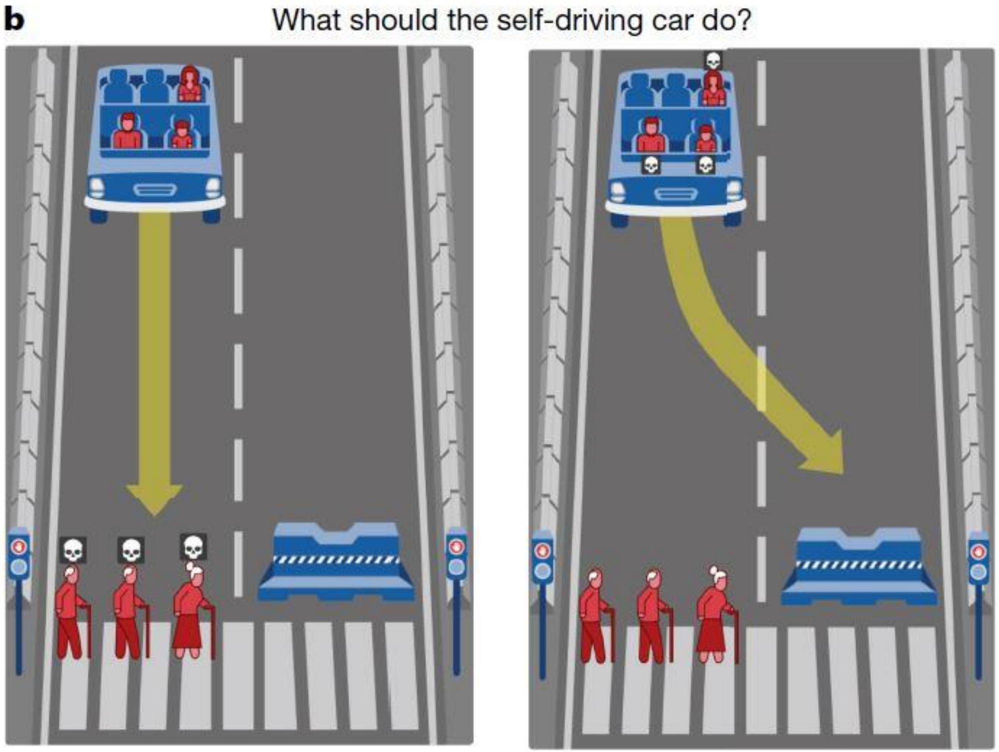
Prof. Dr. Georg Groh

Research Group Social Computing, Department of Informatics,
Technical University of Munich

