AI for the Common Good?

Pitfalls, Challenges, and Ethics Pen-Testing

Bettina Berendt
DTAI, Dept. Computer Science, KU Leuven

https://people.cs.kuleuven.be/~bettina.berendt/

KIAS AI, Ethics & Society Conference,
May 9, 2019, Edmonton, AL

Part 1:
Good for all?!

Part 1 of the talk is based on

Berendt, B. (2019). AI for the Common Good?! Pitfalls, challenges, and Ethics Pen-Testing. *Paladyn. Journal of Behavioral Robotics, 10,* 44-65.

https://www.degruyter.com/view/j/pjbr.2019.10.issue-1/pjbr-2019-0004/pjbr-2019-0004.xml

“Alignment with human values” or “good for some” – not good enough!

Even lethal autonomous weapons are “good” for some!

➔ Typical AI Ethics requirement: *Common* Good (CG)



“Alignment with human values” or “good for some” – not good enough!

Even lethal autonomous weapons are “good” for some!

➔ Typical AI Ethics requirement: *Common* Good (CG)



- Examples from the Asilomar Principles:
  - Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.
  - Shared Benefit: AI technologies should benefit and empower as many people as possible.
  - Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- Similar in Ethics Codes for Comp.Sci., AI by ACM, IEEE, …

How >2000 years of philosophy can help us question Ethical AI

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

How >2000 years of philosophy can help us question Ethical AI

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

Dimensions of / challenges for CG	
Benefit: "utility" or values beyond this? Who is (not) in the *community/society*?	
Substantive vs. procedural	
Communal vs. distributive	
Interests and facilities	

How >2000 years of philosophy can help us question Ethical AI

The Common Good (CG): that which benefits society as a whole

"*that*": The relevant interests and facilities that serve these interests together constitute the CG and serve as a shared standpoint for political deliberation [among the members of a political community].
These facilities include social institutions, rights, human rights.

Dimensions of / challenges for CG	Lead questions for AI for the CG
Benefit: "utility" or values beyond this? Who is (not) in the *community/society*?	Q1: What is the problem?
Substantive vs. procedural	Q2: Who defines the problem?
Communal vs. distributive	Q4: What are important side effects and dynamics?
Interests and facilities	Q3: AI as engineering → a facility → What is the role of knowledge?

An example

Drugs & uses of data science

Who defines the good / the problem?

- Example: What is “the drug problem”?
- And who was involved in formulating it, and under what constraints?



Who defines the good / the problem?

- Example: What is “the drug problem”?
- And who was involved in formulating it, and under what constraints?

- People use drugs
- Poor people use drugs
- Drug users commit crimes
- Drug users become homeless, ill, …
- Drug users die
- There aren’t enough drugs available.



What role does knowledge play?

Imparting knowledge …

- is often *overrated* by AIers: Is it *suitable* to addressing the problem?

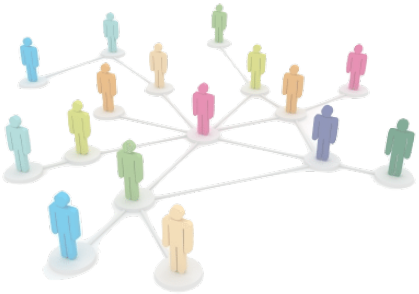




What role does knowledge play?

Imparting knowledge …

- is often *overrated* by AIers: Is it *suitable* to addressing the problem?





- is often *underrated* by AIers (when compared to robots)
  - "War on drugs" (Nixon, 1971)
  - AI recommendations: Current debate on filter bubbles

What role does knowledge play?

Imparting knowledge …

- is often *overrated* by AIers: Is it *suitable* to addressing the problem?





