Ethics for AI Developers

a crash course

6 November 2024

Human-centred Machine Learning course

Giovanni Sileno
g.sileno@uva.nl

University of Amsterdam

Overview

my talk consists of:

- a brief & broad overview on concepts which are relevant to ethics
- a few practical examples of problems that you may encounter as AI designer/developer

Why is this relevant?

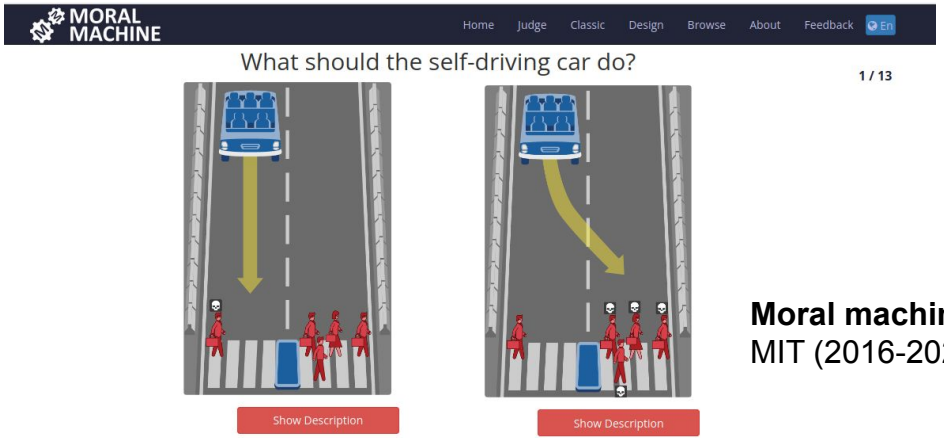
- students going into computational tracks (consequently becoming tech researchers/experts) may have been not explicitly exposed to topics related to ethics

risk of naive if not dangerous misunderstandings

Why is this relevant?

- students going into computational tracks (consequently becoming tech researchers/experts) may have been not explicitly exposed to topics related to ethics

➡ risk of naive if not dangerous misunderstandings



Moral machine, MIT (2016-2020)

Why is this relevant?

- students going into computational tracks (consequently becoming tech researchers/experts) may have been not explicitly exposed to discussions concerning ethics or related topics

➡ risk of naive if not dangerous misunderstandings

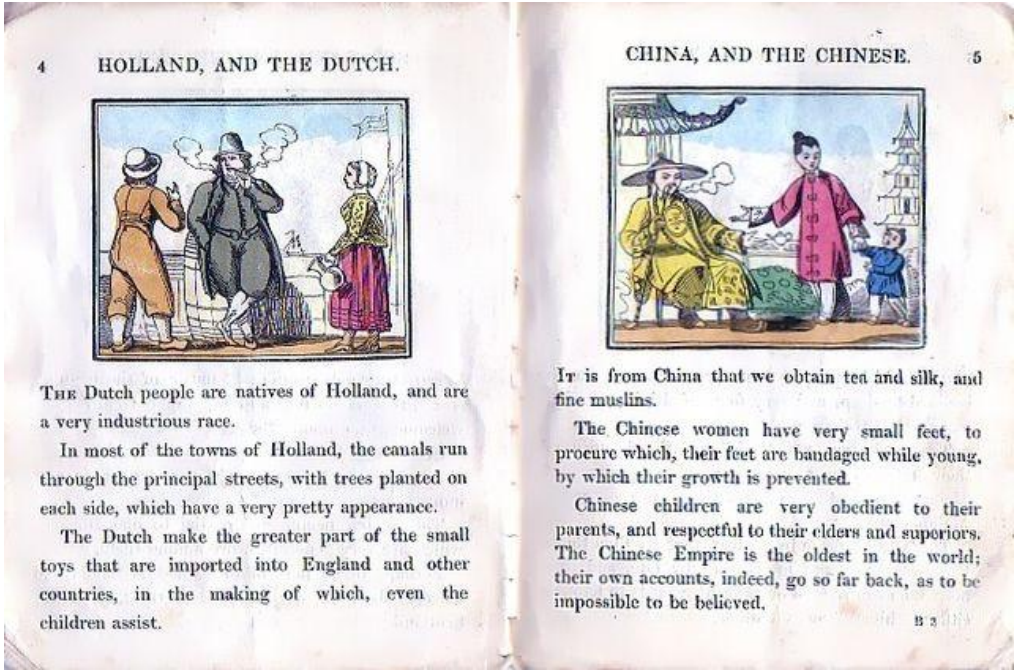
- on the other hand, this is not “rocket science”, these concepts are accessible to all of us **just because we are humans, social beings!**

➡ getting acquainted with ethics increases our *“human capital”*

Morals & co.

Morals & co.

- ***“mores”***, latin word for *social norms*, *costumes*, *habits* of a community



4 HOLLAND, AND THE DUTCH.

THE Dutch people are natives of Holland, and are a very industrious race.

In most of the towns of Holland, the canals run through the principal streets, with trees planted on each side, which have a very pretty appearance.

The Dutch make the greater part of the small toys that are imported into England and other countries, in the making of which, even the children assist.

CHINA, AND THE CHINESE. 5

IT is from China that we obtain tea and silk, and fine muslins.

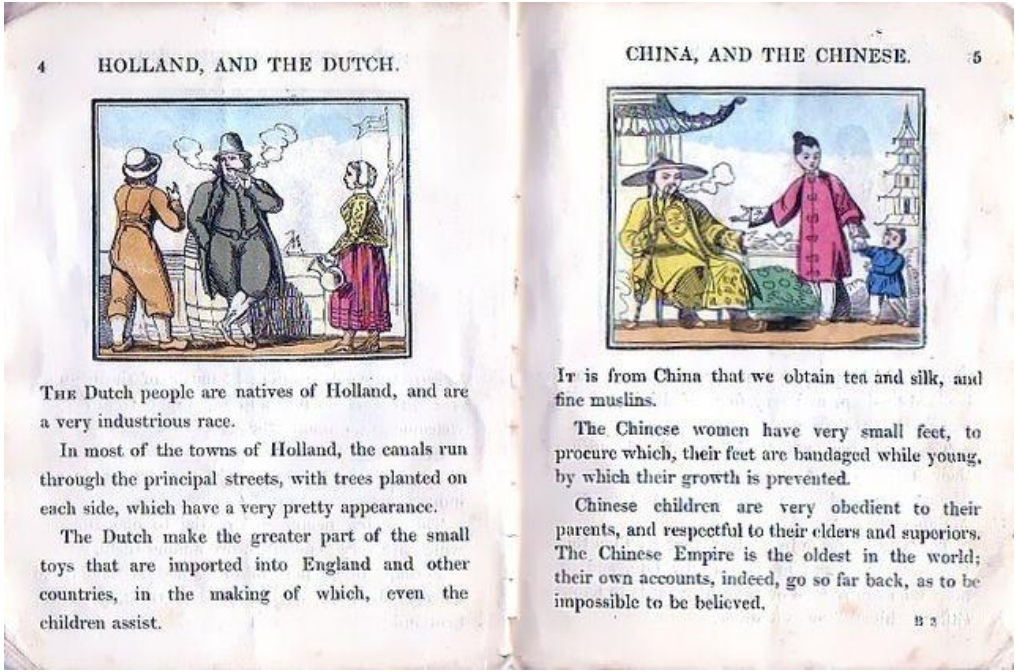
The Chinese women have very small feet, to procure which, their feet are bandaged while young, by which their growth is prevented.

Chinese children are very obedient to their parents, and respectful to their elders and superiors. The Chinese Empire is the oldest in the world; their own accounts, indeed, go so far back, as to be impossible to be believed.

B 3

Morals & co.

- ***"mores"***, latin word for *social norms*, *costumes*, *habits* of a community
- ***"ethos"*** is its greek equivalent (*character)*



4 HOLLAND, AND THE DUTCH.

THE Dutch people are natives of Holland, and are a very industrious race.

In most of the towns of Holland, the canals run through the principal streets, with trees planted on each side, which have a very pretty appearance.

The Dutch make the greater part of the small toys that are imported into England and other countries, in the making of which, even the children assist.

CHINA, AND THE CHINESE. 5

IT is from China that we obtain tea and silk, and fine muslins.

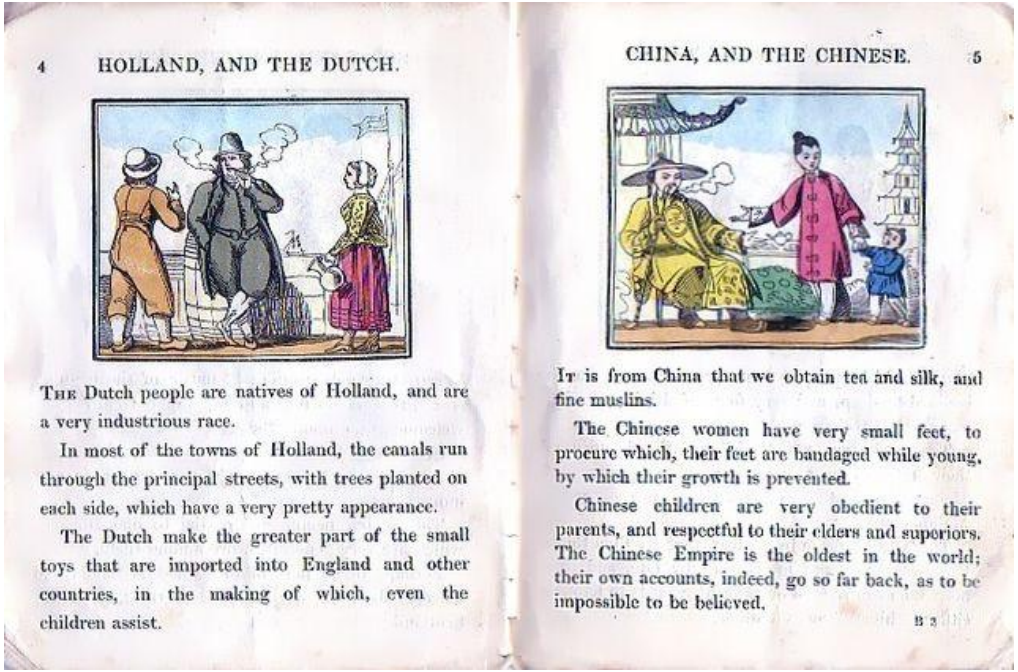
The Chinese women have very small feet, to procure which, their feet are bandaged while young, by which their growth is prevented.