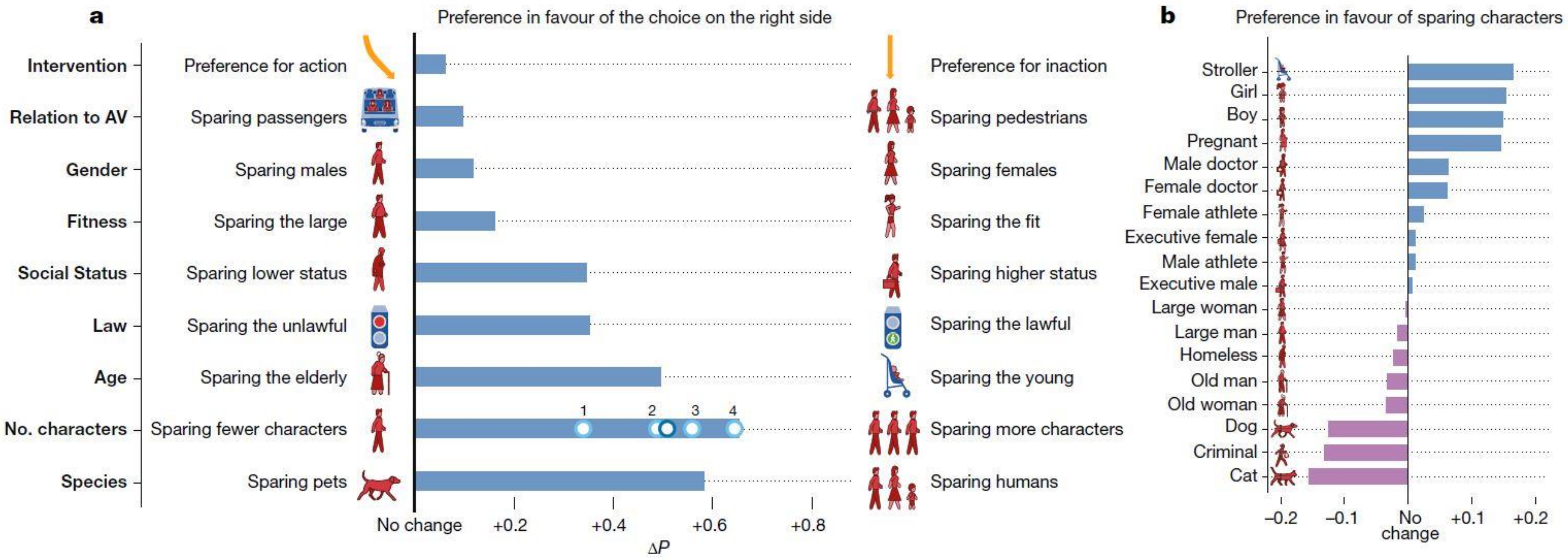
05.05.2021

With thanks to Louis Longin und Johan van der Merwe

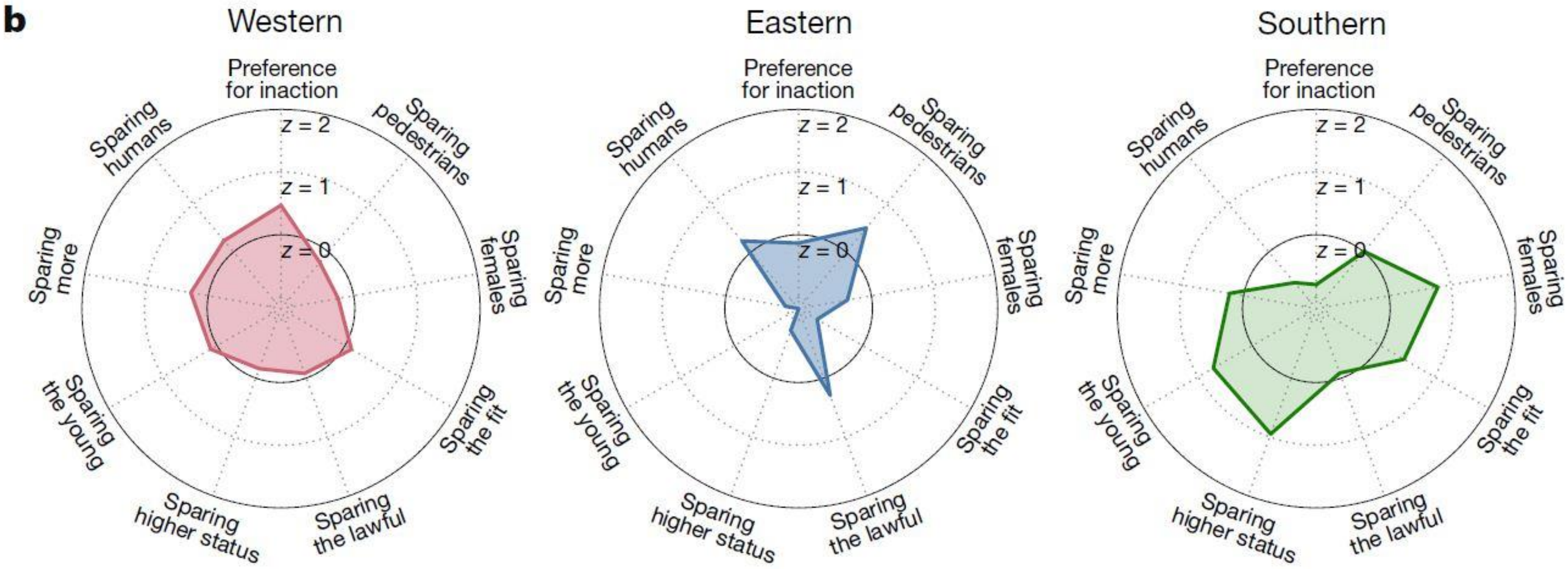




Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59.



Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018).
The moral machine experiment. *Nature*, *563*(7729), 59.



Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59.

Meta-Ethics

Where do our ethical principles come from?

What is the basis of ethics in human interaction?

Normative Ethics

What constitutes moral right or wrong?

How are moral standards established?

Applied Ethics

How do we deal with specific scenarios?

How can different ethical concepts be applied to them?

Meta-Ethics

Metaphysical

Is there morality outside of human interaction?

Psychological

What is the mental basis for acting morally?

e.g. Rationalists vs Anti-Rationalists

Meta-Ethics

Where do our ethical principles come from?

What is the basis of ethics in human interaction?

Normative Ethics

What constitutes moral right or wrong?

How are moral standards established?

Applied Ethics

How do we deal with specific scenarios?

How can different ethical concepts be applied to them?

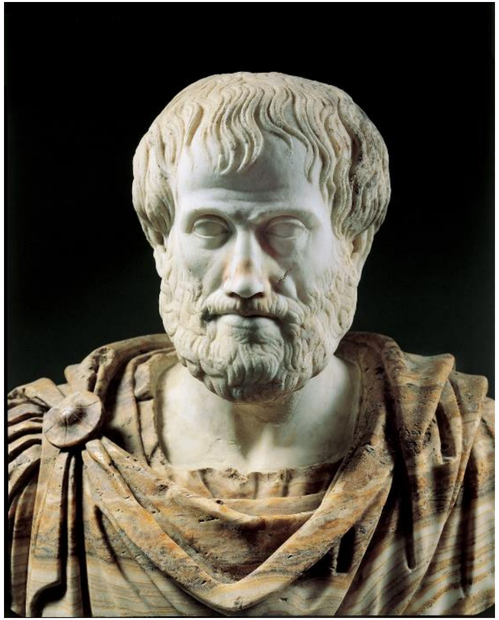
Normative Ethics

Virtue Ethics

Deontology

Consequentialism

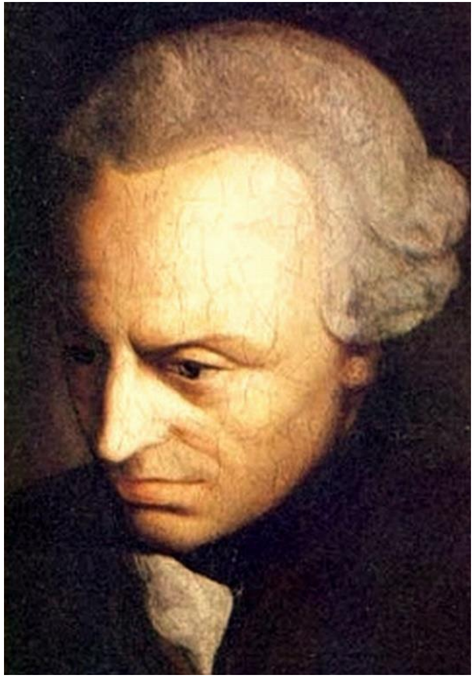
Virtue Ethics



“We are not studying in order to know what virtue is, but to become good, for otherwise there would be no profit in it.”