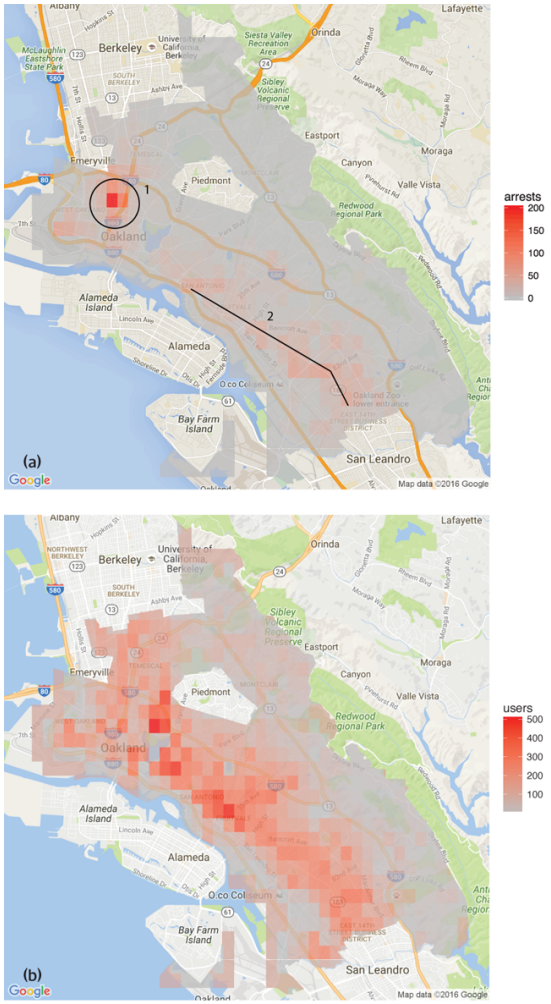
- is often *underrated* by AIers (when compared to robots)
  - "War on drugs" (Nixon, 1971)
  - AI recommendations: Current debate on filter bubbles
- is not always a *good* thing
- may *interfere* with autonomy & with society's duty to care

What are the side-effects?

- Accumulation of personal data
  – privacy violations?

What are the side-effects?

- Accumulation of personal data – privacy violations?
- Conservative bias of machine learning – discrimination?



What are the side-effects?

- Accumulation of personal data – privacy violations?
- Conservative bias of machine learning – discrimination?
- Affirmative bias – reinforcing what you criticised?

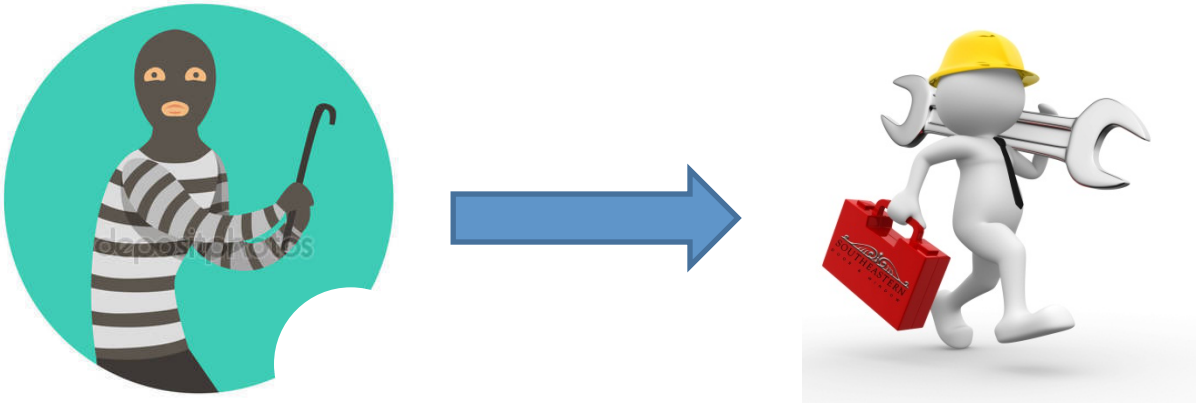




Proposal

What is pen-testing?