Chinese children are very obedient to their parents, and respectful to their elders and superiors. The Chinese Empire is the oldest in the world; their own accounts, indeed, go so far back, as to be impossible to be believed.

B 3

Morals & co.

- ***"mores"***, latin word for *social norms*, *costumes*, *habits* of a community
- ***"ethos"*** is its greek equivalent (*character)*



4 HOLLAND, AND THE DUTCH.

THE Dutch people are natives of Holland, and are a very industrious race.

In most of the towns of Holland, the canals run through the principal streets, with trees planted on each side, which have a very pretty appearance.

The Dutch make the greater part of the small toys that are imported into England and other countries, in the making of which, even the children assist.

CHINA, AND THE CHINESE. 5

IT is from China that we obtain tea and silk, and fine muslins.

The Chinese women have very small feet, to procure which, their feet are bandaged while young, by which their growth is prevented.

Chinese children are very obedient to their parents, and respectful to their elders and superiors. The Chinese Empire is the oldest in the world; their own accounts, indeed, go so far back, as to be impossible to be believed.

B 3

Difference between **descriptive** and **prescriptive** norms!

Morals & co.

- **individuals behave according and against *mores***
  - generally no problem for ***descriptive norms*** (if not for identity matters, and so “qualification”/categorization processes)
  - ***prescriptive norms*** address correct/incorrect behaviour
    ⇒ grounds for *judgment*

Morals & co.

- **individuals behave according and against *mores***
  - for identitarian purposes, ***descriptive norms may become prescriptive!***

punk vs traders: find the difference

Morals & co.

- individuals behave according and against *mores*
- **behaviour is moral** if *justifiable according to a **moral standard*** (morality comes always with an evaluative framework)
- moral judgments are given from a collective standpoint (the one associated to the mores)

adultery for puritans (The Scarlet Letter, Hawthorne)

Whose *mores*?

- *mores* were defined above collectively, but one could apply the same concept at individual level:

  "I have my habits, and my ways to evaluate behaviour as good

Whose *mores*?

- *mores* were defined above collectively, but one could apply the same concept at individual level:

  "I have my habits, and my ways to evaluate behaviour as good

- yet, organizations or communities can be seen as “individuals” (collective agencies)! eg. *at UvA we do like that*.

Normality vs normativity

- descriptive/prescriptive evaluative frameworks can be reinterpreted in ***agentive*** terms:

normality	is (not)	believe (not) to be	the case
normativity	ought (not) to be	desire (not) to be	

Normality vs normativity?

The boundary between descriptive and prescriptive norms is a delicate one.

Suppose you are a shop owner:

- you don't want fraudsters in your shop

Normality vs normativity?

The boundary between descriptive and prescriptive norms is a delicate one.

Suppose you are a shop owner:

- you don't want fraudsters in your shop
- fraudsters typically dress a red tie

Normality vs normativity?

The boundary between descriptive and prescriptive norms is a delicate one.

Suppose you are a shop owner:

- you don't want fraudsters in your shop
- fraudsters typically dress a red tie

➡ you don't want people with red tie in your shop

Normality vs normativity?

The boundary between descriptive and prescriptive norms is a delicate one.

Suppose you are a shop owner:

- you don't want fraudsters in your shop
- fraudsters typically dress a red tie

➡ you don't want people with red tie in your shop

Is this good? is it right? for whom?

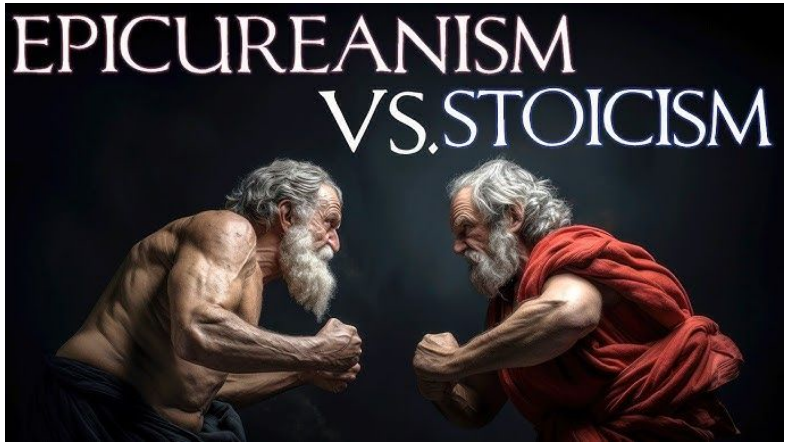
The relation between ethics and morals

Ethics

- In philosophy the word **ethics** was introduced rather to identify a specific branch of philosophical discourse (like *ontology*, *epistemology*, *aestethic*, ...)

Ethics

- In philosophy the word **ethics** was introduced rather to identify a specific branch of philosophical discourse (like *ontology*, *epistemology*, *aestethic*, ...)
- Ethical schools (stoics, epicureans, skeptics, ...) were seen as providing different ***principles*** to define what would be considered to be good or bad



Ethics

- In philosophy the word **ethics** was introduced rather to identify a specific branch of philosophical discourse (like *ontology*, *epistemology*, *aestethic*, ...)
- Ethical schools (stoics, epicureans, skeptics, ...) were seen as providing different ***principles*** to define what would be considered to be good or bad

- There was no “winner”: no ethics school is better than the other

Ethics

- In philosophy the word **ethics** was introduced rather to identify a specific branch of philosophical discourse (like *ontology*, *epistemology*, *aestethic*, ...)
- Ethical schools (stoics, epicureans, skeptics, ...) were seen as providing different ***principles*** to define what would be considered to be good or bad

- There was no “winner”: no ethics school is better than the other
- Rather, individuals are seen as forming their own morality on the basis of mores ***and*** ethics

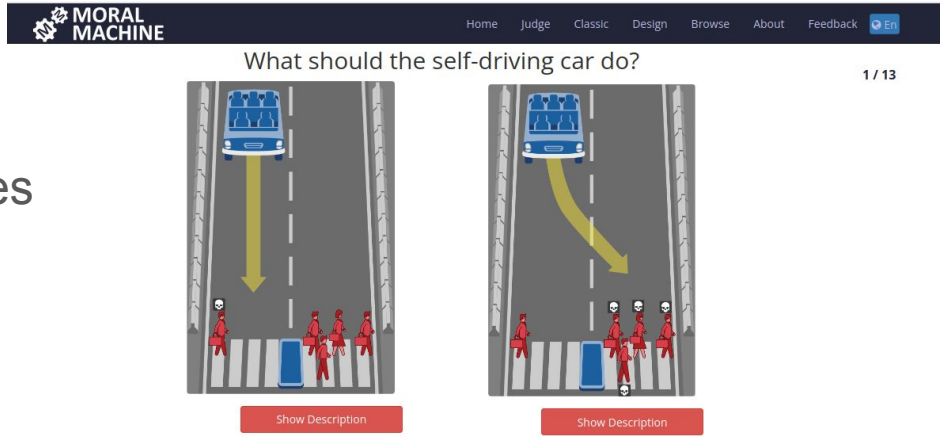
Ethics vs mores

- Yet, a common point of all these schools is that ***one may behave ethically, even going against mores*** (collective morals)

Ethics vs mores

- Yet, a common point of all these schools is that ***one may behave ethically, even going against mores*** (collective morals)

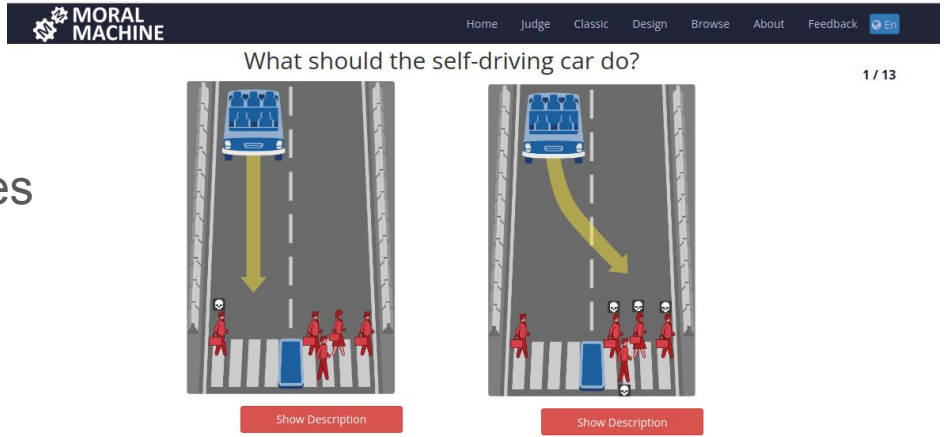
Let's aggregate global preferences over moral dilemmas.
Is the resulting decision-making "moral"? Is it "ethical"?