Deontology

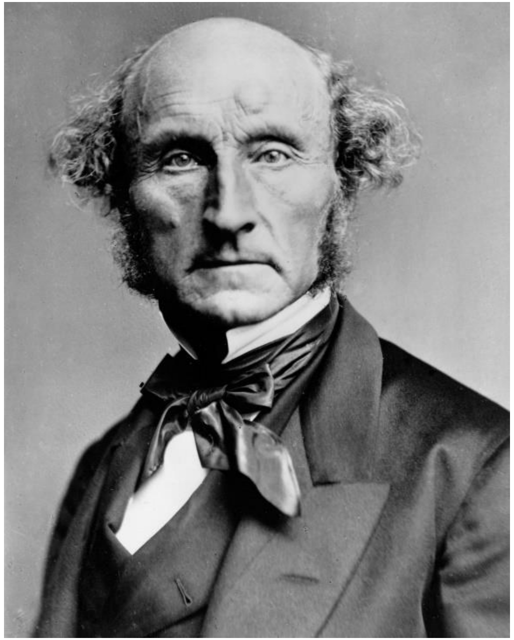
“Act as if the maxims of your action were to become through your will a universal law of nature.”



Consequentialism

Utilitarianism:

“The greatest happiness of the greatest number is the foundation of morals and legislation.”



Meta-Ethics

Where do our ethical principles come from?

What is the basis of ethics in human interaction?

Normative Ethics

What constitutes moral right or wrong?

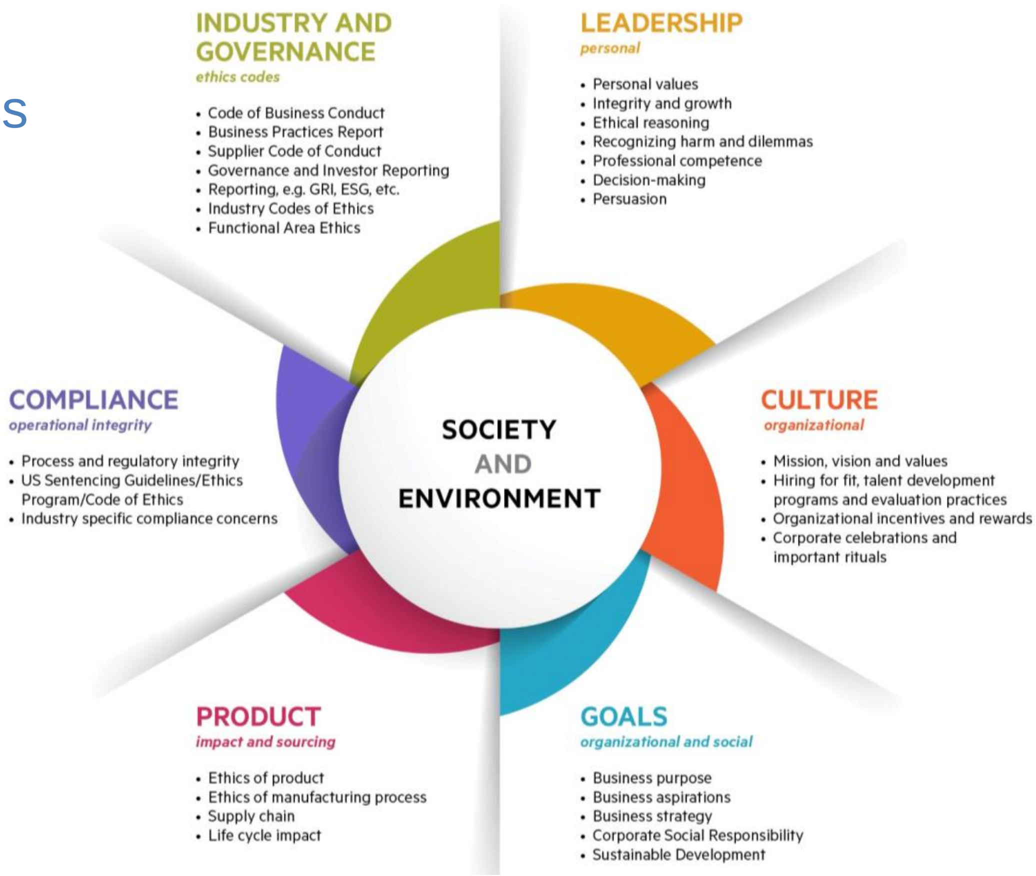
How are moral standards established?

Applied Ethics

How do we deal with specific scenarios?

How can different ethical concepts be applied to them?