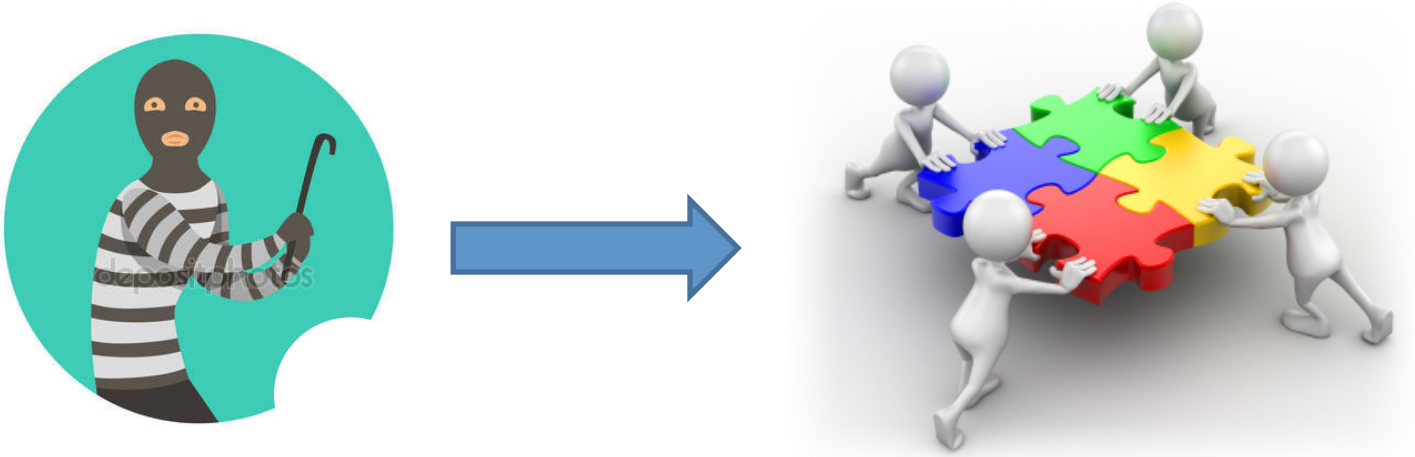
A **penetration test**, colloquially known as a **pen test**, is an authorised simulated attack on a computer system that looks for security weaknesses, potentially gaining access to the system's features and data.



Note: No system is 100% secure. The point is not to pass all pen-tests (this is impossible), the point is to get better through trying.

The idea of ethics pen-testing

An **ethics pen test**, is an authorised simulated attack on a design or artefact that, guided by lead questions, looks for ethics weaknesses, potentially demonstrating how the operation of the design can damage a relevant notion of the Good or Common Good.



Note: No system is 100% good. The point is not to pass all pen-tests (this is impossible), the point is to get better through trying.

Reality check

Reality check: Aren't all these Qs obvious and dealing with them part of standard practices?

- Exploratory study of all 99 hyperlinked contributions to 4 current major venues on AI for Good, Data Science for Social Good, …

- Definition of "the Good": mainly substantive
- Q1: Alternative goals?
- Q2: Different stakeholders, description of how their perspectives and needs were assessed?
- Q3: The role of knowledge?
- Q4: Dynamics?

Reality check: Aren't all these Qs obvious and dealing with them part of standard practices?