Ethics vs mores

- Yet, a common point of all these schools is that ***one may behave ethically, even going against mores*** (collective morals)

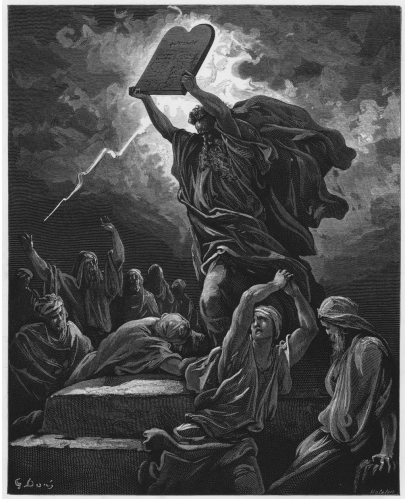
Let's aggregate global preferences over moral dilemmas.
Is the resulting decision-making "moral"? Is it "ethical"?

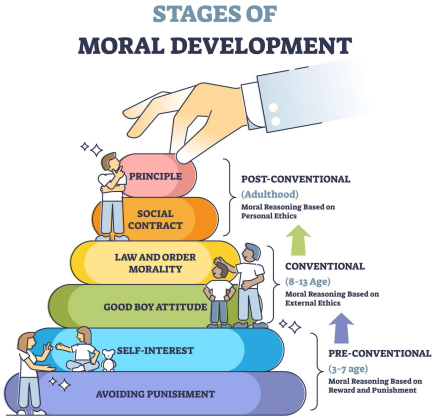


The question of where morality comes from require further investigation…

“Natural” morals…

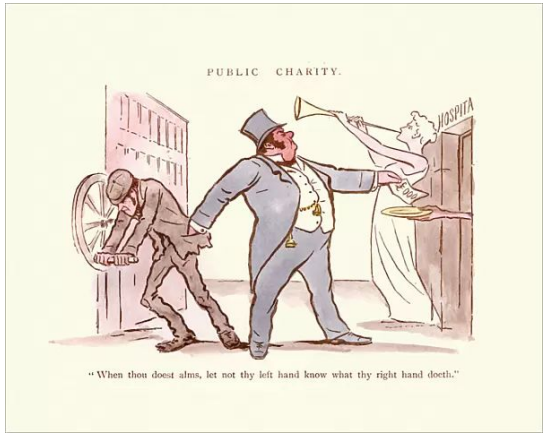
- the simplest way to look at morals is as a **given set** of acceptable and unacceptable behaviours
- examples:
  - eternal truths expressed by religions
  - natural law, as embedded within ourselves
  - innate moral judgments observed in babies





…vs morals as socially constructed…

- morals map to **societal layers** (*niches, strata*, *classes*, *communities*)
- **behaving "well" is source/signal of (good) reputation within a class… but not necessarily in others!**
- see e.g. charity events for higher-classes, or mafia/gang-like practices



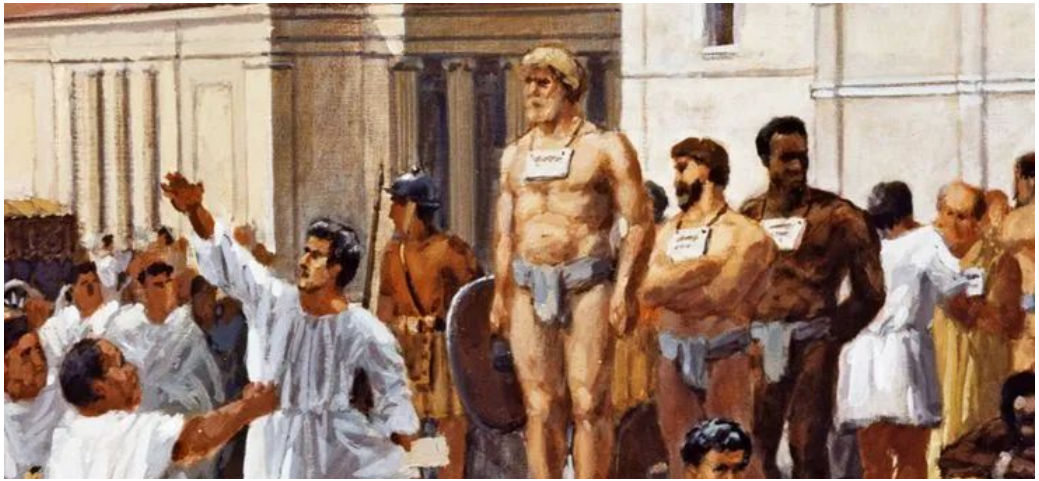
…vs morals as applied ethics…

- another way to look at morals would be as a set of acceptable/unacceptable behaviour ***derived from ethical principles***

(stoics applying emotional detachment)

…vs morals as *materially* constructed!

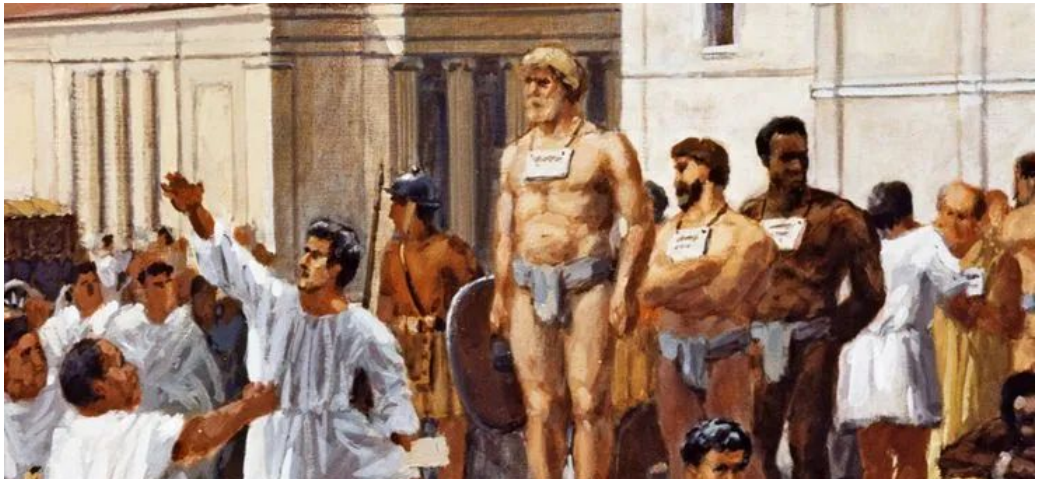
- but then why slavery exists?



- slavery is accepted as long as it sustains (the power of) the people in power.

…vs morals as *materially* constructed!

- but then why slavery exists?



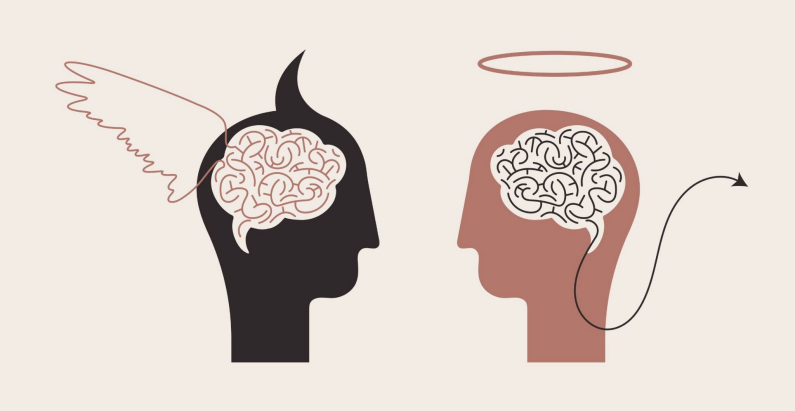
- slavery is accepted as long as it sustains (the power of) the people in power.

→ **morals are socio-economically and historically conditioned**

Norms and values?

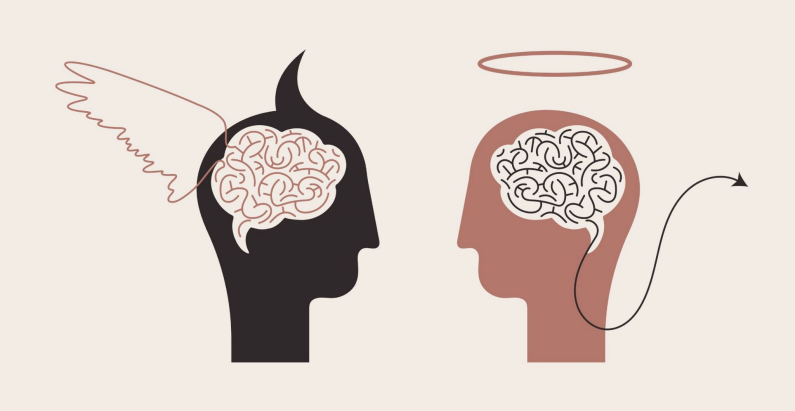
Norms vs values in ethics

- There is a distinction between **right/wrong** (norms) vs **good/bad** (values)



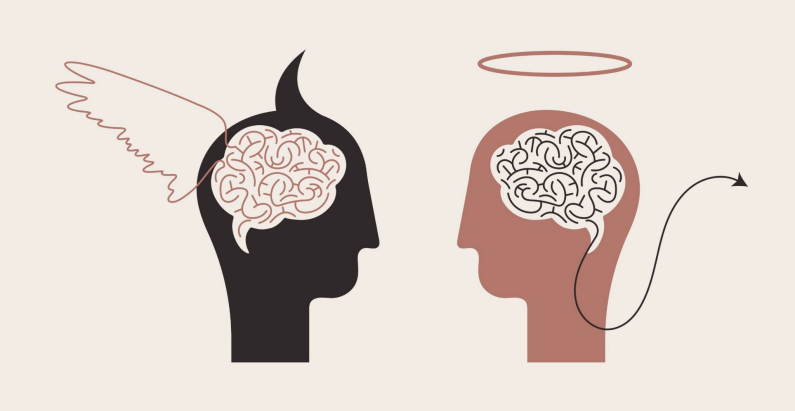
Norms vs values in ethics

- There is a distinction between **right/wrong** (norms) vs **good/bad** (values)
- **Deontology** studies what features make an action right or wrong.