Applied Ethics



INDUSTRY AND GOVERNANCE
ethics codes

- Code of Business Conduct
- Business Practices Report
- Supplier Code of Conduct
- Governance and Investor Reporting
- Reporting, e.g. GRI, ESG, etc.
- Industry Codes of Ethics
- Functional Area Ethics

LEADERSHIP
personal

- Personal values
- Integrity and growth
- Ethical reasoning
- Recognizing harm and dilemmas
- Professional competence
- Decision-making
- Persuasion

COMPLIANCE
operational integrity

- Process and regulatory integrity
- US Sentencing Guidelines/Ethics Program/Code of Ethics
- Industry specific compliance concerns

SOCIETY AND ENVIRONMENT

CULTURE
organizational

- Mission, vision and values
- Hiring for fit, talent development programs and evaluation practices
- Organizational incentives and rewards
- Corporate celebrations and important rituals

PRODUCT
impact and sourcing

- Ethics of product
- Ethics of manufacturing process
- Supply chain
- Life cycle impact

GOALS
organizational and social

- Business purpose
- Business aspirations
- Business strategy
- Corporate Social Responsibility
- Sustainable Development

Applied Ethics

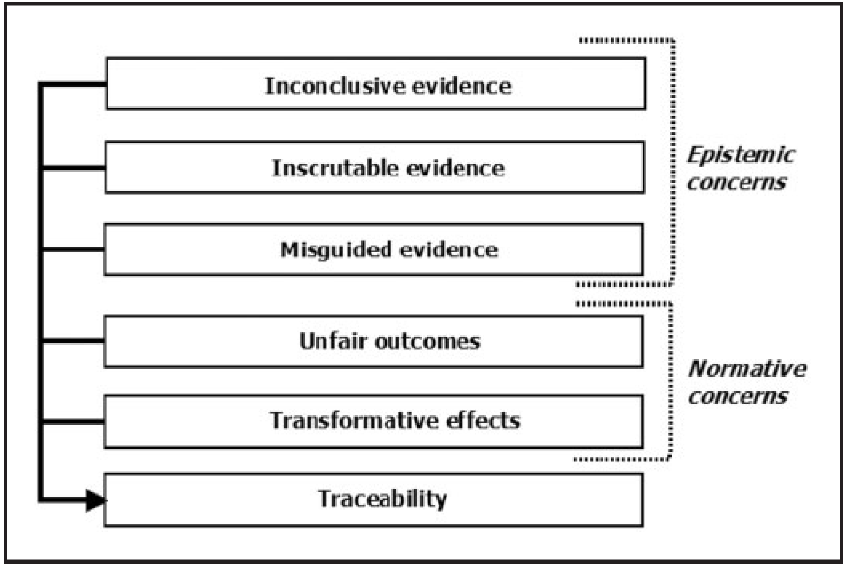
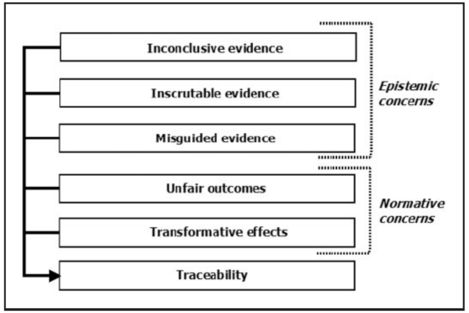


Figure 1. Six types of ethical concerns raised by algorithms.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.



Inconclusive evidence

Inscrutable evidence

Misguided evidence

Epistemic concerns

Unfair outcomes

Transformative effects

Normative concerns

Traceability

Figure 1. Six types of ethical concerns raised by algorithms.

Inconclusive evidence

→ Knowledge produced by ML contains uncertainty

Inscrutable evidence

→ Lack of transparency in algorithms

Misguided evidence

→ The model can’t be uncoupled from training data and its potential inherent problems

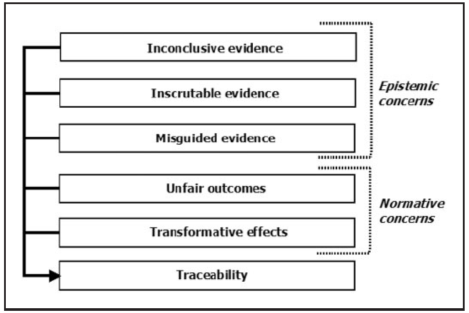


Figure 1. Six types of ethical concerns raised by algorithms.

Unfair outcomes

→ Ethics in ML can often be reduced to recognizing discrimination

Transformative effects

→ There is a feedback loop between algorithmic decisions and the real world

Traceability

→ Cause and effect are hard to follow

Meta-Ethics

Where do our ethical principles come from?

What is the basis of ethics in human interaction?

Normative Ethics

What constitutes moral right or wrong?

How are moral standards established?

Applied Ethics

How do we deal with specific scenarios?

How can different ethical concepts be applied to them?

Ethical Lenses

Which option will produce the most good and do the least harm?
(The Utilitarian Approach)

Which option best respects the rights of all who have a stake?
(The Rights Approach)

Which option treats people equally or proportionately?
(The Justice Approach)

Which option best serves the community as a whole, not just some members?
(The Common Good Approach)

Which option leads me to act as the sort of person I want to be?
(The Virtue Approach)

Awareness and Regulatory Approaches

THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)

for self assessment

ALTAI

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-Discrimination and Fairness
6. Societal and Environmental Wellbeing
7. Accountability