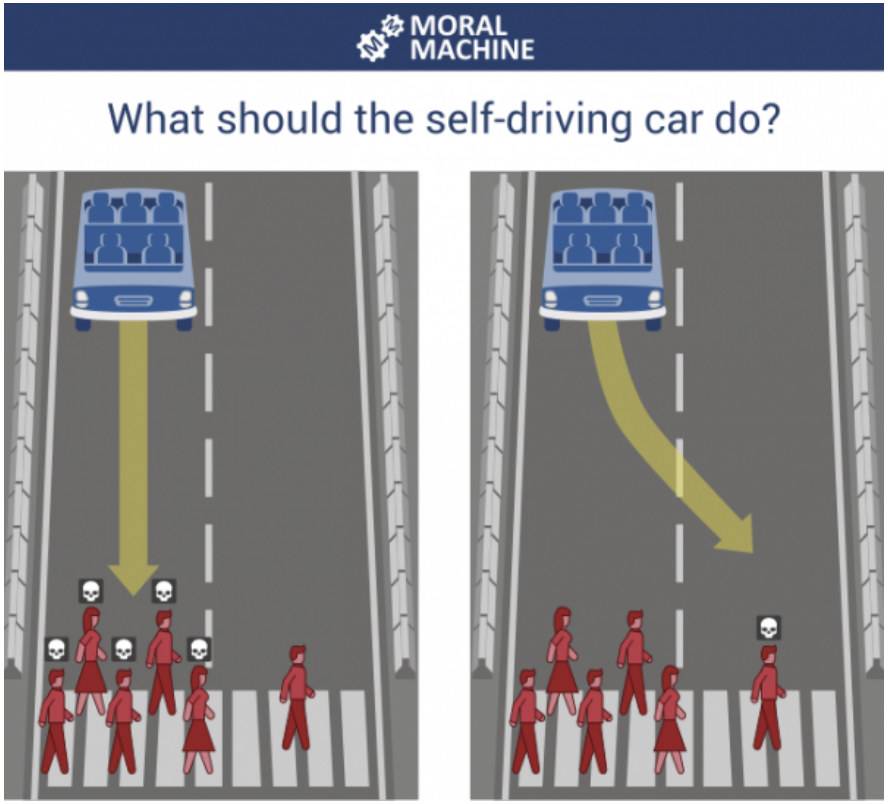
- Exploratory study of all 99 hyperlinked contributions to 4 current major venues on AI for Good, Data Science for Social Good, …

- Definition of "the Good": mainly substantive
- Q1: Alternative goals?
- Q2: Different stakeholders, description of how their perspectives and needs were assessed?
- Q3: The role of knowledge?
- Q4: Dynamics?

} ✘ or in very few con-tribu-tions

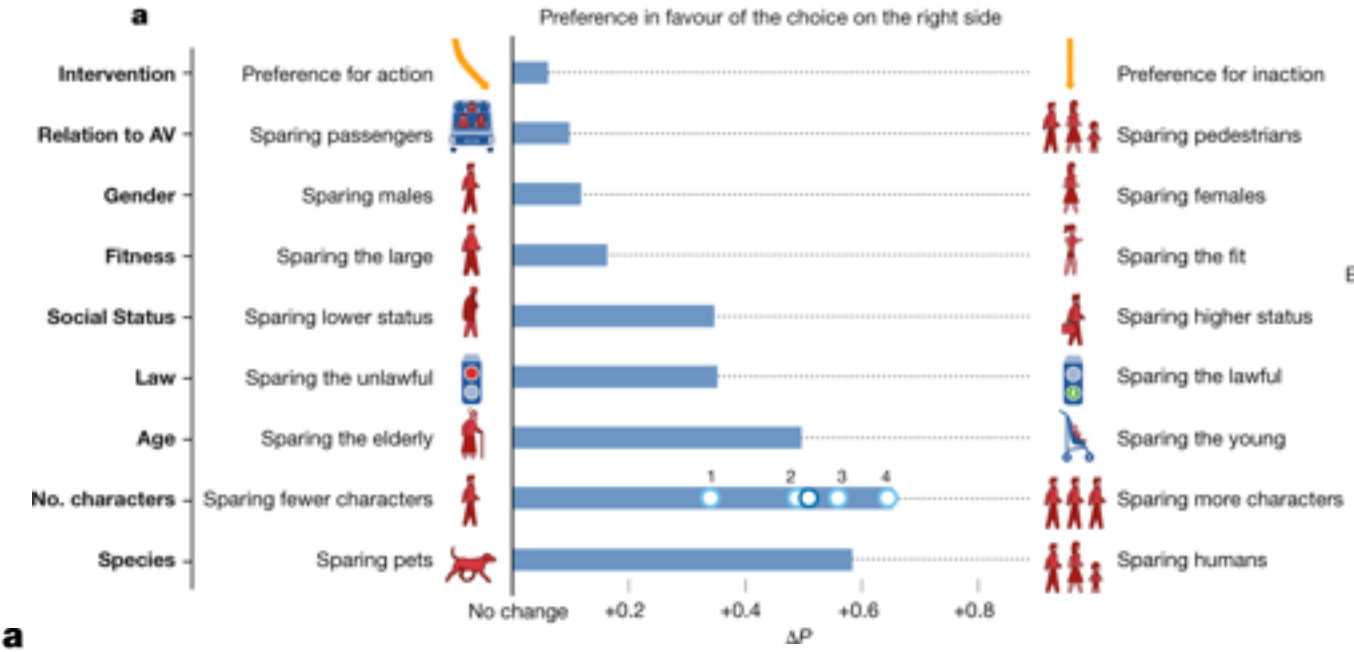
Part 2:
Let’s ethics-pen-test
AI & ethics discussions!