Norms vs values in ethics

- There is a distinction between **right/wrong** (norms) vs **good/bad** (values)
- **Deontology** studies what features make an action right or wrong.
- **Axiology** studies what makes things good (or have value) or bad (or have disvalue or less value) ⇒ **theories of value**



Norms vs values in ethics

- There is a distinction between **right/wrong** (norms) vs **good/bad** (values)
- **Deontology** studies what features make an action right or wrong.
- **Axiology** studies what makes things good (or have value) or bad (or have disvalue or less value) ⇒ **theories of value**

(What comes first?)

Theory of value (philosophy)

Philosophy has long been debating on the distinction between

- **intrinsic values** (something good in itself)
    - pleasure/no pain, appetites, needs, desires?
- **instrumental/final values** (something good for something else)
    - eg. money

Theory of value (economics)

- **Labour**
- Utility (use)
- Exchange
- ...

how much its production costs
cf. **cryptocurrencies** as **blockchain**

Theory of value (economics)

- Labour
- **Utility (use)**
- Exchange
- ...

how much benefit it brings
cf. **software**

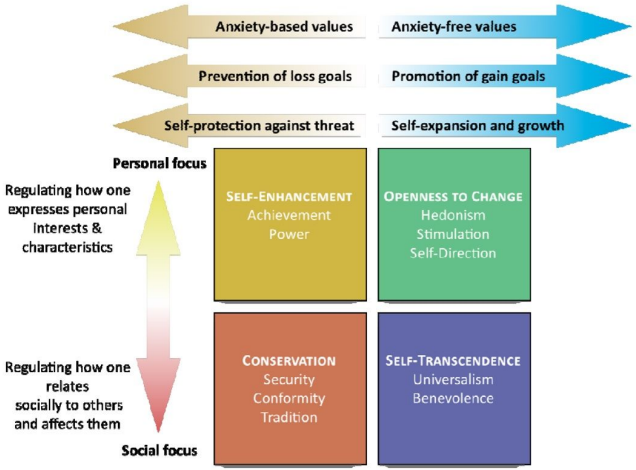
Theory of value (economics)

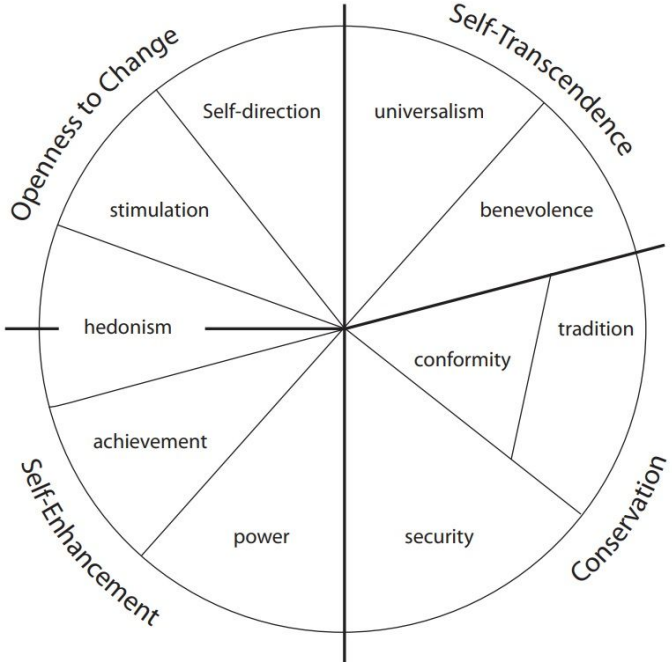
- Labour
- Utility (use)
- **Exchange**
- ...

how much others pay for it
cf. **NFT (non-fungible tokens)**

Theory of value (psychology)

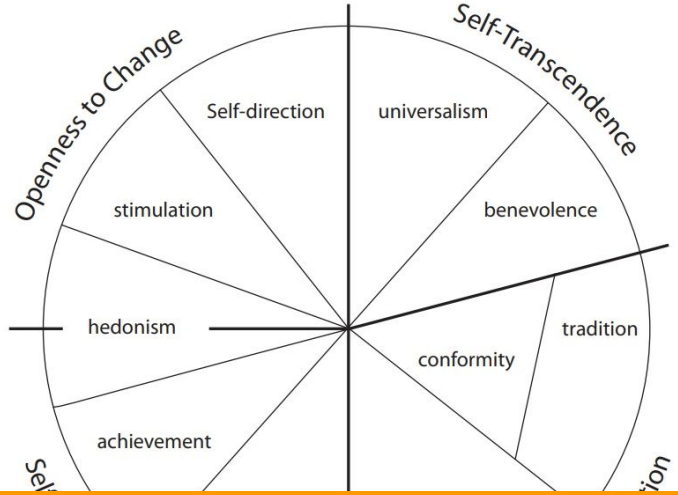
- Values are abstract concepts, but how people express them may be categorized
- eg. Schwartz's value framework





Theory of value (psychology)

- Values are abstract concepts, but how people express them may be categorized
- eg. Schwartz's value framework



several papers are using it to interpret LLMs chatbots.
but what are they capturing?

Theory of value (law)

- Judges embody the law
- For common cases, norms provide sufficient guidance.
- For *hard cases*, values fits into, used for balancing between opposing interests/norms, eg. environmental protection vs economic opportunity

Theory of value (law)

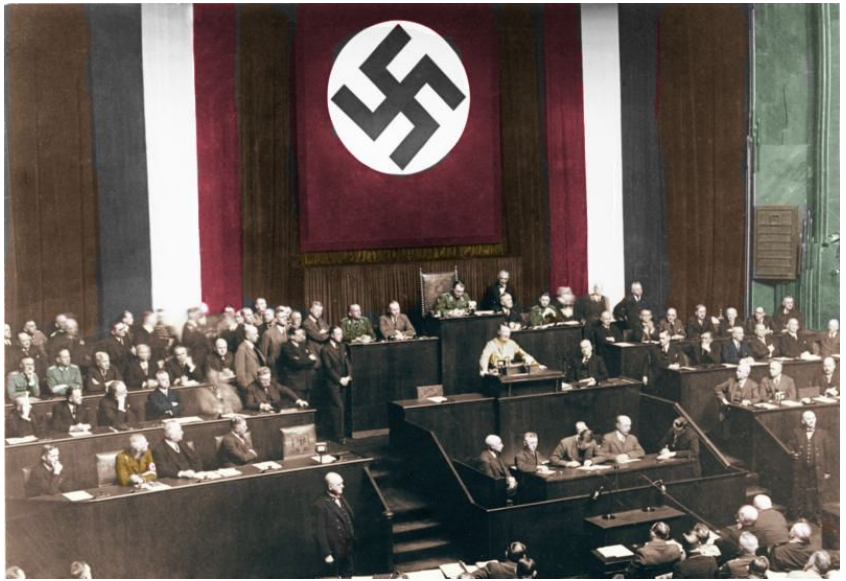
- Increasing reference to human rights and the rule of law in legislation, eg.

  The EU's founding values are 'human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities'; Article 3 Treaty European Union (TEU)

- Talking about values in laws pushes the legislator to take some stance towards them, but cannot guarantee effective application…

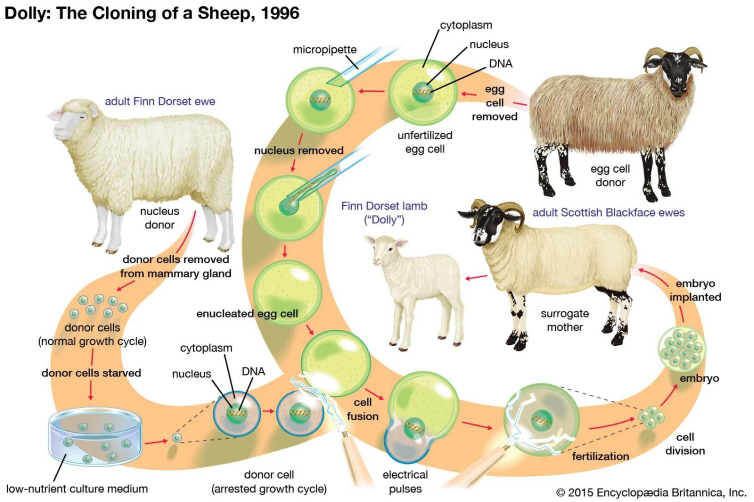
Beyond norms: when laws are written

- There is a distinction between ***state of right*** vs ***state with right/with laws***



Beyond norms: when laws are not there

- When laws are not there (yet), ethics is deemed to play a role (eg. bioethical committees).



Beyond norms: when laws are not there

- When laws are not there (yet), ethics is deemed to play a role (eg. bioethical committees).
- Part of the Ethics in AI track started from a similar perspective.
- It is however argued whether most of the problems are not covered in any case by existing laws (eg. product liability).