AIEI VCIO Model

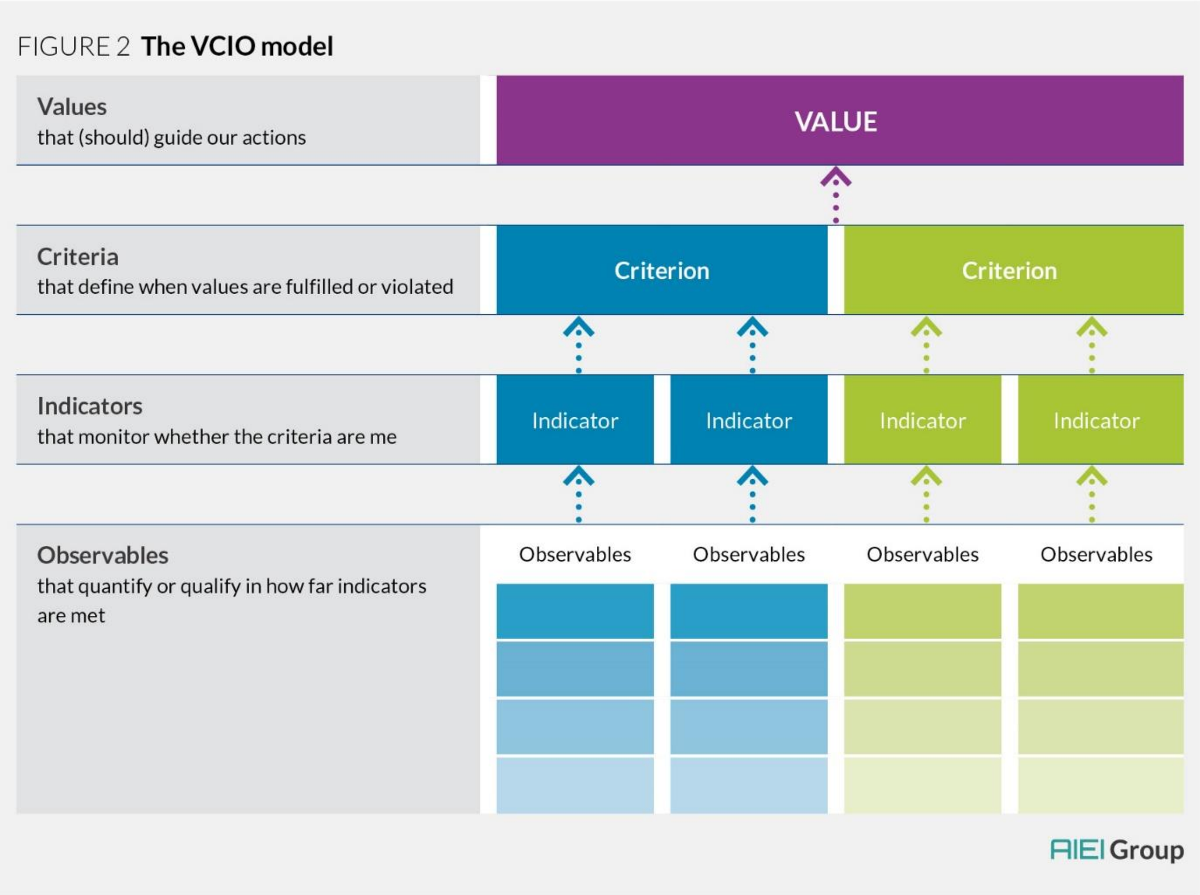


FIGURE 2 **The VCIO model**

Values
that (should) guide our actions

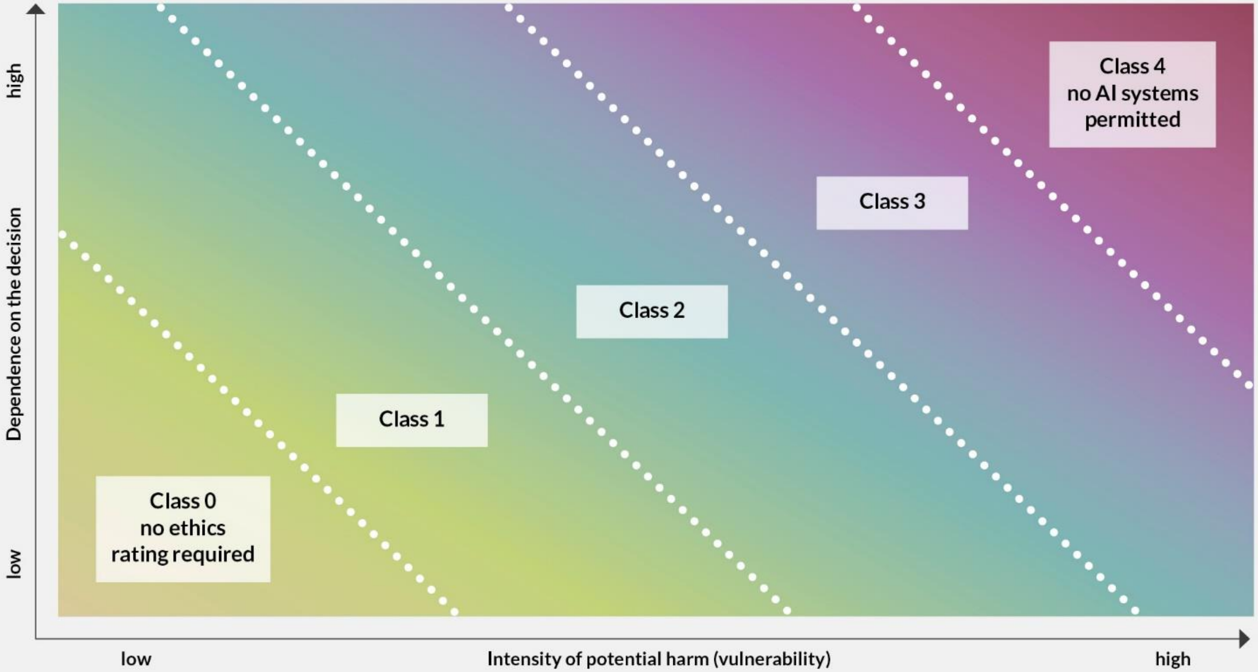
Criteria
that define when values are fulfilled or violated