The Moral Machine Experiment

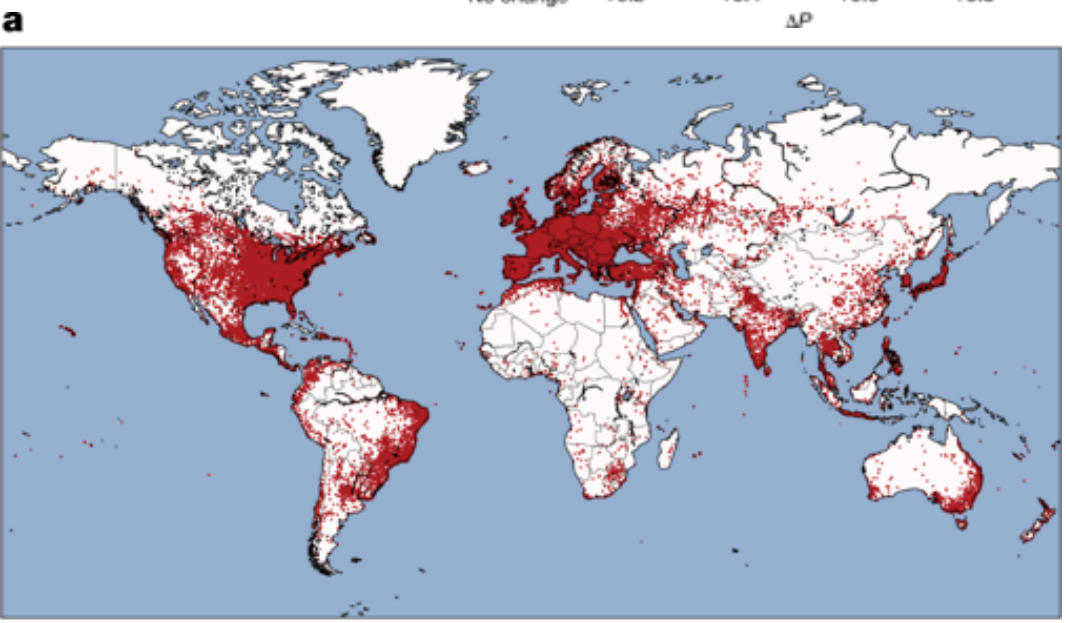


Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563,* 59-64 https://www.nature.com/articles/s41586-018-0637-6