Core components of ethical frameworks

Main ethical constructs

- **deontic**: you perform an action because it is due

Main ethical constructs

- **deontic**: you perform an action because it is due
- **consequentialist**: you pursue an end, and for doing so you perform an action

Main ethical constructs

- **deontic**: you perform an action because it is due
- **consequentialist**: you pursue an end, and for doing so you perform an action
- **virtue ethics**: you behave as a virtuous person would do (exemplary role-models)

Main ethical constructs

- **deontic**: you perform an action because it is due
- **consequentialist**: you pursue an end, and for doing so you perform an action
- **virtue ethics**: you behave as a virtuous person would do (exemplary role-models)
- **discourse ethics**: you behave as it is acceptable (in argumentation) to all those affected by the consequences of your actions

Main ethical constructs

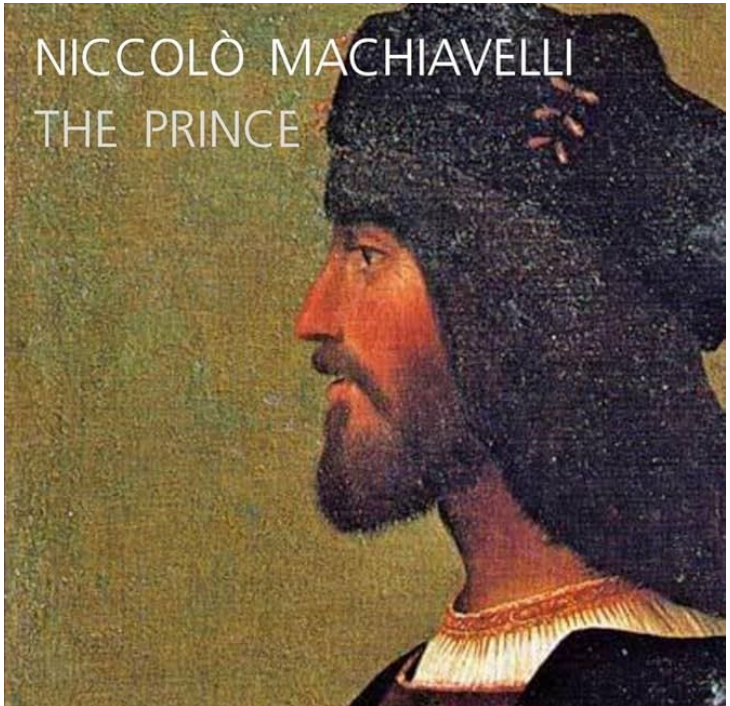
- **deontic**: you perform an action because it is due
- **consequentialist**: you pursue an end, and for doing so you perform an action
- **virtue ethics**: you behave as a virtuous person would do (exemplary role-models)
- **discourse ethics**: you behave as it is acceptable (in argumentation) to all those affected by the consequences of your actions

These express different forms of normativity.

Mingers, John and Walsham, Geoff. Towards Ethical Information Systems: The Contribution of Discourse Ethics. MIS Quarterly, 34 (4) (2010)

Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

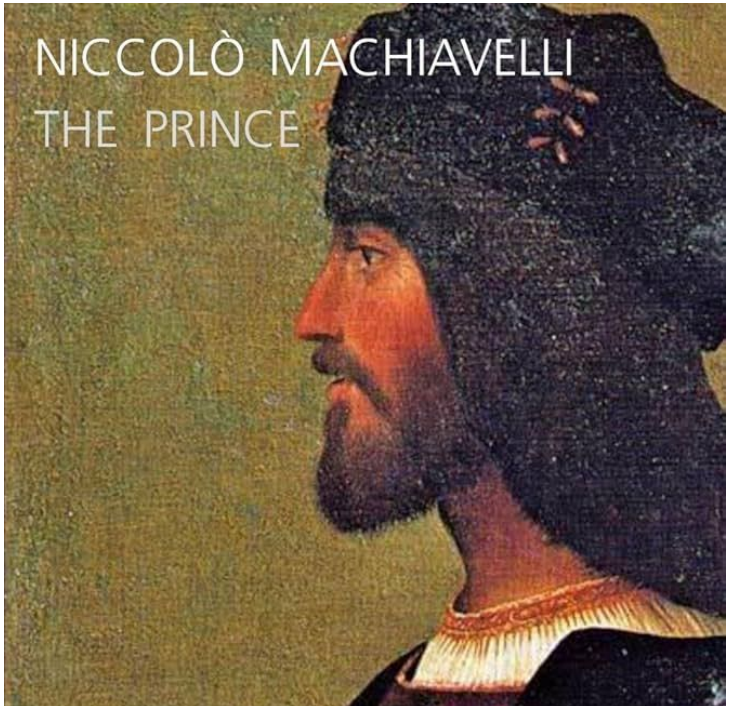


“the ends justify the means”

Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

instrumental reasoning patterns…



“the ends justify the means”

Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

- consequentialism directly connects with **utilitarianism**…
  - (consequences are evaluated in terms of *utility function*)
- …as well as with **optimization** methods:
  - (you *maximize* utility, expressed eg. as an aggregated reward function)

Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

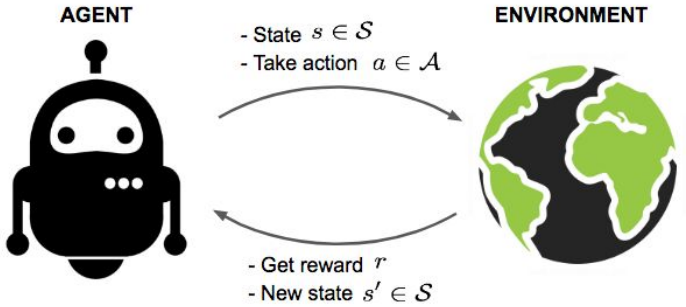
- if the core policy of the agent is seen as its identity, consequentialism can be seen as a *meta-identity*, as it drives a modification of the agent's identity.

Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

- if the core policy of the agent is seen as its identity, consequentialism can be seen as a *meta-identity*, as it drives a modification of the agent's identity.

→ **reinforcement learning** as well as **pure economic rationality** (maximizing profit/minimizing losses) takes a specific normative stance.



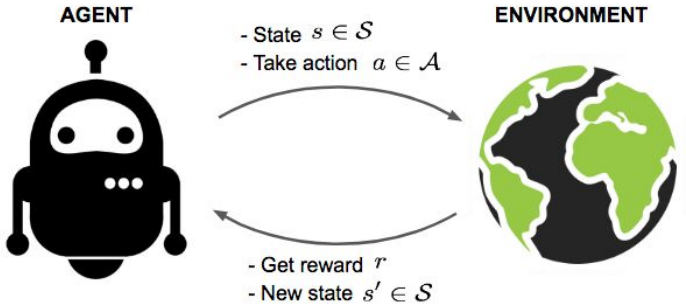
Consequentialism & co.

consequentialist: *you pursue an end, and for doing so you perform an action*

- if the core policy of the agent is seen as its identity, consequentialism can be seen as a *meta-identity*, as it drives a modification of the agent's identity.

→ **reinforcement learning** as well as **pure economic rationality** (maximizing profit/minimizing losses) takes a specific normative stance.

Is this adequate to behave ethically?



An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

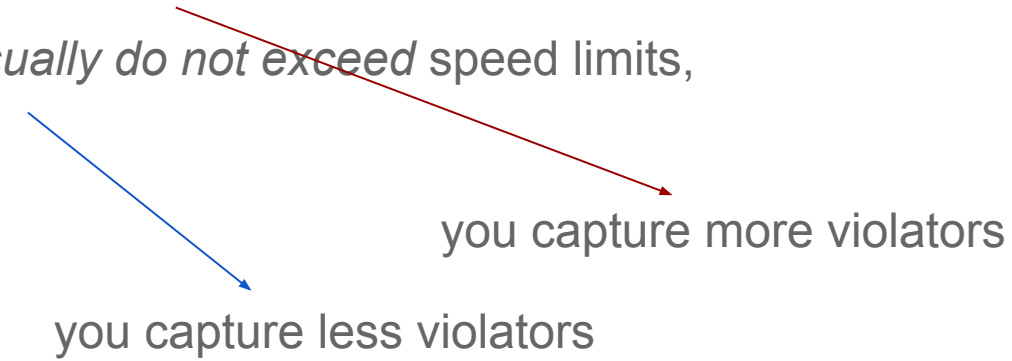
- Where to put the speed cameras?
  - areas in which people *usually exceed* speed limits,
  - areas in which people *usually do not exceed* speed limits,
  - …?

(You have only a limited amount of speed cameras)

An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

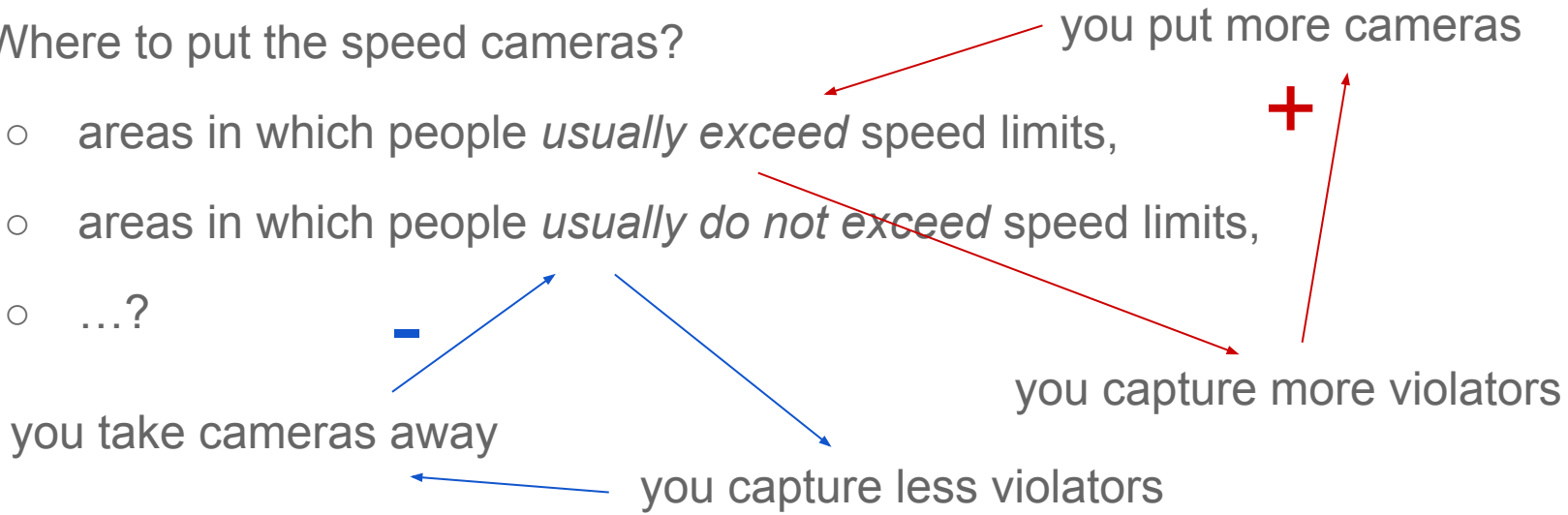
- Where to put the speed cameras?
  - areas in which people *usually exceed* speed limits,
  - areas in which people *usually do not exceed* speed limits,
  - …?