Indicators
that monitor whether the criteria are me

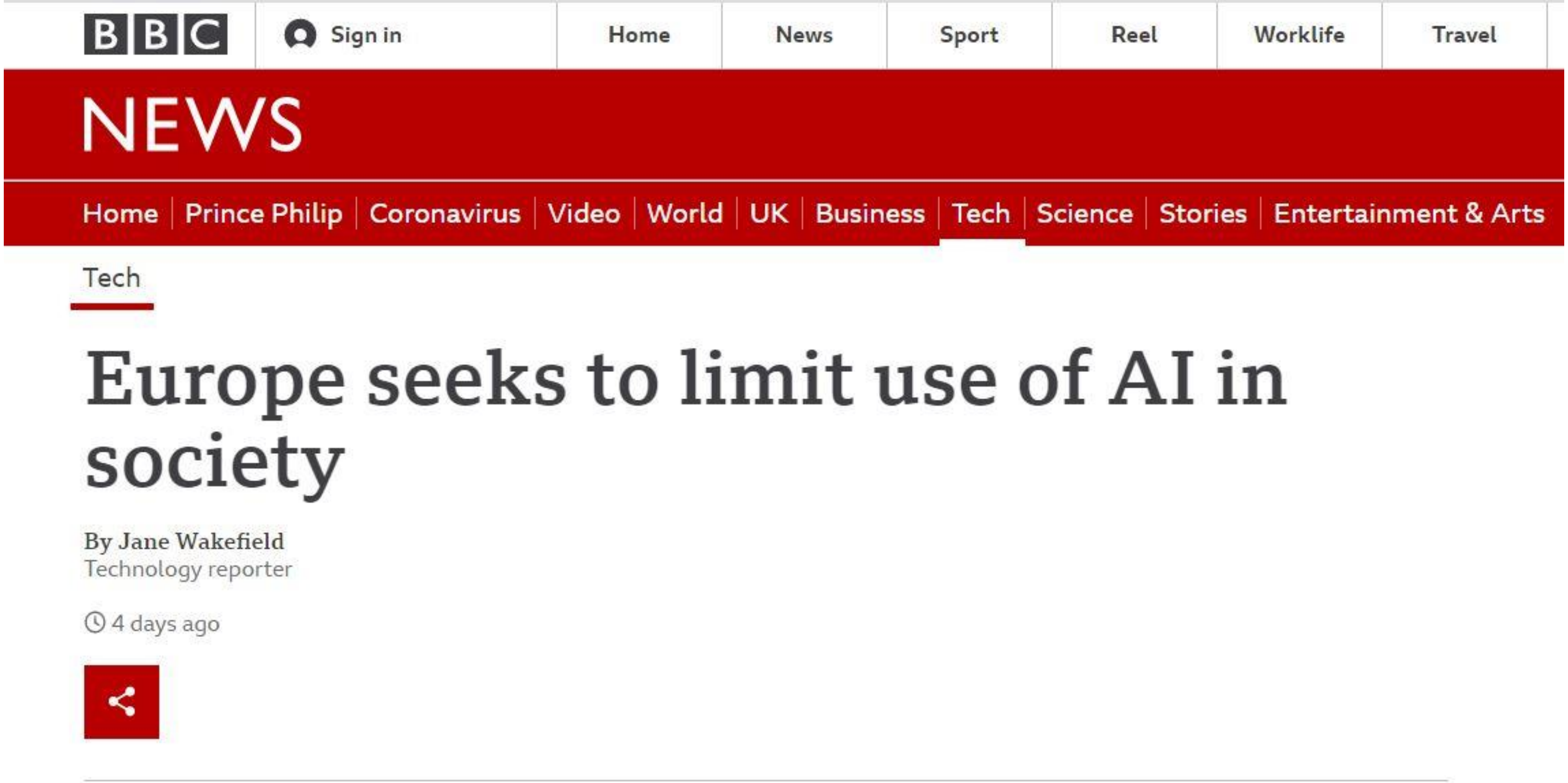
Observables
that quantify or qualify in how far indicators are met

AIEI Risk

FIGURE 7 **Risk matrix with 5 classes of application areas with risk potential ranging from 'no ethics rating required' in class 0 to the prohibition of AI systems in class 4**