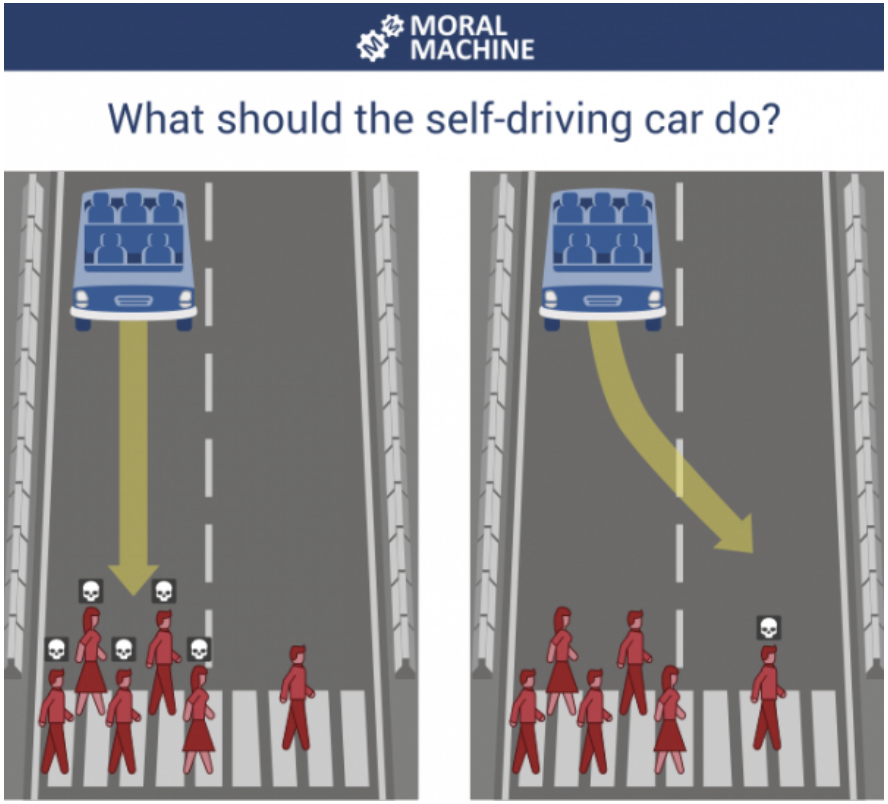
Moral Machine: Participants, (some) results







Questioning the Moral Machine Experiment



Some questions
Q1: What is the problem?
Q2: Who defines the problem? Whose voice is NOT heard?
Q3: What is the role of knowledge?
Q4: What are important side effects and dynamics?
Q5: What is being assumed about the technology? Q6: What is being assumed about the society? Q7: Are there alternatives to these assumed worlds?

Some thoughts on “militant democracy”, or: Why don’t we just crowdsource The Universal Declaration of Human Rights 2.0?

… from the Conclusion of the Awad et al. Moral Machine experiment paper (p. 63):

“Indeed, we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences.

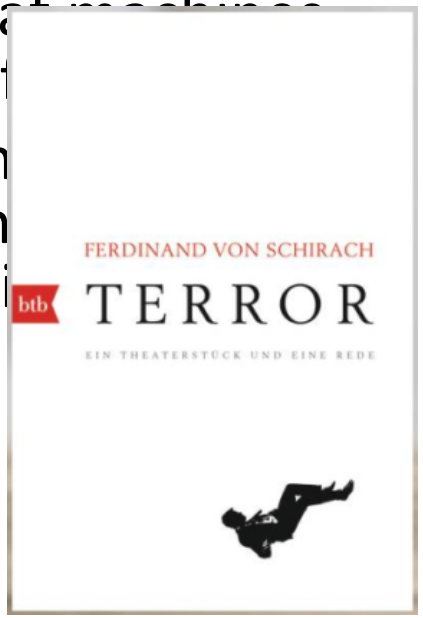
We might not reach universal agreement: even the strongest preferences expressed through the Moral Machine showed substantial cultural variations, and our project builds on a long tradition of investigating cultural variations in ethical judgments.”

Some thoughts on "militant democracy", or: Why don't we just crowdsource The Universal Declaration of Human Rights 2.0?

… from the Conclusion of the Awad et al. Moral Machine experiment paper (p. 63):

"Indeed, we can embrace the challenges of machine ethics as a unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure tha
unlike humans, unerringly follow these moral pref

We might not reach universal agreement: even th
preferences expressed through the Moral Machin
substantial cultural variations, and our project bui
tradition of investigating cultural variations in eth
judgments."



von Schirach, F. (2016). Terror. btb Verlag (eBook)Verlagsgruppe Random House. English version: Terror, Faber & Faber 2017.

Thank you!