An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

- Where to put the speed cameras?
  - areas in which people *usually exceed* speed limits,
  - areas in which people *usually do not exceed* speed limits,
  - …?



An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

- Where to put the speed cameras?
  - areas in which people *usually exceed* speed limits,
  - areas in which people *usually do not exceed* speed limits,

you put more cameras

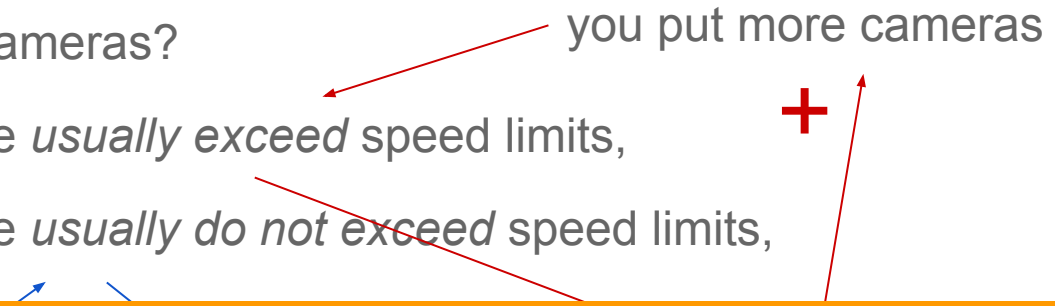
+

This is an example of *self-fulfilling prophecy*:
if we look for what we expect (only) where we expect,
then we'll (only) see what we expect.

An exercise of monitoring design

In your city the number of car accidents is rapidly increasing. You are asked to find who is exceeding the speed limits (to punish violators).

- Where to put the speed cameras?
  - areas in which people *usually exceed* speed limits,
  - areas in which people *usually do not exceed* speed limits,



This is an example of *self-fulfilling prophecy*:
if we look for what we expect (only) where we expect,
then we'll (only) see what we expect.

blindness consequent to blind optimization attitude!

An exercise of monitoring design (2)

You are asked to help the police to identify venues of synthetic drug production (an actual Master IS project a few years ago).

An exercise of monitoring design (2)

You are asked to help the police to identify venues of synthetic drug production (an actual Master IS project a few years ago).

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.
- Agriculture is not rentable at the moment, barn owners may be more lenient in checking who is renting their barn

An exercise of monitoring design (2)

You are asked to help the police to identify venues of synthetic drug production (an actual Master IS project a few years ago).

- Synthetic drug is usually produced in barns rented for a few months, then abandoned, and chemical residuals thrown in the canals.
- Agriculture is not rentable at the moment, barn owners may be more lenient in checking who is renting their barn.

Students: **“Let us build a *risk indicator*: if an area is becoming poorer we may expect barns be rented for drug production”**

what is wrong with this?

Circumstantial vs direct evidence

- One way to reduce these issues is to:
  - strongly focus on **direct evidence** related to the ***modus operandi***, the addressed behaviour.
  - avoid as much as possible using **circumstantial evidence**, co-occurring properties like socio-economic features

Circumstantial vs direct evidence

- One way to reduce these issues is to:
  - strongly focus on **direct evidence** related to the ***modus operandi***, the addressed behaviour.
  - avoid as much as possible using **circumstantial evidence**, co-occurring properties like socio-economic features

- Eventually, students looked at data related to the pattern: barns on rent, chemical residuals in canals, assuming that residuals were thrown not nearby, but also not too far. Triangulating the data they found relevant instances, some of those confirmed. (the project took a 9)

Preventing ethical issues?

You define *behavioural* ***boundaries****,* for instance as deontic directives.

For instance:

- **You just are not allowed to clone humans.** No matter what.
- **You have to use direct evidence for judgments.** No matter what.

Preventing ethical issues?

You define *behavioural* ***boundaries***, for instance in terms of virtue ethics (behave as the virtuous person would do).

For instance:

- **Hippocratic Oath** for doctors https://en.wikipedia.org/wiki/Hippocratic_Oath
- **Archimedian Oath** for engineers https://sefi2024.eu/conference/archimedean-oath/

Preventing ethical issues?

You define *behavioural* ***boundaries***, for instance in terms of virtue ethics (behave as the virtuous person would do).

For instance:

- **Hippocratic Oath** for doctors
  https://en.wikipedia.org/wiki/Hippocratic_Oath
- **Archimedian Oath** for engineers
  https://sefi2024.eu/conference/archimedean-oath/

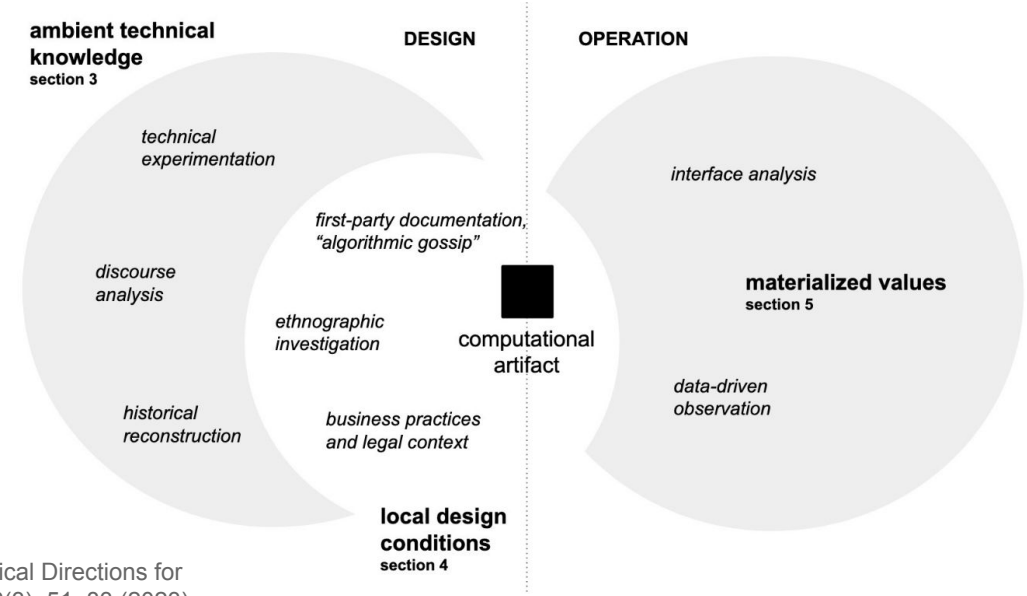
promoting *professional deontology!*

Identifying ethical issues?

- By taking a designer stance, we cover the prescriptive side of developed systems, but looking at the **descriptive side** is important too!

Identifying ethical issues?

- By taking a designer stance, we cover the prescriptive side of developed systems, but looking at the **descriptive side** is important too!
- First, because the reasons why systems behave in certain ways is often inaccessible or not intelligible.
- *Encircling* as a possible option?



Rieder, B., Gordon, G., Sileno, G., Mapping Value(s) in AI: Methodological Directions for Examining Normativity in Complex Technical Systems., Sociologica, 16(3), 51–83 (2023)

Identifying ethical issues?

- By taking the stance of the designer, we cover the prescriptive side of developed systems, but looking at the **descriptive side** is important too!
- Second, because even if systems are accessible, **there is a difference between what systems are designed to do, and what they are doing in practice** (a.k.a. “motivating” vs “revealed” aspects).

We need instruments to evaluate artefacts and systems independently from the official commitments.

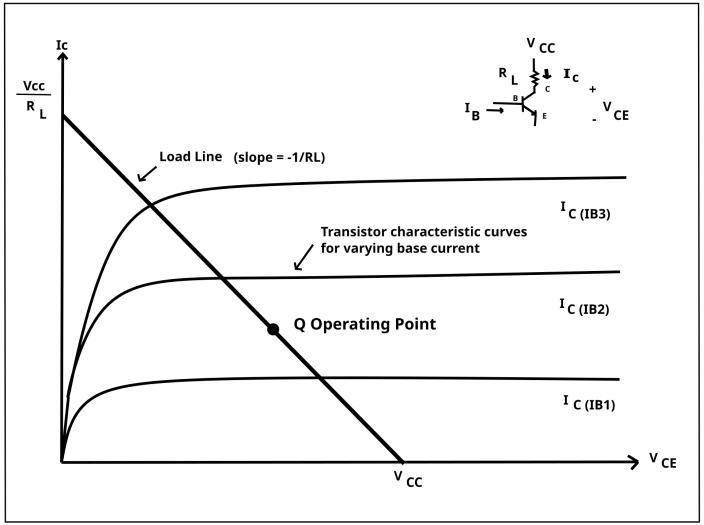
Identifying ethical issues? Several dimensions involved!