Source: Krafft and Zweig 2019

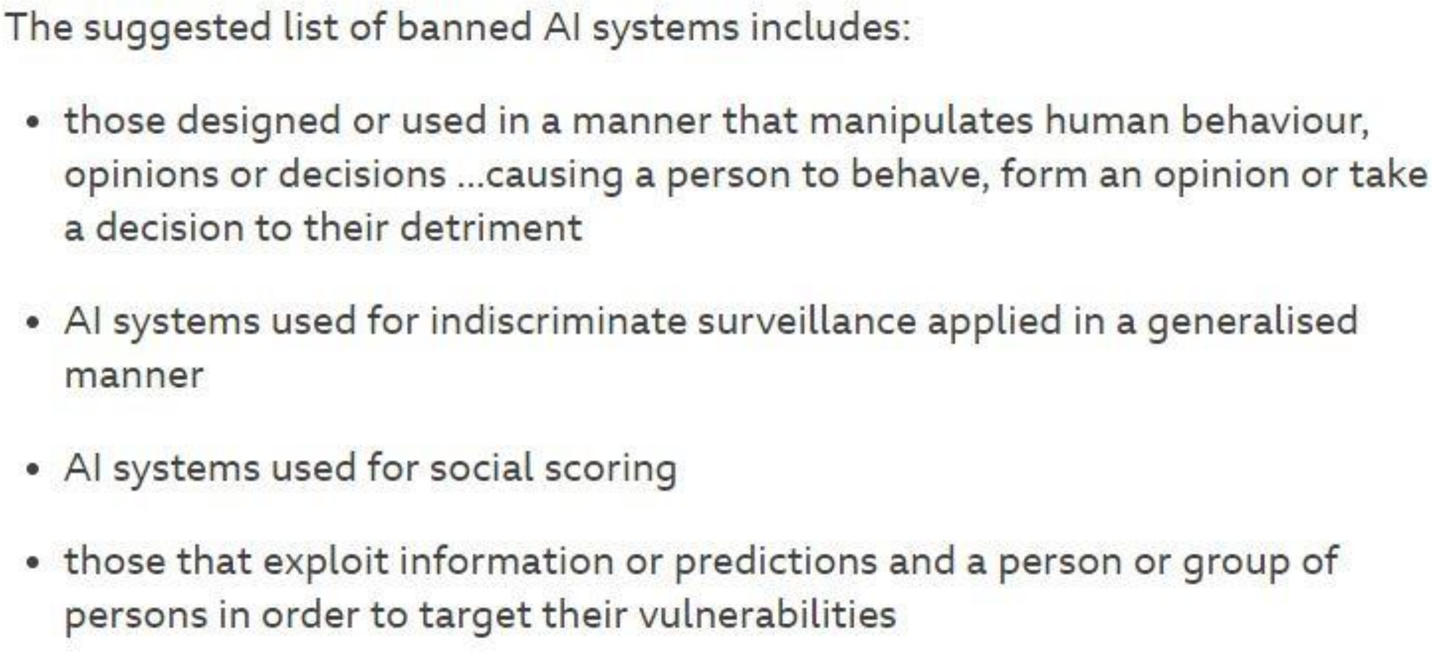


Europe seeks to limit use of AI in society

By Jane Wakefield
Technology reporter

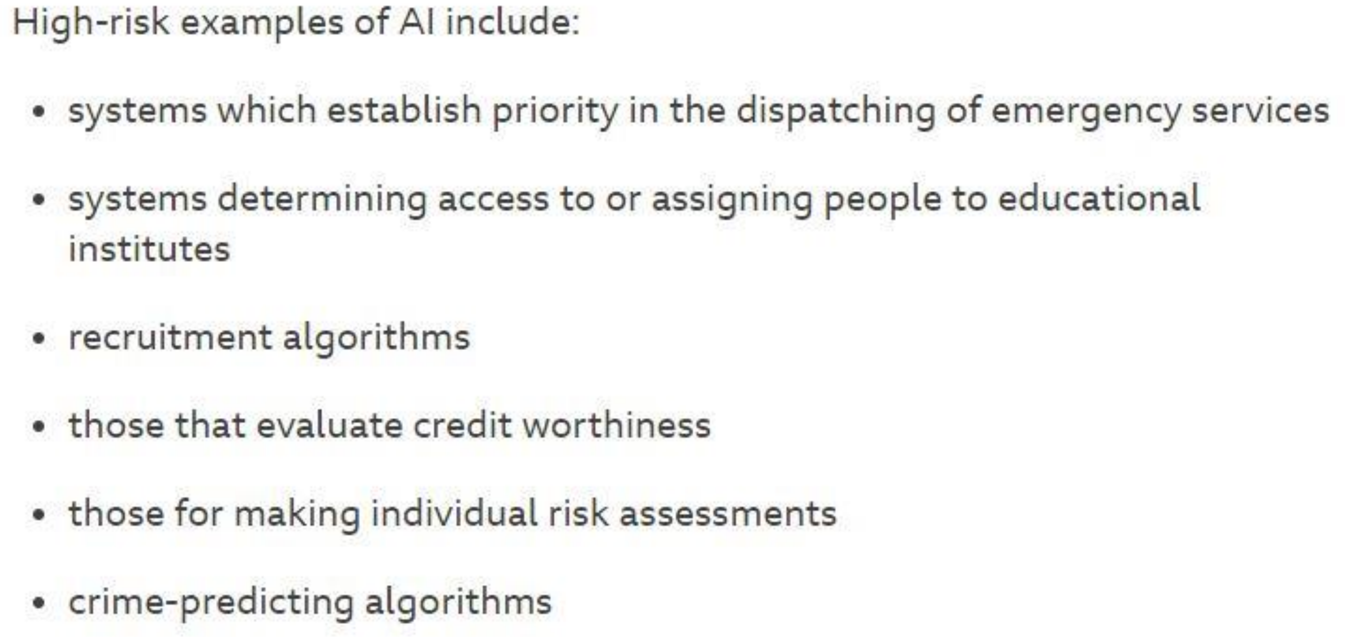
4 days ago

https://www.bbc.com/news/technology-56745730



The suggested list of banned AI systems includes:

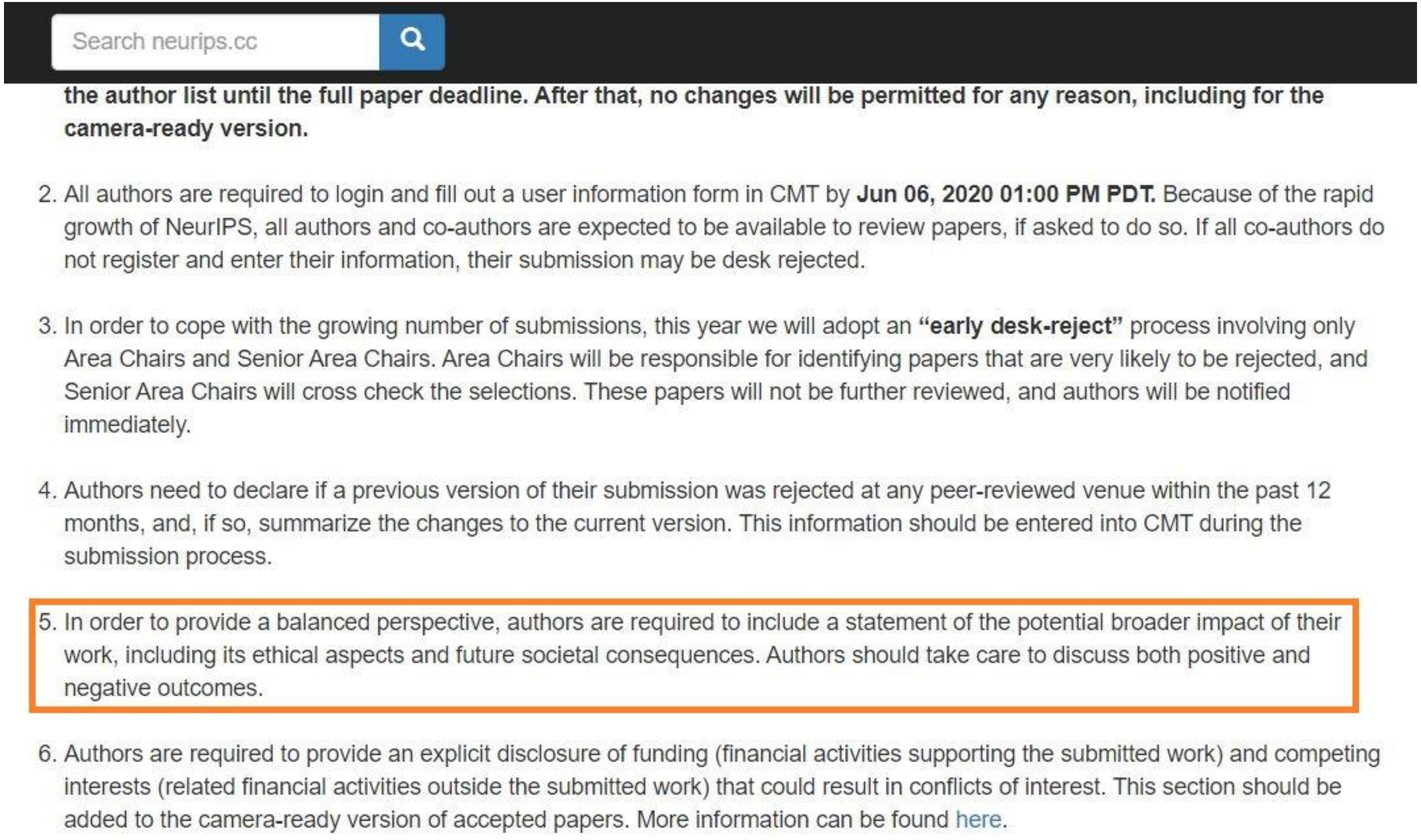
- those designed or used in a manner that manipulates human behaviour, opinions or decisions ...causing a person to behave, form an opinion or take a decision to their detriment
- AI systems used for indiscriminate surveillance applied in a generalised manner
- AI systems used for social scoring
- those that exploit information or predictions and a person or group of persons in order to target their vulnerabilities



High-risk examples of AI include:

- systems which establish priority in the dispatching of emergency services
- systems determining access to or assigning people to educational institutes
- recruitment algorithms
- those that evaluate credit worthiness
- those for making individual risk assessments
- crime-predicting algorithms

Accountability in Science



the author list until the full paper deadline. After that, no changes will be permitted for any reason, including for the camera-ready version.

2. All authors are required to login and fill out a user information form in CMT by **Jun 06, 2020 01:00 PM PDT.** Because of the rapid growth of NeurIPS, all authors and co-authors are expected to be available to review papers, if asked to do so. If all co-authors do not register and enter their information, their submission may be desk rejected.

3. In order to cope with the growing number of submissions, this year we will adopt an **"early desk-reject"** process involving only Area Chairs and Senior Area Chairs. Area Chairs will be responsible for identifying papers that are very likely to be rejected, and Senior Area Chairs will cross check the selections. These papers will not be further reviewed, and authors will be notified immediately.

4. Authors need to declare if a previous version of their submission was rejected at any peer-reviewed venue within the past 12 months, and, if so, summarize the changes to the current version. This information should be entered into CMT during the submission process.

5. In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes.

6. Authors are required to provide an explicit disclosure of funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work) that could result in conflicts of interest. This section should be added to the camera-ready version of accepted papers. More information can be found here.

https://neurips.cc/Conferences/2020/CallForPapers

Issues to look out for

- Privacy
- Consent
- Reliability and Efficacy
- Human Rights
- Fairness and Equity
- Contestability
- Accountability
- Explainability and Transparency

Q&A