- prescriptive vs descriptive ethics
- motivating vs revealed aspects
- non-technical vs technical perspectives

A critical perspective on algorithmic fairness

Bias

- I was personally trained (as an electronic engineer) to see the term bias from a technical point of view: bias as variable (eg. current/tension) or offset that changes the functioning of a device.



(Social) Bias and Fairness

- More commonly, bias is used to denote social discrimination:

  preferring someone over someone else.

- typically to cases of **negative discrimination** (disfavouring/punishing). but in principle it can be applied for **positive discrimination** (favouring/assisting).

(Social) Bias and Fairness

- More commonly, bias is used to denote social discrimination:

  preferring someone over someone else.

- typically to cases of **negative discrimination** (disfavouring/punishing). but in principle it can be applied for **positive discrimination** (favouring/assisting).

affirmative action (“active effort to improve employment, educational, and other opportunities for members of groups that have been subjected to discrimination”) is positive discrimination, and by definition requires discrimination (= *differentiation*)

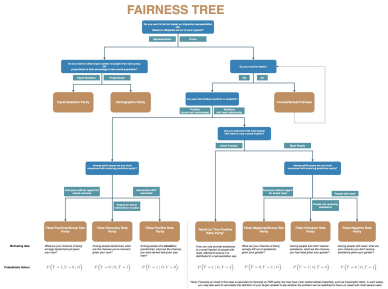
Bias and Algorithmic Fairness

- Particularly in data-centred contexts, bias is measured as **statistical parity**
- Studies in algorithmic fairness have introduced several statistical measures (**demographic parity, equalized odds**, …), often theoretically incompatible between each other.

Bias and Algorithmic Fairness

- Particularly in data-centred contexts, bias is measured as **statistical parity**
- Studies in algorithmic fairness have introduced several statistical measures (**demographic parity, equalized odds**, …), often theoretically incompatible between each other.
- *Which one to choose?*
    - It depends on the specific decision context – eg. how much impact has a false negative error compared to a false positive error.

      see eg. the AEQUITAS fairness tree: datasciencepublicpolicy.org/our-work/tools-guides/aequitas/



Bias and Algorithmic Fairness

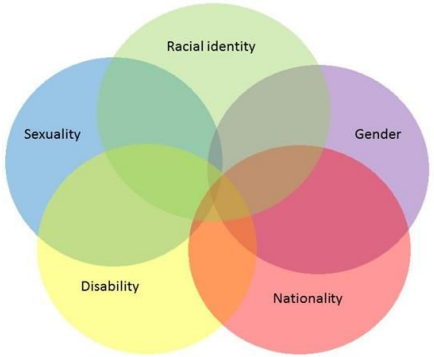
- Particularly in data-centred contexts, bias is measured as **statistical parity**
- Studies in algorithmic fairness have introduced several statistical measures (**demographic parity, equalized odds**, …), often theoretically incompatible between each other.
- *Which one to choose?*
    - It depends on the specific decision context – eg. how much impact has a false negative error compared to a false positive error.

Not in our scope today: I will talk here of two other general problems applying to all these measures.

Parity between whom?

- statistical measures relies on identifying groups
- but which groups? generally protected groups groups defined by sensitive dimensions (eg. sex or gender, race or ethnicity)

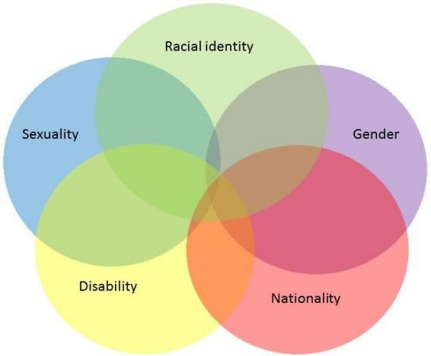
→ problem of ***intersectionality***



Parity between whom?

- statistical measures relies on identifying groups
- but which groups? generally protected groups groups defined by sensitive dimensions (eg. sex or gender, race or ethnicity)

→ problem of ***intersectionality***



Example:

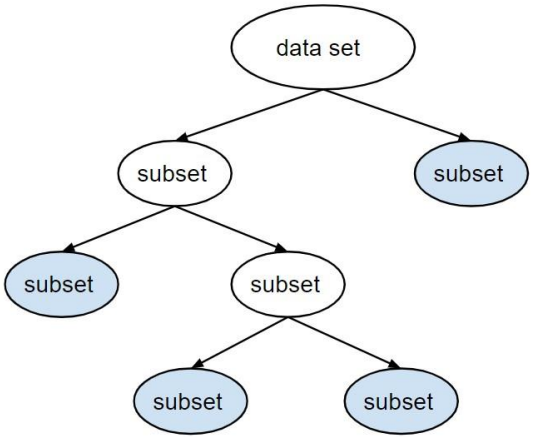
- women vs men
- black women vs black men
- low-class black women vs low-class black men

Which is the right granularity? (problem with affinities to Simpson's paradox!)

Parity between whom?

With a Master IS student we worked on inverting the problem:

- identifying descriptively relevant subgroups in the data space
- check how the algorithm works on these in general, and w.r.t. protected features



Wilms, M., Sileno, G., and Hinda Haned. PEBAM: A Profile-based Evaluation Method for Bias Assessment on Mixed Datasets. Proceedings of KI 2022, 45th German Conference on Artificial Intelligence.

Some results on German credit dataset

– Note: here ML is used to unveil historical (human) bias!

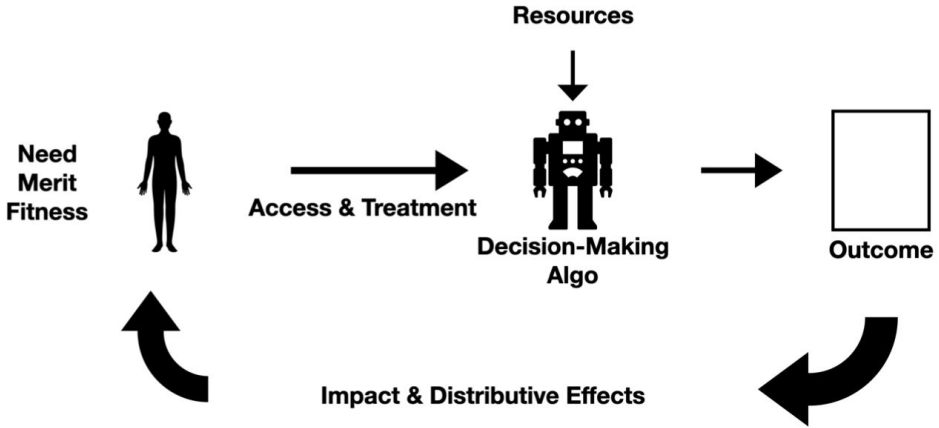
demographic disparity
against women
with low resources
when asking credit
for cars

(slight)
demographic disparity
against women
with low resources
when asking credit
for furniture/equipment

Profile	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Sample
0	26	male	2	rent	moderate	unknown	3577	9	car	859
1	37	male	2	own	unknown	unknown	7409	36	business	868
2	39	male	3	own	little	unknown	6458	18	car	106
3	26	male	2	own	little	little	4370	42	radio/TV	639
4	31	male	2	own	quite rich	unknown	3430	24	radio/TV	19
5	38	female	2	own	unknown	unknown	1240	12	radio/TV	135
6	43	male	1	own	little	little	1344	12	car	929
7	36	male	2	rent	little	little	2799	9	car	586
8	39	male	2	own	little	little	2522	30	radio/TV	239
9	31	male	2	own	little	moderate	1935	24	business	169
10	33	female	2	own	little	little	1131	18	furniture/equipment	166
11	26	male	1	own	little	moderate	625	12	radio/TV	220
12	23	male	2	own	unknown	moderate	1444	15	radio/TV	632
13	42	male	2	own	little	little	4153	18	furniture/equipment	899
14	29	male	2	own	unknown	unknown	3556	15	car	962
15	37	female	2	own	little	moderate	3612	18	furniture/equipment	537
16	27	female	2	own	little	little	2389	18	radio/TV	866
17	26	female	2	rent	little	unknown	1388	9	furniture/equipment	582
18	29	male	2	own	little	unknown	2743	28	radio/TV	426
19	53	male	2	free	little	little	4870	24	car	4
20	36	male	2	own	little	little	1721	15	car	461
21	38	male	2	own	little	unknown	804	12	radio/TV	997
22	29	male	2	own	little	moderate	1103	12	radio/TV	696
23	43	male	2	own	unknown	unknown	2197	24	car	406
24	27	male	2	own	little	little	3552	24	furniture/equipment	558
25	30	male	2	own	little	moderate	1056	18	car	580
26	24	female	2	own	little	moderate	2150	30	car	252
27	34	male	2	own	little	unknown	2759	12	furniture/equipment	452
28	24	female	2	rent	little	little	2124	18	furniture/equipment	761
29	34	male	2	own	little	moderate	5866	36	car	893
30	34	female	2	own	little	unknown	1493	12	radio/TV	638
31	30	female	2	own	little	unknown	1055	18	car	161
32	35	male	2	own	little	unknown	2346	24	car	654
33	35	male	2	own	unknown	unknown	1979	15	radio/TV	625

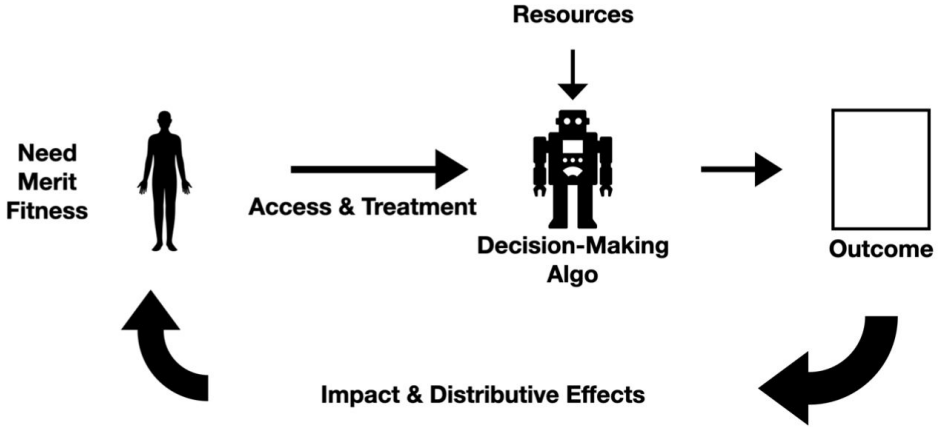
Parity about what?

- Algorithmic approaches overlooks the problem of defining of what fairness should be about (it is just about the decision). In contrast, philosophy has been debating for a long time which dimensions are relevant:



Parity about what?

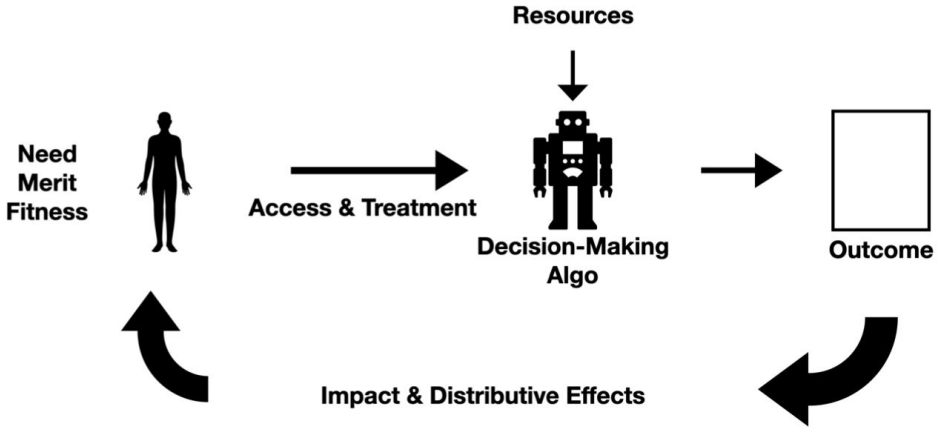
- Algorithmic approaches overlooks the problem of defining of what fairness should be about (it is just about the decision). In contrast, philosophy has been debating for a long time which dimensions are relevant:



- Algorithmic fairness currently applies **short-sighted consequentialist ethics**…

Parity about what?

- Algorithmic approaches overlooks the problem of defining of what fairness should be about (it is just about the decision). In contrast, philosophy has been debating for a long time which dimensions are relevant:



- Algorithmic fairness currently applies **short-sighted consequentialist ethics**… … **and may not bring fairer results** (work of a Master AI student last year).

Conclusions

In sum

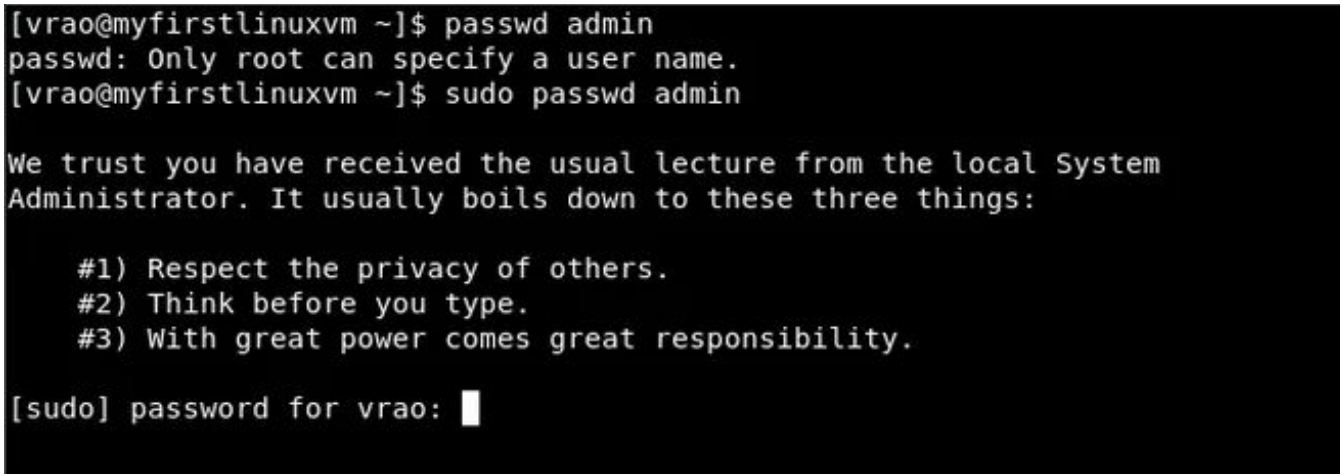
- We looked at the main concepts related to ethics, a few examples of problematic applications, and some insights on algorithmic fairness. Many relevant points were overlooked (eg. privacy, consent, participation).
- These examples demonstrated that ethics does not provide one-fits-all heuristics for ethical concerns. Actually there is no “solution” in **absolute sense**.

In sum

- We looked at the main concepts related to ethics, a few examples of problematic applications, and some insights on algorithmic fairness. Many relevant points were overlooked (eg. privacy, consent, participation).
- These examples demonstrated that ethics does not provide one-fits-all heuristics for ethical concerns. Actually there is no “solution” in **absolute sense**.
- Yet, discussing ethics assists in forming adequate reasons to justify why we accept (or do not accept) an algorithm to work in a certain way.
- This choice is a matter of **individual** and **collective** responsibilities…

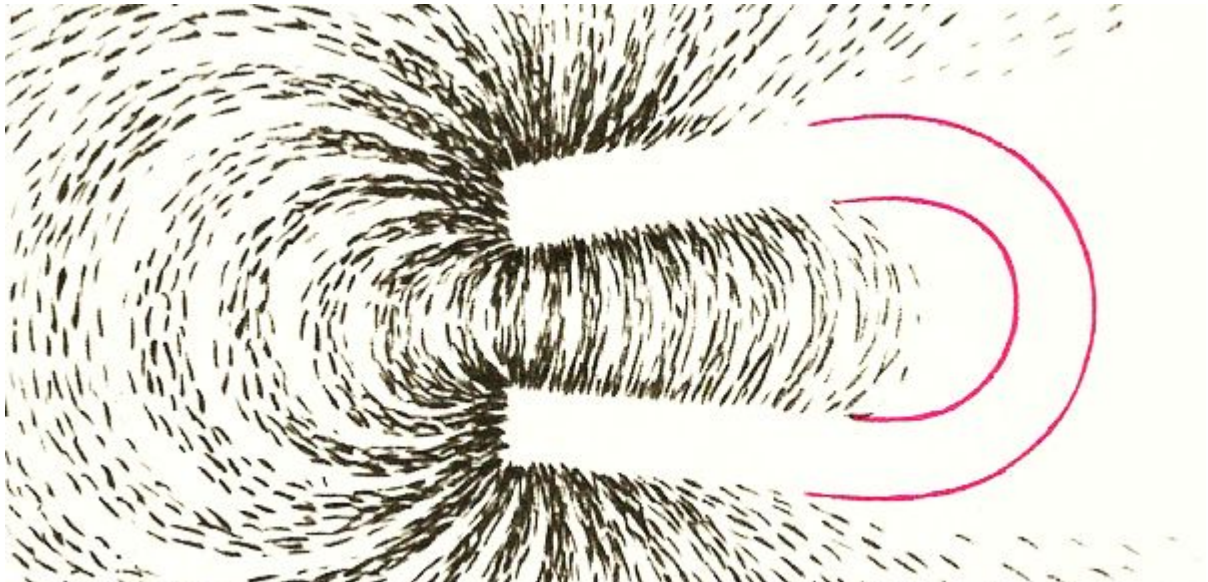
With power comes responsibility

- as AI designer/developer, **you have to consider prescriptive ethics** to spell out which norms/values you want your system to realize



Your actions have consequences

- as AI designer/developer, and as an social actor, **you have to consider descriptive ethics** to assess which norms/values in practice systems realize in the world



Protect your agency: do not be instrumental

- Yet, individual responsibility is only half of the story. Developers are only the last part of the chain.
- Example: *who has responsibility over weapons?*
  - the **designer** who conceived them
  - the **worker** that materially produce them
  - the **company** that sells them
  - the **country** that allows their production and sale.

Protect your agency: do not be instrumental

- Yet, individual responsibility is only half of the story. Developers are only the last part of the chain.
- Addressing individual behaviour only puts shame on people who for any reasons are materially obliged to do it.

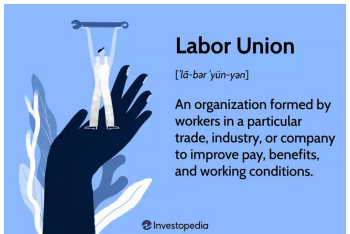
Protect your agency: do not be instrumental

- Yet, individual responsibility is only half of the story. Developers are only the last part of the chain.
- Addressing individual behaviour only puts shame on people who for any reasons are materially obliged to do it.
- *What to do then?*

  Participate to steer who has discourse, political and economic power!

ELECTRONIC FRONTIER FOUNDATION

TECH WORKERS COALITION TWC



Ethics for AI Developers

a crash course

6 November 2024

Human-centred Machine Learning course

Giovanni Sileno
g.sileno@uva.nl

University of Amsterdam